FEDERAL UNIVERSITY OF MINAS GERAIS
INSTITUTE OF BIOLOGICAL SCIENCE
DEPARTMENT OF GENERAL BIOLOGY
GRADUATE PROGRAM OF BIOINFORMATICS

BIOINFORMATICS STRATEGIES FOR IDENTIFICATION OF CANCER BIOMARKERS AND TARGETS IN PATHOGENS ASSOCIATED WITH CANCER

PhD STUDENT: Debmalya Barh

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

DEBMALYA BARH

Bioinformatics strategies for identification of cancer biomarkers and targets in pathogens associated with cancer

PhD STUDENT: Debmalya Barh

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

FEDERAL UNIVERSITY OF MINAS GERAIS

INSTITUTE OF BIOLOGICAL SCIENCES

BELO HORIZONTE - MG
February – 2017

043 Barh, Debmalya.
Bioinformatics strategies for identification of cancer biomarkers and targets in pathogens associated with cancer [manuscrito] / Debmalya Barh. – 2017.

237 f. : il. ; 29,5 cm.

Supervisor: Vasco Ariston de Carvalho Azevedo.
Tese (doutorado) – Federal University of Minas Gerais, Institute of Biological Sciences.

1. Bioinformática - Teses. 2. Marcadores biológicos. 3. Doenças transmissíveis - Teses. 4. Transcriptômica - Teses. 5. MicroRNAs - Teses. I. Azevedo , Vasco Ariston de Carvalho. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE TESE

DEBMALYA BARH

78/2017
entrada
2º/2016
CPF:
703.332.786-23

Às nove horas do dia **20 de fevereiro de 2017**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"BIOINFORMATICS STRATEGIES FOR IDENTIFICATION OF CANCER BIOMARKERS AND TARGETS IN PATHOGENS ASSOCIATED WITH CANCER"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Vasco Ariston de Carvalho Azevedo**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dr. Vasco Ariston de Carvalho Azevedo	UFMG	283.141225-49	APROVADO
Dr. José Miguel Ortega	UFMG	05950126807	Aprovado
Dr. Eduardo Martin Tarazona Santos	UFMG	01249405602	Aprovado
Dr. Siomar de Castro Soares	UFTM	05695182611	APROVADO
Dr. Sandro José de Souza	UFRN	705904099-53	APROVADO
Dr. Raghuvir Krishnaswamy Arni	UNESP	13871039896	APROVADO

Pelas indicações, o candidato foi considerado: APROVADO
O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 20 de fevereiro de 2017.

Dr. Vasco Ariston de Carvalho Azevedo - Orientador _____
Dr. José Miguel Ortega _____
Dr. Eduardo Martin Tarazona Santos _____
Dr. Siomar de Castro Soares _____
Dr. Sandro José de Souza _____
Dr. Raghuvir Krishnaswamy Arni _____

TABLE OF CONTENTS

Dedication

I dedicate this work to my parents Late Purnendu Bhusan Barh & Ms. Mamata Barh, for being my inspiration of life.

ACKNOWLEDGMENT

First and foremost, I would like to thank my supervisor **Prof. Dr. Vasco Ariston de Carvalho Azevedo.** I am grateful for his humble friendship, unconditional cooperation, and committed guidance with his vast wisdom granting me high degree of freedom to conduct and complete the entire work successfully.

I am indebted to my colleagues/ collaborators **Dr. Mukesh Verma** (Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, NCI/ NIH, USA), **Prof. Triantafillos Liloglou** (Department of Molecular and Clinical Cancer Medicine, University of Liverpool, UK), **Dr. Cedric Viero** (Institute of Molecular and Experimental Medicine, Wales Heart Research Institute, School of Medicine, Cardiff University, UK), **Dr. Nidia Leon-Sicairos** and **Adrian Canizalez-Roman (**Unidad de investigacion, Facultad de Medicina, Universidad Autonoma de Sinaloa. Cedros y Sauces, Fraccionamiento Fresnos, Mexico), **Prof. Ranjith Kumavath** (Department of Genomic Science, Central University of Kerala, India), **Prof. Preetam Ghosh** (Department of Computer Science and Center for the Study of Biological Complexity, Virginia Commonwealth University, Virginia, USA), **Prof. Artur Silva** (Instituto de Cieˆncias Biolo´gicas, Universidade Federal do Para´, Brazil), **Dr. Michel Herranz** (Molecular Oncology and Imaging Program, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain), **Dr. Elena Padin-Iruegas** (Medical Oncology Department, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain), and **Dr. Alvaro Ruibal** (Nuclear Medicine Service, Complejo Hospitalario Universitario. Fundación Tejerina. Santiago de Compostela, A Coruña, Spain) for their supports in experiments and analysis.

I am thankful to **Prof. Anil Kumar** (School of Biotechnology, Devi Ahilya University, Indore, India), **Prof. Amarendra Narayan Misra** (Center for Life Sciences, School of Natural Sciences, Central University of Jharkhand, Ranchi, India), **Prof. Kenneth Blum** (University of Florida, College of Medicine, Gainesville, Florida, USA), **Prof. Jan Baumbach** (Computational Biology Group, Department of Mathematics and Computer Science, University of Southern Denmark, Denmark), **Prof. Eric R. Braverman** (Weill-Cornell College of Medicine, Cornell University, New York, USA), **Prof. John K Field** (University of Liverpool, Department of Molecular and Clinical Cancer Medicine, Liverpool, UK) for their various assistances, suggestions, and guidance during the research.

I owe many thanks to my students/ friends **Sandeep Twari, Neha Jain, Amjad Ali, Marriam Bakhtiar, Anderson Rodrigues Santos, Bhanu Kamapantula, Joseph Nalluri, Syed Shah Hassan, Krishnakant Gupta, Siomar de Castro Soares, Sintia Silva de Almeida, Syed Babar Jamal, Rommel Ramos, Edson Luiz Folador,** and **Diego Mariano** for their various helps during this research.

My special thanks to my wife **Priyanka** and daughter **ShauryaShree** and other family members for their all supports during this research and my difficult times.

My deepest gratitude goes to my loving **Parents** for always being beside me and for their love, support, inspiration, and blessings.

LIST OF ABBREVIATIONS

BMI:	Body Mass Index
BPs:	Biological Processes
CAI:	Codon Adaptation Index
CC:	Cellular Component
CFUs:	Colony-Forming Units
CLA:	Caseous Lymphadenitis
CLR:	Context Likelihood of Relatedness
CMNR:	Corynebacterium, Mycobacterium, Nocardia, and Rhodococcus
CNPq:	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COG:	Clusters of Orthologous Groups
Cp:	Corynebacterium pseudotuberculosis
CTD:	Comparative Toxicogenomics Database
DAVID:	Database for Annotation, Visualization, and Integrated Discovery
DC:	Distance Correlation
DEG:	Database of Essential Genes
GENIE3:	Gene Network Inference with Ensemble of Trees
GO:	Gene Ontology
GSEA:	Gene Set Enrichment Analysis
HP-PPIs:	Host–pathogen protein–protein interactions
IEDB:	Immune Epitope Database
KEGG:	Kyoto Encyclopedia of Genes and Genomes
MF:	Molecular Function
miRNA/ miR:	MicroRNA
miRsig:	miRNA Signature
mRNAs:	messenger RNAs
MRNETB:	Maximum Relevance Minimum Redundancy Backward
MVD:	Molegro Virtual Docker
NCBI:	National Center of Biotechnology Information
NSCLC:	Non-Small Cell Lung Cancer
PAIDB:	Pathogenicity Island Database
PATRIC:	PathoSystems Resource Integration Center
PDB:	Protein Data Bank
PPIs:	Protein-Protein Interactions
PSAT:	Prokaryotic Sequence Homology Analysis
QC:	Quality Control
qPCR:	quantitative Real Time Polymerase chain reaction
ROC:	Receiver Operating Characteristic curve
RV:	Reverse Vaccinology
SAVS:	Structure Analysis and Verification Server
SCLC:	Small Cell Lung Cancer
STRING:	Retrieval of Interacting Genes
TFs:	Transcription Factors
TWAS :	The Academy of Sciences for the Developing World, Trieste, Italy

UFMG:	Universidade Federal de Minas Gerais
UniProt:	Universal Protein Resource

ABSTRACT

Cancer and bacterial infectious diseases are major cause of deaths globally and molecular biomarkers are essential tools for screening, diagnosis, prognosis, and therapy of these diseases. Various strategies have been employed to identify biomarkers over years. In this research, five novel bioinformatics strategies have been used to identify biomarkers in human diseases (especially cancer) and genomic targets in human pathogenic bacteria. **(I)** A novel *in silico* reverse-transcriptomics strategy based a panel of sub-type specific lung cancer biomarkers have been identified and validated in patients' blood samples using qPCR. An upregulation of TFPD1, E2F6, IRF1, and HMGA1 + NO expression of SUV39H1, RBL1, and HNRPD in blood sample are characteristics of Adeno and Squamous cell lung carcinomas. E2F6 is found a novel marker in lung cancer. The strategy can be useful in any other complex diseases and can explore novel insight of the disease pathogenesis, identification of early markers, and will be helpful in developing personalized medicine. **(II)** The second method describes "miRegulome"-a manually curated novel miRNA knowledgebase that gives entire regulatory modules of miRNAs and thus provide comprehensive understanding of miRNA regulatory networks and miRNA functions. Exploration of new and novel biological events and discovery of biomarkers and therapeutics can be achieved with high precision using Chemical-disease, miRNA-disease, Gene–disease, and Disease-chemical/miRNA analysis tools that are integrated with miRegulome. **(III)** In third bioinformatics strategy is on a novel computational methodology/pipeline (consensus of six network inference algorithms along with graph theory) for identification of miRNA-miRNA interactions based disease-specific and common miRNA Signatures (miRsig) in cancers or other diseases. The miRsig is powerful enough to identify early deregulated pan-cancer miRNA networks and therefore such miRNAs may be useful as screening or early diagnostic tools in cancer. miRsig can equally be applied in other diseases too. **(IV)** To identify common conserved targets in *M. tuberculosis, C. pseudotuberculosis (Cp*), *C. diphtheriae, C. ulcerans, Y. pestis,* and pathogenic *E. coli,* a novel integrated bioinformatics approach combining protein-protein interactions (PPI), host-pathogen interactions, and subtractive genomics is presented in the forth strategy. Using this method, first time we have developed intra-species PPIs *Cp* strains and acetate kinase (Ack) as a common conserved target for all these pathogens. Piperdardine and Dehydropipernonaline from Piper betel target Ack more effectively than Penicillin and Ceftiofur *in silico* and in *in vitro*, Piperdardine inhibits *E. coli* O157:H7 growth similar to penicillin. **(V)** In the fifth strategy, comparative and subtractive exoproteomics and secretomics in combination with modified reverse vaccinology approach, ompU, uppP and yajC were identified as novel and common conserved targets in 21 *V. cholerae* serotypes. Seven Piper betel compounds show inhibitory effects against these targets in *in silico* and anti- *Vibrio* effects *in vitro*. Although these bacteria are predominantly associated with various infectious diseases, they are also reported to be associated with tumor/ cancer. The *M. tuberculosis* infection increases the risk of lung cancer, and several human miRNAs are deregulated in both lung cancer and pulmonary tuberculosis; The future scope of this research is to develop bioinformatics strategies to identify the common signature associated with both pulmonary tuberculosis and lung cancer so that common cause and common management strategies can be developed against pulmonary tuberculosis and lung cancer.

Keywords: Biomarkers, targets, bioinformatic strategies, genomics, transcriptomics, miRNA, cancer, infectious diseases, *M. tuberculosis*

I. PRESENTATION

I.1 Collaborations

This work has been **conducted under supervision of Prof. Dr. Vasco Ariston de Carvalho Azevedo**, Department of General Biology, Institute of Biological Sciences (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. Other Researchers/Collaborators (National and International) and their respective Institutions, among others, are:

1. **Dr. Mukesh Verma:** Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute (NCI), National Institutes of Health (NIH), 9609 Medical Center Drive, Rockville, MD 20850, USA
2. **Dr. Nidia Leon-Sicairos and Adrian Canizalez-Roman:** Unidad de investigacion, Facultad de Medicina, Universidad Autonoma de Sinaloa. Cedros y Sauces, Fraccionamiento Fresnos, Culiacan Sinaloa 80246, Mexico
3. **Dr. Adrian Canizalez-Roman:** Unidad de investigacion, Facultad de Medicina, Universidad Autonoma de Sinaloa. Cedros y Sauces, Fraccionamiento Fresnos, Culiacan Sinaloa 80246, Mexico
4. **Prof. Ranjith Kumavath:** Department of Genomic Science, School of Biological Sciences, Riverside Transit Campus, Central University of Kerala, Kasaragod, India
5. **Prof. Preetam Ghosh:** Department of Computer Science and Center for the Study of Biological Complexity, Virginia Commonwealth University, 401 West Main Street, Richmond, Virginia 23284-3019, USA
6. **Prof. Artur Silva:** Instituto de Cieˆncias Biolo´gicas, Universidade Federal do Para´, Bele´m, PA, Brazil
7. **Prof. Anil Kumar:** School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India
8. **Prof. Amarendra Narayan Misra:** Center for Life Sciences, School of Natural Sciences, Central University of Jharkhand, Ranchi, Jharkhand State, India
9. **Prof. Kenneth Blum:** University of Florida, College of Medicine, Gainesville, Florida, USA
10. **Prof. Jan Baumbach:** Computational Biology Group, Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark
11. **Prof. Eric R. Braverman:** Weill-Cornell College of Medicine, Cornell University, New York, New York, USA
12. **Prof. John K Field:** University of Liverpool, Department of Molecular and Clinical Cancer Medicine, 200 London Road, Liverpool L3 9TA,UK
13. **Prof. Triantafillos Liloglou:** University of Liverpool, Department of Molecular and Clinical Cancer Medicine, 200 London Road, Liverpool L3 9TA,UK

14. **Dr. Cedric Viero:** Institute of Molecular and Experimental Medicine, Cardiff University, Cardiff CF14 4XN, Wales, UK

15. **Dr. Michel Herranz:** Molecular Oncology and Imaging Program, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain

16. **Dr. Elena Padin-Iruegas:** Medical Oncology Department, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain

17. **Dr. Alvaro Ruibal:** Nuclear Medicine Service, Complejo Hospitalario Universitario. Fundación Tejerina. Santiago de Compostela, A Coruña, Spain

I.2 Thesis Delineation

This thesis is divided into Introduction (based on three review articles), Objectives (general, specific, and future objectives), **five** Chapters (based on five research articles), General conclusions, Future perspectives, Bibliography, and Appendix.

i. The **Introduction** gives the overview and rationale of the research in brief. It comprises of **three articles**. The first article is a book chapter that presents overview, technologies, and strategies associated with molecular biomarkers. The second article is also a book chapter on the challenges, applications, and future perspectives of next-generation molecular markers. The third article is a review paper that provides an overview of genomics approaches towards identification of targets in human bacterial pathogens.

ii. The **Objectives** section provides the general, specific, and future aims of this research.

iii. The **first chapter** presents a research article showing a novel *in silico* reverse-transcriptomics strategy to identify sub-type specific biomarkers in lung cancer and blood based experimental validation of the identified markers. The strategy can be equally applied in any other human disease also. At the end of the chapter, the conclusions form this research are summarized and the media attentions on the research outcomes are provided.

iv. In **second chapter**, the research article describes novel miRegulome knowledgebase along with its integrated analysis tools that can be used to explore miRNA regolome, novel functions of miRNAs, and miRNA based biomarker and therapeutics discovery in any disease. The summary of the outcomes from the research and the media highlights on this work are given at the end of this chapter.

v. The **third** research **chapter** demonstrates miRsig: a consensus-based network inference methodology that can identify pan-cancer miRNA-miRNA interaction and therefore can predict miRNA based early biomarker signatures in cancers. The conclusions form this research are summarized at the end of the chapter.

vi. The **fourth chapter** describes a novel integrated bioinformatics approach that includes PPIs, host-pathogen interactions, subtractive genomics, and pangenomics identify common conserved and novel targets in pathogenic bacteria (*C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis,* and *E. coli*). Further Piper betel compounds were tested if they are potential to target those identified targets. The conclusions form this research and the media highlights are summarized at the end of the chapter.

vii. In **fifth chapter** presents a novel bioinformatics strategy for identification of board spectrum drug and vaccine targets in *V. cholerae* and Piper betel derived phytochemicals are tested *in silico* to target the identified *Vibrio* targets and their anti-*Vibrio* activities were validated *in vitro*. Like the other chapters, the conclusions form this research and the media highlights of both chapters (Chapters four and five) are also given at the end of this chapter.

viii. The **General conclusions** section describes the summary and main outcomes of the entire works represented in this thesis

ix. Under **Future perspectives**, the future extension of the entire work is pointed out with a focus on developing screening or early diagnostic markers for tuberculosis patients who are going to develop lung cancer or *vice versa*.

x. In **Bibliography** section, the references used in Introduction and conclusions and future perspectives sections are listed out.

xi. The **Appendix** section lists out all the other publications, seminar, courses, etc. that are carried out /attended related to the research presented in this thesis

II. OBJECTIVES

II.1 General Objectives

The general aim of this research is to develop novel bioinformatics approaches to identify biomarkers that could be used as diagnostic and target (bacterial pathogens) for effective management of cancer and infectious diseases, respectively.

II.2 Specific Objectives

Cancer

i. Develop bioinformatics strategy (novel *in silico* reverse transcriptomics) to identify lung cancer diagnostic/ screening biomarkers, lung cancer sub-type specific markers selection, and experimental validation of *in silico* identified markers.
ii. Development of an miRNA regulomics knowledge-base and miRNA analytics as a readily available resource for discovery of miRNA, gene, and protein based biomarkers and therapeutics in human diseases, specifically in cancers.
iii. Development of bioinformatics approach towards identification of pan-cancer miRNA-miRNA interaction based miRNA signatures towards screening and early diagnosis of cancers.

Bacterial pathogens

i. Development of an integrated bioinformatics strategy (PPI, host-pathogen interactions, subtractive genomics) to identify common conserved targets in *M. tuberculosis*, *C. pseudotuberculosis, C. diphtheriae, C. ulcerans, Y. pestis,* and *E. coli.*
ii. *In silico* analysis of Piper betel derived phytochemicals to target the identified common conserved targets.
iii. Checking the efficacy of Piper betel compounds as potential antibacterial agents against these pathogens using *E. coli* O157:H7 system.
iv. Identification of broad spectrum drug and vaccine targets using comparative, subtractive, pan-proteomics, and reverse vaccinology strategies applied to exoproteome and secretome of twenty one *V. cholerae* serotypes.
v. Epitope and peptide vaccine design using the identified vaccine targets in *V. cholerae.*
vi. Virtual screening of Piper betel derived phytochemicals against the identified drug targets in *V. cholerae.*
vii. Experimental validation of Anti-*Vibrio* activity of Piper betel derived phytochemicals.

II.3 Future Objectives

The future perspective of this research outcomes is to identify common markers in cancers and infectious diseases especially in lung cancer and tuberculosis.

III. INTRODUCTION

III.1 Preface

Biomarkers are biological factors or molecules that generally indicate a biological state, process, event, or a condition such as a disease state or the cause of the disease. Markers also indicate a biological process, disease process, or a response to a particular therapy. Based on their applications or utility, biomarkers can be classified as diagnostic, prognostic and therapeutic biomarkers and are essential tools for these purposes. Biomarkers can also be classified as imaging biomarkers like X-Ray, ECG, CT, MRI etc. or non-imaging biomarkers such as molecular biomarkers with biophysical or biochemical properties. Such non-imaging molecular biomarkers are DNA variations, chromosomal aberrations, gene/ mRNA and protein expression profiles, microRNA expression and variations, metabolites, enzymes, antigens, and hormones among others [BARH ET AL. 2012].

Cancers and infectious diseases are two major global health problems. According to WHO, approximately 23% deaths across the world are due to various cancers [WHO, 2010] and lung cancer is the leading cause of all cancer related deaths that is estimated to be 1.6 million deaths worldwide [JEMAL ET AL., 2011]. The high incidence of cancer associated deaths is due a combination of lack of effective early diagnostic, prognostic, and therapeutic markers [WANG ET AL., 2010] and gold standard therapeutic regimens [GRANVILLE & DENNIS, 2005]. Further, bacterial infectious agents are also etiological factors for several cancers. For example *Mycoplasma* in prostate cancer [ROGERS, 2011], *Robinsoniella* in pancreatic cancer [SHEN ET AL., 2010], *S. typhi, H. bilis, H. hepaticus*, and *E. coli* in carcinoma of the gallbladder [NATH ET AL., 2010], and *Chlamydia* in cervical cancer. Several reports have documented co-existence of tuberculosis and lung cancer [SKOWRONSKI ET AL., 2015; YU ET AL., 2008; LIANG ET AL., 2009; WU ET AL., YU ET AL., 2011; EL-SHARIF ET AL., 2012] and pulmonary tuberculosis is a risk factor for developing lung cancer [YU ET AL., 2008; LIANG ET AL., 2009; WU ET AL., YU ET AL., 2011]. Pulmonary tuberculosis caused by *M. tuberculosis* is itself a global health problem which is a major causes of death among the infectious diseases and according to WHO 2013 report, it is estimated that 9 million people are infected and 1.5 million are died from tuberculosis in 2012 [GLAZIOU ET AL., 2015]. Apart from these pathogens, *C. pseudotuberculosis, C. diphtheriae, C. ulcerans, Y. pestis, E. coli*, and *V. cholerae* are significantly contribute to the total infectious disease prevalence all over the world. These pathogens are also reported to be associated with tumor or cancer with unknown etiology [**TABLE-1**]. Although, the vaccines and drugs are available against most of these pathogens, the infections remain frequently uncontrolled due to the emerging antibiotic resistance of these pathogens and therefore new and novel targets need to be identified and novel drugs/ vaccines to be developed against these pathogens [MANDAL ET AL., 2011].

With the advent of omics data, the current considerations of molecular biomarkers discovery are commonly based on genomics, proteomics, metabolomics, or transcriptomics approaches. However, in each strategy, bioinformatics is an integral part and gene/protein/miRNA and metabolite based markers have better utilities, sensitivity, and specificity [KULASINGAM & DIAMANDIS, 2008; PLUMP & LUM, 2009; COHEN FREUE ET AL., 2013; ZHANG ET AL., 2015]. Bioinformatics approaches mainly comprise of **(1)** Data-driven approaches, and **(2)** Knowledge-driven approaches. In data-driven approaches, trend analysis, clustering,

classification, visualization, and network analysis are key strategies. On the other hand, in Knowledge-driven approaches, text mining, protein-protein interactions (PPIs), pathway analysis, and gene-set enrichment analysis are mainly followed [Kim ET AL., 2011; MCDERMOTT ET AL., 2013; BAKHTIAR ET AL., 2014]. While the said strategies are effective for biomarker discovery in human diseases; systems approach, comparative and pan-genomics, structural proteomics, and host-pathogen interaction network analysis are generally followed for pathogenic bacterial target identification [ALI ET AL., 2015; RADUSKY ET AL., 2015; DIX ET AL., 2016].

TABLE-1: Infectious bacterial pathogens that are reported to be associated with neoplasms/ cancers

Bacteria Genus	Species/ Strain/ Serogroup	Main infectious disease	Species/ Strain/ Serogroup	Associated neoplasm/ cancer	Reference
Mycobacterium	*tuberculosis*	Tuberculosis	*tuberculosis*	Lung cancer	LIANG ET AL., 2009; SKOWROŃSKI ET AL., 2015
Yersinia	*pestis*	Plague	*enterocolitica,*	Lymphadenitis	ZIŃCZUK ET AL., 2015
Corynebacterium	*diphtheriae*	Diphtheria	*diphtheriae*	Bladder and other cancers	MATTOS-GUARALDI ET AL., 2001; GOMES ET AL., 2009
Corynebacterium	*ulcerans*	Diphtheria-like	*ulcerans*	Skin ulcer	MOORE ET AL., 2015
Corynebacterium	*pseudotuberculosis*	Caseous lymphadenitis	*pseudotuberculosis*	Granulating ulcer	YERUHAM ET AL., 1997; ALMEIDA ET AL., 2016
Escherichia	*coli*	Diarrhea	*coli*	Bladder cancer	EL-MOSALAMY ET AL., 2012
Vibrio	*cholera*	Cholera	*non-O1*	Hepatocellular carcinoma, Pancreatic cancer	CHEN ET AL., 2015

In this research, various novel bioinformatics strategies have been developed to identify gene/ protein and miRNA bases biomarkers in human diseases (especially in cancer) and genomics based targets in pathogenic bacteria. The following three review articles provide the background of this research where the first article presents overview, technologies, and strategies associated with molecular biomarkers; the second article provides the challenges, applications, and future perspectives of next-generation molecular markers; and the third article signifies the overview of genomics approaches towards identification of targets in bacterial pathogens.

III.2 Article I: Literature Review

Molecular Biomarkers: Overview, Technologies, and Strategies

Mukesh Verma, **Debmalya Barh**, Sandeep Tiwari, **Vasco Azevedo**

Book Chapter: Molecular Biology and Biotechnology, 2015; 6th Edition, ISBN: 978-1-84973-795-1, Editors: Ralph Rapley and David Whitehouse, The Royal Society of Chemistry, UK, pp: 365-415

This chapter provides a detail account of biomarkers. It provides overview, type of biomarkers, and omics approaches in biomarkers discovery in brief. Further, the chapter describes in details the genomics, proteomics, transcriptomics, immunomics, metabolomics, epigenomics and miRNomics technologies and strategies -based biomarkers in immune diseases, cardiovascular diseases (CVD), metabolic diseases, infectious diseases, neurological diseases, and cancers.

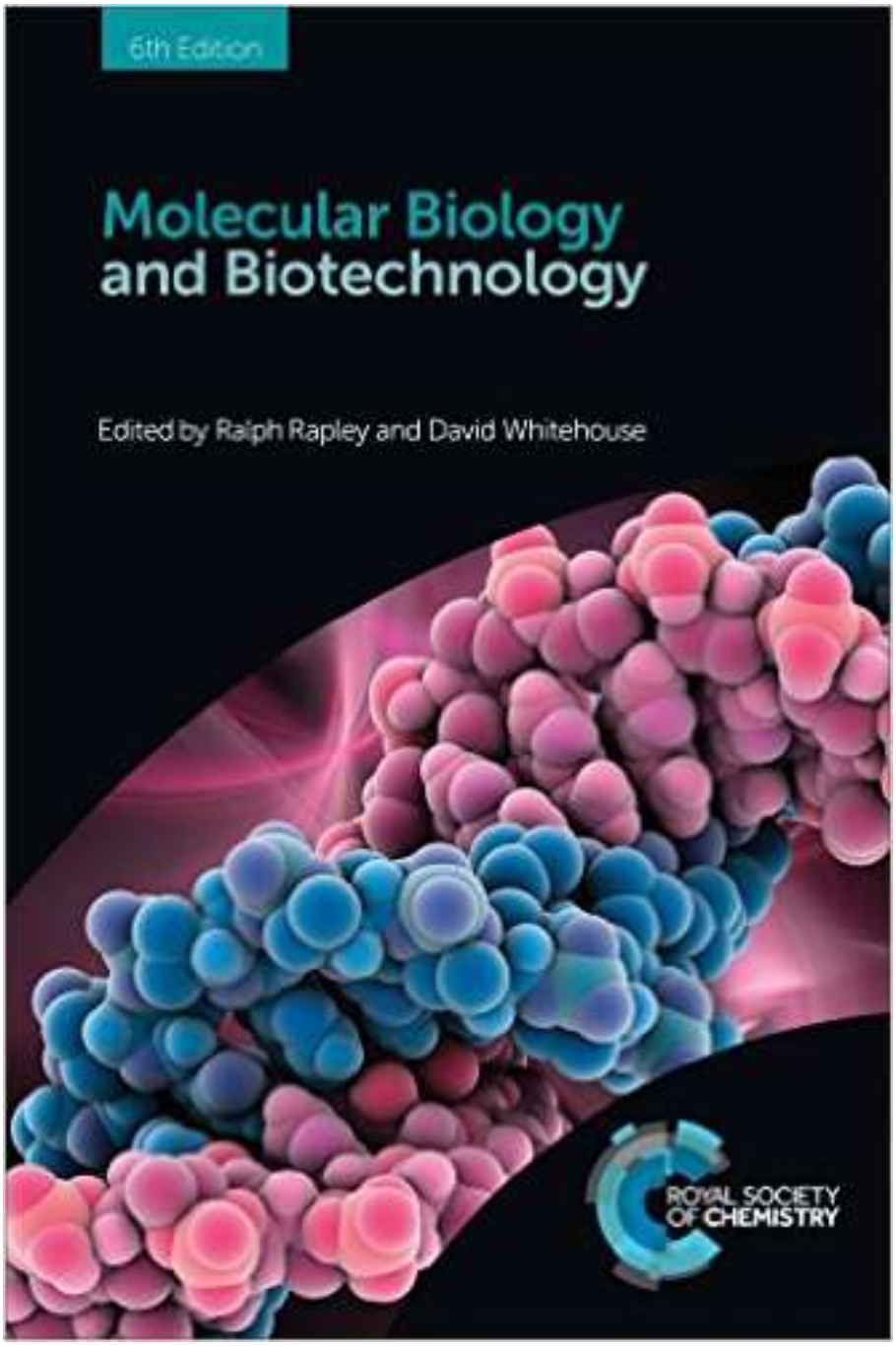
III.3 Article II: Literature Review

Next-generation Molecular Markers: Challenges, Applications, and Future Perspectives

Mukesh Verma, **Debmalya Barh**, Syed Shah Hassan, **Vasco Azevedo**

Book Chapter: Molecular Biology and Biotechnology, 2015; 6th Edition, ISBN: 978-1-84973-795-1, Editors: Ralph Rapley and David Whitehouse, The Royal Society of Chemistry, UK, pp: 416-449

Developing novel and gold standard biomarkers is always a challenge due to several aspects of markers, the target diseases, and the used technologies. Majority of this chapter is on cancer and the chapter deals with cancer biomarker discovery associated with biological and technological/ analytical limitations, clinical and pathologic factors, intellectual properties, and health economy factors. The chapter also presents the next-generation molecular markers and their applications in diagnostic screening, early diagnosis, risk assessment, prognosis, drug response/ pharmacogenetics, and therapeutics in various diseases including immune diseases, cardiovascular diseases (CVDs), metabolic diseases, infectious diseases, nneurological diseases, and cancers. At the end of the chapter, the recent trends and future directions of next-generation biomarkers are also discussed.



Molecular Biology and Biotechnology, 6th Edition
Editors: Ralph Rapley and David Whitehouse,
Publisher: The Royal Society of Chemistry, UK
Published: 03 Dec 2014,
Print ISBN: 978-1-84973-795-1

Amazon Best Sellers Rank: #2,966,876 in Books (As on 20 Jan, 20117)

CHAPTER 13

Molecular Biomarkers: Overview, Technologies, and Strategies

MUKESH VERMA,[a] DEBMALYA BARH,*[b] SANDEEP TIWARI[b,c] AND VASCO AC AZEVEDO[c]

[a] Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute (NCI), National Institutes of Health (NIH), 9609 Medical Center Drive, Rockville, MD 20850, USA; [b] Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal-721172, India; [c] Departamento de Biologia Geral, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Minas Gerais, Brazil
*Email: dr.barh@gmail.com

13.1 INTRODUCTION

Biomarkers can be specific cells, molecules, images, genes, gene products, enzymes, hormones, or receptors.[1] The term ''biomarkers'' is sometimes replaced by ''actionable biomarkers'' to describe biomarkers that can inform that clinical practice can be applied for disease diagnosis, prognosis, and treatment. Most of such actionable biomarkers exist in cancer although in other diseases progress has also been made in the last few years. Sometime entirely different kinds of biomarkers exist, for example, temperature is a biomarker of abnormal physiology of the body (fever) possibly due to infection, and blood pressure is a biomarker for stroke. To evaluate the current status of the biomarker field in disease diagnosis and therapy, we analyzed published literature starting from 1985 until April 2013 and summarized

Molecular Biology and Biotechnology, 6th Edition
Edited by Ralph Rapley and David Whitehouse

Published by the Royal Society of Chemistry, www.rsc.org

Table 13.1 Publications of biomarker studies indicating investigator interest in the field.

Topic	Number of Publications
Biomarker	601401
Biomarker and diagnosis	373653
Biomarker and therapy	173756
Biomarker and cancer	215529
Biomarker and neurological disorders	46330
Biomarker and cardiovascular disorders	57744
Biomarker and metabolic disorders	521
Biomarker and immunologic disorders	73417
Biomarker and infectious agents related diseases	159
Biomarkers of viruses in diseases	11445
Biomarkers of bacteria in diseases	14538
Biomarker and breast cancer and epigenetics	45
Biomarker and breast cancer and methylation	381
Biomarker and breast cancer and histone	247
Biomarker and breast cancer and microRNA	209
Biomarker and breast cancer and proteomics	419
Biomarker and breast cancer and imaging	1094

The analysis was PubMed-based, and references up to April 2013 were considered.

in Table 13.1. The total number of publications in the biomarker field was 601 401. The number of publications in the biomarker diagnosis field was almost double those in therapeutics. This seems reasonable because more biomarkers exist which can be used for disease diagnosis but it takes time to find biomarkers for therapies. Metabolomics disease biomarkers are fewer compared to cardiovascular disease markers.[2] Another important point in such meta-analysis is the proper use of terms for the analysis. For example, when we used ''infectious agent biomarker'', the number of publications was much less than using the term ''bacterial biomarkers'' or ''viral biomarker'' (see Table 13.1). In some diseases, a single biomarker is sufficient to diagnose a disease whereas in other diseases multiple biomarkers are needed for diagnosis. The utility of multiple biomarkers is determined based on their combined sensitivity and specificity, which should be superior to a single marker. In this chapter biomarkers of immune diseases, cardiovascular diseases (CVDs), metabolic diseases, infectious diseases, neurological diseases, and cancer are discussed (see Tables 13.2–13.7).

13.2 OVERVIEW OF BIOMARKERS

Biomarkers are often used in clinics where they are derived from biological fluids that are easily available. Biomarkers show characteristic biological properties that can be detected and measured in parts of the body such as organs or body fluid (blood, urine, saliva, nipple aspirate, pleural lavage, pancreatic juice).[1,3] They represent the normal or disease state of the body

and can be used to follow up treatment in the disease state. A number of assays have been developed to identify biomarkers. These assays include immunohistochemistry, imaging, gene constitution (amplification, mutation, and rearrangement), gene and protein expression analysis such as single gene or protein expression, methylation and histone profiling, miRNA polymorphism and miRNA profiling.[3–6] Biomarkers play a major role in medicinal biology.[4–7] In medicine, a biomarker generally refers to a specific protein concentration in blood that may in turn reflect the presence, progression or severity of a disease and guide the treatment.[8] Biomarkers have been used for more than half a century but their application has increased remarkably since the 21st century. A diseased state indicates a structural change in the proteins or enzymes. The physiological changes between the normal and diseased state is compared to look out for biomarkers. This is because the gene and protein expression profiles along with the metabolic expression profiles change in a diseased state. The researchers study the up-regulated and down-regulated genes, proteins and metabolites and understand the genetic patterns closely associated with a particular type of disease that leads to the discovery of a biomarker. Once identified, the biomarkers are validated in a large number of samples and possibly at two different laboratories. Microarrays—RNA, DNA, protein or antibody—play an important role in analyzing a biomarker. The comparison of plasma protein concentration levels in normal and diseased states is often used in biomarker analysis. For example, various new forms of glycoproteins are formed during glycosylation wherein polysaccharides or sugars are added to the polypeptides (proteins). Any abnormal concentration of glycoproteins can act as a biomarker for various diseases such as muscular dystrophy or acute chronic inflammation. In case of pancreatic cancer, RNAase-1 is used as a biomarker for disease diagnosis. An obvious alteration in the pattern of glycolysation of the enzyme was observed in the urine and blood serum in tumorous pancreatic cells. In molecular terms, a biomarker is the subset of markers that can be discovered using various omics or imaging technologies. We have selected biomarkers in cancer, cardiovascular disease, neurological disorders, infectious agents related diseases, metabolic diseases and immune diseases to cover wide-spectrum of diseases and up to date information about biomarkers associated with these diseases and disorders.

13.3 TYPES OF BIOMARKERS

13.3.1 Based on Utility

Physicians and scientist use various types of biomarkers to study human diseases either to track the progression of a disease, or to detect the effect of a drug. The use of biomarkers in the diagnosis of infections, genetic disorders and cancer is well known. The expansion of lab technology and growth of molecular biology increases the feasibility of biomarkers that are

technically advanced. Biomarkers play a very important role in personalized medicine as they facilitate the combination of therapeutics with diagnostics. In contrast, big pharmaceutical and biotechnology companies use biomarkers as a drug discovery tool. Biomarkers are important in drug development as they help to determine the pharmacodynamic effects of the drug being developed and assess the safety and the efficacy of the drug. Safer drugs with better efficacy can be developed in a cost effective manner by using biomarkers.

Biomarkers include tools and technologies that can be used in prediction, progression and outcome (survival, recurrence, multiple diseases) of a disease. Biomarkers can be classified based on detection techniques (imaging and non-imaging), properties (molecular, cellular, and nuclear), and utility (diagnosis, prognosis and therapy). Because of the latest high-throughput omics technologies in genomics, epigenomics, proteomics, metabolomics, and transcrptomics, a large number of potential biomarkers have been identified that can be utilized in various ways for disease management and patient care.[2,3] Biomarkers can be classified as diagnostic, prognostic, and therapeutic biomarkers based on their utility.[8] Early detection of biomarkers are used to detect a particular disease in its early stage. Diagnostic biomarkers are those, which are used to identify the presence or absence of a disease or diagnose a disease. Predictive biomarkers are those that are present prior to an event occurring and that predict the outcome and the efficacy of the drug in a treatment. Prognostic biomarkers give valid information about the outcome without using a therapy. They are used to determine how such a disease may develop (etiology of the disease) in an individual and increase survival of the patient. Disease prognosis biomarkers are related to measures of a disease state. Efficacy biomarkers reflect the effects of a particular treatment using a specific drug. Since they correlate with the desired clinical outcomes, they can be used to obtain provisional regulatory approval of a drug. Surrogate biomarkers are used to measure the clinical outcomes. Toxicity biomarkers, as their name reflects, measure the toxicity of drugs or interventions. Target biomarkers reflect the presence of a specific drug target and indicate the drug target interaction and outcome. Lastly, pharmacodynamic biomarkers (pharmacogenomics and pharmacoepigenomics) belong to the category of biomarkers that are used in drug development for intervention and/or treatment.

13.3.2 Based on Diagnosis Approaches

Based on diagnostic approaches, biomarkers can also be classified as imaging biomarkers or non-imaging biomarkers. Imaging biomarkers are those biomarkers which are detectable by using imaging techniques, such as X-ray (mammograms), electrocardiogram (ECG), ultrasound imaging, computed tomography (CT) and CT scanning, magnetic resonance imaging (MRI) and quantum dots.[9] Imaging biomarkers give reproducible results

and have tremendous potential in diagnosis as well as therapy and prognosis.

Non-imaging biomarkers can be molecular biomarkers with biophysical properties. They may include nucleic acid based biomarkers such as gene mutations or polymorphisms, alterations in copy number (copy number variance or CNVs), changes in mitochondrial genome, chromosomal aberrations, gene and protein expression profile, microRNA, metabolites, enzymes, antigens, hormones, *etc.*[1,4–7] Serum levels of IL-27, IL-29, IL-31, BALF were high in lung cancer patients.[10] Higher levels of miR-21 and lower levels of miR-451 and miR-485 were observed in lung cancer.[11]

If a combination of biomarkers (biomarker profiles) can detect a disease early, it provides an opportunity for developing intervention and therapeutic approaches.[3] For example, if biomarkers can discriminate pre-rheumatoid arthritis subjects from normal subjects, intervention and therapeutic approaches can be applied in high-risk individuals because such drugs exist which can treat rheumatoid arthritis. Biomarkers are also useful in detecting secondary cancers (recurrence).

13.3.3 Based on Therapy

Biomarkers are used to follow up therapy and their levels indicate the host's response to therapy. Due the recent development in genomics, it is possible to make therapy personalized. The identification of biomarkers represents a fundamental medical advance that can lead to an improved understanding of a disease, and holds the potential to define surrogate diagnostic and prognostic end points.

Understanding biomarkers in a complex system such as central nervous system (CNS) is extremely valuable to develop therapeutics in brain and neurological disorders. Brain parenchyma is a highly complex microvasuclar structure and it undergoes a variety of changes during tumor formation. Neo-angiogenesis starts in tumors and if anti-angiognesis therapy is planned for such brain tumors, biomarkers should be known which can be used to follow therapy. Some biomarkers in this category are vascular endothelial growth factor (VEGF), phosphatidylinositol glycan biosynthesis class F protein (PIGF), angiopoitin-1, and integrin.[12] In Alzheimer's disease, glutamate levels are increased (in the hippocampus region of the brain) after treatment with galantamine.[13] Therefore, glutamate can be considered a therapeutic biomarker of Alzheimer's disease. Galantamine is a cholinesterase inhibitor that is generally used in the treatment of this disease.[14]

An ideal biomarker should indicate whether a treatment is non-toxic and effective. Molecular biomarkers are required to predict the likelihood of an individual tumor's responsiveness or of toxicity in normal organs and to advise optimized treatments with improved efficacy at reduced side effects for each cancer patient. Biomarkers with prognostic value concerning treatment response and patient survival can then be used as targets to

develop optimized drugs. For example: (i) chemo-selective treatment of tumors with 9p21 deletion by L-alanosine; (ii) treatment of multidrug-resistant P-glycoprotein-expressing tumor cells by non-cross-resistant natural products or by inhibitors of P-glycoprotein to overcome multidrug resistance; and (iii) natural products that inhibit the epidermal growth factor receptor (EGFR) in EGFR-over expressing tumor cells.[15]

In cancer cells, cyclic AMP dependent protein kinase (PKA) is secreted into the conditioned medium. This PKA, designated as extracellular protein kinase A (ECPKA), is markedly up-regulated in the sera of patients with cancer. The currently available tumor biomarkers are based on the antigen determination method and lack specificity and sensitivity. An ECPKA auto-antibody detection method for a universal biomarker that detects cancer of various cell types has been reported. The receiver-operating characteristic (ROC) plot showed that autoantibody enzyme immunoassay exhibited 90% sensitivity and 88% specificity, whereas the enzymatic assay exhibited 83% sensitivity and 80% specificity. These results show that the autoantibody method distinguished between patients with cancer and controls better than the antigen method could. Serum biomarker measurement in body fluid immuno assays has been the most widely used approach, generally of established tumor-associated markers as carcinoembryonic antigen (CEA), alpha feto protein (AFP), human chorionic gonadotropin (hCG), prostate specific antigen (PSA), and carcinoembryonic antigen 125 (CA125). These biomarkers have low specificity and sensitivity; therefore, they are either not used in screening and diagnosis or used in combination with other biomarkers. The limitations of the presently available serum tumor biomarkers, based on the antigen determination method, indicate the need for other means of screening.

Plasma and urinary concentrations of two members of the vascular endothelial growth factor (VEGF) family and their receptors as potential response and toxicity biomarkers of bevacizumab with neoadjuvant chemoradiation in patients with localized rectal cancer were evaluated by different groups of investigators.[16] Of all biomarkers, pretreatment plasma sVEGFR-1—an endogenous blocker of VEGF—and PlGF—and a factor linked with vascular normalization—were associated with both primary tumor regression and the development of adverse events after neoadjuvant bevacizumab and chemoradiation. Plasma sVEGFR-1 should be further evaluated to establish as a potential biomarker to stratify patients in future studies of bevacizumab and/or cytotoxics in the neoadjuvant setting.

13.3.4 Imaging Biomarkers and Non-Imaging Biomarkers

Mammography is the process of using low-energy X-rays to examine human breast cancer and is used as a screening and diagnostic tool. Mammography is applied in clinic for breast cancer screening and has helped in reducing breast cancer mortality.[17] We should, however, keep in mind that the radiation exposure associated with mammography has a potential health risk.[18,19]

Imaging is an enabling scientific discipline combining advanced technology and complex computational and analytic methods to provide unique ability to extract spatially and temporally defined information from humans.[20,21] It allows us to investigate intact biological system (without isolating samples or taking biopsies) across the spectrum from sub-cellular to macroscopic and from discovery to clinical decision making. By this technology early breast cancer is detected *via* characteristic masses and/or microcalcification. Thus, mammography is considered as a non-invasive biomarker of cancer diagnosis. For the average woman between the age of 50 to 74 years, mammography is recommended every two years. This helps in avoiding unnecessary surgery, treatment, and anxiety. On a cautionary note, mammography has a false-negative rate of approximately 10% due to dense tissues obscuring the cancer and also due to the fact that the appearance of cancer on the mammogram has a large overlap with the appearance of normal tissues.[22] Quantum dots technology (nanotechnology) is also based on imaging and has been successfully used for cancer diagnosis.[23]

Another technology called positron emission tomography (PET) scan was used to evaluate the treatment response in breast cancer patients.[24] Although reasonable success could be achieved, the main problem with imaging technologies (along with health hazards of multiple exposures) is that tumor heterogeneity interferes with interpretation of results and a combination of other biomarkers and patient related information is needed to infer any clinical value. It is re-emphasized here that multiple biomarkers should be used to achieve high sensitivity and specificity (discussed below in other sections of this article).

13.3.5 Invasive and Non-invasive Biomarkers

Tissues are the best source of material to assay early detection cancer biomarkers because they represent true expression of biomarkers during disease development.[9,10] However, tissue collection is a noninvasive procedure and it is difficult to get healthy tissue for comparison. Therefore, such biomarkers are preferred which can be assayed in samples collected non-invasively.[10] Biofluids (urine, blood, sputum) and exfoliated cells are good examples of noninvasive source of biomarkers for early diagnosis of the disease. After identifying biomarkers, the assay and the biomarker have to be approved by the Food and Drug Administration (FDA) so that these biomarkers can be assayed in clinical samples.[25] The FDA has provided guidelines in this direction (www.fda.gov). If biomarkers, assays or devices are planned for clinical use in patient samples, they should be reviewed by the FDA's Center for Devices and Radiological Health (CDRH) for their ability to analytically measure the biomarker.[26] Biomarkers and devices for quantification are expected to yield equivalent results. Biomarkers should have passed analytical and clinical validation tests specified by the FDA. Analytical validity in this context is defined as the ability of an assay to accurately and reliably measure the analyte in the laboratory as well as in the

clinical sample. Clinical validation requires the detection or prediction of the associated disease in specimens from targeted patient. Biomarker qualification by the FDA enables collaboration among stakeholders, reduces costs for individual stakeholders and provides biomarkers that are useful for the general public and private parties.

Sometimes biomarkers fail to show reasonable sensitivity and specificity when validation is conducted.[27] Although noninvasive biomarkers are the best for breast cancer detection but when these biomarkers are validated in large number of samples, some of them do not show reasonable sensitivity and specificity.[28] The traditional treatment options for breast cancer are radiation, chemicals, and surgery (lumpectomy, quadendroctemy, mastectomy). Surgery is usually combined with adjuvant therapy (hormonal and/or chemical therapy). Chemicals used for therapy have considerable toxicity and hormonal treatment also have long lasting adverse effects. Surviving patients generally have poor quality of life. Furthermore, resistance to chemotherapy is another problem observed in breast cancer patients.[29,30] Pharmacogenomics is an area of research which may provide some useful information in these cases.[31] By applying diagnostic tests and knowing the genetic background of an individual, personalized treatments are possible.

Noninvasive biomarkers may also help in guiding the type of therapy that is more beneficial for patients. For example, in cases of breast cancer, women with ductal carcinoma *in situ* (DCIS), the treatment is by tamoxifen instead of aromatase inhibitor.[32] However, in early invasive stage, as judged by a panel of biomarkers, aromatase inhibitor proved better for treatment than tamoxiphen. The use of hormonal therapy changed with the age of the patient and tumor characteristics. Most of these characteristics correlated better with early stage than DCIS. This research led to the development of prevention strategies. Now endocrine therapy is used for preventing new primary breast cancers and invasive recurrence for women with DCIS or early invasive breast cancer. The dose used was higher at the early stage but decreased with the age of the patient.

Epigenomic biomarkers have enormous potential for clinical implication in cancer diagnosis and prognosis.[3] Because of the availability of genome-wide methylation, histone, and miRNA analysis technologies, and our rapidly accumulating knowledge regarding epigenome, the translation of findings discussed in this article may be possible in near future. Epigenetic biomarkers may also help in identifying patients who will benefit from the therapy and will not develop resistance to drugs and ultimately increase survival.[33] Recently developed drugs for disease treatment are based on specific pathways and may be useful for those individuals where those pathways are altered. This approach can be designed for personalized medicine and precision medicine. Epigenetic biomarkers may help in such approaches. All potential treatment biomarkers for diseases discussed in this chapter have by no means been exhausted, and it is expected that additional high penetrance biomarkers will be identified.

13.4 OMICS APPROACHES IN BIOMARKER DISCOVERY

Omics is an emerging and exciting area in the field of science and medicine.[4,5,9,34] In the post-genome era, efforts are focused on biomarker discovery and the early diagnosis of different diseases through the application of various omics technologies.[3] Numerous promising developments have been elucidated using omics (transcriptomics, proteomics, metabonomics/metabolomics, peptidomics, glycomics, phosphoproteomics or lipidomics on tissue samples and body fluids) in different diseases.[32,34] The development of high-throughput technologies that permit the solution of deciphering a disease from higher dimensionality may provide a knowledge base, which changes the face of the disease biology, etiology, and therapeutics. The omics technology that has driven these new areas of research consists of DNA and protein microarrays, mass spectrometry and a number of other instruments that enable high-throughput analyses at relatively low cost.[1,34] High-throughput omic technologies are being established and validated to create better predictive models for diagnosis, prognosis and therapy, to identify and characterize key signaling networks and to find new targets for drug development.[34]

13.4.1 Immune Diseases

Various applications of omics technology include studying changes in tissue specific protein expression in normal and disease samples, evaluating changes in immune response of the proteins in disease states, translating results from the laboratory to bed side delivery of patient care, and developing an artificial neural network (ANN) to distinguish high risk individuals from normal subjects. Interferons and cytokines were identified as a big group of biomarkers in immune diseases.[35] Interferons belong to three families: I, II, and III, with multiple subtypes. In infections and chronic inflammatory diseases, interferons and their stimulating genes could be utilized to follow up outcome and survival. Multiple sclerosis is an immune disease with the characteristic feature of deregulated T lymphocyte apoptosis.[36] Biomarkers identified during MS development are CASP8AP2, IL-23, CD36, ITGAL, OLR1, RNASEL, RTN4RL2, and THBS1.[36] Selected biomarkers are shown in Table 13.2.

13.4.2 Cardiovascular Diseases (CVDs)

Cardiovascular diseases may arise due to psychosocial stress.[37] CVD biomarkers, adhesion and proinflammatory molecules (IL-6, other cytokines, C-reactive proteins, and fibrinogen), and pathogens were evaluated to assess their contribution in socioeconomic position (SEP) and CVDs.[37] Selected biomarkers are shown in Table 13.3.

13.4.3 Metabolic Diseases, Metabolomics and Proteomics

A multitude of factors should be considered in selecting the specific technology to be adopted for metabolomics studies.[2] Two major technologies,

Table 13.2 Biomarkers and their characteristics in immune diseases.

Disease	Biomarker	Characteristics
Autoimmune disease (rheumatoid arthritis)	Antigen arrays	Can be used for disease stratification and planning treatment strategies.[73]
Autoimmune disease (rheumatoid arthritis, systemic and cutaneous lupus erythematosus, SLE)	Annexin-I	Annexin-1 exerts its anti-inflammatory effect by suppressing the generation of inflammatory mediators and anti-annexin I antibodies are present in rheumatoid arthritis and SLE patients.[74]
Autoimmune disease (systemic sclerosis, systemic and cutaneous lupus erythematosus, SLE), polymyositis (PM), dermatomyositis (DM), multiple sclerosis (MS)	Prolactin (PRL), ferritin, vitamin D, tumor marker tissue polypeptide antigen (TPA)	High levels of PRL, ferritin, vitamin D, and TPA were observed in SLE, PM, DM, and MS although levels of these markers were different in different diseases.[76]
Rheumatoid arthritis (RA)	IL-17 and Wnt/beta katenin	In rheumatoid arthritis (RA) bone loss occurs.[99] Transcriptomics approaches were applied to identify genes and pathways that contributed to bone loss in RA patients. Results indicated IL-17 and Wnt/beta katenin as transcriptional biomarkers of RA.[100]
Multiple sclerosis (MS)	CASP8AP2, IL-23, CD36, ITGAL, OLR1, RNASEL, RTN4RL2, and THBS1	In infections and chronic inflammatory diseases, interferons and their stimulating genes can be utilized to follow up outcome and survival. Multiple sclerosis is an immune disease with characteristic feature of deregulated T lymphocyte apoptosis. Biomarkers identified during MS development were CASP8AP2, IL-23, CD36, ITGAL, OLR1, RNASEL, RTN4RL2, and THBS1.[36]
Asthma	Nitric oxide	Measurements of fractional excretion of nitric oxide can be used as a biomarker for diagnosing asthma.[161]
Rheumatoid arthritis (RA)	IL-17 and Wnt/beta katenin	In rheumatoid arthritis (RA) bone loss occurs. Transcriptomics approaches were applied to identify genes and pathways, which contributed to bone loss in RA patients. Results indicated IL-17 and Wnt/beta katenin as transcriptional biomarkers of RA.[100]

Autism	Increased number of mast cells and serotonin in urine	High number of mast cells and serotonin in urine are observed in autism patients. Mast cells produce inflammatory cytokines in large amounts in autism patients and the possibility of autoimmunity has been proposed.[110]
Multiple sclerosis (MS)	HLA DRB1-1*5 haplotype, SHP-1 hypermethylation	MS is an inflammatory disease in which the fatty myelin sheaths around the axons of the brain and spinal cord are damaged. Female specific MS is associated with HLA DRB1-1*5 haplotype.[162] SHP-1 hypermathylation is a biomarker for multiple sclerosis.[144]
Asthma	Arginine kinase, sarcoplasmic calcium binding proteins, and tropomycin	Proteomics approaches identified allergenic proteins, based on their reactivity to patient's sera, using tandem mass spectrometry.[160] The most significant allergens identified were arginine kinase, sarcoplasmic calcium binding proteins, and tropomycin that can be used for screening.
Rheumatoid arthritis (RA)	Aggrecan fragments, C-propeptide of type II collagen	In rheumatoid arthritis (RA) inflammation of joint synovium is persistent, which leads to erosion of cartilage and bone. A biomarker, aggrecan fragments, was identified which was detected in synovial fluid (SF). Levels of aggregant fragments (structural components of cartilage) were higher in RA patients than in healthy individuals. Another biomarker was C-propeptide of type II collagen and its levels were directly proportional to the rate of collagen synthesis. Cartilage oligomeric matrix protein (COMP) is also a component of cartilage and it can be measured in serum and SF.
Multiple sclerosis (MS)	Neurofilament light protein (NFL), Glial fibrillary acidic protein (GFAP)	MS is associated with autoimmune-mediated inflammation of the central nervous system and may lead to demyelination and axonal damage. Biomarker neurofilament light protein (NFL) is a cytoskeleton component in large myelinated axon and it is released into the cerebrospinal fluid (CSF) and can be used to determine the level of axonal damage. Glial fibrillary acidic protein (GFAP) is another biomarker for MS.

Table 13.2 (*Continued*)

Disease	Biomarker	Characteristics
Rheumatoid arthritis and systemic and cutaneous lupus erythematosus (SLE)	Annexin I, II and V	Annexin-I exerted its anti-inflammatory effect by suppressing the generation of inflammatory mediators and anti-annexin I antibodies were present in rheumatoid arthritis and systemic and cutaneous lupus erythematosus (SLE) patients.[74] Annexin II and V were involved in coagulation cascade because they had affinity towards phospholipids.
Autoimmune disease (systemic sclerosis, systemic and cutaneous lupus erythematosus, SLE), polymyositis (PM), dermatomyositis (DM), multiple sclerosis (MS)	Prolactin (PRL), ferritin, vitamin D	Prolactin (PRL), ferritin, vitamin D, tumor biomarker tissue polypeptide antigen (TPA) levels were higher in patients with autoimmune disease (systemic sclerosis, systemic and cutaneous lupus erythematosus, SLE), polymyositis (PM), dermatomyositis (DM), multiple sclerosis (MS).[76] Granin family is a group of acidic proteins present in the secretory granules of a wide variety of endocrine, neuronal, and neuroendocrine cells. Few important granins, chromogranin A and B, secretogrannin II, HISL-19 antigen, NESP55, ProSAAS should be studied further for their clinical implication.[77]
Autism	Serotonin, inflammatory cytokines	High number of mast cells and serotonin in urine were observed in autism patients. Mast cells produce inflammatory cytokines in large amount in autism patients and the possibility of autoimmunity was proposed.[110]

Asthma	Glutathione S transferase M1 (GSTM1)	In one meta-analysis glutathione S transferase M1 (GSTM1) biomarker was identified which helped in screening a cohort of children who were at high risk of developing asthma.[163] In a population-based study C-reactive protein, fibrinogen, and interleukin-6 were found useful biomarkers for screening.[164]
RA	CTX-1, CTX-II, C-telopeptide I and II	Biomarkers of bone and cartilage turnover are collagen C-telopeptides I and II, which are predictors of structural damage in RA patients. C-terminal cross-linked telopeptide of type I collagen (CTX-I) could be measured in serum or urine after it was released during bone resorption. Bone erosions and osteoporosis both occurred as a consequence of RA and CTX-I levels could be used for prognosis. On the other hand, increased CTX-II levels were associated with rapid progression of joint damage. For clinical implication, investigators of these studies recommended that new prognostic biomarkers should supply information beyond that provided by risk factors in the prediction of disease outcome.[165]

Note: For the same disease more than one row are mentioned because biomarkers shown in column 2 were identified in different studies.

Table 13.3 Biomarkers and Their Characteristics in Cardiovascular Disease (CVD).

Disease	Biomarker	Characteristics
Cardiovascular disease (CVD)	C-reactive protein	In a risk prediction model of cardiovascular disease (CVD) C-reactive protein levels were implemented.[8]
Cardiovascular disease (CVD)	Blood pressure	Biomarkers studies of CVD prediction in elderly indicate that mortality and cardiovascular events are dependent on low peripheral pulse pressure not on high blood pressure.[166]
Cardiovascular disease (CVD)	Lipoprotein associated phospholipase A2 and antiphosphorylcholine IgM	In cardiovascular disease (CVD), the relative contribution of genetic and environmental effect was studied on two inflammatory biomarkers, lipoprotein associated phospholipase A2 and antiphosphorylcholine IgM, in a Swedish population.[167]
Cardiovascular disease (CVD)	TNF alpha, IL-6, plasma vitamin E concentrations, total and LDL cholesterol, and antioxidant profiles	Biomarkers discussed in the diagnostic section can be used for follow up of the treatment. Dietary intervention of CVD by fish oil (salmon, herring, and pompus) and other nutrients has been demonstrated in a number of studies. Some of the participants had higher levels of triacylglycerolaemia. Biomarkers TNF alpha and IL-6 were reduced and level of adiponectin increased in the treated arm. Thus TNF alpha, IL-6, and adiponectin were used as therapeutic biomarkers. In another study, argon oil supplement reduced plasma levels of lipids and antioxidant status. Therapeutic biomarkers used in this study were plasma vitamin E concentrations, total and LDL cholesterol, and antioxidant profiles.[168]
Cardiovascular disease (CVD)	Atherosclerosis	A very well characterized biomarker of CVD.
Cardiovascular disease (CVD)	IL-6, other cytokines, C-reactive proteins, and fibrinogen	CVD biomarkers, adhesion and proinflammatory molecules (IL-6, other cytokines, C-reactive proteins, and fibrinogen), and pathogens were evaluated to assess their contribution in socioeconomic position (SEP) and CVD.[37]
Atherosclerosis	Osteoprotegerin (OPG)	Osteoprotegerin (OPG) is such a marker which is independently associated not only with risk factors of atherosclerosis but also with subclinical peripheral atherosclerosis and clinical atherosclerosis and is recommended as a prognostic biomarker for ischemic heart disease and ischemic stroke.[78]

Hypertension	Methylated ADD1 gene	In hypertension, global methylation profiling and individual gene methylation status (ADD1 gene) have been used for diagnosis and outcome.[146]
Atherosclerosis	Hypermethylation of monoamine oxidae A (MAOA)	Atherosclerosis can be diagnosed when other biomarkers are combined with epigenetic biomarkers, such as hypermethylation of monoamine oxide A (MAOA).[145]
Atherosclerosis (early diagnosis)	Carotid intima media thickness (IMT), circulating oxidized low density lipoprotein (LDL), and flow-mediated dialation (FMD)	For early diagnosis of carotid atherosclerosis for which obesity is the risk factor, the early biomarkers are carotid intima media thickness (IMT) and circulating oxidized low-density lipoprotein.[169] To evaluate endothelium status, flow-mediated dialation (FMD) and IMT have been used as early markers of atherosclerosis in patients with nonalcoholic fatty liver disease (NAFLD).[170]
Atherosclerosis	CARD8, Ephs, ephrins	Atherosclerosis is a chronic inflammatory disease of the vessel wall. The gene CARD8, which codes proteins involved in innate immunity in atherosclerosis patients, is an excellent biomarker for atherosclerosis. Inflammatory markers also over expressed in this disease. Inflammosomes produce interleukin 1 beta in response to cholesterol crystal accumulation in macrophages. Ephs and ephrins also were proposed as biomarkers of atherosclerosis.[108]
Atherosclerosis	Hypermethylation of monoamine oxidae A (MAOA) and ADD1	Atherosclerosis can be diagnosed when other biomarkers are combined with epigenetic biomarkers, such as hypermethylation of monoamine oxidae A (MAOA). Along with gene-specific methylation biomarkers, global DNA methylation biomarkers were identified for atherosclerosis.[145] Similarly in other CVDs, such as hypertension, global methylation profiling and individual gene methylation status (ADD1 gene) were used for diagnosis and outcome.[146,147]

Note: For the same disease more than one row is mentioned because biomarkers shown in column 2 were identified in different studies.

mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR), are generally considered as they can measure hundreds to thousands of unique chemical entities. MS is highly sensitive and has the capacity to detect metabolites with concentrations in the picomole range and above, requires small biospecimen volumes, enables metabolites to be individually identified and quantified, and is well-suitable for use in a high-throughput mode. However, MS requires expensive consumables, has relatively lower analytical reproducibility, poorly represents highly polar metabolites when using standard chromatography protocols, and requires more complex software and algorithms for routine data analysis.[2] In contrast, NMR allows for the comprehensive generation of metabolite profiles by a single nondestructive method, is fully automated with high-throughput capacity, inherently quantitative, and highly suitable for metabolite structure elucidation, and has very high analytical reproducibility with a well-established mathematical and statistical toolbox. The disadvantages to using NMR include its relative insensitivity in detecting metabolites with concentrations in the micromole range and below. Furthermore, its validity is dependent on the quality of sample collection and handling, as well as the available metadata. Before applying either of these technologies to population-based studies, investigators must consider the advantages and disadvantages in relation to the design and aims of the study.

Quantitative proteomics can be used for the identification of disease biomarkers that could be used for early detection, serve as therapeutic targets, or monitor response to treatment. Several quantitative proteomics tools are currently available to study differential expression of proteins in a variety of types of samples. Two-dimensional gel electrophoresis (2-DE), which was classically used for proteomic profiling, has been coupled to fluorescence labeling for differential proteomics. Isotope labeling methods such as stable isotope labeling with amino acids in cell culture (SILAC), isotope-coded affinity tagging (ICAT), isobaric tags for relative and absolute quantitation (iTRAQ), and ^{18}O labeling have all been used in quantitative approaches for identification of biomarkers. In addition, heavy isotope labeled peptides can be used to obtain absolute quantitative data. Label-free methods for quantitative proteomics, which have the potential of replacing isotope-labeling strategies, are becoming popular. Other emerging technologies such as protein microarrays have the potential for providing additional opportunities for biomarker identification. Selected biomarkers are shown in Table 13.4.

13.4.4 Infectious Diseases

Bacterial infection occurs in pneumonia. Community acquired pneumonia is very common and for proper diagnosis and treatment the other biomarker which showed promise was procalcitonin.[38] C-reactive protein and inflammatory biomarkers could be used in combination with procalcitonin to make intelligent clinical decisions. Selected biomarkers are shown in Table 13.5.

Table 13.4 Biomarkers and their characteristics in metabolic diseases.

Disease	Biomarker	Characteristics
Diabetes	Alanine amino transferase (ALT), gamma glutamyl transferase (GGT), triglyceride, plasminogen activator inhibitor (PAI-1) antigen, ferritin, C-reactive protein (CRP), sex-hormone binding globulin (SHBG)	Biomarkers of risk for diabetes.
Diabetes	IL-6	Inflammation marker of diabetes (genetic studies on association of IL-6 with diabetes have not been completed yet).
Diabetes	Adiponectin	Some groups of investigators consider adiponectin as an excellent biomarker of diabetes pathogenesis.[171]
Diabetes	Insulin resistance (IR) and blood glucose levels	The most studies biomarkers which are used in clinic routinely.
Diabetes	Hemoglobin A1c	For the management of diabetes hemoglobin A1c (HbA1c) is used which is considered as a reliable indicator of glycemic control. In most of the clinical studies in diabetes, HBA1c biomarker is used to determine the glucose control.
Diabetes	1,5-anhydroglucitol (1,5 A)	Circulating biomarker used to measure hyperglycemic condition
Metabolic Diseases	Alpha-hydroxybutyrate (αHB) and linolyl-glycerophosphocholine (L-GPC)	Alpha-hydroxybutyrate (αHB) and linolyl-glycerophosphocholine (L-GPC) were identified as biomarkers of insulin resistance and glucose intolerance in a large population study of more than 1000 participants from the Relationship between Insulin sensitivity and Cardiovascular Disease (RISC) study.[2]
Chronic kidney disease	Proteomic profiling	Differentially expressed peaks in the spectra are future biomarkers for chronic kidney disease.[172]
Diabetes	Proteomic profiling	Isobaric tags for relative and absolute quantitation (iTRAQ) in combination with two-dimensional gel electrophoresis technologies identified a group of proteins, which were associated with diabetes, pancreatitis, and/or pancreatic cancer.[80]

Table 13.4 (*Continued*)

Disease	Biomarker	Characteristics
Diabetes	RCAN1, perilipin A and G0/G1 switch gene, G0S2	Glucose response gene, RCAN1, was over expressed whereas perilipin A and G0/G1 switch gene, G0S2, were under expressed in diabetes-2.[102,103]
Diabetes (early diagnosis)	Dysglycemia, alphahydroxy-butyrate, linoleyolglycerophos-phocholine, advanced glycation end products (AGE), albumin excretion rate (AR) and SNPS in epidermal growth factor gene intron 2	A number of markers have been described for early diagnosis of diabetes. Few of them are dysglycemia, alphahydroxy-butyrate, linoleyolglycerophos-phocholine, advanced glycation end products (AGE), albumin excretion rate (AR) and SNPS in epidermal growth factor gene intron 2.[2,173,174]

Note: For the same disease more than one row are mentioned because biomarkers shown in column 2 were identified in different studies.

13.4.5 Neurological Diseases

For the diagnosis of seizures, carbohydrate deficient transferrin (CDT) was evaluated and results indicated that this biomarker alone could not diagnose the disease but contribute in better diagnosis when combined with other biomarkers such as GGT, ASAT, ALAT, and ASAT/ALAT ratio.[39] A novel missense mutation at position 134 T to A resulting in amino acid change at codon V45E was identified as a biomarker for Norrie disease (ND) which is a rare X-linked disorder characterized by congenital blindness and sometime mental retardation, and deafness.[40] A novel mutation in ATP7B gene was used as a diagnostic biomarker for neurological impairment in Wilson's disease.[41] Selected biomarkers are shown in Table 13.6.

13.4.6 Cancer

Proteomic profiling of samples from cancer and healthy individuals has identified disease-associated profile, especially in Matrix Associated Laser Ionization/Desorption Time-of-Flight Mass Spectrometry (MALDI TOF MS) in colorectal and other cancers which can be used for diagnosis either alone or in combination with other biomarkers.[42] Methylation biomarkers were useful in diagnosis of almost all major cancers.[7,9,43] Defects in the mismatch repair (MMR) genes such as MLH1, MSH2, or MSH6 or methylation of the MLH1 promoter led to erroneous replication of segments of simple nucleotide repeats which contributed to microsatellite instability (MSI). MSI increases the risk of cancer occurrence. However, MSI is uncommon in cancers of the breast as compared to some other cancers such as colorectal carcinoma (CRC). It was observed that prognosis in MSI-positive breast cancer patients was worse than that of patients with MSI-negative tumors. Selected biomarkers are shown in Table 13.7.

Table 13.5 Biomarkers and their characteristics in infectious diseases.

Disease	Biomarker	Characteristics
Infectious diarrhea	Calprotectin	In cases of infectious diarrheal, fecal calprotectin is a good prediction marker.
Fungal infection	1,3-beta-D-glucan (BG)	1,3-beta-D-glucan (BG) can be used as a biomarker in invasive fungal infections (especially infections involving *Candida*) in patients undergoing treatment of candedemia with anidulafungin
Community acquired pneumonia (CAP)	Procalcitonin, C-reactive protein, inflammatory cytokines, and bacterial infection	Bacterial infection occurs in pneumonia. Community acquired pneumonia (CAP) is very common and for proper diagnosis and treatment the other biomarker which showed promise was procalcitonin.[38] C-reactive protein and inflammatory biomarkers can be used in combination with procalcitonin to make intelligent clinical decisions.
Ulcer diseases and gastritis (and gastric cancer)	*H. pylori*	*H. pylori* is the most common chronic bacterial infection in humans. This bacteria is involved not only in gastric cancer but in ulcer diseases and gastritis also. Its synergistic gastrotoxic interaction with non-steroidal anti-inflammatory drugs, and association with atherosclerotic events is a matter of concern.
Diarrhea in AIDS patients	Tubuloreticulin inclusions (TRIs)	Diarrhea in AIDS patients was treated with specific medications and therapeutic response was measured by levels of tubuloreticulin inclusions (TRIs).
Pneumonia	Procalcitonin (PCT)	For pneumonia therapy on more than 100 patients, biomarker procalcitonin (PCT) levels were very useful and this biomarker has been recommended for future prognosis. Sometime detecting bacteria alone is not sufficient to design therapy.[175]
HIV/AIDS	CRC5-delta32 mutation	In one study conducted in Georgia, where the prevalence of HIV/AIDS is high, polymorphism of CCR5 gene was studied.[50] More than 100 subjects were enrolled in this study. Results identified CRC5-delta32 mutation as a marker of the disease in this population.

Table 13.5 (*Continued*)

Disease	Biomarker	Characteristics
HIV/AIDS	Mitochondrial DNA mutations	Mitochondrial DNA mutations are also biomarkers for AIDS, especially in those individuals who are undergoing therapy.[51]
Liver cancer (HCC)	p16 and c-myc hypermethylation	In early liver cancer, where risk factors are infection by HBV and HCV, clustered DNA methylation changes in polycomb repressor target genes were observed. P16 and c-myc hypermethylation also suggested initiation of hepatocellular carcinoma.[127,128]
Tuberculosis	Vitamin D receptor (VDR) methylation	In tuberculosis, vitamin D receptor (VDR) methylation indicated high risk of developing the disease.[149] VDR gene encodes a transcription factor that alters calcium homeostasis and immune function.
Tuberculosis (early diagnosis)	Region-of-Difference-1 (RD-1) gene product and sputum cytokine levels	Region-of-Difference-1 (RD-1) gene product and sputum cytokine levels are considered a biomarker for early detection of tuberculosis.[176,177]
Hepatocellular carcinoma (early diagnosis)	HSP70, CAP2, glypican 3 and glutamine synthase	Hepatocellular carcinoma (HCC) involves infectious agents and the early diagnostic markers for HCC are HSP70, CAP2, glypican 3 and glutamine synthase.[178] These markers were identified based on gene expression analysis of HCC samples.
Hepatocellular carcinoma (HCC)	HBV, HCV, methylation of polycomb repressor, hypermethylation of p16 and c-myc,	In early liver cancer, where risk factors are infection by HBV and HCV, clustered DNA methylation changes in polycomb repressor target genes were observed. P16 and c-myc hypermethylation also suggested initiation of hepatocellular carcinoma.[127,128]
Tuberculosis	Vitamin D receptor (VDR) methylation	In tuberculosis, vitamin D receptor (VDR) methylation indicated high risk of developing the disease.[149] VDR gene encodes a transcription factor, which alters calcium homeostasis and immune function.

Note: For the same disease more than one row are mentioned because biomarkers shown in column 2 were identified in different studies.

Table 13.6 Biomarkers and their characteristics in neurological diseases.

Disease	Biomarker	Characteristics
Neurological disorders	Granins	The granin family is a group of acidic proteins present in the secretory granules of a wide variety of endocrine, neuronal, and neuroendocrine cells. A few important granins, chromogranin A and B, secretogrannin II, HISL-19 antigen, NESP55, ProSAAS should be studied further for their clinical implication.[77]
Neurological disorders	Microvesicles	Microvesicles (MVs) are used as biomarkers of neurological disorders because their release is increased in these diseases.[83]
Neurological disorders (Alzheimer's disease)	Glutamate	In Alzheimer's disease, glutamate levels increased (in the hippocampus region of the brain) after the treatment with galantamine.[13] Therefore, glutamate can be considered a therapeutic biomarker of Alzheimer's disease.[14]
Neurological disorders (seizures)	Carbohydrate deficient transferrin (CDT)	For the diagnosis of seizures, carbohydrate deficient transferrin (CDT) was evaluated and results indicated that this marker alone couldn't diagnose the disease but contribute in better diagnosis when combined with other biomarkers such as GGT, ASAT, ALAT, and ASAT/ALAT ratio.[39]
Norrie disease (ND) (X-linked disorder)	Missense mutation at position 134 T to A	A novel missense mutation at position 134 T to A resulting in amino acid change at codon V45E was identified as a biomarker for Norrie disease (ND) which is a rare X-linked disorder characterized by congenital blindness and sometime mental retardation, and deafness.[40]
Wilson's disease	Mutation in ATP7B gene	A novel mutation in ATP7B gene can be used as a diagnostic marker for neurological impairment in Wilson's disease.[41]
Alzheimer's disease	CXR4 and CCR3	Genes involved in inflammation and immune system regulatory pathways were identified when profiling of Alzheimer and non-Alzheimer's samples was conducted.[116] Results also indicated a role for chemokines and their receptors (CXR4 and CCR3) in patient samples and these two receptors may be considered biomarkers for the Alzheimer's disease.
Alzheimer's disease	Altered methylation of Alu, Line-1, and SAT-alpha sequences	In Alzheimer disease altered methylation levels of repeat sequences (Alu, Line-1, and SAT-alpha) were observed. This modification induces genomic instability, which contributes to disease initiation and progression.[150]
Autism	MECP2 hypermethylation	Epigenetic regulation in autism was also studied and locus specific hypermethylation of MECP2 was observed.[151]

Note: For the same disease more than one row is mentioned because biomarkers shown in column 2 were identified in different studies.

Table 13.7 Biomarkers and Their Characteristics in Cancer.

Disease	Biomarker	Characteristics
Apoptosis (cancer)	Annexin XI	Anti-annexin XI has been reported in different cancers.[74]
Oral cancer	Mutation in exon 4, codon 63 of the p53 gene	Antibodies against abnormal p53 were found in saliva and serum of oral cancer patients.[53]
Gastric cancer	Metabolites of glycolysis, fatty-acid beta oxidation, and cholesterol and amino acid metabolism	In gastric cancer, alterations in metabolites of glycolysis, fatty-acid beta-oxidation, and cholesterol and amino acid metabolism were observed.[122] These metabolites can be used to follow gastric cancer development and treatment response.
Pancreatic cancer	Serum levels of antibodies against periodontal bacteria *P. gingivalis*	In pancreatic cancer the incidence and mortality rates are the same and the survival is only five years after the diagnosis of the disease. Plasma and serum of pancreatic cancer contains antibodies against periodontal bacteria *P. gingivalia* and their level is associated with pancreatic cancer.[84]
Pancreatic cancer	C-reactive proteins, interleukin-6 (IL6), and soluble receptor of tumor necrosis factor alpha	Inflammatory biomarkers, C-reactive proteins, interleukin-6 (IL6), and soluble receptor of tumor necrosis factor alpha are being used in identifying pancreatic cancer patients.[85]
Ovarian cancer	CA 125, Osteopontin, Kllikrein 6, B7-H4, spondin 2, and DcR3	B7-H4, spondin 2, and DcR3 were identified as early biomarkers in ovarian cancer.[179]
Prostate cancer	Prostate-specific antigen, Alpha methyl CoA-racemase	Results have not been validated for Alpha methyl CoA-racemase.[180]
Colon cancer	APC, CDKNA	Expression analysis was used to identify these biomarkers.[181]
Lung cancer	EGFR, KRAS	Multiple investigators have identified these biomarkers.[182,183]
Breast cancer	BRCA-1, BRCA-2, Let-7	These biomarkers have been used in clinical samples.[184]
Acute myeloid leukemia, Chemotherapy resistant AML, Acute lymphoblastic leukemia, Acute non-promyelocytic leukemia	FLT3, DAPK1, hPer3, DNMT3A, repeat sequences LINE-1	Several investigators have identified these markers. Several methylation markers of AML have also been reported.[185–187]
Renal cell carcinoma (Kidney cancer)	APAF-1, DAPK-1	Hypermathylation of APAF-1 and DAPK-1
Breast cancer	SNAPs in in 1p11.2, 2q35, 3p, 5p12, 8q24, 10q23, 13, 14q24.1, and 16q regions	In GWAS study, SNPs identified in this study were primary located in 1p11.2, 2q35, 3p, 5p12, 8q24, 10q23, 13, 14q24.1, and 16q regions.

Pancreatic cancer	*P. gingivalia*	Plasma and serum of pancreatic cancer contained antibodies against periodontal bacteria *P. gingivalia* and their level is associated with pancreatic cancer.[84]
Pancreatic cancer	Inflammatory biomarkers, C-reactive proteins, interleukin-6 (IL6), and soluble receptor of tumor necrosis factor alpha, cell cycle regulatory proteins (cyclin D1 and Ki67), glycolytic enzyme lactate dehydrogenase (LDH), matrix metalloproteinases (MMPs)	Inflammatory biomarkers, C-reactive proteins, interleukin-6 (IL6), and soluble receptor of tumor necrosis factor alpha also were used in identifying pancreatic cancer patients.[85] Some biomarkers identified include increased salivary levels of cell cycle regulatory proteins (cyclin D1 and Ki67), glycolytic enzyme lactate dehydrogenase (LDH), matrix metalloproteinases (MMPs) and reduction in DNA repair enzyme (8-oxoquanine DNA glycosylase) and mapsin in oral cancer patients.[86]
Hepatocellular carcinoma (HCC)	hTERT, alpha-fetoprotein (AFP), des-gamma-carboxy prothrombin (DCP)	A highly sensitive method for serum human telomerase reverse transcriptase (hTERT) mRNA for hepatocellular carcinoma (HCC) was reported.[87] Alpha-fetoprotein (AFP) and des-gamma-carboxy prothrombin (DCP) were found to be good markers for HCC.
Retinoblastoma	Trimethylation of H4K20	Retinoblastoma levels were lower whenever trimethylation of H4K20 was present. A correlation with the tumor stage and grade was also established based on these histone biomarkers.[188]
Breast cancer	HDAC1	Another interesting study reported quantitative expression of HDAC1 and its correlation with breast cancer patient's age, lymph node status, tumor size and her2/neu negative, ER and PR positive status.[146]
Kidney cancer	VHL, MET, FLCN, fumarate hydratase, succinate dehydrogenase, TSC1, TSC2, TFE3	Altered expression of VHL, MET, FLCN, fumarate hydratase, succinate dehydrogenase, TSC1, TSC2, and TFE3 genes in kidney cancer was observed.[189]
Lung cancer	IL-27, IL-29, IL-31, BALF	Serum levels of IL-27, IL-29, IL-31, BALF are high in lung cancer patients.[10]
Lung cancer	miR-21, miR-485, miR-451	Higher levels of miR-21 and lower levels of miR-451 and miR-485 were observed in lung cancer.[11]
Lung cancer	CYFRA 21-1	Increased CYFRA 21-1 levels were observed in samples from lung cancer patients.[190]

Note: For the same disease more than one row are mentioned because biomarkers shown in column 2 were identified in different studies.

13.5 TECHNOLOGIES AND STRATEGIES FOR MOLECULAR BIOMARKER DISCOVERY

Technologies and strategies developed for cancer diagnosis can also be applied for other diseases.[4–6] The detection and treatment of cancer is greatly facilitated by omics technologies. For example, genomics analysis provides clues for gene regulation and gene knockdown for cancer management. The approval of Mammaprint and Oncotype DX indicates that multiplex diagnostic biomarker sets are becoming feasible. The microRNA field in human cancers has opened a new avenue for cancer researchers. Some therapeutic drugs targeting DNA methylation and histone deacetylation are currently undergoing keen studies. Proteomics also plays an important role in cancer biomarker discovery and quantitative proteome-disease relationships provide a mean for connectivity analysis. Fluorescent dye enables a more reliable and quantitative analysis and is facilitated by the progress of biochip and cytomics. The huge amount of information collected by multiparameter single cell flow or slide-based cytometry measurements serves to investigate the molecular behavior of cancer cell populations. Metabolite profiling is such a field which can be applied to cancer and other diseases.

13.5.1 Genomics Based Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

Discovery and validation of novel disease-associated biomarkers remain a crucial goal of future patient care. Advanced genomic technologies, such as SNP array and next generation sequencing, help shape the genome and epigenome landscapes. Genome wide association studies (GWAS) as a powerful approach to identify common, low penetrance disease loci have been conducted in several types of diseases such as diabetes and cancer and have identified many novel associated loci, confirming that susceptibility to these diseases is polygenic. Though the creation of risk profiles from combinations of susceptible SNPs is not yet clinically applicable, future, large scale GWAS holds great promise for the individualized screening and prevention. Epigenomic biomarkers like DNA methylation have emerged as highly promising biomarkers and are actively studied in multiple diseases. Validated as being associated with disease risk or drug response, some DNA methylation biomarkers are being transferred into clinical use.[34] Discovery of the genes and pathways mutated in diseased states, especially through large scale genome wide sequencing, has provided key insights into the mechanisms underlying disease process and suggested new candidate biomarkers for diagnosis, clinical intervention as well as prognosis. The comprehensive landscapes of cancer genome point out the convergence of mutations onto pathways that govern the course of disease development and indicate that rather than seeking genomics and epigenomics alterations of specific mutated genes, the combination with dynamic transcriptomics,

proteomics and metabonomics of the downstream mediators or key nodal points may be preferable for future disease biomarker discovery.

Most of the genomic biomarkers are mutations in genes or small nucleotide polymorphisms (generally present in the noncoding region).

13.5.1.1 Immune Diseases. Among immune diseases, autoimmune diseases are well characterized. These diseases result when the immune system goes awry and recognizes self-tissues as foreign. The major contribution to these diseases is by autoantibodies and autoreactive cellular responses, which ultimately contribute to the ongoing autoimmune disease process. During the process, inflammatory enzymes are recruited to the affected organ and tissues degrading proteolytic enzymes are released. Therefore, identifying biomarkers that may detect the process early is very significant. An ideal biomarker for these diseases should reach abnormal levels (either higher or lower levels compared to control) in conjunction with disease development, should fluctuate in relation to disease severity and should normalize after treatment. Three groups of biomarkers are being characterized in these diseases. First, degradation products arising from destruction of the affected tissue; second, enzymes that plays a role in tissue degradation, and; third, cytokines and other proteins associated with immune system activation and the inflammatory response. In rheumatoid arthritis (RA), persistent inflammation of joint synovium occurs, which leads to erosion of cartilage and bone. A biomarker, aggrecan fragments, was identified which was detected in synovial fluid (SF). Levels of aggregant fragments (structural components of cartilage) were higher in RA patients than in healthy individuals. Another biomarker was C-propeptide of type II collagen and its levels were directly proportional to the rate of collagen synthesis. Cartilage oligomeric matrix protein (COMP) is also a component of cartilage and it can be measured in serum and SF. Multiple sclerosis (MS) is associated with autoimmune-mediated inflammation of the central nervous system and may lead to demyelination and axonal damage. Biomarker neurofilament light protein (NFL) is a cytoskeleton component in large myelinated axon and it is released into the cerebrospinal fluid (CSF) and can be used to determine the level of axonal damage. Glial fibrillary acidic protein (GFAP) is another biomarker for MS.

A genetic link with HLA-DR4 and related allotypes of MHC class II and T cell associated protein PTPN22 with rheumatoid arthritis was established.[44] The presence of autoantibodies to IgGFc (rheumatoid factor) and antibodies to citrullinated peptide (ACPA) also indicated the presence of disease.

In rheumatoid arthritis, SNPS were reported in several genes which were used as biomarkers for disease diagnosis. These genes include PTPN22, IL23R, TRAF1, CTLA4, IRF5, STAT4, CCR6, and PAD14.[45]

13.5.1.2 Cardiovascular Diseases (CVDs). Pulmonary arterial hypertension (PAH) is influenced by genetic background and may be useful in guiding

therapy of the disease.[46] In this disease, increase in blood pressure in the pulmonary artery and pulmonary vein occurs which leads to shortness of breath, dizziness and fainting. GWAS studies indicated that other factors (rare exonic mutations, epigenetic phenomena, and interaction with environmental factors) might also contribute in the development of this disease.[47]

13.5.1.3 Metabolic Diseases, Metabolomics and Proteomics. Like other omics approaches, metabolomics also takes the advantage of non-targeted approach for identifying disease associated biomarkers. The technology has the advantage of using minimum amount of sample and no prior knowledge of the substances to be analyzed by nuclear magnetic resonance (NMR) or mass spectrometry (MS).[34] The complete process includes the acquisition of the experimental data, the multivariate statistical analysis, and the projection of the profiles, which is the acquired information, to construct the patient map (or phenotype). Main diseases where metabolomics has been applied are rheumatoid arthritis, spondyloarthritis, systemic lupus erythrematosus, and osteoarthritis.[48] In spondyloarthritis, association of HLA-A, B, and HLA-DR gene expression was studied and population specific alterations were reported.[49]

13.5.1.4 Infectious Diseases. Regarding genomic biomarkers in infectious diseases, we have selected AIDS (although the topic is so big that we cannot cover in this article). In one study conducted in Georgia, where the prevalence of HIV/AIDS is high, polymorphism of CCR5 gene was studied.[50] More than 100 subjects were enrolled in this study. Results identified CRC5-delta32 mutation as a biomarker of the disease in this population. Mitochondrial mutations could also be used as biomarkers for AIDS, especially in those individuals undergoing therapy.[51]

13.5.1.5 Neurologic Diseases. Because of the anatomical location of the nervous system, it is difficult to get biomarkers of the neurological diseases. The accessibility of the affected organ is a challenge and therefore surrogate biomarkers are generally used to identify and follow these diseases. Gene expression analysis, mutations, and SNPs were the common approaches to identify neurological diseases. Compared to healthy subjects, altered gene expression profiling was observed in schizophrenia and Alzheimer's patient samples.[52]

13.5.1.6 Cancer. Here we describe the specific example of breast cancer with reference to genomic approaches in population science for screening, which may result in early detection of the disease and ultimately long survival. Mortality from breast cancer is very high worldwide. More than half of breast cancer cases occur in Western countries. The cost of treatment is

higher when breast cancer is detected late in its development; therefore, detecting this cancer early is the key to success. Mammography has been successful in reducing mortality from this cancer, but it is an expensive technique. The occurrence of breast cancer in the general population can be explained by inherited genetic susceptibility, somatic changes, effects of endogenous and exogenous environments, and interaction of these factors (especially gene–environment interactions). In case of oral cancer, exon 4, codon 63 of the p53 gene was mutated in salivary DNA in patients.[53] Autoantibodies against abnormal p53 were reported in saliva and serum of these patients.

Inherited genes for breast cancer susceptibility can be low- or high-penetrance genes; the few genes with allelic variants that confer a high degree of risk to the individual are known as high-penetrance genes. Other genes confer a small to moderate degree of breast cancer risk to the individual and are known as low-penetrance genes. Relatively few individuals in the population carry risk-increasing genotypes at the loci where high-penetrance genes act; therefore, the population-attributable risk is low. On the other hand, the low-penetrance genes are not associated with syndromic or Mendelian patterns but are associated with sporadic breast cancer. The allelic variation of low-penetrance genes is relatively high, and large breast cancer populations carry low-penetrance genes. Different investigators identified low- and high-penetrance genes in breast cancer in a number of populations. The BRCA1 gene was the first gene identified to represent susceptibility to hereditary breast cancer and later on BRCA2 (located on 17q21) was also confirmed for breast cancer and ovarian cancer.[54] To identify breast cancer associated genetic biomarkers a number of cohorts with exposure and lifestyle data and other details of the participants were used.[55] One of those cohorts, the Collaborative Oncological Gene Environment Study (COGS), which is a large scale genotyping cohort funded by European Commission was utilized to identify disease associated biomarkers. More than 150 000 samples were genotyped in this study. Familial based high penetrance susceptibility genes were identified first and then low penetrance genes by association studies.[56,57] Carriers of such genes and SNPs predispose to breast cancer. A panel of 70 genes were able to predict breast cancer prognosis.[58] Genomic markers include small nucleotide polymorphisms (SNPs), mutations, additions and deletions, recombinations, and change in copy number (altered CNVs).[59,60]

GWAS were conducted by different groups in different cohorts to identify breast cancer susceptibility genes which may be useful for breast cancer screening of high risk populations.[61,62] In one such study, genotyping of 2702 women of European ancestry with invasive breast cancer and 5726 controls was conducted.[61] SNPs identified in this study were primary located in 1p11.2, 2q35, 3p, 5p12, 8q24, 10q23, 13, 14q24.1, and 16q regions. Genes affected by these SNPs are involved in regulation of actin cytoskeleton, glycan degradation, alpha linoleic metabolism, circadian rhythm, and drug metabolism.

13.5.2 Proteomics Based Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

Proteomics technologies have emerged as a useful tool in the discovery of diagnostic biomarkers and substantial technological advances in proteomics and related computational science have been made in the last decade.[1,63] These advances overcome in part the complexity and heterogeneity of the human proteome, permitting the quantitative analysis and identification of protein changes associated with tumor development. With the advent of new and improved proteomic technologies, it is possible to discover new biomarkers for the early detection and treatment monitor of different diseases. The contribution of the Human Proteomic Organization is valuable in this regard because this organization has provided the dataset of normal healthy human proteomic profiles which can be used as a reference dataset for identifying disease-associated proteomic changes (http://www.hupo.org/). Protein biomarkers are more related to disease phenotype and are more targetable for therapy in comparison with transcriptomic or genomic biomarkers. Proteomics provides a powerful tool to investigate potential biomarkers in diseases due to its high sensitivity, precise characterization of their interaction, and ability to detect functionally significant post-translational modifications. Proteomic biomarkers have been identified in blood (serum and plasma) as well as in tissue samples by applying approaches such as nuclear magnetic resonance spectroscopy (NMR), mass spectrometry (MS), two-dimensional gel electrophoresis and immunoprecipitation. In one study, investigators identified circulating proteomic biomarkers from different stages of breast cancer using an innovative strategy employing high sensitivity label-free proteomics. The approach was MS based and provided semi-quantitative results and could be applied in preclinical and clinical studies. Furthermore, breast cancer patient serum was analyzed by bidimensional nanoUPLC tandem nano ESI-MS to identify breast cancer biomarkers which are differentially expressed at early stages of cancer development.[64] Higher GRHL3 expression and lower levels of TNF alpha were reported during early stage of the diseases whereas PMS2 expression was high in advanced stages of the disease. These results were validated in a different set of patients although the number of participants was low. These investigators plan to evaluate the impact of such markers in determining survival rates of patients and recurrence of breast cancer or other cancers.

Proteomics with the recent advances in mass spectrometry is considered as a powerful analytical method for deciphering proteins expressions alterations as a function of disease progression.[63,65] Proteomics based analyses of breast serum and tissue lysates have resulted in the finding of a number of potential tumor biomarkers providing, therefore, a basis for a better understanding of the breast-cancer development and progression, and eventually serving as diagnostic and prognostic markers.[64] Probably the most widely used proteomic technology is the identification of alterations in

protein expression between two different samples through comparative two-dimensional gel electrophoresis (2-DE) which provides high-resolution separation of proteins and offers a powerful method for their identification and characterization.[66]

Proteomic analysis is an essential component to explain the information contained in genomic sequences in terms of the structure, function, and control of biological processes and pathways.[67] The proteome reflects the cellular state or the external conditions encountered by a cell. In addition, proteomic analysis is a genome-wide assay to differentiate distinct cellular states and to determine the molecular mechanisms that control them.[68] Infection-associated proteomic biomarkers have also been characterized.[69] High-throughput proteomic methodologies have the potential to revolutionize protein biomarker discovery and to allow for multiple proteins biomarkers to be assayed simultaneously. With the significant advances in 2-DE and mass spectrometry (MS), protein biomarker discovery has become one of the central applications of proteomics.[67]

13.5.2.1 Immune Diseases. Antigen arrays are valuable for profiling autoantibodies in diverse rheumatic autoimmune diseases and can be composed of proteins, peptides, protein complexes, glycoproteins, sugar nucleic acids, and lipids. T and B cells can be isolated from disease subjects and used for disease stratification, which ultimately help in designing treatment and disease management of autoimmune diseases (such as rheumatoid arthritis or RA).[70] The proteomic profile might suggest whether the disease was aggressive or not. Treatment can be selected based on the aggressiveness of the disease.[71] Less than two-thirds of all individuals with rheumatoid arthritis had an adequate response to anti-TNF therapy. Biomarkers could identify those individuals who were less likely to respond to this therapy.[70,72] This would not only reduce the cost associated with therapy but also avoid unwanted adverse reactions due to anti-TNF therapy among non-responders. Technologies such as mass cytometry, peptide and protein arrays and BCR (b cell receptor) and TCR (T cell receptor) sequencing might prove useful to improve the management of autoimmune diseases and represent the state of art in analyzing cells, soluble proteins and genes.[73]

Another group of biomarkers which was characterized for autoimmune diseases was annexins, a group of 12 highly conserved proteins which regulate cell cycle. Their abnormal expression was associated with disease development.[74] Annexin-I exerts its anti-inflammatory effect by suppressing the generation of inflammatory mediators and anti-annexin I antibodies are present in rheumatoid arthritis and systemic and cutaneous lupus erythematosus (SLE) patients.[75] Annexin II and V are involved in coagulation cascade because they have affinity towards phospholipids. Prolactin (PRL), ferritin, vitamin D, tumor biomarker tissue polypeptide antigen (TPA) levels are higher in patients with autoimmune disease (systemic sclerosis, systemic

and cutaneous lupus erythematosus, SLE), polymyositis (PM), dermatomyositis (DM), multiple sclerosis (MS).[76] The granin family is a group of acidic proteins present in the secretory granules of a wide variety of endocrine, neuronal, and neuroendocrine cells. A few important granins, chromogranin A and B, secretogrannin II, HISL-19 antigen, NESP55, ProSAAS should be studied further for their clinical implication.[77]

13.5.2.2 Cardiovascular Diseases (CVDs). Atherosclerosis is the main cause of cardiovascular diseases. Diagnosis of subclinical atherosclerosis is a clinical challenge. Furthermore, determining the extent of atherosclerosis aggressiveness in individual patients is a challenge too. Plasma osteoprotegerin (OPG) turned out to be an excellent proteomic biomarker which could detect preclinical atherosclerosis, as validated in a case-control study.[78] Proteomic profiling in circulating cells and plasma extracellular vesicles could distinguish populations at high risk of developing atherosclerosis.[79]

13.5.2.3 Metabolic Diseases. Isobaric tags for relative and absolute quantitation (iTRAQ) in combination with two-dimensional gel electrophoresis technologies identified a group of proteins which were associated with either diabetes, pancreatitis, and/or pancreatic cancer.[80]

13.5.2.4 Infectious Diseases. HIV-associated neurodegenerative disease progression and treatment response could be measured by following levels of biomarkers complement C3, soluble superoxide dismutase and prostaglandin synthase.[81] Although Infectious disease research had focused more on HIV-related diseases, HIV infection had its effects on the central nervous system (CNS) as this lentivirus could infect brain cells. CNS dysfunction then led to a group of cognitive and behavior changes (called HIV-associated neurocognitive disorders or neuroAIDS) which serve as biomarkers.[82]

13.5.2.5 Neurological Diseases. Because of the anatomical location of nervous system, it is difficult to get biomarkers of neurological disorders. The accessibility of the affected organ is a challenge and therefore surrogate biomarkers are generally used to identify and follow neurological disorders. Microvesicles (MVs) have been used as biomarkers of neurological disorders because their release is increased in these diseases.[83] MVs originate from exosomes (which are derived from endothelial cells) and could be found in plasma or serum. MVs are linked to neurological pathologies with a vascular or ischemic pathogenic component (sometimes used in diagnosis and follow up of strokes). In another disease, multiple sclerosis, MVs of oligodendroglial origin were reported in the cerebrospinal fluid (CSF).[83] MVs detection should be explored further to gain pathogenic

information, identify therapeutic targets, and select specific biomarkers for neurological diseases.

13.5.2.6 Cancer. In pancreatic cancer, the incidence and mortality rates are the same and the survival is only five years after the diagnosis of the disease. Plasma and serum of pancreatic cancer patients contain antibodies against periodontal bacteria *P. gingivalia* and their level is associated with pancreatic cancer.[84] Inflammatory biomarkers, C-reactive proteins, interleukin-6 (IL6), and soluble receptor of tumor necrosis factor alpha have also been used in identifying pancreatic cancer patients.[85] Some biomarkers identified include increased salivary levels of cell cycle regulatory proteins (cyclin D1 and Ki67), glycolytic enzyme lactate dehydrogenase (LDH), matrix metalloproteinases (MMPs) and reduction in DNA repair enzyme (8-oxoquanine DNA glycosylase) and mapsin in oral cancer patients.[86]

A highly sensitive method for serum human telomerase reverse transcriptase (hTERT) mRNA for hepatocellular carcinoma (HCC) was reported.[87] Alpha-fetoprotein (AFP) and des-gamma-carboxy prothrombin (DCP) were found to be good markers for HCC. This group also verified the significance of hTERT mRNA in a large scale multi-centered trial. hTERT mRNA was demonstrated to be independently correlated with clinical parameters: tumor size and tumor differentiation. hTERT mRNA proved to be superior to AFP, AFP-L3, and DCP in the diagnosis and underwent an indisputable change in response to therapy. The detection rate of small HCC by hTERTmRNA was superior to the other biomarkers.[87]

To identify the potential biomarkers involved in Hepatocellular carcinoma (HCC) carcinogenesis, a comparative proteomics approach was utilized to identify the differentially expressed proteins in the serum of 10 HCC patients and 10 controls. A total of 12 significantly altered proteins were identified by mass spectrometry. Of the 12 proteins identified, HSP90 was one of the most significantly altered proteins and its over-expression in the serum of 20 HCC patients was confirmed using ELISA analysis. The observations suggest that HSP90 might be a potential biomarker for early diagnosis, prognosis, and monitoring in the therapy of HCC.[88]

Proteomic analysis with 2-DE and MS was used to identify other potential serum markers for breast cancer.[89,90] Protein extracts expressed in the serum of breast cancer patients after depletion of high abundance proteins were compared to sera from healthy women using proteomic approaches. By comparing 2-DE profiles between tumor and non-tumor samples and using MALDI-TOF mass spectrometry of their trypsinized fragments, the identification of two proteins of interest, haptoglobin precursor and alpha-1-antitrypsin precursor, was observed.[93] Separation and analysis of proteins from cells, tissue samples and breast tumor biopsies proved very successful in identifying novel biomarkers.[91] Using proteomic approaches, 26 immunoreactive proteins (antigens) against which sera from newly diagnosed patients with infiltrating ductal carcinomas exhibited reactivity were detected.[89,92]

Among these antigens, peroxiredoxin-2 (Prx-2) belongs to a family of thiol-specific antioxidant proteins that control intracellular H_2O_2 by reducing reactive oxygen species (ROS) issued from free radicals. Such proteins might have an important role and protect the breast tumor cells against oxidative injury and modulate cell proliferation and apoptosis of malignant cells.[93]

Proteomic approaches are useful to identify protein-protein interaction and in one study estrogen receptor alpha and its interaction with a number of transcription factors was characterized which resulted in clinically useful information about breast cancer therapeutics.[94] Laser-capture microdissected breast cancer and normal tissue cells were analyzed by mass spectrometry to identify proteomic profiles associated with breast cancer.[95] In another study, glyoxalase-1 was found to be expressed in breast cancer.[89] This protein was involved in detoxification of methylglyoxal which is a cytotoxic product of glycolysis. Further analysis of tissue microarray indicated correlation of glyoxalase-1 with tumor grade. Based on results from reversed phase protein array, a model was created to predict pathological response in patients receiving neoadjuvant taxane and anthracyclin taxane based systemic therapy, thus indicating translational significance of proteomic biomarkers in breast cancer.[96]

13.5.3 Transcriptomics Based Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

Transcriptomics is the study of how our genes are regulated and expressed in different biological settings. All expressed genes can be quantitatively measured in a tissue at a given time (and this science is termed transcriptomics).[97] Over the last decade, microarray technology based transcriptomic analysis has contributed enormously to our understanding of the molecular basis of a number of diseases. Gene expression profiling offers an unparalleled opportunity to develop biomarkers that are useful in diagnosis and prognosis and in helping to achieve the goal of individualized treatment. However, the limitations of the technology and the danger of inappropriate experimental processes should not be underestimated. In clinical settings, blood transcriptomics of Alzheimer's patients undergoing treatment with EHT 0202 has been conducted.[98] Transcriptomic analysis indicated activation of pathways associated with CNS disorders, diabetes, inflammation, and autoimmunity. Treatment resulted in deactivation of these pathways in patients. Thus transcriptomic biomarker profiling could be used in disease prognosis. Such studies would help us identifying those patient populations who would respond to treatment, and also identifying those individuals who would likely to have recurrence of the disease.

13.5.3.1 Immune Diseases. In rheumatoid arthritis (RA) bone loss occurs.[99] Transcriptomic approaches were applied to identify genes and

pathways which contributed to bone loss in RA patients. Results indicated IL-17 and Wnt/beta katenin as transcriptional biomarkers of RA.[100]

13.5.3.2 Cardiovascular Diseases (CVDs). Atherosclerosis is a pathological process in which the walls of large arteries thicken and lose elasticity as a result of the growth of atheromatous lesions. Transcriptomics based gene expression analysis did not identify specific genes but pathways (especially inflammation related pathways) were identified which were associated with atherosclerosis development.[101] The analysis was conducted in monocytes and macrophages.

13.5.3.3 Metabolic Diseases. Most gene expression profiling identified diabetes but in a few cases individual gene expression (also called candidate gene expression approach) also provided information about disease-associated transcriptomics. Glucose response gene, RCAN1, was over expressed whereas perilipin A and G0/G1 switch gene, G0S2, were under expressed in diabetes-2.[102,103]

13.5.3.4 Infectious Diseases. Based on transcriptomics, a regression model was developed which could predict the onset of ventilator-associated pneumonia.[104] In another study, transcriptomics identified a gene-expression pattern which helped in identifying patients who were at high risk of developing trachibronchitis or pneumonia.[105]

13.5.3.5 Neurological Diseases. In the neurobiology field, neuronal development, function, and the subsequent degeneration of the brain are still serious problems. There is a need to find better targets for developing therapeutic intervention by identifying new biomarkers by applying transcriptomics and other approaches.[106] A high-throughput high-content screening approach is needed to go from genes to gene-networks.

13.5.3.6 Cancer. To implicate a combination of omics technologies, one investigator performed transcriptomics and metabolomics in breast cancer samples and identified a disease-associated profile, which could be used for diagnosis purposes.[107] Transcriptomics identified genes and pathways whereas the functional significance was evaluated by metabolite characterization.

13.5.4 Immunomics Based Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

The huge amount of immunological information hidden in the plasma could be better revealed by combining the characterization of antibody

binding target epitomes with improved estimation of effector functions triggered by these binding events. Functional immune profiles can be generated characterizing general immune responsiveness by designing arrays with information about epitope collections from different antibody targets. Immunomics was implicated in searching and identifying proteins of interest in the case of breast or colorectal cancers.[108] Two approaches were developed at their laboratory:[109] the top down SERological Proteome Analysis (SERPA) and the bottom up MAPPing (Multiple Affinity Protein Profiling) strategies. The first one relied on two dimensional electrophoresis (2 DE), immunoblotting, image analysis and mass spectrometry. The second approach dealt with the use of two dimensional immuno affinity chromatography, enzymatic digestion of the antigens, and analyses by tandem mass spectrometry. Using immunoinformatics approach, putative T- and B- cell epitopes of capsid proteins were identified which were conserved in existing serotypes.[109]

13.5.4.1 Immune Diseases. Although autism is a neurological disorder, its regulation involves autoimmunity. This disease is characterized by impaired social interaction and verbal and non-verbal communication. A high number of mast cells and serotonin in urine were observed in autism patients. Mast cells produce inflammatory cytokines in large amount in autism patients and the possibility of autoimmunity was proposed.[110]

13.5.4.2 Cardiovascular Diseases (CVDs). Atherosclerosis is a chronic inflammatory disease of the vessel wall. The gene CARD8, which codes proteins involved in innate immunity in atherosclerosis patients, is an excellent biomarker for atherosclerosis. Inflammatory markers also over expressed in this disease. Inflammosomes produce interleukin 1beta in response to cholesterol crystal accumulation in macrophages. Ephs and ephrins also were proposed as biomarkers of atherosclerosis.[111]

13.5.4.3 Metabolic Diseases, Metabolomics and Proteomics. Chronic low grade inflammation contributes to the pathogenesis of insulin resistance. In diabetes innate immune cells accumulate in metabolic tissues and release inflammatory cytokines, especially IL-1beta and TNF. One investigator proposed cell mediated immunity in diabetes. Regulation of type 1 and type 2 diabetes by immune system occurred differently in diabetic patients under study.[112,113]

13.5.4.4 Infectious Diseases. Tuberculosis (TB) is a common and lethal infectious disease caused by various strains of mycobacteria. The lung is the primary organ which is infected by the bacteria although any part of the body can be infected. For several years antibiotics were able to treat TB but in the past few years antibiotic resistant mycobacteria have been reported. To address this point, the use of anti-mycobacteria antibodies,

enhancing the Th1 protective responses by using mycobacterial antigen, or increasing Th1 cytokines, was evaluated and promising results were obtained.[114] Childhood deficiency of vitamin D was proposed to be a susceptible biomarker of TB.[115]

13.5.4.5 Neurological Diseases. Genes involved in inflammation and immune system regulatory pathways were identified when profiling of Alzheimer's and non-Alzheimer's samples was conducted.[116] Results also indicated a role for chemokines and their receptors (CXR4 and CCR3) in patient samples; and these two receptors may be considered biomarkers for the Alzheimer's disease. Another biomarker in the same disease was amyloid beta which is expressed at higher levels in patients compared to age and sex matched controls.

13.5.4.6 Cancer. Genetic variation in innate immunity and inflammation pathway associated with lung cancer risk were proposed in different studies. Some investigators proposed T cell mediated immunity in lung cancer using HLA-DR telomerase derived epitopes. In cervical cancer, cell mediated immunity is the main mechanism of cancer development. Infection-associated cancers, infectious agents and their epitopes were considered excellent biomarkers which could be used for diagnostic purposes.[75,117,118]

13.5.5 Metabolomics Based Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

Metabolomics is the study of small molecules of both endogenous and exogenous origin, such as metabolic substrates and their products, lipids, small peptides, vitamins and other protein cofactors, generated by metabolism,[2,119] which are downstream from genes. This approach has received more attention in recent years as an ideal methodology to unravel signals closer to the culmination of the disease process. The compounds identified through metabolomic profiling represent a range of intermediate metabolic pathways that may serve as biomarkers of exposure, susceptibility or disease.[120] Therefore, it is a valuable approach for deciphering metabolic outcomes with a phenotypic change. Inherited metabolic disorders can be measured by following a set of metabolomic biomarkers by directly measuring metabolites using LC-MS or NMR. A few examples of these diseases are: amino acid metabolism (phenylketonuria, maple syrup urine disease, tyrosinemia I, argininemia, homocystinuria, ornithine transcarbamylase deficiency, and nonketotic hyperglycemia), organic acidurias, and mitochondrial defects.[121]

Risk prediction for diabetes and CVD is difficult although a number of biomarkers have been identified (see Table 13.4). So far insulin resistance (IR) and atherosclerosis are the only promising biomarkers.

In gastric cancer, alterations in metabolites of glycolysis, fatty-acid beta-oxidation, and cholesterol and amino acid metabolism were observed.[122] These metabolites can be used to follow gastric cancer development and treatment response.

13.5.5.1 Immune Diseases. Diagnosis of asthma is sometime difficult and respiratory symptoms alone are not sufficient for diagnosing this disease. Measurements of fractional excretion of nitric oxide can be used as a biomarker for diagnosing asthma. There is a considerable effect of the environment on health from pollution, climate change, and epigenetic influences, underlining the importance of understanding gene-environment interactions in the pathogenesis of asthma and response to treatment. Metabolomic and proteomic approaches can be applied to determining levels of nitric oxide.[81,123]

13.5.5.2 Cardiovascular Diseases (CVDs). Metabolites characterized from biofluids of CVD patients serve as biomarkers for CVDs, such as atherosclerosis. Dyslipidemia in HIV patients puts them at high risk of developing CVDs. In another study, intestinal microbiodata and metabolites were suggested to contribute in CVD development.[124,125]

13.5.5.3 Metabolic Diseases, Metabolomics and Proteomics. Serum metabolites from healthy and diabetes-2 patients showed different profiles and disease-associated profiles were suggested biomarkers for diabetes diagnosis. In a large population study diabetes associated metabolites were identified which were generated by seven genes.[126,127]

13.5.5.4 Infectious Diseases. Antibiotic resistance is a common feature of TB where mycobacteria become resistant to antibiotic treatment. GC-MS based technologies were used to identify metabolites, especially metabolites from fatty acid metabolism, from wild type mycobacteria infected and mutant infected samples. These observations laid the groundwork for developing therapeutics for TB treatment. In a recent study, the metabolomic adaptation of bacteria in the host was also evaluated.[128]

13.5.5.5 Neurological Diseases. An epileptic seizure is a transient symptom of abnormal excessive or synchronous neuronal activity in the brain. Errors of metabolism result in abnormal levels of metabolites at the neonatal stage of development resulting in neonatal seizures. One group of investigators demonstrated altered metabolism of astrocytes contributing to seizures.[93,129]

13.5.5.6 Cancer. GC-MS based metabolomics in serum identified colorectal cancer associated biomarkers. Metabolites were successfully used to detect bladder, breast, and pancreatic cancer. Here this is emphasized

that pancreatic cancer is such a disease where early detection markers are sought because the incidence and mortality rates are almost the same, therefore, metabolomic biomarkers should be explored for this fetal disease.[130,131]

13.5.6 Epigenomics and miRNOMICS Based Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

In addition to genetic code, human cells contain an additional regulatory level predominating the genetic code; this is called epigenetic code.[34] This involves altered gene expression without changing the genomic structure.[33] Due to different chromatin status, condensed or relaxed, the same genetic variants might be associated with different phenotypes, depending on the environment, life-style, and exposure. A rapidly growing number of genes with epigenetic regulation altering their expression by chromatin-remodeling (condensation and relaxation) have been identified.[7,34,43,132] Methylation of cytosines in DNA, histone modifications, alterations of non-coding RNAs (especially miroRNAs) are the mechanisms involved in chromatin remodeling.

The term epigenome is used to define a cell's overall epigenetic state. Basic biological properties of DNA-segments such as gene density, replication timing, and recombination are linked to their GC content. The promoter region is rich in CpG content. A genomic region of about 0.4 kb with more than 50% GC content is called a CpG island. In mammals, CpG islands typically are 200–300 bp long. Promoters of tissue specific genes that are situated within CpG islands are generally unmethylated but during the disease development, these CpG sites start getting methylated. Cytosine methylation can regulate gene expression by hindering the association of some transcriptional factors with their cognate DNA recognition sequences, or methyl CpG binding protein (MBP) can bind to methylated cytosines mediating a repressive signal, or MBPs can interact with chromatin forming proteins modifying the surrounding chromatin, linking DNA methylation with chromatin modification. DNA methylation at position five of cytosine is conducted by DNA methyltransferases (DNMTs). These enzymes are needed for methylation initiation and methylation maintenance.[34]

Alterations due to epigenetic mechanisms can be stably passed over numerous cycles of cell division. A selected epigenetic alteration can be inherited from one generation to another generation.[133] Cancer specific methylation alterations are hallmark of different cancers. Alteration in methylation may cause genomic instability, genomic alterations and change in gene expression.[34,134] A systematic approach to determine epigenetic changes in tumor development may lead to identify biomarkers needed for cancer diagnosis. It has been suggested that an integration of genome and hypermethylome might provide insight into major pathways of cancer development which in turn might help us identify new biomarkers of cancer

diagnosis and prognosis.[135] Methylation and microRNA (miR) alterations are the main biomarkers which can be assayed easily in samples non-invasively.[136] The finding that monozygotic twins are epigenetically indistinguishable early in life but with age exhibit substantial differences in the epigenome indicates that environmentally determined alterations in a cell's epigenetic marks are responsible for an altered epigenome.[137] Examples exist to show that environmental factors influence disease initiation, progression, and development.[138,139]

MicroRNAs (miRs), small RNA molecules of approximately 22 nucleotides, have been shown to be up or down regulated in specific cell types and disease states. These molecules have become recognized as one of the major regulatory gatekeepers of coding genes in the human genome. The structure, nomenclature, mechanism of action, technologies used for miR detection, and associations of miRs with human cancer have been evaluated by a number of investigators.[140,141] miRs are produced in a tissue-specific manner, and changes in miR within a tissue type can be correlated with disease status.[142] miRs regulate mRNA translation and their degradation. Among a number of regulators of gene expression, miRs are the key regulators. Tissue specific miRs have been reported by different groups.[28] These RNAs are of small size with distinct stem and loop structure.[143] A number of miRs can be isolated in circulation. Because of their small size and stability (due to secondary structure), these circulating miRs provide a rich source of diagnostic biomarkers.

13.5.6.1 Immune Diseases. Genetic and environmental factors contribute in multiple sclerosis (MS). However, recent research indicate role of epigenetics in MS development. MS is an inflammatory disease in which the fatty myelin sheaths around the axons of the brain and spinal cord are damaged. Female specific MS is associated with HLA DRB1-1*5 haplotype. SHP-1 hypermethylation is a biomarker for multiple sclerosis.[144]

13.5.6.2 Cardiovascular Diseases (CVDs). Atherosclerosis can be diagnosed when other biomarkers are combined with epigenetic biomarkers, such as hypermethylation of monoamine oxide A (MAOA). Along with gene-specific methylation biomarkers, global DNA methylation biomarkers were identified for atherosclerosis.[145] Similarly in other CVDs, such as hypertension, global methylation profiling and individual gene methylation status (ADD1 gene) were used for diagnosis and outcome.[146]

13.5.6.3 Metabolic Diseases, Metabolomics and Proteomics. A combination of biomarkers, genetic and epigenetics, was used for diabetes diagnosis using mutations in duodenal homeobox-1 (PDX-1) and its methylation levels. Hypermethylation sites were reported at multiple sites distributed

throughout the genome. At least one investigator reported a combination of histone and methylation biomarkers for diabetes diagnosis.[147,148]

13.5.6.4 Infectious Diseases. In early liver cancer, where risk factors are infection by HBV and HCV, clustered DNA methylation changes in polycomb repressor target genes were observed. P16 and c-myc hypermethylation also suggested initiation of hepatocellular carcinoma. In tuberculosis, vitamin D receptor (VDR) methylation indicated high risk of developing the disease. VDR gene encodes a transcription factor which alters calcium homeostasis and immune function.[149]

13.5.6.5 Neurological Diseases. In Alzheimer's disease, altered methylation levels of repeat sequences (Alu, Line-1, and SAT-alpha) were observed. This modification induced genomic instability that contributed to disease imitation and progression.[150] Epigenetic regulation in autism was also studied and locus specific hypermethylation of MECP2 was observed.[151]

13.5.6.6 Cancer. Both histone modifications and DNA methylation were observed in different cancers, especially breast cancer. When epigentic profiling of MCF-7, MDA-MB-231, and MDA-MB-231 (S30) was followed, decreased trimethylation of H4K20 and hyperacetylation of H4 was observed. Concomitant to a decrease in trimethylation, lower levels of Suv4-20h2 histone methyl transferase were also observed. The effect was more in MDA-MB-231 compared to other cells which suggested that differential expression of histone modifications could represent aggressiveness of the disease. In another study, HDAC6 (one of the histone acetyl transferase) responded to estrogen treatment.[152] Retinoblastoma levels were lower whenever trimethylation of H4K20 was present. A correlation with the tumor stage and grade was also established based on these histone biomarkers. Another interesting study reported quantitative expression of HDAC1 and its correlation with breast cancer patient's age, lymph node status, tumor size and her2/neu negative, ER and PR positive status.[153]

Cancer cells accumulate abnormal DNA methylation patterns that result in malignant phenotypes. The genomic distribution of methylation is not well understood. In this direction, a number of genome-wide association studies have been conducted so that cancer risk associated biomarkers can be identified. Using methylated DNA immunoprecipitation combined with high-throughput sequencing (MeDIP-seq), levels of methylation were compared in samples from normal and cancer cells and global hypomethylation was observed in cancer samples, especially in the CpG rich regions. The location of these CpG rich regions was not related with the transcription start sites of various genes. Using this approach, the methylation patterns during epithelial to mesenchymal transition also was evaluated and used for disease stratification.[154]

In breast cancer, methyl acceptance capacity in malignant breast tissues was approximately 2–3 fold greater compared with matched controls. However, the variation in methyl acceptance capacity among patients varied a lot.[155] Quantitative analysis for 5meC levels showed a substantial decrease compared with normal tissues. Levels of hypomethylation in BRCA1 and BRCA2 cancers were slightly lower but significant.[156] Genome-wide hypomethylation correlated with satellite sequence hypomethylation. Specific regions (Sa2 coding) on chromosome 1 and sat-alpha were specifically hypomethylated.[157] On chromosome 5, the region containing the coding sequence of SATr-1 also showed hypomethylation.

The tissue concentrations of specific miRs have been associated with tumor invasiveness, metastatic potential, and other clinical characteristics for several types of cancers, including chronic lymphocytic leukemia, and breast, colorectal, hepatic, lung, pancreatic, and prostate cancers. By targeting and controlling the expression of mRNA, miRs can control highly complex signal-transduction pathways and other biological pathways. The biologic roles of miRs in cancer suggest a correlation with prognosis and therapeutic outcome. Further investigation of these roles may lead to new approaches for the categorization, diagnosis, and treatment of human cancers. Frequent dysregulation of miR in malignancy highlights the study of molecular factors upstream of gene expression following the extensive investigation on elucidating the important role of miR in carcinogenesis. For example, esophageal carcinogenesis is a multi-stage process, involving a variety of changes in gene expression and physiological structure change. Recent innovation in miRs profiling technology have shed new light on the pathology of esophageal carcinoma (EC), and also showed great potential for exploring novel biomarkers for both EC diagnosis and treatment. A thorough review of the role of miRs in EC, addressing miR functions, their putative role as oncogenes or tumor suppressors and their potential target genes has been explored by different investigators.[158]

In inflammatory breast cancer cells, more than 300 miRs were evaluated for their association with the disease.[159] The most promising miRs were miR-29a, miR-30b, miR-342-5p, and 520a-5p. The functional analysis of these miRs revealed their role in cell proliferation and signal transduction pathways. These markers could be useful to identify inflammatory breast cancer cells. The promoter regions of miR coding region were evaluated and several miR promoters were found hypermethylated, especially those of miR-31, miR-130a, miR-let7a-3/let 7-b, miR-155, and miR-137.[142] In one example, an advantage of using miRs for detecting cancer is demonstrated due to their stability even in fixed tissues.[143] miR-155 predicted prognosis of triple negative breast cancer (higher miR-155 expression correlated with higher angiogenesis and aggressiveness).[160]

13.6 CONCLUSION

In the last few decades, biomarkers have played a very significant biological role in disease diagnosis and therapy and consequently numerous studies

have been published so far. Comparatively, most of the work has been carried out on biomarker diagnosis than their usage for therapies where single and multiple biomarkers have been used for diagnosis. In this chapter, a number of important biomarkers for a broad-range of disorders have been described in detail with a further focus on updating informations in this field. Furthermore, different criteria and approaches for the selection and specification of different types of biomarkers have also been described, which are very helpful in ameliorating our knowledge in understanding and distinguishing these biomarkers from one another. The cutting-edge NGS technologies and the OMICS approaches have already contributed a vital role in biomarkers discovery in various diseases and they further provide a platform and an insight into the robustness and progress of this field of knowledge.

REFERENCES

1. M. Verma, J. Kagan, D. Sidransky and S. Srivastava, *Nat. Rev. Cancer*, 2003, **3**, 789–795.
2. E. Ferrannini, A. Natali, S. Camastra, M. Nannipieri, A. Mari, K. P. Adam, M. V. Milburn, G. Kastenmuller, J. Adamski, T. Tuomi, V. Lyssenko, L. Groop and W. E. Gall, *Diabetes*, 2013, **62**, 1730–1737.
3. M. Verma, *Curr. Genomics*, 2012, **13**, 308–313.
4. M. Verma and S. Srivastava, *Recent Results Cancer Res.*, 2003, **163**, 72–84, discussion 264–266.
5. M. Verma, G. L. Wright, Jr., S. M. Hanash, R. Gopal-Srivastava and S. Srivastava, *Ann. N. Y. Acad. Sci.*, 2001, **945**, 103–115.
6. S. Srivastava, M. Verma and D. E. Henson, *Clin. Cancer Res.*, 2001, **7**, 1118–1126.
7. M. Verma, *Methods Mol. Biol.*, 2012, **863**, 467–480.
8. S. A. Peters, F. L. Visseren and D. E. Grobbee, *Nat. Rev. Cardiol.*, 2013, **10**, 12–14.
9. C. W. Peng, Q. Tian, G. F. Yang, M. Fang, Z. L. Zhang, J. Peng, Y. Li and D. W. Pang, *Biomaterials*, 2012, **33**, 5742–5752.
10. W. Naumnik, B. Naumnik, K. Niewiarowska, M. Ossolinska and E. Chyczewska, *Exp. Oncol.*, 2012, **34**, 348–353.
11. C. C. Solomides, B. J. Evans, J. M. Navenot, R. Vadigepalli, S. C. Peiper and Z. X. Wang, *Acta Cytol.*, 2012, **56**, 645–654.
12. A. Di Ieva, *Microvasc. Res.*, 2010, **80**, 522–533.
13. J. Penner, R. Rupsingh, M. Smith, J. L. Wells, M. J. Borrie and R. Bartha, *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, 2010, **34**, 104–110.
14. R. Rupsingh, M. Borrie, M. Smith, J. L. Wells and R. Bartha, *Neurobiol. Aging*, 2011, **32**, 802–810.
15. T. Efferth, *Planta Med.*, 2010, **76**, 1143–1154.
16. D. G. Duda, C. G. Willett, M. Ancukiewicz, E. di Tomaso, M. Shah, B. G. Czito, R. Bentley, M. Poleski, G. Y. Lauwers, M. Carroll, D. Tyler,

C. Mantyh, P. Shellito, J. W. Clark and R. K. Jain, *Oncologist*, 2010, **15**, 577–583.
17. H. Xu, C. Chen, C. M. Liu, J. Peng, Y. Li, Z. L. Zhang and H. W. Tang, *Guangpuxue Yu Guangpufenxi*, 2009, **29**, 3216–3219.
18. S. J. Otto, J. Fracheboud, A. L. Verbeek, R. Boer, J. C. Reijerink-Verheij, J. D. Otten, M. J. Broeders and H. J. de Koning, *Cancer Epidemiol., Biomarkers Prev.*, 2012, **21**, 66–73.
19. H. Allahverdipour, M. Asghari-Jafarabadi and A. Emami, *Women Health*, 2011, **51**, 204–219.
20. M. Heijblom, J. M. Klaase, F. M. van den Engh, T. G. van Leeuwen, W. Steenbergen and S. Manohar, *Technol. Cancer Res. Treat.*, 2011, **10**, 607–623.
21. M. Sentis, *Breast Cancer Res. Treat.*, 2010, **123**(Suppl 1), 11–13.
22. R. Luckmann, *ACP Journal Club*, 2005, **142**, 23.
23. H. Zhang, D. Yee and C. Wang, *Nanomedicine (London, U. K.)*, 2008, **3**, 83–91.
24. L. M. Kenny, A. Al-Nahhas and E. O. Aboagye, *Nucl. Med. Commun.*, 2011, **32**, 333–335.
25. R. Lieberman, *Am. J. Phytomed. Clin. Ther.*, 2012, **19**, 395–396.
26. J. L. Parker, N. Lushina, P. S. Bal, T. Petrella, R. Dent and G. Lopes, *Breast Cancer Res. Treat.*, 2012, **136**, 179–185.
27. C. S. Zhu, P. F. Pinsky, D. W. Cramer, D. F. Ransohoff, P. Hartge, R. M. Pfeiffer, N. Urban, G. Mor, R. C. Bast, Jr., L. E. Moore, A. E. Lokshin, M. W. McIntosh, S. J. Skates, A. Vitonis, Z. Zhang, D. C. Ward, J. T. Symanowski, A. Lomakin, E. T. Fung, P. M. Sluss, N. Scholler, K. H. Lu, A. M. Marrangoni, C. Patriotis, S. Srivastava, S. S. Buys and C. D. Berg, *Cancer Prev. Res.*, 2011, **4**, 375–383.
28. K. Tjensvoll, K. N. Svendsen, J. M. Reuben, S. Oltedal, B. Gilje, R. Smaaland and O. Nordgard, *Biomarkers*, 2012, **17**, 463–470.
29. S. Knappskog, R. Chrisanthar, E. Lokkevik, G. Anker, B. Ostenstad, S. Lundgren, T. Risberg, I. Mjaaland, B. Leirvaag, H. Miletic and P. E. Lonning, *Breast Cancer Res.*, 2012, **14**, R47.
30. E. Rivera and H. Gomez, *Breast Cancer Res.*, 2010, **12**(Suppl 2), S2.
31. W. G. Newman and D. Flockhart, *Pharmacogenomics*, 2012, **13**, 629–631.
32. B. A. Virnig, M. T. Torchia, S. L. Jarosek, S. Durham and T. M. Tuttle, in *Data Points Publication Series*, Agency for Healthcare Research and Quality (US), Rockville (MD), 2011.
33. M. Verma, *Curr. Opin. Clin. Nutr. Metab. Care*, 2013, **16**, 376–384.
34. M. Verma, M. J. Khoury and J. P. Ioannidis, *Cancer Epidemiol., Biomarkers Prev.*, 2013, **22**, 189–200.
35. M. Lucas and S. Gaudieri, *Biomarkers Med.*, 2012, **6**, 133–135.
36. C. Ferrandi, F. Richard, P. Tavano, E. Hauben, V. Barbie, J. P. Gotteland, B. Greco, M. Fortunato, M. F. Mariani, R. Furlan, G. Comi, G. Martino and P. F. Zaratin, *Mult. Scler.*, 2011, **17**, 43–56.
37. A. E. Aiello and G. A. Kaplan, *Biodemography Soc. Biol.*, 2009, **55**, 178–205.
38. G. Lippi, T. Meschi and G. Cervellin, *Eur. J. Intern. Med.*, 2011, **22**, 460–465.

39. G. Brathen, K. S. Bjerve, E. Brodtkorb and G. Bovim, *J. Neurol., Neurosurg. Psychiatry*, 2000, **68**, 342–348.
40. D. Lev, Y. Weigl, M. Hasan, E. Gak, M. Davidovich, C. Vinkler, E. Leshinsky-Silver, T. Lerman-Sagie and N. Watemberg, *Am. J. Med. Genet., Part A*, 2007, **143A**, 921–924.
41. N. Elleuch, I. Feki, E. Turki, M. I. Miladi, A. Boukhris, M. Damak, C. Mhiri, E. Chappuis and F. Woimant, *Rev. Neurol.*, 2010, **166**, 550–552.
42. D. Zhu, J. Wang, L. Ren, Y. Li, B. Xu, Y. Wei, Y. Zhong, X. Yu, S. Zhai, J. Xu and X. Qin, *J. Cell. Biochem.*, 2013, **114**, 448–455.
43. M. Verma, *Antioxid. Redox Signaling*, 2012, **17**, 355–364.
44. S. Viatte, D. Plant and S. Raychaudhuri, *Nat. Rev. Rheumatol.*, 2013, **9**, 141–153.
45. J. Kurko, T. Besenyei, J. Laki, T. T. Glant, K. Mikecz and Z. Szekanecz, *Clin. Rev. Allergy Immunol.*, 2013, **45**(2), 170–179.
46. B. P. Smith, D. H. Best and C. G. Elliott, *Hear Fail. Clin.*, 2012, **8**, 319–330.
47. J. Simino, D. C. Rao and B. I. Freedman, *Curr. Opin. Nephrol. Hypertens.*, 2012, **21**, 500–507.
48. G. Massawi, P. Hickling, D. Hilton and C. Patterson, *Rheumatology (Oxford, U. K.)*, 2003, **42**, 1012–1014.
49. N. Mahfoudh, M. Siala, M. Rihl, A. Kammoun, F. Frikha, H. Fourati, M. Younes, R. Gdoura, L. Gaddour, F. Hakim, Z. Bahloul, S. Baklouti, N. Bargaoui, S. Sellami, A. Hammami and H. Makni, *Clin. Rheumatol.*, 2011, **30**, 1069–1073.
50. G. Kamkamidze, C. Capoulade-Metay, M. Butsashvili, Y. Dudoit, O. Chubinishvili, P. Debre and L. Theodorou, *Georgian Medical News*, 2005, **118**, 74–79.
51. B. J. Grady, D. C. Samuels, G. K. Robbins, D. Selph, J. A. Canter, R. B. Pollard, D. W. Haas, R. Shafer, S. A. Kalams, D. G. Murdock, M. D. Ritchie and T. Hulgan, *J. Acquired Immune Defic. Syndr.*, 2011, **58**, 363–370.
52. J. C. Mar, N. A. Matigian, A. Mackay-Sim, G. D. Mellick, C. M. Sue, P. A. Silburn, J. J. McGrath, J. Quackenbush and C. A. Wells, *PLoS Genetics*, 2011, **7**, e1002207.
53. F. D. Shah, R. Begum, B. N. Vajaria, K. R. Patel, J. B. Patel, S. N. Shukla and P. S. Patel, *Indian J. Clin. Biochem.*, 2011, **26**, 326–334.
54. R. Kirk, *Nat. Rev. Clin. Oncol.*, 2011, **8**, 383.
55. P. Hall, *J. Intern. Med.*, 2012, **271**, 318–320.
56. A. C. Antoniou, A. B. Spurdle, O. M. Sinilnikova, S. Healey, K. A. Pooley, R. K. Schmutzler, B. Versmold, C. Engel, A. Meindl, N. Arnold, W. Hofmann, C. Sutter, D. Niederacher, H. Deissler, T. Caldes, K. Kampjarvi, H. Nevanlinna, J. Simard, J. Beesley, X. Chen, S. L. Neuhausen, T. R. Rebbeck, T. Wagner, H. T. Lynch, C. Isaacs, J. Weitzel, P. A. Ganz, M. B. Daly, G. Tomlinson, O. I. Olopade, J. L. Blum, F. J. Couch, P. Peterlongo, S. Manoukian, M. Barile, P. Radice, C. I. Szabo, L. H. Pereira, M. H. Greene, G. Rennert, F. Lejbkowicz, O. Barnett-

Griness, I. L. Andrulis, H. Ozcelik, A. M. Gerdes, M. A. Caligo, Y. Laitman, B. Kaufman, R. Milgrom, E. Friedman, S. M. Domchek, K. L. Nathanson, A. Osorio, G. Llort, R. L. Milne, J. Benitez, U. Hamann, F. B. Hogervorst, P. Manders, M. J. Ligtenberg, A. M. van den Ouweland, S. Peock, M. Cook, R. Platte, D. G. Evans, R. Eeles, G. Pichert, C. Chu, D. Eccles, R. Davidson, F. Douglas, A. K. Godwin, L. Barjhoux, S. Mazoyer, H. Sobol, V. Bourdon, F. Eisinger, A. Chompret, C. Capoulade, B. Bressac-de Paillerets, G. M. Lenoir, M. Gauthier-Villars, C. Houdayer, D. Stoppa-Lyonnet, G. Chenevix-Trench and D. F. Easton, *Am. J. Hum. Genet.*, 2008, **82**, 937–948.

57. S. A. Gayther, H. Song, S. J. Ramus, S. K. Kjaer, A. S. Whittemore, L. Quaye, J. Tyrer, D. Shadforth, E. Hogdall, C. Hogdall, J. Blaeker, R. DiCioccio, V. McGuire, P. M. Webb, J. Beesley, A. C. Green, D. C. Whiteman, M. T. Goodman, G. Lurie, M. E. Carney, F. Modugno, R. B. Ness, R. P. Edwards, K. B. Moysich, E. L. Goode, F. J. Couch, J. M. Cunningham, T. A. Sellers, A. H. Wu, M. C. Pike, E. S. Iversen, J. R. Marks, M. Garcia-Closas, L. Brinton, J. Lissowska, B. Peplonska, D. F. Easton, I. Jacobs, B. A. Ponder, J. Schildkraut, C. L. Pearce, G. Chenevix-Trench, A. Berchuck and P. D. Pharoah, *Cancer Res.*, 2007, **67**, 3027–3035.
58. P. D. Pharoah and C. Caldas, *Nat. Rev. Clin. Oncol.*, 2010, **7**, 615–616.
59. A. A. Ponomareva, E. Rykova, N. V. Cherdyntseva, E. L. Choinzonov, P. P. Laktionov and V. V. Vlasov, *Mol. Biol. (Mosk.)*, 2011, **45**, 203–217.
60. V. R. Adams and R. D. Harvey, *Am. J. Health-Syst. Pharm.*, 2010, **67**, S3–S9, quiz S15–S16.
61. J. Li, K. Humphreys, T. Heikkinen, K. Aittomaki, C. Blomqvist, P. D. Pharoah, A. M. Dunning, S. Ahmed, M. J. Hooning, J. W. Martens, A. M. van den Ouweland, L. Alfredsson, A. Palotie, L. Peltonen-Palotie, A. Irwanto, H. Q. Low, G. H. Teoh, A. Thalamuthu, D. F. Easton, H. Nevanlinna, J. Liu, K. Czene and P. Hall, *Breast Cancer Res. Treat.*, 2011, **126**, 717–727.
62. C. Turnbull, S. Ahmed, J. Morrison, D. Pernet, A. Renwick, M. Maranian, S. Seal, M. Ghoussaini, S. Hines, C. S. Healey, D. Hughes, M. Warren-Perry, W. Tapper, D. Eccles, D. G. Evans, M. Hooning, M. Schutte, A. van den Ouweland, R. Houlston, G. Ross, C. Langford, P. D. Pharoah, M. R. Stratton, A. M. Dunning, N. Rahman and D. F. Easton, *Nat. Genet.*, 2010, **42**, 504–507.
63. P. R. Srinivas, M. Verma, Y. Zhao and S. Srivastava, *Clin. Chem.*, 2002, **48**, 1160–1169.
64. C. Panis, L. Pizzatti, A. C. Herrera, R. Cecchini and E. Abdelhay, *Cancer Lett.*, 2013, **330**, 57–66.
65. M. Verma, *Methods Mol. Biol.*, 2009, **471**, 197–215.
66. K. Na, M. J. Lee, H. J. Jeong, H. Kim and Y. K. Paik, *Methods Mol. Biol.*, 2012, **854**, 223–237.
67. E. S. Baker, T. Liu, V. A. Petyuk, K. E. Burnum-Johnson, Y. M. Ibrahim, G. A. Anderson and R. D. Smith, *Genome Med.*, 2012, **4**, 63.

68. N. G. Anderson, *Clin. Chem.*, 2010, **56**, 154–160.
69. W. Rozek, J. Horning, J. Anderson and P. Ciborowski, *Proteomics: Clin. Appl.*, 2008, **2**, 1498–1507.
70. E. J. Toonen, C. Gilissen, B. Franke, W. Kievit, A. M. Eijsbouts, A. A. den Broeder, S. V. van Reijmersdal, J. A. Veltman, H. Scheffer, T. R. Radstake, P. L. van Riel, P. Barrera and M. J. Coenen, *PloS One*, 2012, **7**, e33199.
71. P. Emery, *Rheumatology (Oxford, U. K.)*, 2012, **51**(Suppl 5), v22–v30.
72. E. Solau-Gervais, C. Prudhomme, P. Philippe, A. Duhamel, C. Dupont-Creteur, J. L. Legrand, E. Houvenagel and R. M. Flipo, *Jt., Bone, Spine*, 2012, **79**, 281–284.
73. H. T. Maecker, T. M. Lindstrom, W. H. Robinson, P. J. Utz, M. Hale, S. D. Boyd, S. S. Shen-Orr and C. G. Fathman, *Nat. Rev. Rheumatol.*, 2012, **8**, 317–328.
74. L. Iaccarino, A. Ghirardello, M. Canova, M. Zen, S. Bettio, L. Nalotto, L. Punzi and A. Doria, *Autoimmun. Rev.*, 2011, **10**, 553–558.
75. N. Leung, *Med. J. Malaysia*, 2005, **60**(Suppl B), 63–66.
76. H. Orbach, G. Zandman-Goddard, H. Amital, V. Barak, Z. Szekanecz, G. Szucs, K. Danko, E. Nagy, T. Csepany, J. F. Carvalho, A. Doria and Y. Shoenfeld, *Ann. N. Y. Acad. Sci.*, 2007, **1109**, 385–400.
77. A. Bartolomucci, G. M. Pasinetti and S. R. Salton, *Neuroscience*, 2010, **170**, 289–297.
78. R. Mogelvang, S. H. Pedersen, A. Flyvbjerg, M. Bjerre, A. Z. Iversen, S. Galatius, J. Frystyk and J. S. Jensen, *Am. J. Cardiol.*, 2012, **109**, 515–520.
79. O. B. Bleijerveld, Y. N. Zhang, S. Beldar, I. E. Hoefer, S. K. Sze, G. Pasterkamp and D. P. de Kleijn, *Proteomics: Clin. Appl.*, 2013, **7**(7–8), 490–503.
80. W. S. Wang, X. H. Liu, L. X. Liu, D. Y. Jin, P. Y. Yang and X. L. Wang, *J. Proteomics*, 2013, **84C**, 52–60.
81. A. J. Apter, *J. Allergy Clin. Immunol.*, 2010, **125**, 79–84.
82. G. Pendyala and H. S. Fox, *Genome Med.*, 2010, **2**, 22.
83. E. Colombo, B. Borgiani, C. Verderio and R. Furlan, *Front. Physiol*, 2012, **3**, 63.
84. D. S. Michaud, J. Izard, C. S. Wilhelm-Benartzi, D. H. You, V. A. Grote, A. Tjonneland, C. C. Dahm, K. Overvad, M. Jenab, V. Fedirko, M. C. Boutron-Ruault, F. Clavel-Chapelon, A. Racine, R. Kaaks, H. Boeing, J. Foerster, A. Trichopoulou, P. Lagiou, D. Trichopoulos, C. Sacerdote, S. Sieri, D. Palli, R. Tumino, S. Panico, P. D. Siersema, P. H. Peeters, E. Lund, A. Barricarte, J. M. Huerta, E. Molina-Montes, M. Dorronsoro, J. R. Quiros, E. J. Duell, W. Ye, M. Sund, B. Lindkvist, D. Johansen, K. T. Khaw, N. Wareham, R. C. Travis, P. Vineis, H. B. Bueno-de-Mesquita and E. Riboli, *Gut*, 2012, **62**(12), 1764–1770.
85. V. A. Grote, R. Kaaks, A. Nieters, A. Tjonneland, J. Halkjaer, K. Overvad, M. R. Skjelbo Nielsen, M. C. Boutron-Ruault, F. Clavel-Chapelon, A. Racine, B. Teucher, S. Becker, T. Pischon, H. Boeing, A. Trichopoulou, C. Cassapa, V. Stratigakou, D. Palli, V. Krogh, R. Tumino, P. Vineis,

S. Panico, L. Rodriguez, E. J. Duell, M. J. Sanchez, M. Dorronsoro, C. Navarro, A. B. Gurrea, P. D. Siersema, P. H. Peeters, W. Ye, M. Sund, B. Lindkvist, D. Johansen, K. T. Khaw, N. Wareham, N. E. Allen, R. C. Travis, V. Fedirko, M. Jenab, D. S. Michaud, S. C. Chuang, D. Romaguera, H. B. Bueno-de-Mesquita and S. Rohrmann, *Br. J. Cancer*, 2012, **106**, 1866–1874.

86. S. Tejpar, M. Bertagnolli, F. Bosman, H. J. Lenz, L. Garraway, F. Waldman, R. Warren, A. Bild, D. Collins-Brennan, H. Hahn, D. P. Harkin, R. Kennedy, M. Ilyas, H. Morreau, V. Proutski, C. Swanton, I. Tomlinson, M. Delorenzi, R. Fiocca, E. Van Cutsem and A. Roth, *Oncologist*, 2010, **15**, 390–404.
87. N. Miura, Y. Osaki, M. Nagashima, M. Kohno, K. Yorozu, K. Shomori, T. Kanbe, K. Oyama, Y. Kishimoto, S. Maruyama, E. Noma, Y. Horie, M. Kudo, S. Sakaguchi, Y. Hirooka, H. Ito, H. Kawasaki, J. Hasegawa and G. Shiota, *BMC Gastroenterol.*, 2010, **10**, 46.
88. Y. Sun, Z. Zang, X. Xu, Z. Zhang, L. Zhong, W. Zan, Y. Zhao and L. Sun, *Int. J. Mol. Sci.*, 2010, **11**, 1423–1433.
89. B. Hamrita, K. Chahed, M. Trimeche, C. L. Guillier, P. Hammann, A. Chaieb, S. Korbi and L. Chouchane, *Clin. Chim. Acta*, 2009, **404**, 111–118.
90. M. A. Fonseca-Sanchez, S. Rodriguez Cuevas, G. Mendoza-Hernandez, V. Bautista-Pina, E. Arechaga Ocampo, A. Hidalgo Miranda, V. Quintanar Jurado, L. A. Marchat, E. Alvarez-Sanchez, C. Perez Plasencia and C. Lopez-Camarillo, *Int. J. Oncol.*, 2012, **41**, 670–680.
91. B. Hamrita, H. Ben Nasr, P. Hammann, L. Kuhn, A. Ben Anes, S. Dimassi, A. Chaieb, H. Khairi and K. Chahed, *Ann. Biol. Clin.*, 2012, **70**, 553–565.
92. B. Hamrita, H. B. Nasr, K. Chahed and L. Chouchane, *Gulf J. Oncolog.*, 2011, **1**, 36–44.
93. Y. M. Chung, Y. D. Yoo, J. K. Park, Y. T. Kim and H. J. Kim, *Anticancer Res.*, 2001, **21**, 1129–1133.
94. F. Cirillo, G. Nassa, R. Tarallo, C. Stellato, M. R. De Filippo, C. Ambrosino, M. Baumann, T. A. Nyman and A. Weisz, *J. Proteome Res.*, 2013, **12**, 421–431.
95. N. Q. Liu, R. B. Braakman, C. Stingl, T. M. Luider, J. W. Martens, J. A. Foekens and A. Umar, *J. Mammary Gland Biol. Neoplasia*, 2012, **17**, 155–164.
96. A. M. Gonzalez-Angulo, B. T. Hennessy, F. Meric-Bernstam, A. Sahin, W. Liu, Z. Ju, M. S. Carey, S. Myhre, C. Speers, L. Deng, R. Broaddus, A. Lluch, S. Aparicio, P. Brown, L. Pusztai, W. F. Symmans, J. Alsner, J. Overgaard, A. L. Borresen-Dale, G. N. Hortobagyi, K. R. Coombes and G. B. Mills, *Clin. Proteomics*, 2011, **8**, 11.
97. D. M. Pedrotty, M. P. Morley and T. P. Cappola, *Prog. Cardiovasc. Dis.*, 2012, **55**, 64–69.
98. L. Desire, E. Blondiaux, J. Carriere, R. Haddad, O. Sol, P. Fehlbaum-Beurdeley, R. Einstein, W. Zhou and M. P. Pando, *J. Alzheimer's Dis.*, 2013, **34**, 469–483.

99. C. Q. Yi, C. H. Ma, Z. P. Xie, Y. Cao, G. Q. Zhang, X. K. Zhou and Z. Q. Liu, *GMR, Genet. Mol. Res.*, 2013, **12**.
100. J. Caetano-Lopes, A. Rodrigues, A. Lopes, A. C. Vale, M. A. Pitts-Kiefer, B. Vidal, I. P. Perpetuo, J. Monteiro, Y. T. Konttinen, M. F. Vaz, A. Nazarian, H. Canhao and J. E. Fonseca, *Clin. Rev. Allergy Immunol.*, 2013.
101. J. C. Laguna and M. Alegret, *Pharmacogenomics*, 2012, **13**, 477–495.
102. H. Peiris, R. Raghupathi, C. F. Jessup, M. P. Zanin, D. Mohanasundaram, K. D. Mackenzie, T. Chataway, J. N. Clarke, J. Brealey, P. T. Coates, M. A. Pritchard and D. J. Keating, *Endocrinology*, 2012, **153**, 5212–5221.
103. T. S. Nielsen, U. Kampmann, R. R. Nielsen, N. Jessen, L. Orskov, S. B. Pedersen, J. O. Jorgensen, S. Lund and N. Moller, *J. Clin. Endocrinol. Metab.*, 2012, **97**, E1348–1352.
104. J. M. Swanson, G. C. Wood, L. Xu, L. E. Tang, B. Meibohm, R. Homayouni, M. A. Croce and T. C. Fabian, *PloS One*, 2012, **7**, e42065.
105. I. Martin-Loeches, E. Papiol, R. Almansa, G. Lopez-Campos, J. F. Bermejo-Martin and J. Rello, *Medicina Intensiva*, 2012, **36**, 257–263.
106. S. Jain and P. Heutink, *Neuron*, 2010, **68**, 207–217.
107. E. Borgan, B. Sitter, O. C. Lingjaerde, H. Johnsen, S. Lundgren, T. F. Bathen, T. Sorlie, A. L. Borresen-Dale and I. S. Gribbestad, *BMC Cancer*, 2010, **10**, 628.
108. J. Hardouin, J. P. Lasserre, L. Sylvius, R. Joubert-Caron and M. Caron, *Ann. N. Y. Acad. Sci.*, 2007, **1107**, 223–230.
109. M. R. Amin, M. S. Siddiqui, D. Ahmed, F. Ahmed and A. Hossain, *Int. J. Bioinf. Res. Appl.*, 2011, **7**, 287–298.
110. M. Careaga and P. Ashwood, *Methods Mol. Biol.*, 2012, **934**, 219–240.
111. S. D. Funk and A. W. Orr, *Pharmacol. Res.*, 2013, **67**, 42–52.
112. J. I. Odegaard and A. Chawla, *Cold Spring Harbor Perspect. Med.*, 2012, **2**, a007724.
113. H. S. Kim and M. S. Lee, *Curr. Mol. Med.*, 2009, **9**, 30–44.
114. M. Gonzalez-Juarrero, *Immunotherapy*, 2012, **4**, 187–199.
115. A. J. Battersby, B. Kampmann and S. Burl, *Clin. Dev. Immunol.*, 2012, **2012**, 430972.
116. A. T. Weeraratna, A. Kalehua, I. Deleon, D. Bertak, G. Maher, M. S. Wade, A. Lustig, K. G. Becker, W. Wood, 3rd, D. G. Walker, T. G. Beach and D. D. Taub, *Exp. Cell Res.*, 2007, **313**, 450–461.
117. R. P. Young and R. J. Hopkins, *Cancer*, 2013, **119**, 1761.
118. Y. Godet, E. Fabre, M. Dosset, M. Lamuraglia, E. Levionnois, P. Ravel, N. Benhamouda, A. Cazes, F. Le Pimpec-Barthes, B. Gaugler, P. Langlade-Demoyen, X. Pivot, P. Saas, B. Maillere, E. Tartour, C. Borg and O. Adotevi, *Clin. Cancer Res.*, 2012, **18**, 2943–2953.
119. H. Wang, V. K. Tso, C. M. Slupsky and R. N. Fedorak, *Future Oncol.*, 2010, **6**, 1395–1406.
120. Y. He, Z. Yu, I. Giegling, L. Xie, A. M. Hartmann, C. Prehn, J. Adamski, R. Kahn, Y. Li, T. Illig, R. Wang-Sattler and D. Rujescu, *Transl. Psychiatry*, 2012, **2**, e149.

121. H. Janeckova, K. Hron, P. Wojtowicz, E. Hlidkova, A. Baresova, D. Friedecky, L. Zidkova, P. Hornik, D. Behulova, D. Prochazkova, H. Vinohradska, K. Peskova, P. Bruheim, V. Smolka, S. Stastna and T. Adam, *J. Chrom. A*, 2012, **1226**, 11–17.
122. H. Song, L. Wang, H. L. Liu, X. B. Wu, H. S. Wang, Z. H. Liu, Y. Li, D. C. Diao, H. L. Chen and J. S. Peng, *Oncol. Rep.*, 2011, **26**, 431–438.
123. L. Pedersen, J. Elers and V. Backer, *Phys. Sportsmed.*, 2011, **39**, 163–171.
124. H. Rose, H. Low, E. Dewar, M. Bukrinsky, J. Hoy, A. Dart and D. Sviridov, *Atherosclerosis*, 2013, **229**(1), 206–211.
125. R. A. Koeth, Z. Wang, B. S. Levison, J. A. Buffa, E. Org, B. T. Sheehy, E. B. Britt, X. Fu, Y. Wu, L. Li, J. D. Smith, J. A. Didonato, J. Chen, H. Li, G. D. Wu, J. D. Lewis, M. Warrier, J. M. Brown, R. M. Krauss, W. H. Tang, F. D. Bushman, A. J. Lusis and S. L. Hazen, *Nat. Med.*, 2013, **19**, 576–585.
126. N. Friedrich, *J. Endocrinol.*, 2012, **215**, 29–42.
127. R. Wang-Sattler, Z. Yu, C. Herder, A. C. Messias, A. Floegel, Y. He, K. Heim, M. Campillos, C. Holzapfel, B. Thorand, H. Grallert, T. Xu, E. Bader, C. Huth, K. Mittelstrass, A. Doring, C. Meisinger, C. Gieger, C. Prehn, W. Roemisch-Margl, M. Carstensen, L. Xie, H. Yamanaka-Okumura, G. Xing, U. Ceglarek, J. Thiery, G. Giani, H. Lickert, X. Lin, Y. Li, H. Boeing, H. G. Joost, M. H. de Angelis, W. Rathmann, K. Suhre, H. Prokisch, A. Peters, T. Meitinger, M. Roden, H. E. Wichmann, T. Pischon, J. Adamski and T. Illig, *Mol. Syst. Biol.*, 2012, **8**, 615.
128. I. du Preez and T. du Loots, *OMICS*, 2012, **16**, 596–603.
129. C. Ficicioglu and D. Bearden, *Pediatr. Neur.*, 2011, **45**, 283–291.
130. T. Kobayashi, S. Nishiumi, A. Ikeda, T. Yoshie, A. Sakai, A. Matsubara, Y. Izumi, H. Tsumura, M. Tsuda, H. Nishisaki, N. Hayashi, S. Kawano, Y. Fujiwara, H. Minami, T. Takenawa, T. Azuma and M. Yoshida, *Cancer Epidemiol., Biomarkers Prev.*, 2013, **22**, 571–579.
131. M. Verma, *Technol. Cancer Res. Treat.*, 2005, **4**, 295–301.
132. S. Khare and M. Verma, *Methods Mol. Biol. (Clifton, N.J.)*, 2012, **863**, 177.
133. M. D. Anway and M. K. Skinner, *Endocrinology*, 2006, **147**, S43–S49.
134. F. Fang, S. Turcan, A. Rimner, A. Kaufman, D. Giri, L. G. Morris, R. Shen, V. Seshan, Q. Mo, A. Heguy, S. B. Baylin, N. Ahuja, A. Viale, J. Massague, L. Norton, L. T. Vahdat, M. E. Moynahan and T. A. Chan, *Sci. Transl. Med.*, 2011, **3**, 75ra25.
135. J. M. Yi, M. Dhir, L. Van Neste, S. R. Downing, J. Jeschke, S. C. Glockner, M. de Freitas Calmon, C. M. Hooker, J. M. Funes, C. Boshoff, K. M. Smits, M. van Engeland, M. P. Weijenberg, C. A. Iacobuzio-Donahue, J. G. Herman, K. E. Schuebel, S. B. Baylin and N. Ahuja, *Clin. Cancer Res.*, 2011, **17**, 1535–1545.
136. L. Yu, N. W. Todd, L. Xing, Y. Xie, H. Zhang, Z. Liu, H. Fang, J. Zhang, R. L. Katz and F. Jiang, *Int. J. Cancer*, 2010, **127**(12), 2870–2878.
137. A. Harder, S. Titze, L. Herbst, T. Harder, K. Guse, S. Tinschert, D. Kaufmann, T. Rosenbaum, V. F. Mautner, E. Windt, U. Wahllander-Danek, K. Wimmer, S. Mundlos and H. Peters, *Twin Res. Hum. Genet.*, 2010, **13**, 582–594.

138. J. Ashley-Martin, J. VanLeeuwen, A. Cribb, P. Andreou and J. R. Guernsey, *Int. J. Environ. Res. Public Health*, 2012, **9**, 1846–1858.
139. J. Qiu, R. Yang, Y. Rao, Y. Du and F. W. Kalembo, *PloS One*, 2012, **7**, e36497.
140. M. Alshalalfa, *Adv. Bioinf.*, 2012, **2012**, 839837.
141. K. W. Chang, S. Y. Kao, Y. H. Wu, M. M. Tsai, H. F. Tu, C. J. Liu, M. T. Lui and S. C. Lin, *Oral Oncol.*, 2013, **49**, 27–33.
142. N. Nishida, M. Nagahara, T. Sato, K. Mimori, T. Sudo, F. Tanaka, K. Shibata, H. Ishii, K. Sugihara, Y. Doki and M. Mori, *Clin. Cancer Res.*, 2012, **18**, 3054–3070.
143. P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin and M. Tewari, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 10513–10518.
144. C. Kumagai, B. Kalman, F. A. Middleton, T. Vyshkina and P. T. Massa, *J. Neuroimmunol.*, 2012, **246**, 51–57.
145. J. Zhao, C. W. Forsberg, J. Goldberg, N. L. Smith and V. Vaccarino, *BMC Med. Genet.*, 2012, **13**, 100.
146. L. N. Zhang, P. P. Liu, L. Wang, F. Yuan, L. Xu, Y. Xin, L. J. Fei, Q. L. Zhong, Y. Huang, L. M. Hao, X. J. Qiu, Y. Le, M. Ye and S. Duan, *PloS One*, 2013, **8**, e63455.
147. V. K. Rakyan, H. Beyan, T. A. Down, M. I. Hawa, S. Maslau, D. Aden, A. Daunay, F. Busato, C. A. Mein, B. Manfras, K. R. Dias, C. G. Bell, J. Tost, B. O. Boehm, S. Beck and R. D. Leslie, *PLoS Genetics*, 2011, **B**, e1002300.
148. L. M. Villeneuve, M. A. Reddy, L. L. Lanting, M. Wang, L. Meng and R. Natarajan, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 9047–9052.
149. C. Andraos, G. Koorsen, J. C. Knight and L. Bornman, *Hum. Immunol.*, 2011, **72**, 262–268.
150. V. Bollati, D. Galimberti, L. Pergoli, E. Dalla Valle, F. Barretta, F. Cortini, E. Scarpini, P. A. Bertazzi and A. Baccarelli, *Brain, Behav., Immun.*, 2011, **25**, 1078–1083.
151. R. P. Nagarajan, K. A. Patzel, M. Martin, D. H. Yasui, S. E. Swanberg, I. Hertz-Picciotto, R. L. Hansen, J. Van de Water, I. N. Pessah, R. Jiang, W. P. Robinson and J. M. LaSalle, *Autism Res.*, 2008, **1**, 169–178.
152. S. Saji, M. Kawakami, S. Hayashi, N. Yoshida, M. Hirose, S. Horiguchi, A. Itoh, N. Funata, S. L. Schreiber, M. Yoshida and M. Toi, *Oncogene*, 2005, **24**, 4531–4539.
153. Z. Zhang, H. Yamashita, T. Toyama, H. Sugiura, Y. Ando, K. Mita, M. Hamaguchi, Y. Hara, S. Kobayashi and H. Iwase, *Breast Cancer Res. Treat.*, 2005, **94**, 11–16.
154. Y. Ruike, Y. Imanaka, F. Sato, K. Shimizu and G. Tsujimoto, Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing, *BMC Genomics*, 2010, **11**, 137.

155. J. Soares, A. E. Pinto, C. V. Cunha, S. Andre, I. Barao, J. M. Sousa and M. Cravo, *Cancer*, 1999, **85**, 112–118.
156. K. Jackson, M. C. Yu, K. Arakawa, E. Fiala, B. Youn, H. Fiegl, E. Muller-Holzner, M. Widschwendter and M. Ehrlich, *Cancer Biol. Ther.*, 2004, **3**, 1225–1231.
157. A. Narayan, W. Ji, X. Y. Zhang, A. Marrogi, J. R. Graff, S. B. Baylin and M. Ehrlich, *Int. J. Cancer*, 1998, **77**, 833–838.
158. S. L. Zhou and L. D. Wang, *World J. Gastroenterol.*, 2010, **16**, 2348–2354.
159. I. Van der Auwera, R. Limame, P. van Dam, P. B. Vermeulen, L. Y. Dirix and S. J. Van Laere, *Br. J. Cancer*, 2010, **103**, 532–541.
160. A. M. Abdel Rahman, S. D. Kamath, S. Gagne, A. L. Lopata and R. Helleur, *J. Proteome Res.*, 2013, **12**, 647–656.
161. T. L. Mertz, *J. Am. Osteopath. Assoc.*, 2011, **111**(S27–S29), S32.
162. M. J. Chao, S. V. Ramagopalan, B. M. Herrera, M. R. Lincoln, D. A. Dyment, A. D. Sadovnick and G. C. Ebers, *Hum. Mol. Genet.*, 2009, **18**, 261–266.
163. F. Li, S. Li, H. Chang, Y. Nie, L. Zeng, X. Zhang and Y. Wang, *Genet. Test. Mol. Biomarkers*, 2013, **17**(9), 656–661.
164. D. R. Taylor, *Thorax*, 2009, **64**, 261–264.
165. M. G. Tektonidou and M. M. Ward, *Nat. Rev. Rheumatol.*, 2011, **7**, 708–717.
166. G. B. Lim, *Nat. Rev. Cardiol.*, 2012, **9**, 672.
167. I. Rahman, R. Atout, N. L. Pedersen, U. de Faire, J. Frostegard, E. Ninio, A. M. Bennet and P. K. Magnusson, *Atherosclerosis*, 2011, **218**, 117–122.
168. S. Sour, M. Belarbi, D. Khaldi, N. Benmansour, N. Sari, A. Nani, F. Chemat and F. Visioli, *Br. J. Nutr.*, 2012, **107**, 1800–1805.
169. I. Okur, L. Tumer, F. S. Ezgu, E. Yesilkaya, A. Aral, S. O. Oktar, A. Bideci and A. Hasanoglu, *J. Pediatr. Endocrinol. Metab.*, 2013, **26**(7–8), 657–662.
170. M. Kucukazman, N. Ata, B. Yavuz, K. Dal, O. Sen, O. S. Deveci, K. Agladioglu, A. O. Yeniova, Y. Nazligul and D. T. Ertugrul, *Eur. J. Gastroenterol. Hepatol.*, 2013, **25**, 147–151.
171. S. G. Wannamethee, P. H. Whincup, L. Lennon and N. Sattar, *Arch. Intern. Med.*, 2007, **167**, 1510–1517.
172. W. Mullen, C. Delles and H. Mischak, *Curr. Opin. Nephrol. Hypertens.*, 2011, **20**, 654–661.
173. *Duke Medicine Health News*, 2012, **18**, 4–5.
174. K. Nakanishi and C. Watanabe, *Clin. Chim. Acta*, 2009, **402**, 171–175.
175. A. Maisel, S. X. Neath, J. Landsberg, C. Mueller, R. M. Nowak, W. F. Peacock, P. Ponikowski, M. Mockel, C. Hogan, A. H. Wu, M. Richards, P. Clopton, G. S. Filippatos, S. Di Somma, I. Anand, L. L. Ng, L. B. Daniels, R. H. Christenson, M. Potocki, J. McCord, G. Terracciano, O. Hartmann, A. Bergmann, N. G. Morgenthaler and S. D. Anker, *Eur. J. Heart Failure*, 2012, **14**, 278–286.
176. D. P. Dosanjh, M. Bakir, K. A. Millington, A. Soysal, Y. Aslan, S. Efee, J. J. Deeks and A. Lalvani, *PloS One*, 2011, **6**, e28754.

177. R. Ribeiro-Rodrigues, T. Resende, Co, J. L. Johnson, F. Ribeiro, M. Palaci, R. T. Sa, E. L. Maciel, F. E. Pereira Lima, V. Dettoni, Z. Toossi, W. H. Boom, R. Dietze, J. J. Ellner and C. S. Hirsch, *Clin. Diagn. Lab. Immunol.*, 2002, **9**, 818–823.
178. M. Sakamoto, *J. Gastroenterol.*, 2009, **44**(Suppl 19), 108–111.
179. I. Simon, Y. Liu, K. L. Krall, N. Urban, R. L. Wolfert, N. W. Kim and M. W. McIntosh, *Gynecol. Oncol.*, 2007, **106**, 112–118.
180. J. N. Mubiru, A. J. Valente and D. A. Troyer, *The Prostate*, 2005, **65**, 117–123.
181. D. J. Birnbaum, S. Laibe, A. Ferrari, A. Lagarde, A. J. Fabre, G. Monges, D. Birnbaum and S. Olschwang, *Transl. Oncol.*, 2012, **5**, 72–76.
182. F. Al Dayel, *J. Infect. Public Health*, 2012, **5**(Suppl 1), S31–S34.
183. H. Takeda, N. Takigawa, K. Ohashi, D. Minami, I. Kataoka, E. Ichihara, N. Ochi, M. Tanimoto and K. Kiura, *Exp. Cell Res.*, 2013, **319**, 417–423.
184. S. Bernholtz, Y. Laitman, B. Kaufman, S. Shimon-Paluch and E. Friedman, *Breast Cancer Res. Treat.*, 2012, **132**, 669–673.
185. M. Gonen, Z. Sun, M. E. Figueroa, J. P. Patel, O. Abdel-Wahab, J. Racevskis, R. P. Ketterling, H. Fernandez, J. M. Rowe, M. S. Tallman, A. Melnick, R. L. Levine and E. Paietta, *Blood*, 2012, **120**, 2297–2306.
186. R. Claus, B. Hackanson, A. R. Poetsch, M. Zucknick, M. Sonnet, N. Blagitko-Dorfs, J. Hiller, S. Wilop, T. H. Brummendorf, O. Galm, U. Platzbecker, J. C. Byrd, K. Dohner, H. Dohner, M. Lubbert and C. Plass, *Int. J. Cancer*, 2012, **131**, E138–E142.
187. P. A. Ho, M. A. Kutny, T. A. Alonzo, R. B. Gerbing, J. Joaquin, S. C. Raimondi, A. S. Gamis and S. Meshinchi, *Pediatr. Blood Cancer*, 2011, **57**, 204–209.
188. C. A. Krusche, P. Wulfing, C. Kersting, A. Vloet, W. Bocker, L. Kiesel, H. M. Beier and J. Alfer, *Breast Cancer Res. Treat.*, 2005, **90**, 15–23.
189. W. M. Linehan, *Genome Res.*, 2012, **22**, 2089–2100.
190. J. Hur, H. J. Lee, J. E. Nam, Y. J. Kim, Y. J. Hong, H. Y. Kim, S. K. Kim, J. Chang, J. H. Kim, K. Y. Chung, H. S. Lee and B. W. Choi, *BMC Cancer*, 2012, **12**, 392.

CHAPTER 14

Next-generation Molecular Markers: Challenges, Applications, and Future Perspectives

MUKESH VERMA,[a] DEBMALYA BARH,*[b] SYED SHAH HASSAN[c] AND VASCO AC AZEVEDO[c]

[a] Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute (NCI), National Institutes of Health (NIH), 9609 Medical Center Drive, Rockville, MD 20850, USA; [b] Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal-721172, India; [c] Departamento de Biologia Geral, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Minas Gerais, Brazil
*Email: dr.barh@gmail.com

14.1 INTRODUCTION

14.1.1 Biological Limitations in Cancer Biomarker Discovery

Metabolomics represent the phenotypic state of a disease.[1] Here we describe the challenges in translational research when metabolomic biomarkers are used for disease diagnosis, risk assessment and prognosis. One barrier is a paucity of commercially available standards for identification and quantification of metabolites for humans, as well as data comparisons across studies.[2] Particular consideration should be given to the quantity needed and which compounds should be included. Additionally, the lack of carefully

Molecular Biology and Biotechnology, 6th Edition
Edited by Ralph Rapley and David Whitehouse

Published by the Royal Society of Chemistry, www.rsc.org

selected, well annotated, and easily accessible reference samples greatly limits investigations. Although the standard National Institute of Standards and Technology (NIST) pooled plasma reference sets are available, they are of limited value in determining the individual metabolite variation. Standards for other biological media, such as urine, are needed for investigating the most physiologically plausible pathways that best reflect the etiology of disease. In the case of epidemiologic consortia, different laboratories may be involved in sample analyses; and in turn, inter-laboratory comparisons are problematic without reference samples. An opportunity surrounds the use of archived samples from population-based studies to gain insights into optimizing collection and storage protocols for different media for sample integrity. In addition to high quality samples, quantitative robustness requires provision for quality controls, pooled references, and standard reference materials to control for instrumentation variability drift and allow for comparisons across laboratories. Well-standardized protocols for sample collection, storage, and analysis are also needed.

Cancer cells show progressive heterogeneity at cellular and molecular level. Transient expression of various intermediated components (proteins/ genes/ metabolites *etc.*) during intermediate stages of carcinogenesis affects detection of biomarkers resulting false positive or negative. Similarly, biomarkers may get affected in response to therapy and other internal physiological and pathological factors including age of the subject. Other influential factors in biomarker variation include food habit, nutrition, and, life style.

14.1.2 Clinical and Pathologic Factors in Cancer Biomarker Discovery

Although a considerable amount of knowledge has been obtained in understanding disease biology and identification of biomarkers which can detect a disease, implication of that information in clinic is still challenging. Clinical validation is the main hurdle in the process. Just to make our point clear we are describing an example in cancer. In one case control study of the European Prospective Investigation into Cancer and nutrition (EPIC) where more than 300 breast cancer patients and matched controls were tested for breast cancer over a period of three years using a panel of eight serum markers (osteopontin, haptoglobin, cancer antigen 15-3, carcinoembryonic antigen, cancer antigen-125, prolactin, cancer antigen 19-9, and alpha-fetoprotein), very low specificity (50%) and sensitivity (50%) were observed.[3] This may be due to different subtypes of breast cancer present in collected samples. Such epidemiologic studies should select a broader target set of potential biomarkers which could be enabled by antibody array technologies where profiles of up to 100 antibodies can be followed simultaneously. Making different groups, based on subtypes of cancer based on the status of hormone receptors (estrogen and progesterone), might also be helpful.

Although current biomarkers can differentiate normal versus cancer condition; more accurate detection of each stages of cancer such as

initiation, progression, metastasis, early and aggressive stages, and recurrence are required to be developed.

14.1.3 Analytical Limitations in Cancer Biomarker Discovery

Analytical limitation exists in different biomarkers mainly because they have not been validated. Another area that needs progress/attention is the cost and high throughput. Systematic and adequate progress has not been made is the application of biomarkers in clinic. Proper analytical and clinical validation of early markers has not been achieved and the number of biomarkers approved by the FDA for clinical use is very small.[4] Clinical validation of identified biomarkers especially is the key challenge in the field. Here is one example to emphasize our point. The National Cancer Institute has developed guidelines for the analytical and clinical validation of biomarkers but none of the biomarkers have been validated to date.[5] Integration of genomic and proteomic markers with epigenetic markers may help us subtyping different cancers and cancer stages (and this is true for other diseases also).[6] Values of biomarkers may be different in different tissues. For example, results of methylation profiling from blood and tissues are different. Koestler *et al.* conducted a systematic epigenome-wide methylation analysis and demonstrated that shifts in leukocyte subpopulations might account for a considerable proportion of variability in these patterns.[7] Multiplexing of biomarkers may reduce false positive results in screening studies where intention is to identify populations which are at high risk of developing a disease. Quantitative imaging data storage and maintenance have their own challenges as we discussed above. Whether miR expression is localized in a specific part of the diseased tissue has to be carefully evaluated.[2] In a tissue biopsy the local concentration (number of miRs) may be low or high. Determining the accurate level of miRs is very critical.

Although opportunities exist for the use of metabolomic profiling in population-based research, multiple challenges prevent the proper integration of these data into epidemiologic studies for meaningful interpretation.[1,2] Therefore, epidemiologists should strive to understand the principles of metabolomics to better determine and apply their appropriate uses. In addition, it is essential for metabolite profiles to be validated both for analytical purposes and clinical use as biomarkers. Some of the primary challenges to population-based studies include the incorporation of this technology in the initial study design, identifying appropriate sample collection protocols and quality control methods, and the selection of analytical approaches and quantification techniques. New strategies will likely be necessary for combining data from different analytical platforms to allow generalizability of data and interpretation. Furthermore, the need to develop better methods for analyzing large amounts of data remains a critical barrier, including improvement of statistical and bioinformatics methods for data analysis.

Unfortunately, metabolomic profiling is used far too frequently *post hoc* in epidemiology studies. Investigators should strive for well-designed, prospective studies that would establish causal effect, as well as temporal

changes. In any case, causality is difficult to establish with a clear biological explanation in association studies, even with particularly well-designed studies. Therefore, it is critical to improve methodologies for integration of metabolomics data with other data, such as genomic and proteomic, to help understand the functional and temporal relationship between a biomarker and an effect.

In general, the concentrations of biomarkers are very low in biospecimens collected from patients therefore quantification of markers requires highly sensitive assays.[2] Standard procedures, reference materials, and quality controls need to be strictly followed to ensure accuracy as well as reproducibility of the assay for a given biomarker. Unfortunately, such good manufacturing/laboratory practice (GMP/GLP) and quality control are not followed in most cases in biomarker discovery.

14.1.4 Intellectual Property in Cancer Biomarker Discovery

The National Institutes of Health (NIH) provides an opportunity to investigators to keep IP rights although the NIH supports the project. Leaving regulatory, financial, intellectual property, and cultural issues aside, developing a diagnostic biomarker often requires expertise or patients that its discoverer may not possess. We know very well that discovery, development, and validation of a biomarker is costly and no assurance that a project will succeed. Therefore, investors are often not interested in investing in such projects. However, investors may agree if they are assured a good return on their investment by means of patent protection of novel biomarkers of high sensitivity, efficacy, and clinical utility.

14.1.5 Health Economy Factors in Cancer Biomarker Discovery

For a better treatment outcome, the procedures should be inexpensive. Most of the novel and useful biomarkers and assay methods are patent protected. Therefore, the cost of diagnostic and prognostic assays using such biomarkers is very high and usually not affordable by common people. Cost effective assays need to be developed to meet the needs of common people.

14.2 NEXT-GENERATION MOLECULAR MARKERS AND THEIR APPLICATIONS

Affinity reagents are considered the most significant tools for biomarker discovery. Affimers are a type of affinity reagents, which are relatively small (13 kDa), non-posttranslationally modified, biophysically stable protein scaffolds.[8] They contain three variable regions into which distinct peptides are inserted. This structure then binds to proteins and other molecules the way an antigen and antibody binds. So far 20 000 Affimer arrays have been tested in serum of RA patients and elevated levels of biomarkers such as C-reactive protein have been identified along with more than 20 new biomarkers. The detection and treatment of different diseases is greatly facilitated by the omics technologies. For example, in cancer, genomics

analysis provides clue for gene regulation and gene knockdown for cancer management. Discovery of the involvement of microRNAs in human cancers has opened a new page for cancer researchers.[9,10] The latest technologies are Mammaprint and Oncotype DX, which are being used in clinics confirming the feasibility of multiplexing of techniques.[11,12] Some therapeutic drugs, called epigenetic drugs, target DNA methylation and histone deacetylation in solid tumors and leukemia. These drugs work well with PARP inhibitors.[13] Proteomics also plays an important role in cancer biomarker discovery and quantitative proteome–disease relationships provide a means for connectivity analysis.[14] Fluorescent dye enables a more reliable analysis and it facilitates the progress of biochip and cytomics. The huge amount of information collected by multiparameter single cell flow or slide-based cytometry measurements serves to investigate the molecular behavior of cancer cell populations. Diseases selected in this article are very appropriate for metabolite profiling owing to their unique biochemical properties.

In Figure 14.1, we have presented a schematic approach to demonstrate factors contributing to disease development and approaches to isolate clinical samples and analyze biomarkers for disease detection, diagnosis, prognosis, and drug response. Table 14.1 highlights key biomarkers in different diseases and their application in detection, diagnosis, screening, prognosis, and treatment response.

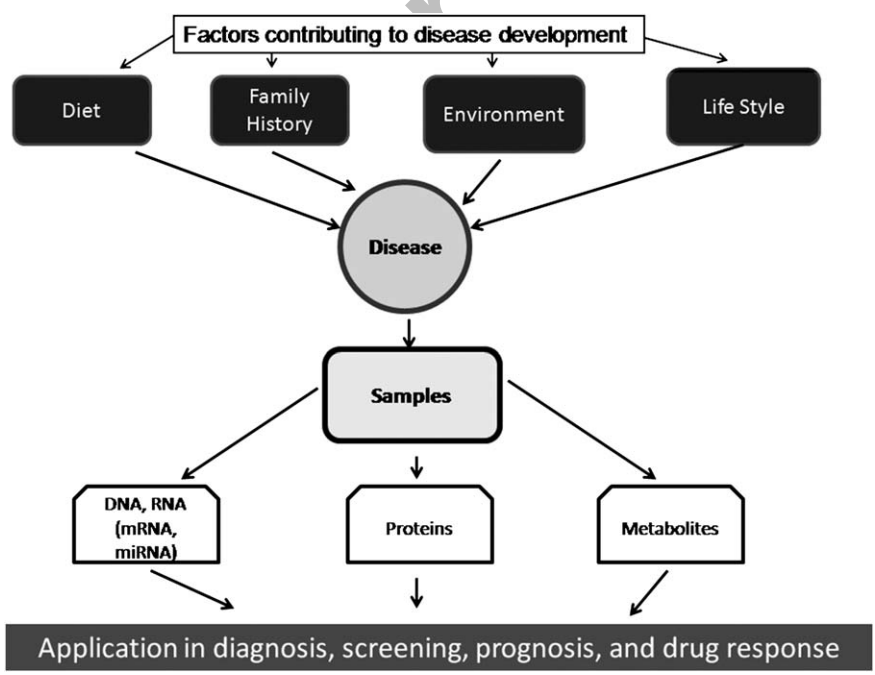


Figure 14.1 Schematic diagram to demonstrate factors contributing to disease development and approaches to isolate clinical samples and analyze biomarkers for disease detection, diagnosis, prognosis, and drug response.

Table 14.1 Next-Generation Molecular Markers for Diagnosis Screening, Early Diagnosis, Risk Assessment, and Prognosis.

Category	Diseases	Markers
Diagnostic Screening		
	Immunologic Diseases	Arginine kinase, sarcoplasmic calcium binding proteins, tropomycin.[15]
	Cardiovascular disease (CVD)	Lipoprotein associated phospholipase A2, antiphosphorylcholine IgM.[16]
	Metabolic Diseases	Glycine, glutamine, glycerophasphatidyl choline.[18]
	Infectious Diseases	Hepatitis B virus (HBV), hepatitis C virus (HCV), certain strains of the human papillomavirus (HPV), Epstein–Barr virus (EBV), human immunodeficiency virus type 1 (HIV-1), human T-cell lymphotropic virus type-1 (HTLV-1), and the gram-negative bacterium *Helicobacter pylori* (*H. pylori*).[24]
	Neurologic Diseases	Proteomic profiling of serum (multiple biomarkers).[29]
	Cancer	Gene promoter hypermethylation in death-associated protein kinase 1 (*DAPK1*), *p16* and *RASSF1A*1 could be used as biomarkers for detection of HNSCCs; prostate-specific antigen (PSA) for prostate cancer; serum markers (PAP, tPSA, fPSA, proPSA, PSAD, PSAV, PSADT, EPCA, and EPCA-2), tissue markers (AMACR, methylated GSTP1, and the TMPRSS2-ETS gene rearrangement), and a urine marker (DD3PCA3/UPM-3) for prostate cancer.[35–42] Serum tumor markers like carcinoembryonic antigen (CEA) and cancer antigen (CA) CA 15.3, CA 27.29 for breast and ovarian cancer.[44]
		Using saliva as the starting material, mRNAs specific for esophageal, head and neck, oral and lung cancer also were identified.[48–50] MSI biomarkers were reported in a number of cancers, including bladder, colon, esophageal, and skin cancer.[54–56]
Early Diagnosis		
	Immunologic Diseases	Glutathione S transferase M1 (GSTM1) biomarker was identified which helped in screening a cohort of children who were at high risk of developing asthma;[73] C-reactive protein, fibrinogen, and interleukin-6 for screening.[74]
	Cardiovascular disease (CVD)	Flow-mediated dialation (FMD) and IMT for atherosclerosis.[60]
	Metabolic Diseases	Dysglycemia, alphahydroxybutyrate, linoleyolglycerophosphocholine, advanced glycation end products (AGE), albumin excretion rate (AR) and SNPS in epidermal growth factor gene intron 2.[1,62,63]

Table 14.1 (*Continued*)

Category	Diseases	Markers
	Infectious Diseases	Region-of-Difference-1 (RD-1) gene product and sputum cytokine levels for early detection of tuberculosis;[64] HSP70, CAP2, glypican 3 and glutamine synthase[65] for HCC.
	Neurologic Diseases	Beta amyloid for early detection of Alzheimer's disease.[66]
	Cancer	B7-H4, spondin 2, and DcR3 were identified as early biomarkers in ovarian cancer.[68–70]
Risk Assessment (screening)		
	Immunologic Diseases	C-reactive protein, fibrinogen, and interleukin-6.[74]
	Cardiovascular disease (CVD)	C-reactive protein.[75]
	Metabolic Diseases	Alanine amino transferase (ALT), gamma glutamyl transferase (GGT), triglyceride, plasminogen activator inhibitor (PAI-1) antigen, ferritin, C-reactive protein (CRP), sex-hormone binding globulin (SHBG), ALT and GGT for diabetes.[78]
	Infectious Diseases	HIV infection for AIDS;[80,81] gamma interferon, p2×7 polymorphism, and mycobacteria antigens for tuberculosis.[82,83]
	Neurologic Diseases	ABCA1 gene polymorphism, and apolypoprotein E genotyping for Alzheimer's disease.[84–86]
	Cancer	Methylation levels of genes NKX-25, CLSTN1, SPOCK2, SLC16A12, DPYS, and NSE1 for screening prostate, colon, and breast cancer.[88]
Prognostic Markers		
	Immunologic Diseases	C-telopeptides I and II (CTX-1 and CTX-2) for rheumatoid arthritis (RA).[92]
	Cardiovascular disease (CVD)	Osteoprotegerin (OPG) for atherosclerosis.[95]
	Metabolic Diseases	Alpha-hydroxybutyrate (Alpha-HB) and linolyl-glycerophosphocholine (L-GPC) for insulin resistance (IR) and glucose intolerance (GI).[1]
	Infectious Diseases	Fecal calprotectin in infectious diarrhea.[96] In a case control study, both bacterial and viral infection was involved. Main bacteria were *Salmonella, Campylobacter, Yersinia*, and *Shigella*, and viruses were rotavirus, norovirus, and adenovirus.
	Neurologic Diseases	MRI and proteomic based biomarkers in spinal fluid for Alzheimer's disease.[97]
	Cancer	Comparative genomic hybridization (CHG) identified complex genetic variants associated with adverse prognosis in cancer;[98] cytogenetics biomarkers for breast, head and neck, lung, liver, and ovarian cancers.[99–102]

14.2.1 Diagnostic Screening

14.2.1.1 Immunologic Diseases. Asthma involves inflammation, hyper responsiveness, bronchoconstriction, and symptoms (episodic breathlessness, wheeze, cough, tightness of the chest, and shortness of breath). Proteomics approaches identified allergenic proteins, based on their reactivity to patients' sera, using tandem mass spectrometry.[15] The most significant allergens identified were arginine kinase, sarcoplasmic calcium binding proteins, and tropomycin, which can be used for screening. This study also emphasized that MS analysis is a sensitive and accurate tool in identifying and quantifying aerosolized allergens.

14.2.1.2 Cardiovascular Disease (CVD). In cardiovascular disease (CVD), the relative contribution of genetic and environmental effect was studied on two inflammatory biomarkers, lipoprotein associated phospholipase A2 and antiphosphorylcholine IgM, in a Swedish population.[16] Results indicated that lipoprotein associated phospholipase A2 had low heritability and higher environmental regulation. Therefore, for diagnostic screening of CVD, biomarkers should be selected based on the context (family history, exposure history, lifestyle, *etc*.). Non-invasive technologies transcarnial Doppler, magnetic resonance and computed tomography were used for the diagnosis of intracranial atherosclerosis.[17] In these cases images were considered as biomarkers.

14.2.1.3 Metabolic Diseases. Metabolomic profiles can be used as biomarkers and can indicate the development of obesity, which in turn affects diabetes, cardiovascular disease, liver disease, renal disease and selected cancers. Metabolomic analysis is a valid and powerful tool that helps us understands the mechanism underlying different diseases. Biomarkers identified in obesity included glycine, glutamine, and glycerophasphatidyl choline in serum.[18]

14.2.1.4 Infectious Diseases. Potential applications of biomarkers in infectious diseases include distinguishing bacterial from nonbacterial infection, monitoring response to treatment and predicting survival (outcome). Periodontal disease, where aerobic and anaerobic bacteria infect gums, was proposed as a biomarker and risk factor for CVDs, especially abnormal function of progenitor endothelial cells.[19] Another group demonstrated an association of the oral microbiome with gastrointestinal cancer.[20] In such situations oral microbiodata was considered a biomarker for gastrointestinal cancer. Procalcitonin (PCT), a proinflammatory biomarker, is an excellent early detection biomarker of invasive bacterial infection in febrile children evaluated in emergency departments.[21]

Infectious agents are associated with at least 15% of cancers.[22] Chronic infections are the second most preventable cause of cancer,[23] and about 18% of the global cancer burden has been attributed to infectious agents.[24]

The presence of an infectious agent in tumor tissue is not sufficient to establish it as a causal agent; but the agent is termed carcinogenic if evidence from epidemiologic, clinical, and biologic studies suggests a strong cancer etiology.[25] The International Agency for Research on Cancer (IARC) has identified seven major infectious agents as carcinogenic: hepatitis B virus (HBV), hepatitis C virus (HCV), certain strains of the human papillomavirus (HPV), Epstein–Barr virus (EBV), human immunodeficiency virus type 1 (HIV-1), human T-cell lymphotropic virus type-1 (HTLV-1), and the gram-negative bacterium *Helicobacter pylori* (*H. pylori*).[24] Aflatoxin B1 (AFB1), the product of the common mold *Aspergillus flavus*, was also considered in this category because the IARC has identified AFB1 as a chemical liver carcinogen.[26] Hepatitis C infected individuals can be distinguished from healthy people based on metabolomic profiling (which can be considered a biomarker for screening).[27] It was demonstrated that IL-18 levels in serum could be used as a biomarker for acute Epstein–Barr virus (EBV) infection.[28]

14.2.1.5 Neurologic Diseases. Metabolomics based technologies (liquid chromatography quadrupole time-of-flight mass spectrometry) and chemometrics were used to identify Alzheimer disease (AD) specific profiling.[29] Samples from healthy age-matched controls were also used to identify new biomarkers (profiles). The prediction value for AD was 94–97% in this metabolomic study, which was considered quite remarkable for early detection of AD. The confirmation of these results was done in another set of patients and control and 100% diagnosis was achieved.[29] Such studies are significant to understand etiology, pathophysiology, and treatment of degenerative brain disorders.[30]

14.2.1.6 Cancer. In this section we discuss different tumor types and next generation biomarkers identified in different tumor types. Cancer is a genetic and epigenetic disease and methylation biomarkers have been used in diagnostic screening of cancer.[31,32] Methylation changes in the precarcinoma stage could be used as biomarkers for diagnosis in cancer if these epigenetic changes are not present in normal cells.[32,33] With the development of high throughput next generation sequencing, it has become easier to detect methylation changes qualitatively as well as quantitatively and recently global methylation profiling has also been accomplished.[34] Gene promoter hypermethylation in death-associated protein kinase 1 (*DAPK1*), *p16* and *RASSF1A*1 could be used as biomarkers for detection of HNSCCs. The development of biomarkers for prostate cancer screening, detection, and prognostication revolutionized the management of this disease. During the progression of cancer, it was observed that the levels of certain proteins were elevated. These abnormally increased molecules could be used as biomarkers for gaining an insight on the course of the disease. Prostate-specific antigen (PSA) is a useful, though not specific, biomarker for detecting prostate cancer. Serum markers (PAP, tPSA, fPSA, proPSA, PSAD, PSAV, PSADT, EPCA, and

EPCA-2), tissue markers (AMACR, methylated GSTP1, and the TMPRSS2-ETS gene rearrangement), and a urine marker (DD3PCA3/UPM-3) of prostate cancer are very well characterized biomarkers.[35–42] Disease-specific protein biomarkers also were identified[43] and characterization of such biomarkers in body fluids could aid in the early detection of cancer and help in monitoring cancer progression. It has become easier to identify cancer specific markers with the advancement in the available technologies. In the case of breast cancer, some of the serum tumor markers like carcinoembryonic antigen (CEA) and cancer antigen (CA) CA 15.3, CA 27.29 were not confirmed to be sensitive for early detection; however, their high levels did reflect disease progression and recurrence.[44] Mammoglobin and MASPIN (mammary serine protease inhibitor) were shown to be diagnostic biomarkers.[45] The early detection of circulating breast cancer cells by morphologic methods is currently being challenged by ultrasensitive proteomic and PCR-based methods often enhanced by immunomagnetic bead-based cell capture.[46] The role of MASPIN in diagnosis and prognosis of a variety of cancers is being explored.[47] It has been reported that low or absent MASPIN cytoplasmic expression was frequently observed in oral carcinomas with lymph node metastasis.

Using saliva as the starting material, mRNAs specific for esophageal, head and neck, oral and lung cancer were also identified.[48–50] Seven biomarkers showed 3.5 fold increases in their levels in oral SCC patients compared to healthy individuals. Further validation of values is needed before implementing in clinics.

Microsatellites are repeated sequences of DNA.[51,53] When these repeats are shortened, this is called microsatellite instability (MSI). This stage arises in regions where the mutation rate is very high, mainly due to defective DNA genes such as mismatch repair gene. MSI biomarkers were reported in a number of cancers, including bladder, colon, esophageal, and skin cancer.[54–56] As high as 30% of HNSCC patients had MSI [52] suggesting that MSI is a good biomarker for HNSCC. A number of studies evaluated the potential of detecting these markers in tumor samples as well as in other biospecimens and found these biomarkers useful for diagnosis and prognosis.[57]

14.2.2 Early Diagnosis

14.2.2.1 Immune Diseases. Airways inflammation starts soon after inception of exposure to allergens in asthma. Inflammatory cytokines along with arginine kinase, sarcoplasmic calcium binding proteins are the most practical biomarkers for early diagnosis of asthma.[15,58]

14.2.2.2 Cardiovascular Diseases (CVDs). For early diagnosis of carotid atherosclerosis for which obesity is the risk factor, the early biomarkers are carotid intima media thickness (IMT) and circulating oxidized low density lipoprotein.[59] To evaluate endothelium status, flow-mediated

dialation (FMD) and IMT are used as early biomarkers of atherosclerosis in patients with nonalcoholic fatty liver disease (NAFLD).[60]

14.2.2.3 Metabolic Diseases. A number of markers have been described for early diagnosis of diabetes.[1,61] They include dysglycemia, alphahydroxybutyrate, linoleyolglycerophosphocholine, advanced glycation end products (AGE), albumin excretion rate (AR) and SNPS in epidermal growth factor gene intron 2.[1,62,63]

14.2.2.4 Infectious Diseases. Region-of-Difference-1 (RD-1) gene product and sputum cytokine levels are considered a biomarker for early detection of tuberculosis.[64] Hepatocellular carcinoma (HCC) involves infectious agents and the early diagnostic biomarkers for HCC are HSP70, CAP2, glypican 3 and glutamine synthase.[65] These biomarkers were identified based on gene expression analysis of HCC samples.

14.2.2.5 Neurological Diseases. A bioinformatic approach has helped in identifying early diagnosis biomarkers for Alzheimer's disease. Based on data from patients and matched controls, the Artificial Neural Network (ANN) identified beta amyloid cascade as a potential biomarker for early detection of Alzheimer's disease.[66]

14.2.2.6 Cancer. The key to successful application of biomarkers is to detect the disease early so that a variety of treatment approaches can be applied.[67] Initially it was thought that genetic changes arise first at the time of initiation of a disease but genome-wide profiling suggests that epigenetic changes occur much earlier than genomic changes.[33] Technologies exist to follow up these changes of early detection. B7-H4, spondin 2, and DcR3 were identified as early biomarkers in ovarian cancer although other investigators have proposed other groups of biomarkers.[68–70]

14.2.3 Risk Assessment (Screening)

Genome-wide association studies (GWAS) are extremely powerful in identifying new low-penetrance SNPs (biomarkers) which may have therapeutic implications.[71] Identification of common low-susceptibility alleles is useful because it provides possible insight into the mechanisms of tumor biology in cases of cancer and identifies high risk individuals.[72]

Associations do not necessarily mean causality; the potential for confounding and reverse causality should always be kept in mind. It will be beneficial to predict complications, in case of diabetes in particular and other diseases in general, and determine subgroups that may be responsive to therapy. Clinical implications should not be exaggerated while characterizing prediction biomarkers.

14.2.3.1 Immune Diseases. In one meta-analysis glutathione S transferase M1 (GSTM1) biomarker was identified which helped in screening a

cohort of children who were at high risk of developing asthma.[73] In a population based study C-reactive protein, fibrinogen, and interleukin-6 were found useful biomarkers for screening.[74]

14.2.3.2 Cardiovascular Diseases (CVDs). In a risk prediction model of cardiovascular disease, C-reactive protein levels were implemented.[75] However, after completion of studies, it was suggested that evaluation of the potential impact of CRP levels would require studies to quantify the effects of additional CRP assessment on medical decision making. Biomarker studies of CVD prediction in elderly patients indicated that mortality and cardiovascular events were dependent on low peripheral pulse pressure not on high blood pressure.[76] Genetic variants in CVD were studied to evaluate their association with the disease but with limited success.[77]

14.2.3.3 Metabolic Diseases. For diabetes, alanine amino transferase (ALT), gamma glutamyl transferase (GGT), triglyceride, plasminogen activator inhibitor (PAI-1) antigen, ferritin, C-reactive protein (CRP), and sex-hormone binding globulin (SHBG) were identified as prediction markers after a large study "West of Scotland Coronary Prevention Study (WOSCOPS) was conducted. Results from this study were validated in a different population and two biomarkers, ALT and GGT, looked very promising for risk assessment.[78] CRP was not causally related to insulin resistance or obesity. IL-6 upregulation was also associated with diabetes. Early diagnosis of chronic kidney diseases and atherosclerosis in the same subjects could be accomplished by carotid ultra sound technology.[79]

14.2.3.4 Infectious Diseases. HIV/AIDS screening involves behavior biomarkers as well as molecular omics biomarkers and a number of investigators have reported the screening strategies and their outcome in different populations.[80,81] In tuberculosis, gamma interferon, p2×7 polymorphism, and mycobacteria antigens were used as biomarkers for risk assessment and screening.[82,83]

14.2.3.5 Neurological Diseases. The most common biomarkers used for risk assessment of Alzheimer's disease were ABCA1 gene polymorphism, and apolypoprotein E genotyping; although a combination of markers was also used, synergism was lacking.[84–86]

14.2.3.6 Cancer. Transcriptomic and miRNAs biomarkers were used for screening colorectal cancer.[87] For screening prostate, colon, and breast cancer, methylation levels of genes NKX-25, CLSTN1, SPOCK2, SLC16A12, DPYS, and NSE1 were used.[88] Malignant melanoma is one of the most aggressive types of tumor. Because malignant melanoma is difficult to treat once it has metastasized, early detection and treatment are essential. The search for reliable biomarkers of early-stage melanoma, therefore, has received much attention. By using an approach of screening tumor antigens

and their auto-antibodies, bullous pemphigoid antigen 1 (BPAG1) was identified as a melanoma antigen recognized by its auto-antibody when anti-BPAG1 auto-antibodies were detected in melanoma patients at both early and advanced stages of disease.[89]

14.2.4 Prognostic Markers

Prognostic biomarkers should be tightly linked to the outcome so that they can be used as surrogate measures of efficacy and treatment response. Prognostic markers can be defined as factors that can predict an outcome in the absence of systemic therapy or predict an outcome different from patients who are devoid of the biomarker, despite empiric therapy (initiating treatment before confirming diagnosis).[90] It is not known whether there is any association between ovarian cyst and estrogen levels during tamoxifen use. A very well designed study was conducted where breast cancer prognostic markers were utilized to evaluate the effect of tamoxifen use in premenopausal women with ovarian cyst and results indicated an association.[91] Hence, prognostic markers can be utilized to classify patients into appropriate groups for treatments.

14.2.4.1 Immune Diseases. Biomarkers of bone and cartilage turnover are collagen C-telopeptides I and II, which are predictors of structural damage in RA patients. C-terminal cross-linked telopeptide of type I collagen (CTX-I) could be measured in serum or urine after it was released during bone resorption. Bone erosions and osteoporosis both occur as a consequence of RA and CTX-I levels and could be used for prognosis. On the other hand, increased CTX-II levels were associated with rapid progression of joint damage. For clinical implication, investigators of these studies recommended that new prognostic biomarkers should supply information beyond that provided by risk factors in the prediction of disease outcome.[92]

European countries restrict anti-tumor necrosis factor therapy for RA and prescription is provided based on evaluation of inflammatory markers, C-reactive protein levels and general health of a patient.[93] Genomic variants of C-reactive proteins play a major role in therapeutics of RA and can be used as biomarkers.[94]

14.2.4.2 Cardiovascular Diseases (CVDs). Osteoprotegerin (OPG) is such a biomarker, which is independently associated not only with risk factors of atherosclerosis but also with subclinical peripheral atherosclerosis and clinical atherosclerosis and is recommended as a prognostic biomarker for ischemic heart disease and ischemic stroke.[95]

14.2.4.3 Metabolic Diseases. Alpha-hydroxybutyrate (Alpha-HB) and linolyl-glycerophosphocholine (L-GPC) were identified as biomarkers of insulin resistance (IR) and glucose intolerance (GI) in a large population study of

more than 1000 participants from the Relationship between Insulin sensitivity and Cardiovascular Disease (RISC) study.[1] AHB correlated positively and L-GPC negatively with both diseases (IR and GI) indicating that AHB was a positive predictor and L-GPC, a negative predictor independent of family history of diabetes, sex, age, fasting glucose, and BMI.

14.2.4.4 Infectious Diseases. Diarrhea in children may involve infection. In cases of infectious diarrhea, fecal calprotectin is a good prediction marker.[96] In a case control study, both bacterial and viral infection was involved. The main bacteria were *Salmonella, Campylobacter, Yersinia*, and *Shigella*, and viruses were rotavirus, norovirus, and adenovirus. These studies suggested that fecal calprotectin could be used as a noninvasive biomarker in management of children with infectious diarrhea.

14.2.4.5 Neurological Diseases. A few Alzheimer's diagnosis biomarkers are suitable for prognosis also. One investigator combined MRI and spinal fluid based biomarkers and observed better results than using a single biomarker for prognosis.[97] This investigation was performed in different sites of the brain (for MRI) and spinal fluid biomarker values were used as a reference standard.

14.2.4.6 Cancer. Despite huge efforts into research for studying new biological prognostic markers, only a few out of several hundreds have progressed to clinical use. Comparative genomic hybridization (CHG) identified complex genetic variants associated with adverse prognosis in cancer.[98] Cytogenetics biomarkers turned out to be very useful biomarkers for breast, head and neck, lung, liver, and ovarian cancers.[99–102] In case of HNSCCs, loss of heterozygosity (LOH) on distal arm of 18q was reported to be associated with poorer survival.[103] About 70% of colorectal cancers showed allelic deletions in chromosomes 18q and 17p. The *p53* gene located on 17p is mutated in about 40 to 60% of CRCs and demonstrated association with prognosis and prediction of the disease. CRC patients with chromosome 18q loss showed worse disease-free and overall survival. Markers like Ki-67 staining detecting cell proliferation were significantly correlated with breast cancer outcome. One-fifth of breast cancer patients showed amplification or over-expression of cyclin D1 (*PRAD1* or bcl-1). Germline prognostic markers of bladder cancer were identified and characterized.[104] Stage II and III CRC patients with high microsatellite instability have been reported to show improved survival and better relapse-free survival as compared to microsatellite stable (MSS) patients.

14.2.5 Drug Response

Systems immunology, based on mathematical and computational models, has the potential to predict treatment response.[105] High throughput ''omics''

technologies generate vast amount of data which may help in identifying immunological processes at high resolution in disease development. Using a systems immunology approach it is possible to construct causal relationships between complex molecular processes and specific disease-associated phenotypes.[106] Drug response and pharmacogenomics in the context of biomarkers in different diseases is shown in Table 14.2.

14.2.5.1 Immune Diseases. The prevalence of autism is 1 in 80 in the USA. Evidence based treatments are practiced in autism[107] although sometimes drug treatment (mGluR antagonist and GABA agonist) is also recommended based on the severity and stage of the disease.[108] The etiology of autism is not completely understood. However, synaptic maturation and plasticity in the pathogenesis of autism spectrum disorder resulting in imbalance of excitation and inhibition has been observed. The drugs mentioned above control excessive excitement.

14.2.5.2 Cardiovascular Diseases (CVDs). Nutritional foods and their biological food components alter cardiovascular disease status and there is very well established evidence of disease modifying effects of these compounds with anti-inflammatory and antioxidant effects in CVD patients. Reduced total LDL cholesterol levels were observed when polyphenolic compounds were given to patients because bioactive phytochemicals play an important therapeutic role in attenuating oxidative damage.

14.2.5.3 Metabolic Diseases. The key players in measuring drug response in type II diabetes mellitus (T2DM) are glycemic index (GI), glucose response curves (GRCs) and daily mean plasma glucose (DMPG). In a trial, a variety of foods were supplied to participants undergoing treatment with oral antidiabetic drugs (OADs).[109] Promising results were obtained in this pilot study and validation of these observations in large population is planned.

14.2.5.4 Infectious Diseases. 1,3-Beta-D-glucan (BG) can be used as a biomarker in invasive fungal infections (especially infections involving *Candida*) in patients undergoing treatment of candedemia with anidulafungin.[110] In those patients who were given anidulafungin followed by fluconazole/voriconazole therapy, decreasing BG concentrations reflected success of the treatment.

In unipolar depression, transcriptomic biomarker approaches were used to identify responders of treatment.[111] The main player in the process was tumor necrosis factor in the inflammatory cytokine pathway. These investigators evaluated genetic background of participants and suggested that SNPs rs1126757 in IL11and rs 7801617 in IL6 play major role in responding

Table 14.2 Drug Response and Pharmacogenomics in the Context of Biomarkers.

Category	Diseases	Markers
Drug Response		
	Immunologic Diseases	Evidence based treatments are practiced in autism [107] although sometime drug treatment (mGluR antagonist and GABA agonist) is also recommended based on the severity and stage of the disease.[108]
	Cardiovascular disease (CVD)	Reduced total LDL cholesterol levels were observed when polyphenolic compounds were given to patients because bioactive phytochemicals play an important therapeutic role in attenuating oxidative damage. Nutritional foods and their biological food components alter cardiovascular disease status and have anti-inflammatory and antioxidant effects in CVD patients.
	Metabolic Diseases	In a trial, a variety of foods were supplied to participants undergoing treatment with oral antidiabetic drugs (OADs).[109] The key players in measuring drug response in type II diabetes mellitus (T2DM) are glycemic index (GI), glucose response curves (GRCs) and daily mean plasma glucose (DMPG).
	Infectious Diseases	In those patients who were given anidulafungin followed by fluconazole/voriconazole therapy, decreasing BG concentrations reflected success of the treatment. 1,3-Beta-D-glucan (BG) can be used as a biomarker in invasive fungal infections (especially infections involving *Candida*) in patients undergoing treatment of candedemia with anidulafungin.[110]
	Neurologic Diseases	For the treatment of schizophrenia, ionotropic glutamate receptors are targeted.[112] Recent pharmacologic strategies have focused on improving cognition by drugs.[113]
	Cancer	Biomarkers of drug response may be new or the same which are diagnostic or prognostic biomarkers depending on the biology and etiology of the type of cancer studied. In different kinds of cancers pharmacological response of drugs is different and a variety of factors (genetic background, life style, other diseases in the same person, BMI *etc*.) contribute to drug response.[114] Development of epigenetic drugs showed promise in selected cancer treatment.[115]

Table 14.2 (*Continued*)

Category	Diseases	Markers
Pharmacogenomics		
	Immunologic Diseases	The response of infliximab was favorable in a group of patients where genes regulated by TNF were active compared to other participants of the study indicating a strong evidence of role that genomic background plays in responding to pharmacological drugs.[120]
		Pharmacogenomics of rheumatoid arthritis (RA) has been studied by several investigators[120,121] although RA is a highly heterogeneous disease in all aspects including response to therapy.
	Cardiovascular disease (CVD)	Infliximab treatment was provided to rheumatoid arthritis patients to reduce levels of TNF and results were evaluated based on the genomic background of the participants.[120] Host response due to inflammatory diseases such as arthritis depends on the genomic background and pharmacological agents give response based on the genomic susceptibility and some other yet unidentified factors.[120–122]
	Metabolic Diseases	Genetic background was considered in treatment of type 2 diabetes and an association of ADIPOR2 gene variants with CVDs and type 2 diabetes risk in individuals with defective glucose tolerance (conducted in Finnish population) was observed.[123] Additionally, genetic predisposition and nongenetic risk factors of thioazolidine-related edema in type 2 individuals, the pharmacogenomics of metaformin were also observed.[124]
	Infectious Diseases	Genetic variants at least at three loci, NAT2, CYP2E1, and GSTM1, were involved in pharmacogenomics. Pharmacogenomics of anti-TB drug related hepatotoxicity was studied to understand involvement of genetic background in response to treatment in TB patients.[125]
	Neurologic Diseases	Genetic variants were identified in Alzheimer's disease and demonstrated treatment response.[126, 127] In a recent study of an Italian cohort, response to cholinesaterse inhibitors was evaluated in Alzheimer's patients using 48 SNPs.[126] Results indicated association of two SNPs with the response to treatment.
	Cancer	Promising results were obtained in colon, gastric, breast, ovarian, GI tract and lung cancer.[129–131] Polymorphism in miR encoding genes also was evaluated for its implication in pharmacogenomics.[128]

to antidepressants. This was an excellent example where genomics and transcriptomics biomarkers were used to follow up treatment.

H. pylori is the most common chronic bacterial infection in humans. This bacteria is involved not only in gastric cancer but also in ulcer diseases and gastritis. Its synergistic gastrotoxic interaction with non-steroidal anti-inflammatory drugs and association with atherosclerotic events is a matter of concern. Transmission of *H. pylori* is through oral ingestion, mainly within families in early childhood. Therefore, treatment and prevention approaches are more appropriate for children. One-week proton pump inhibitor based triple therapy with clarithromycin and either amoxicillin or metronidazole is the most common therapy.

14.2.5.5 Neurological Diseases. Schizophrenia is a chronic brain disorder and affects approximately 2.5 million Americans and more than 24 million people worldwide. For the treatment of schizophrenia, ionotropic glutamate receptors are targeted.[112] Recent pharmacologic strategies have focused on improving cognition with drugs.[113] The consensus among investigators in the schizophrenia field is that physiological and pharmacological approaches should be combined for this disease for better outcome of the treatment.

14.2.5.6 Cancer. In different kinds of cancers pharmacological response of drugs is different and a variety of factors (genetic background, life style, other diseases in the same person, BMI *etc.*) contribute to drug response.[114] Biomarkers of drug response may be new or the same which are diagnostic or prognostic biomarkers depending on the biology and etiology of the type of cancer studied. Personalized medicine is recommended for cancer because cancer is a heterogeneous disease and the rate of recurrence after the treatment is high for some cancer types. Development of epigenetic drugs showed promise in selected cancer treatment.[115]

14.2.6 Pharmacogenomics

Pharmacogenomics is the study of interindividual genetic variability that plays a significant role in treatment response and toxicity due to the drug. Based on a better understanding of biological variability, attempts are being made to customize treatment at the personal level.[116] In pharmacogenomic approaches the following points are considered critical: pharmacokinetic related genes and phenotypes, pharmacodynamic targets, genes and products, risk of disease related with metabolomic cycle, physiological variations, and environment interaction. When targeted agents matched with tumor molecular aberrations were implied in a phase I clinical trial, encouraging results were observed.[117] Since pharmacogenomics science is new, success is not guaranteed, as happened in case of gastric cancer where efficacy of Trastuzumab was checked in patients with advanced gastric cancer and the drug was not effective.[118] Other investigators have identified challenges in

the field including poor coordinated diagnostic testing and current models of implication in disease stratification.[119]

14.2.6.1 Immune Diseases. Compared to other diseases, the pharmacogenomics of rheumatoid arthritis (RA) has been studied by several investigators.[120,121] Although RA is a highly heterogeneous disease in all aspects including response to therapy. The response of infliximab was favorable in a group of patients where genes regulated by TNF were active compared to other participants of the study indicating strong evidence of the role that genomic background plays in responding to pharmacological drugs.[120]

14.2.6.2 Cardiovascular Diseases (CVDs). Host response due to inflammatory diseases such as arthritis depends on the genomic background and pharmacological agents give response based on the genomic susceptibility and some other yet unidentified factors.[120–122] Infliximab treatment was provided to rheumatoid arthritis patients to reduce levels of TNF and results were evaluated based on the genomic background of the participants.[120] Pharmacogenomics field in CVDs is still in infancy and next few years research should focus on accurately identify, from available therapies, which is the optimal therapy for any particular individual.[121]

14.2.6.3 Metabolic Diseases. In few studies the role of genetic background in treatment of type 2 diabetes was demonstrated, such as association of ADIPOR2 gene variants with CVDs and type 2 diabetes risk in individuals with defective glucose tolerance (conducted in the Finnish population),[123] genetic predisposition and nongenetic risk factors of thioazolidine-related edema in type 2 individuals, the pharmacogenomics of metaformin.[124] Validation in large number of participants is still awaited.

14.2.6.4 Infectious Diseases. Pharmacogenomics of anti-TB drug related hepatotoxicity was studied to understand the involvement of genetic background in response to treatment in TB patients.[125] Genetic variants at least at three loci, NAT2, CYP2E1, and GSTM1, were involved in pharmacogenomics. This study was conducted only in one population and it should be validated in populations with different ethnic backgrounds.

14.2.6.5 Neurological Diseases. Due to its serious effects on day to day life, Alzheimer's pharmacogenomics has been studied by a number of groups in the last decade and genetic variants were identified which were associated with treatment response.[126,127] In a recent study of an Italian cohort, response to cholinesaterse inhibitors was evaluated in Alzheimer's

patients using 48 SNPs.[126] Results indicated association of two SNPs with the response to treatment.

14.2.6.6 Cancer. The National Cancer Institute has identified pharmacoepidemiology related to pharmaceutical use and cancer risk, recurrence and survival as one of the priority areas of research. Polymorphism in miR encoding genes also was evaluated for its implication in pharmacogenomics.[128] Although a considerable amount of new-targeted agents have been designed based on cancer biology, challenges and gaps exist between pharmacogenomics knowledge and clinical application. Promising results were obtained in colon, gastric, breast, ovarian, GI tract and lung cancer.[129–131]

14.2.7 Therapeutic Biomarkers in Immune Diseases, Cardiovascular Diseases (CVD), Metabolic Diseases, Infectious Diseases, Neurological Diseases, and Cancer

Since there is no measurable disease while undergoing treatment, generally therapeutic biomarkers are difficult to study. Another key point is the adverse reaction of patients due to treatment. Such parameters are not needed for prognostic biomarkers; therefore, it is easier to identify prognostic biomarkers than therapeutic biomarkers. In the following section, examples of therapeutic biomarkers in different diseases are described.

14.2.7.1 Immune Diseases. In the case of multiple sclerosis (MS), treatment with disease-modifying therapies (DMT) with ability to prevent axonal damage resulted in reduction of progression of the disease from clinically isolated syndrome (CIS).[132] In another study, inclusion of information about Max RNA during interferon treatment of MS resulted in a better response than interferon only treatment.[133] Other investigators also observed similar results.[134–135]

14.2.7.2 Cardiovascular Diseases (CVDs). Biomarkers discussed in the diagnostic section can be used for follow up of the treatment. Dietary intervention of CVD by fish oil (salmon, herring, and pompano) and other nutrients was demonstrated in a number of studies.[136] Some of the participants had higher levels of triacylglycerolaemia. Biomarkers TNFalpha and IL-6 were reduced and the level of adiponectin increased in the treated arm. Thus TNFalpha, IL-6, and adiponectin were used as therapeutic biomarkers. In another study, argon oil supplement reduced plasma levels of lipids and antioxidant status.[137] Therapeutic biomarkers used in this study were plasma vitamin E concentrations, total and LDL cholesterol, and antioxidant profiles.

14.2.7.3 Metabolic Diseases. For the management of diabetes hemoglobin A1c (HbA1c) is used which is considered a reliable indicator

of glycemic control. In most of the clinical studies in diabetes, HBA1c biomarker is used to determine the glucose control.

14.2.7.4 Infectious Diseases. Diarrhea in AIDS patients was treated with specific medications and therapeutic response was measured by levels of tubuloreticulin inclusions (TRIs).[138] Although this study may be considered an isolated study TRI has the potential to be a therapeutic biomarker after larger studies are conducted. For pneumonia therapy of more than 100 patients, biomarker procalcitonin (PCT) levels were very useful and this biomarker has been recommended for future prognosis.[139]

14.2.7.5 Neurological Diseases. In Parkinson's disease, glial cell-line derived neurotrophic factor (GDNF) and family of ligands (GDFLs) were used as biomarkers to follow up therapy.[140]

14.2.7.6 Cancer. Several cancer biomarkers are currently used to develop targeted anti-cancer drugs. Topomerase I inhibition in colorectal cancer is one example of targeted therapeutics and utilization of biomarkers.[141] In leukemia, an antibody therapeutics approach was applied, using rituximab, and therapeutic biomarkers Ki67 and PIM1 were followed for their response.[142] Results indicated that Ki63 was an independent biomarker. In another study CD20 was used as a biomarker to evaluate the therapeutic potential of rituximab in B-cell lymphoma.[143] In B-cell non-Hodgkin lymphoma (NHL), treatment with the same agent, interleukin-6 plasma levels were followed to evaluate drug response.[144] Eradication of NHL was achieved with a monoclonal antibody therapy combining rituximab with a blocking anti-CD40 antibody.[145] Other examples of treatment are trastuzumab for breast cancer (target p185neu), gemtuzumab for AML (target CD33), ibritumomab for AML (target CD2090Y), edrecolomab for colorectal cancer (target EpCAM), tositumomab for NHL (target CD20), and cetuximab for colorectal cancer (target EGFR). Cisplatin is a widely used chemotherapeutic in head and neck cancer patients. However, the effect of the drug is limited by the resistance observed in these patients; and it has been seen that hypermethylation in some of the genes responsible for cytotoxicity causes the resistance, for example the *S100P* gene. The above description indicates that biological markers that can predict therapeutic outcomes enable guiding the choice of treatment; and depending on the levels of the biomarkers it could be estimated if the patient would respond to a particular therapy or not and finally an appropriate treatment regimen can be selected. In addition, prediction of drug response helps in reducing the treatment cost.

14.3 RECENT TRENDS AND FUTURE DIRECTIONS

Significant progress in a variety of biomarkers has been made in the last two decades. Metabolomic and epigenomic biomarkers are relatively new and

have shown promising results in disease diagnosis and prognosis. The monitoring of relative changes in metabolomic profiles in predisposed versus healthy individuals may help identify unique metabolites involved in disease processes.[146] These profiles have been used to predict the risk of diabetes,[147] cardiovascular disease,[148] and lung cancer,[149] diagnose prostate cancer,[150] differentiate benign and malignant ovarian tumors,[151] and identify biomarkers of Crohn's disease.[7] Such shifts also may identify diagnostic biomarkers, which could provide insights into strategies for disease prevention and be used to monitor the response to treatment. In recent years, metabolomics and other post-GWAS platforms, such as proteomics and transcriptomics, have undergone rapid improvement in both their reliability and throughput; as such, it may be an appropriate time for their use in epidemiologic studies.[152] Although metabolomic profiling has been used in some larger-scale population studies,[153] the number of published reports to date remains small. If the successes of genomics and transcriptomics in epidemiology are reliable indicators, there is a large, yet unexplored, potential for metabolomics to contribute to public health research. So far the assessment of geneotypes of candidate biomarkers in metabolomic diseases and CVD in blood samples has not improved prediction of these diseases. Probably multiple markers (transcriptiomics, epigenomics, metabolomics, genomics) should be tested for better prediction of these diseases. Prognostication is one of the promising area in translation of experimental research into clinical practice. In this approach, patterns of altered gene expression in tumors are used to construct classifiers instead of standard indices such as Nottingham Prognostic Index, Adjuvant Online, and Predict.[154]

Multiplexing of biomarkers may reduce false positive results in screening studies where intention is to identify populations which are at high risk of developing a disease.[2] Quantitative imaging data storage and maintenance have their own challenges as we discussed above. In case of miR biomarkers, whether miR expression is localized in a specific part of the tissue has to be carefully evaluated. In a tissue biopsy the local concentration (number of miRs) may be low or high. Determining the accurate level of miRs is very critical.

It has been seen that a combination of genes measured concurrently imparts better information about the clinical effect than a single gene. Oncotype DX™ is the first multigene predicting test of prognosis for breast cancer patients receiving anti-estrogen therapy (Genomic Health, USA). The FDA has also approved the test MammaPrint that predicts relapse in breast cancer patients by analyzing the activity of 70 genes. In case of prostate cancer, the PSA (Prostate-specific antigen) test has been approved, where PSA levels help in detecting prostate cancer and also predict a recurrence in patients suffering from the disease. Alpha-fetoprotein (AFP) has been approved for the diagnosis and monitoring of patients with non-seminoma testicular cancer.

Significant challenges exist regarding IP issues. At times tests are known but their associations with diseases are new. Sometimes patent belongs to smaller companies for new tests but their clinical utility can only be developed with larger companies in large clinical trials. In general, every situation represents unique opportunities and IP related issues can only be solved in a case by case manner.

The presence of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her2, or ERBB2) is used for the clinical and pathological classification of breast cancer.[3] Generally, ER-positive (+) and PR+ are indicators of good prognosis, and Her2+ is an indicator of bad prognosis. In addition, ER-negative (−), PR−, and Her2− (also called triple-negative) status is considered to be an indicator of poor prognosis. Basal cells exhibit triple-negative features. On the basis of oncologic pathway activity analysis, up to 18 subtypes of breast cancer have been suggested.[4] However, the implications of this information for clinical practice remain to be determined. Furthermore, many prognostic gene expression signatures that dichotomize patient populations into treatment-responsive and nonresponsive groups lack specificity.[155,156] Additional biomarkers are needed that are better prognostic indicators than hormone receptor status, and a better understanding of the genetic characteristics of patients is needed to improve current clinical practice. Ideally, a method for preoperative molecular profiling should be developed that can guide treatment strategies.

In recent years metabolites of biofluids have been analyzed for their potential in cancer diagnosis and follow up of treatment. Especially urine analysis for the routine monitoring of metabolomic disorders has attracted a reasonable amount of interest among scientists because the procedure can be done easily, noninvasively and repeatedly for a large number of samples with high precision. Generally volatile organic metabolites (VOM) get enriched in urine and their analysis is easy.[157] The advantage of adopting metabolomic approach lies in the fact that metabolites are much more stable than RNA and proteins and their levels predict those pathways which are affected during disease development. In one small study with urine from controls and breast cancer, VOMs were identified which were differentially expressed in patients only.[158] Higher levels of 4-carene, 3-heptanone, 1,2,4-trimethylbenzene, 2-methoxythiophene, and phenol, and lower levels of dimethyl sulfides were observed in breast cancer patients. Urine metabolites have also been used for diagnosis of colon, lung, liver, and prostate cancer.[157]

Characterizing metabolomic pathways helps in making treatment decisions.[159] Circulating biomarkers, especially diabetes biomarkers discussed in this article, should be used for disease stratification of patients followed by completion of questionnaires such as: (i) how can the group be best defined in terms of biomarker levels? (ii) cause of the unexpected results; (iii) underlying genetic trait; and (iv) will one group respond better than other? This approach may lead us in the direction of personalized medicine. We have summarized potential challenges and research opportunities in

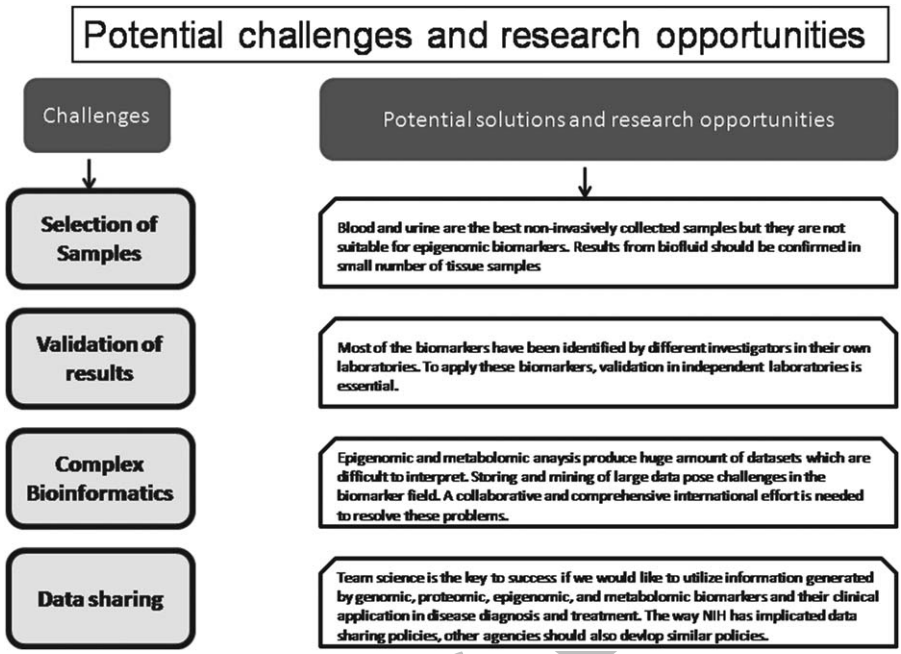


Figure 14.2 Potential challenges and research opportunities in utilizing biomarker information in disease detection, diagnosis, prognosis, and drug response.

utilizing biomarker information in disease detection, diagnosis, prognosis, and drug response in Figure 14.2.

14.4 CONCLUSIONS

Biomarker screening tests face the challenge in transition of tests from the research level in the laboratory to their beneficial use in the clinic. The biomarkers need to be highly specific and sensitive for the detection in clinical samples in order to avoid false positive tests that could lead to misdiagnosis and wrong therapeutic selection. An ideal biomarker should be used for screening in a large population study to validate the effectiveness of the screening and should be proven effective in populations with different genetic background. It has been seen that many biomarkers that correlate with disease statistically may not be useful in the clinic. It is advisable to study patient populations with diversity because the biomarkers might show different responses in different populations.

Considerable amount of knowledge has been obtained in understanding cancer biology and identification of biomarkers which can detect cancer but implication of that information in clinic is still challenging.[160] Clinical validation is the main hurdle in the process. In one case control study of the

Prospect-EPIC (European Prospective Investigation into Cancer and nutrition) where more than 300 breast cancer patients and matched controls were tested for breast cancer over a period of three years using a panel of eight serum biomarkers (osteopontin, haptoglobin, cancer antigen 15-3, carcinoembryonic antigen, cancer antigen-125, prolactin, cancer antigen 19-9, and alpha-fetoprotein), very low specificity (50%) and sensitivity (50%) was observed.[1]This may be due to different subtypes of breast cancer in collected samples. Such epidemiologic studies should select a broader target set of potential biomarkers which could be enabled by antibody array technologies where profiles of up to 100 antibodies can be followed simultaneously. Making different groups, based on the status of hormone receptors (estrogen and progesterone), might also be helpful.

In case of cancer, the need for identification and characterization of early cancer diagnostic biomarkers is high because cancer is a heterogeneous disease and the patient's individual molecular profiling due to tumor microenvironment determines the disease development and response to treatment.[2,5] The tumor microenvironment is affected by several factors including epigenetic factors of the cell.

Reasonable progress has been made in clinic in some cases where biomarkers lead to better efficacy, less toxicity, better diagnosis and predictable prognosis. However, this has been possible only in a few select disease conditions. We anticipate that such success will soon be replicated in several critical diseases. There is an apparent need for new biomarkers and the upcoming technologies promise the development of new biomarkers, which would change the course of disease detection and management. This will be a gain for the medical field with improved patient care and better clinical outcome.

The main areas which need progress/attention are the cost and high throughput. Another area where scope for further progress remains is the application of biomarkers in clinic. Proper analytical and clinical validation of early biomarkers has not been achieved. Clinical validation of identified biomarkers is especially the key challenge in the field. The National Cancer Institute has developed guidelines for the analytical and clinical validation of biomarkers but none of the biomarkers has been validated to date.[161] Integration of genomic and proteomic biomarkers with epigenetic biomarkers may help us subtyping disease stages.[6] Many times results of methylation profiling from blood and tissues are different. Koestler *et al.* conducted a systematic epigenome-wide methylation analysis and demonstrated that shifts in leukocyte subpopulations might account for a considerable proportion of variability in these patterns.[7] The location of miRs in any specific organ should be carefully determined. In a tissue biopsy the local concentration (number of miRs) may be low or high. Determining the accurate level of miRs is very critical.

Association studies are extremely powerful in identifying low-penetrance new SNPs (biomarkers) which may have therapeutic implications. Identification of common low-susceptibility alleles is useful because it provides

possible insight into the mechanisms of tumor biology and identify high risk individuals. Since genotyping is not expensive these days, the information from such studies can be utilized in personalized medicine by targeted primary and secondary prevention.

Although individual omics techniques have generated a large number of potential biomarkers for any given cancer; due to heterogeneity of carcinogenesis process, an integrated approach is required to gather all potential biomarkers and to identify key biomarkers for diagnosis, prognosis and therapy.

We emphasize that considerable progress has been made in disease-associated biomarkers which can be used for the complete spectrum of different diseases, from risk assessment to follow up survival. Information discussed in this article may be useful in developing new intervention and therapeutic targets.

REFERENCES

1. E. Ferrannini, A. Natali, S. Camastra, M. Nannipieri, A. Mari, K. P. Adam, M. V. Milburn, G. Kastenmuller, J. Adamski, T. Tuomi, V. Lyssenko, L. Groop and W. E. Gall, *Diabetes*, 2013, **62**, 1730–1737.
2. M. Verma, M. J. Khoury and J. P. Ioannidis, *Cancer Epidemiol., Biomarkers Prev.*, 2013, **22**, 189–200.
3. A. W. Opstal-van Winden, W. Rodenburg, J. L. Pennings, C. T. van Oostrom, J. H. Beijnen, P. H. Peeters, C. H. van Gils and A. de Vries, *Int. J. Mol. Sci.*, 2012, **13**, 13587–13604.
4. C. S. Zhu, P. F. Pinsky, D. W. Cramer, D. F. Ransohoff, P. Hartge, R. M. Pfeiffer, N. Urban, G. Mor, R. C. Bast, Jr., L. E. Moore, A. E. Lokshin, M. W. McIntosh, S. J. Skates, A. Vitonis, Z. Zhang, D. C. Ward, J. T. Symanowski, A. Lomakin, E. T. Fung, P. M. Sluss, N. Scholler, K. H. Lu, A. M. Marrangoni, C. Patriotis, S. Srivastava, S. S. Buys and C. D. Berg, *Cancer Prev. Res.*, 2011, **4**, 375–383.
5. M. Verma, *Methods Mol. Biol.*, 2012, **863**, 467–480.
6. J. M. Yi, M. Dhir, L. Van Neste, S. R. Downing, J. Jeschke, S. C. Glockner, M. de Freitas Calmon, C. M. Hooker, J. M. Funes, C. Boshoff, K. M. Smits, M. van Engeland, M. P. Weijenberg, C. A. Iacobuzio-Donahue, J. G. Herman, K. E. Schuebel, S. B. Baylin and N. Ahuja, *Clin. Cancer Res.*, 2011, **17**, 1535–1545.
7. D. C. Koestler, C. J. Marsit, B. C. Christensen, W. Accomando, S. M. Langevin, E. A. Houseman, H. H. Nelson, M. R. Karagas, J. K. Wiencke and K. T. Kelsey, *Cancer Epidemiol., Biomarkers Prev.*, 2012, **21**, 1293–1302.
8. A. Johnson, Q. Song, P. Ko Ferrigno, P. R. Bueno and J. J. Davis, *Anal. Chem.*, 2012, **84**, 6553–6560.
9. W. Gao, J. Xu and Y. Q. Shu, *Expert Rev. Respir. Med.*, 2011, **5**, 699–709.

10. C. A. Castaneda, M. T. Agullo-Ortuno, J. A. Fresno Vara, H. Cortes-Funes, H. L. Gomez and E. Ciruelos, *Expert Rev. Anticancer Ther.*, 2011, **11**, 1265–1275.
11. J. E. Joh, N. N. Esposito, J. V. Kiluk, C. Laronga, M. C. Lee, L. Loftus, H. Soliman, J. C. Boughey, C. Reynolds, T. J. Lawton, P. I. Acs, L. Gordan and G. Acs, *Oncologist*, 2011, **16**, 1520–1526.
12. E. A. Slodkowska and J. S. Ross, *Expert Rev. Mol. Diagn.*, 2009, **9**, 417–422.
13. B. A. Teicher, *Curr. Opin. Pharmacol.*, 2010, **10**, 397–404.
14. J. Pang, W. P. Liu, X. P. Liu, L. Y. Li, Y. Q. Fang, Q. P. Sun, S. J. Liu, M. T. Li, Z. L. Su and X. Gao, *J. Proteome Res.*, 2010, **9**, 216–226.
15. A. M. Abdel Rahman, S. D. Kamath, S. Gagne, A. L. Lopata and R. Helleur, *J. Proteome Res.*, 2013, **12**, 647–656.
16. I. Rahman, R. Atout, N. L. Pedersen, U. de Faire, J. Frostegard, E. Ninio, A. M. Bennet and P. K. Magnusson, *Atherosclerosis*, 2011, **218**, 117–122.
17. G. Rebovich, E. J. Duffis and L. R. Caplan, *Expert Opin. Med. Diagn.*, 2010, **4**, 267–279.
18. A. Zhang, H. Sun and X. Wang, *Obes. Rev.*, 2012, **14**(4), 344–349.
19. X. Li, H. F. Tse and L. J. Jin, *J. Dent. Res.*, 2011, **90**, 1062–1069.
20. J. Ahn, C. Y. Chen and R. B. Hayes, *Cancer, Causes Control,*, 2012, **23**, 399–404.
21. C. Luaces-Cubells, S. Mintegi, J. J. Garcia-Garcia, E. Astobiza, R. Garrido-Romero, J. Velasco-Rodriguez and J. Benito, *Pediatr. Infect. Dis. J.*, 2012, **31**, 645–647.
22. M. Verma, *Ann. N. Y. Acad. Sci.*, 2003, **983**, 170–180.
23. H. Kuper, H. O. Adami and D. Trichopoulos, *J. Intern. Med.*, 2000, **248**, 171–183.
24. D. M. Parkin, *Int. J. Cancer*, 2006, **118**, 3030–3044.
25. H. zur Hausen, *Eur. J. Cancer*, 1999, **35**, 1174–1181.
26. C. P. Wild and P. C. Turner, *Mutagenesis*, 2002, **17**, 471–481.
27. H. Sun, A. Zhang, G. Yan, C. Piao, W. Li, C. Sun, X. Wu, X. Li, Y. Chen and X. Wang, *Mol. Cell. Proteomics*, 2013, **12**, 710–719.
28. F. L. van de Veerdonk, P. C. Wever, M. H. Hermans, R. Fijnheer, L. A. Joosten, J. W. van der Meer, M. G. Netea and P. M. Schneeberger, *J. Infect. Dis.*, 2012, **206**, 197–201.
29. S. F. Graham, O. P. Chevallier, D. Roberts, C. Holscher, C. T. Elliott and B. D. Green, *Anal. Chem.*, 2013, **85**, 1803–1811.
30. P. Vanek, O. Bradac, P. DeLacy, K. Saur, T. Belsan and V. Benes, *Spine*, 2012, **37**, 1645–1651.
31. M. Verma, *Curr. Opin. Clin. Nutr. Metab. Care*, 2013, **16**(4), 376–384.
32. B. K. Dunn, M. Verma and A. Umar, *Ann. N. Y. Acad. Sci.*, 2003, **983**, 1–4.
33. M. Verma, *Curr. Genomics*, 2012, **13**, 308–313.
34. M. Faryna, C. Konermann, S. Aulmann, J. L. Bermejo, M. Brugger, S. Diederichs, J. Rom, D. Weichenhan, R. Claus, M. Rehli, P. Schirmacher, H. P. Sinn, C. Plass and C. Gerhauser, *FASEB J.*, 2012, **26**, 4937–4950.

35. B. Poudel, A. Mittal, R. Shrestha, A. K. Nepal and P. S. Shukla, *Asian Pac. J. Cancer Prev.*, 2012, **13**, 2149–2152.
36. I. Casanova-Salas, J. Rubio-Briones, A. Fernandez-Serra and J. A. Lopez-Guerrero, *Clin. Transl. Oncol.*, 2012, **14**, 803–811.
37. L. C. Soliman, Y. Hui, A. K. Hewavitharana and D. D. Chen, *J. Chrom. A*, 2012, **1267**, 162–169.
38. L. Murphy and R. W. Watson, *Nat. Rev. Urol.*, 2012, **9**, 464–472.
39. R. Kuner, J. C. Brase, H. Sultmann and D. Wuttig, *Methods*, 2013, **59**, 132–137.
40. S. K. Martin, T. B. Vaughan, T. Atkinson, H. Zhu and N. Kyprianou, *Oncol. Rep.*, 2012, **28**, 409–417.
41. L. Ng, N. Karunasinghe, C. S. Benjamin and L. R. Ferguson, *N. Z. Med. J.*, 2012, **125**, 59–86.
42. J. R. Prensner, M. A. Rubin, J. T. Wei and A. M. Chinnaiyan, *Sci. Transl. Med.*, 2012, **4**, 127rv123.
43. S. Souchelnytskyi, M. Lomnytska, A. Dubrovska, U. Hellman and N. Volodko, *Proteomics*, 2006, **6**(Suppl 2), 65–68.
44. P. D. Hayes, D. A. Payne, N. J. Evans, M. M. Thompson, N. J. London, P. R. Bell and A. R. Naylor, *Eur. J. Vasc. Endovasc.*, 2003, **26**, 665–669.
45. M. P. Endsley and M. Zhang, *Methods Enzymol.*, 2011, **499**, 149–165.
46. E. S. Lianidou, A. Markou and A. Strati, *Cancer Metastasis Rev.*, 2012, **31**, 663–671.
47. A. Alvarez Secord, K. M. Darcy, A. Hutson, Z. Huang, P. S. Lee, E. L. Jewell, L. J. Havrilesky, M. Markman, F. Muggia and S. K. Murphy, *Gynecol. Oncol.*, 2011, **123**, 314–319.
48. Z. J. Xie, G. Chen, X. C. Zhang, D. F. Li, J. Huang and Z. J. Li, *Asian Pac. J. Cancer Prev.*, 2012, **13**, 6145–6149.
49. M. E. Arellano-Garcia, S. Hu, J. Wang, B. Henson, H. Zhou, D. Chia and D. T. Wong, *Oral Dis.*, 2008, **14**, 705–712.
50. C. A. Righini, F. de Fraipont, J. F. Timsit, C. Faure, E. Brambilla, E. Reyt and M. C. Favrot, *Clin. Cancer Res.*, 2007, **13**, 1179–1185.
51. Z. Yalniz, S. Demokan, Y. Suoglu, M. Ulusan and N. Dalay, *Mol. Biol. Rep.*, 2010, **37**, 3541–3545.
52. L. L. Gleich, J. Wang, J. L. Gluckman and C. M. Fenoglio-Preiser, *ORL J. Otorhinolaryngol. Relat. Spec.*, 2003, **65**, 193–198.
53. G. Cuda, A. Gallelli, A. Nistico, P. Tassone, V. Barbieri, P. S. Tagliaferri, F. S. Costanzo, C. M. Tranfa and S. Venuta, *Lung Cancer*, 2000, **30**, 211–214.
54. Z. Q. Yuan, B. Legendre, D. Q. Cai, J. Cao, J. Zhu and T. K. Weber, *Pathology*, 2009, **41**, 393–394.
55. A. I. Shemirani, M. M. Haghighi, S. M. Zadeh, S. R. Fatemi, M. Y. Taleghani, N. Zali, Z. Akbari, S. M. Kashfi and M. R. Zali, *Asian Pac. J. Cancer Prev.*, 2011, **12**, 2101–2104.
56. H. Danaee, H. H. Nelson, M. R. Karagas, A. R. Schned, T. D. Ashok, T. Hirao, A. E. Perry and K. T. Kelsey, *Oncogene*, 2002, **21**, 4894–4899.

57. S. Mourah, O. Cussenot, V. Vimont, F. Desgrandchamps, P. Teillac, B. Cochant-Priollet, A. Le Duc, J. Fiet and H. Soliman, *Int. J. Cancer*, 1998, **79**, 629–633.
58. U. Wahn, R. L. Bergmann and R. Nickel, *Clin. Exp. Allergy*, 1998, **28**(Suppl 1), 20–21discussion 32–26discussion 32–26.
59. I. Okur, L. Tumer, F. S. Ezgu, E. Yesilkaya, A. Aral, S. O. Oktar, A. Bideci and A. Hasanoglu, *J. Pediatr. Endocrinol. Metab.*, 2013, 1–6.
60. M. Kucukazman, N. Ata, B. Yavuz, K. Dal, O. Sen, O. S. Deveci, K. Agladioglu, A. O. Yeniova, Y. Nazligul and D. T. Ertugrul, *Eur. J. Gastroenterol. Hepatol.*, 2013, **25**, 147–151.
61. *Duke Medicine Health News*, 2012, 18, 4–5.
62. E. F. Kern, P. Erhard, W. Sun, S. Genuth and M. F. Weiss, *Am. J. Kidney Dis.*, 2010, **55**, 824–834.
63. A. E. Mehta, *Cleve. Clin. J. Med.*, 1995, **62**, 210–211.
64. D. P. Dosanjh, M. Bakir, K. A. Millington, A. Soysal, Y. Aslan, S. Efee, J. J. Deeks and A. Lalvani, *PloS One*, 2011, **6**, e28754.
65. M. Sakamoto, *J. Gastroenterol.*, 2009, **44**(Suppl 19), 108–111.
66. M. Di Luca, E. Grossi, B. Borroni, M. Zimmermann, E. Marcello, F. Colciaghi, F. Gardoni, M. Intraligi, A. Padovani and M. Buscema, *J. Transl. Med.*, 2005, **3**, 30.
67. M. Verma, G. L. Wright, Jr., S. M. Hanash, R. Gopal-Srivastava and S. Srivastava, *Ann. N. Y. Acad. Sci.*, 2001, **945**, 103–115.
68. I. Simon, Y. Liu, K. L. Krall, N. Urban, R. L. Wolfert, N. W. Kim and M. W. McIntosh, *Gynecol. Oncol.*, 2007, **106**, 112–118.
69. V. Kashuba, A. A. Dmitriev, G. S. Krasnov, T. Pavlova, I. Ignatjev, V. V. Gordiyuk, A. V. Gerashchenko, E. A. Braga, S. P. Yenamandra, M. Lerman, V. N. Senchenko and E. Zabarovsky, *Int. J. Mol. Sci.*, 2012, **13**, 13352–13377.
70. J. Ren, H. Cai, Y. Li, X. Zhang, Z. Liu, J. S. Wang, Y. L. Hwa, Y. Zhang, Y. Yang and S. W. Jiang, *Expert Rev. Mol. Diagn.*, 2010, **10**, 787–798.
71. S. Lindstrom, F. Schumacher, A. Siddiq, R. C. Travis, D. Campa, S. I. Berndt, W. R. Diver, G. Severi, N. Allen, G. Andriole, B. Bueno-de-Mesquita, S. J. Chanock, D. Crawford, J. M. Gaziano, G. G. Giles, E. Giovannucci, C. Guo, C. A. Haiman, R. B. Hayes, J. Halkjaer, D. J. Hunter, M. Johansson, R. Kaaks, L. N. Kolonel, C. Navarro, E. Riboli, C. Sacerdote, M. Stampfer, D. O. Stram, M. J. Thun, D. Trichopoulos, J. Virtamo, S. J. Weinstein, M. Yeager, B. Henderson, J. Ma, L. Le Marchand, D. Albanes and P. Kraft, *PloS One*, 2011, **6**, e17142.
72. K. K. Tsilidis, R. C. Travis, P. N. Appleby, N. E. Allen, S. Lindstrom, F. R. Schumacher, D. Cox, A. W. Hsing, J. Ma, G. Severi, D. Albanes, J. Virtamo, H. Boeing, H. B. Bueno-de-Mesquita, M. Johansson, J. R. Quiros, E. Riboli, A. Siddiq, A. Tjonneland, D. Trichopoulos, R. Tumino, J. M. Gaziano, E. Giovannucci, D. J. Hunter, P. Kraft, M. J. Stampfer, G. G. Giles, G. L. Andriole, S. I. Berndt, S. J. Chanock, R. B. Hayes and T. J. Key, *Am. J. Epidemiol.*, 2012, **175**, 926–935.

73. F. Li, S. Li, H. Chang, Y. Nie, L. Zeng, X. Zhang and Y. Wang, *Genet. Test. Mol. Biomarkers*, 2013.
74. D. R. Taylor, *Thorax*, 2009, **64**, 261–264.
75. S. A. Peters, F. L. Visseren and D. E. Grobbee, *Nat. Rev. Cardiol.*, 2013, **10**, 12–14.
76. G. B. Lim, *Nat. Rev. Cardiol.*, 2012, **9**, 672.
77. J. Madden, C. M. Williams, P. C. Calder, G. Lietz, E. A. Miles, H. Cordell, J. C. Mathers and A. M. Minihane, *Annu. Rev. Nutr.*, 2011, **31**, 203–234.
78. N. Sattar, O. Scherbakova, I. Ford, D. S. O'Reilly, A. Stanley, E. Forrest, P. W. Macfarlane, C. J. Packard, S. M. Cobbe and J. Shepherd, *Diabetes*, 2004, **53**, 2855–2860.
79. A. Betriu-Bars and E. Fernandez-Giraldez, *Nefrologia*, 2012, **32**, 7–11.
80. Y. Shiferaw, A. Alemu, A. Girma, A. Getahun, A. Kassa, A. Gashaw, T. Teklu and B. Gelaw, *BMC Res. Notes*, 2011, **4**, 505.
81. G. B. Gerbi, T. Habtemariam, B. Tameru, D. Nganwa and V. Robnett, *AIDS Care*, 2012, **24**, 331–339.
82. T. Oni, H. P. Gideon, N. Bangani, R. Tsekela, R. Seldon, K. Wood, K. A. Wilkinson, R. T. Goliath, T. H. Ottenhoff and R. J. Wilkinson, *Clin. Vaccine Immunol.*, 2012, **19**, 1243–1247.
83. M. F. Humblet, M. Gilbert, M. Govaerts, M. Fauville-Dufaux, K. Walravens and C. Saegerman, *J. Clin. Microbiol.*, 2010, **48**, 2802–2808.
84. X. F. Wang, Y. W. Cao, Z. Z. Feng, D. Fu, Y. S. Ma, F. Zhang, X. X. Jiang and Y. C. Shao, *Mol. Biol. Rep.*, 2013, **40**, 779–785.
85. M. Silvestrini, G. Viticchi, C. Altamura, S. Luzzi, C. Balucani and F. Vernieri, *J. Alzheimer's Dis.*, 2012, **32**, 689–698.
86. H. M. Schipper, *Alzheimer's Dementia*, 2011, **7**, e118–e123.
87. F. E. Ahmed, P. Vos, S. iJames, D. T. Lysle, R. R. Allison, G. Flake, D. R. Sinar, W. Naziri, S. P. Marcuard and R. Pennington, *Cancer Genomics Proteomics*, 2007, **4**, 1–20.
88. W. Chung, B. Kwabi-Addo, M. Ittmann, J. Jelinek, L. Shen, Y. Yu and J. P. Issa, *PloS One*, 2008, **3**, e2079.
89. T. Shimbo, A. Tanemura, T. Yamazaki, K. Tamai, I. Katayama and Y. Kaneda, *PloS One*, 2010, **5**, e10566.
90. D. J. Sargent, M. B. Resnick, M. O. Meyers, A. Goldar-Najafi, T. Clancy, S. Gill, G. O. Siemons, Q. Shi, B. M. Bot, T. T. Wu, G. Beaudry, J. F. Haince and Y. Fradet, *Ann. Surg. Oncol.*, 2011, **18**, 3261–3270.
91. W. Han, H. Kim, S. Y. Ku, S. H. Kim, Y. M. Choi, J. G. Kim and S. Y. Moon, *Gynecol. Endocrinol.*, 2013, **29**, 16–19.
92. M. G. Tektonidou and M. M. Ward, *Nat. Rev. Rheumatol.*, 2011, **7**, 708–717.
93. D. Plant, I. Ibrahim, M. Lunt, S. Eyre, E. Flynn, K. L. Hyrich, A. W. Morgan, A. G. Wilson, J. D. Isaacs and A. Barton, *Arthritis Res. Ther.*, 2012, **14**, R214.
94. I. Ibrahim, S. A. Owen and A. Barton, *Expert Rev. Clin. Immunol.*, 2012, **8**, 509–511.

95. R. Mogelvang, S. H. Pedersen, A. Flyvbjerg, M. Bjerre, A. Z. Iversen, S. Galatius, J. Frystyk and J. S. Jensen, *Am. J. Cardiol.*, 2012, **109**, 515–520.
96. C. C. Chen, J. L. Huang, C. J. Chang and M. S. Kong, *J. Pediatr. Gastroenterol. Nutr.*, 2012, **55**, 541–547.
97. K. B. Walhovd, A. M. Fjell, J. Brewer, L. K. McEvoy, C. Fennema-Notestine, D. J. Hagler, Jr., R. G. Jennings, D. Karow and A. M. Dale, *AJNR. American journal of neuroradiology*, 2010, **31**, 347–354.
98. K. A. Kolquist, R. A. Schultz, A. Furrow, T. C. Brown, J. Y. Han, L. J. Campbell, M. Wall, M. L. Slovak, L. G. Shaffer and B. C. Ballif, *Cancer Genet*, 2011, **204**, 603–628.
99. A. S. Patel, A. L. Hawkins and C. A. Griffin, *Curr. Opin. Oncol.*, 2000, **12**, 62–67.
100. J. M. Cowan, *Otolaryngol. Clin. North Am.*, 1992, **25**, 1073–1087.
101. Z. Gibas and L. Gibas, *Cancer Genet. Cytogenet.*, 1997, **95**, 108–115.
102. D. Geleick, H. Muller, A. Matter, J. Torhorst and U. Regenass, *Cancer Genet. Cytogenet.*, 1990, **46**, 217–229.
103. R. P. Pearlstein, M. S. Benninger, T. E. Carey, R. J. Zarbo, F. X. Torres, B. A. Rybicki and D. L. Dyke, *Genes, Chromosomes Cancer*, 1998, **21**, 333–339.
104. D. W. Chang, J. Gu and X. Wu, *Urol. Oncol.*, 2012, **30**, 524–532.
105. B. Ludewig, J. V. Stein, J. Sharpe, L. Cervantes-Barragan, V. Thiel and G. Bocharov, *Eur. J. Immunol.*, 2012, **42**, 3116–3125.
106. V. Narang, J. Decraene, S. Y. Wong, B. S. Aiswarya, A. R. Wasem, S. R. Leong and A. Gouaillard, *Immunol. Res.*, 2012, **53**, 251–265.
107. J. D. Bregman, *J. Am. Acad Child. Adolesc. Psychiatry*, 2012, **51**, 1113–1115.
108. L. M. Oberman, *Expert Opin. Invest. Drugs*, 2012, **21**, 1819–1825.
109. P. Karolina, R. Chlup, Z. Jana, K. D. Kohnert, P. Kudlova, J. Bartek, M. Nakladalova, B. Doubravova and P. Seckar, *J. Diabetes Sci. Technol.*, 2010, **4**, 983–992.
110. S. Jaijakul, J. A. Vazquez, R. N. Swanson and L. Ostrosky-Zeichner, *Clin. Infect. Dis.*, 2012, **55**, 521–526.
111. T. R. Powell, L. C. Schalkwyk, A. L. Heffernan, G. Breen, T. Lawrence, T. Price, A. E. Farmer, K. J. Aitchison, I. W. Craig, A. Danese, C. Lewis, P. McGuffin, R. Uher, K. E. Tansey and U. M. D'Souza, *Eur. Neuropsychopharmacol.*, 2012, **23**(9), 1105–1114.
112. R. E. McCullumsmith, J. Hammond, A. Funk and J. H. Meador-Woodruff, *Curr. Pharm. Biotechnol.*, 2012, **13**, 1535–1542.
113. H. M. Ibrahim and C. A. Tamminga, *Curr. Pharm. Biotechnol.*, 2012, **13**, 1587–1594.
114. J. Mullenders, W. von der Saal, M. M. van Dongen, U. Reiff, R. van Willigen, R. L. Beijersbergen, G. Tiefenthaler, C. Klein and R. Bernards, *Clin. Cancer Res.*, 2009, **15**, 5811–5819.
115. S. Maier, C. Dahlstroem, C. Haefliger, A. Plum and C. Piepenbrock, *Am. J. PharmacoGenomicse*, 2005, **5**, 223–232.

116. D. B. Longley, W. L. Allen and P. G. Johnston, *Biochim. Biophys. Acta*, 2006, **1766**, 184–196.
117. A. M. Tsimberidou, N. G. Iskander, D. S. Hong, J. J. Wheler, G. S. Falchook, S. Fu, S. Piha-Paul, A. Naing, F. Janku, R. Luthra, Y. Ye, S. Wen, D. Berry and R. Kurzrock, *Clin. Cancer Res.*, 2012, **18**, 6373–6383.
118. H. Wong and T. Yau, *Oncologist*, 2012, **17**, 346–358.
119. C. B. Weldon, J. R. Trosman, W. J. Gradishar, A. B. Benson, 3rd and J. C. Schink, *J. Oncol. Pract.*, 2012, **8**, e24–e31.
120. L. G. van Baarsen, C. A. Wijbrandts, D. M. Gerlag, F. Rustenburg, T. C. van der Pouw Kraan, B. A. Dijkmans, P. P. Tak and C. L. Verweij, *Genes Immun.*, 2010, **11**, 622–629.
121. S. Marsal and A. Julia, *Pharmacogenomics*, 2010, **11**, 617–619.
122. M. P. Grimaldi, S. Vasto, C. R. Balistreri, D. di Carlo, M. Caruso, E. Incalcaterra, D. Lio, C. Caruso and G. Candore, *Ann. N. Y. Acad. Sci.*, 2007, **1100**, 123–131.
123. N. Siitonen, L. Pulkkinen, J. Lindstrom, M. Kolehmainen, U. Schwab, J. G. Eriksson, P. Ilanne-Parikka, S. Keinanen-Kiukaanniemi, J. Tuomilehto and M. Uusitupa, *Cardiovasc. Diabetol.*, 2011, **10**, 83.
124. G. Ragia and V. G. Manolopoulos, *Pharmacogenomics*, 2012, **13**, 261–264.
125. P. D. Roy, M. Majumder and B. Roy, *Pharmacogenomics*, 2008, **9**, 311–321.
126. F. Martinelli-Boneschi, G. Giacalone, G. Magnani, G. Biella, E. Coppi, R. Santangelo, P. Brambilla, F. Esposito, S. Lupoli, F. Clerici, L. Benussi, R. Ghidoni, D. Galimberti, R. Squitti, A. Confaloni, G. Bruno, S. Pichler, M. Mayhaus, M. Riemenschneider, C. Mariani, G. Comi, E. Scarpini, G. Binetti, G. Forloni, M. Franceschi and D. Albani, *Neurobiol. Aging*, 2013, **34**, 1711.e7–1711.e13.
127. F. Listi, C. Caruso, D. Lio, G. Colonna-Romano, M. Chiappelli, F. Licastro and G. Candore, *J. Alzheimer's Dis.*, 2010, **19**, 551–557.
128. E. Dreussi, P. Biason, G. Toffoli and E. Cecchin, *Pharmacogenomics*, 2012, **13**, 1635–1650.
129. D. T. Merrick, *Chest*, 2012, **141**, 1377–1378.
130. C. Justenhoven, O. Obazee and H. Brauch, *Pharmacogenomics*, 2012, **13**, 659–675.
131. M. Nishiyama and H. Eguchi, *Adv. Drug Delivery Rev.*, 2009, **61**, 402–407.
132. T. Kohriyama, *Rinsho Shinkeigaku*, 2011, **51**, 179–187.
133. S. Malucchi, F. Gilli, M. Caldano, F. Marnetto, P. Valentino, L. Granieri, A. Sala, M. Capobianco and A. Bertolotto, *Neurology*, 2008, **70**, 1119–1127.
134. C. A. Braun Hashemi, Y. C. Zang, J. A. Arbona, J. A. Bauerle, M. L. Frazer, H. Lee, L. Flury, E. S. Moore, M. C. Kolar, R. Y. Washington and O. J. Kolar, *Mult. Scler.*, 2006, **12**, 652–658.
135. A. Miller, L. Glass-Marmor, M. Abraham, I. Grossman, S. Shapiro and Y. Galboiz, *Clin. Neurol. Neurosurg.*, 2004, **106**, 249–254.
136. J. Zhang, C. Wang, L. Li, Q. Man, L. Meng, P. Song, L. Froyland and Z. Y. Du, *Br. J. Nutr.*, 2012, **108**, 1455–1465.

137. S. Sour, M. Belarbi, D. Khaldi, N. Benmansour, N. Sari, A. Nani, F. Chemat and F. Visioli, *Br. J. Nutr.*, 2012, **107**, 1800–1805.
138. L. Ozick, P. Chander, A. Agarwal and A. Soni, *American J. Gastroenterol.*, 1989, **84**, 195–197.
139. A. Maisel, S. X. Neath, J. Landsberg, C. Mueller, R. M. Nowak, W. F. Peacock, P. Ponikowski, M. Mockel, C. Hogan, A. H. Wu, M. Richards, P. Clopton, G. S. Filippatos, S. Di Somma, I. Anand, L. L. Ng, L. B. Daniels, R. H. Christenson, M. Potocki, J. McCord, G. Terracciano, O. Hartmann, A. Bergmann, N. G. Morgenthaler and S. D. Anker, *Eur. J. Heart Failure*, 2012, **14**, 278–286.
140. F. P. Manfredsson, M. S. Okun and R. J. Mandel, *Curr. Gene Ther.*, 2009, **9**, 375–388.
141. D. C. Gilbert, A. J. Chalmers and S. F. El-Khamisy, *Br. J. Cancer*, 2012, **106**, 18–24.
142. E. D. Hsi, S. H. Jung, R. Lai, J. L. Johnson, J. R. Cook, D. Jones, S. Devos, B. D. Cheson, L. E. Damon and J. Said, *Leuk. Lymphoma*, 2008, **49**, 2081–2090.
143. P. C. Tsai, F. J. Hernandez-Ilizaliturri, N. Bangia, S. H. Olejniczak and M. S. Czuczman, *Clin. Cancer Res.*, 2012, **18**, 1039–1050.
144. M. Giachelia, M. T. Voso, M. C. Tisi, M. Martini, V. Bozzoli, G. Massini, F. D'Alo, L. M. Larocca, G. Leone and S. Hohaus, *Leuk. Lymphoma*, 2012, **53**, 411–416.
145. M. P. Chao, A. A. Alizadeh, C. Tang, J. H. Myklebust, B. Varghese, S. Gill, M. Jan, A. C. Cha, C. K. Chan, B. T. Tan, C. Y. Park, F. Zhao, H. E. Kohrt, R. Malumbres, J. Briones, R. D. Gascoyne, I. S. Lossos, R. Levy, I. L. Weissman and R. Majeti, *Cell*, 2010, **142**, 699–713.
146. J. N. Sampson, S. M. Boca, X. O. Shu, R. Z. Stolzenberg-Solomon, C. E. Matthews, A. W. Hsing, Y. T. Tan, B. T. Ji, W. H. Chow, Q. Cai, D. K. Liu, G. Yang, Y. B. Xiang, W. Zheng, R. Sinha, A. J. Cross and S. C. Moore, *Cancer Epidemiol., Biomarkers Prev.*, 2013.
147. X. Zhao, J. Fritsche, J. Wang, J. Chen, K. Rittig, P. Schmitt-Kopplin, A. Fritsche, H. U. Haring, E. D. Schleicher, G. Xu and R. Lehmann, *Metabolomics*, 2010, **6**, 362–374.
148. E. P. Rhee and R. E. Gerszten, *Clin. Chem.*, 2012, **58**, 139–147.
149. J. M. Yuan, Y. T. Gao, S. E. Murphy, S. G. Carmella, R. Wang, Y. Zhong, K. A. Moy, A. B. Davis, L. Tao, M. Chen, S. Han, H. H. Nelson, M. C. Yu and S. S. Hecht, *Cancer Res.*, 2011, **71**, 6749–6757.
150. C. Abate-Shen and M. M. Shen, *Nature*, 2009, **457**, 799–800.
151. T. Zhang, X. Wu, C. Ke, M. Yin, Z. Li, L. Fan, W. Zhang, H. Zhang, F. Zhao, X. Zhou, G. Lou and K. Li, *J. Proteome Res.*, 2013, **12**, 505–512.
152. M. Chadeau-Hyam, T. M. Ebbels, I. J. Brown, Q. Chan, J. Stamler, C. C. Huang, M. L. Daviglus, H. Ueshima, L. Zhao, E. Holmes, J. K. Nicholson, P. Elliott and M. De Iorio, *J Proteome Res.*, 2010, **9**, 4620–4627.
153. C. Gieger, L. Geistlinger, E. Altmaier, M. Hrabe de Angelis, F. Kronenberg, T. Meitinger, H. W. Mewes, H. E. Wichmann, K. M.

Weinberger, J. Adamski, T. Illig and K. Suhre, *PLoS Genetics*, 2008, **4**, e1000282.
154. R. W. Blamey, I. O. Ellis, S. E. Pinder, A. H. Lee, R. D. Macmillan, D. A. Morgan, J. F. Robertson, M. J. Mitchell, G. R. Ball, J. L. Haybittle and C. W. Elston, *Eur. J. Cancer*, 2007, **43**, 1548–1555.
155. S. C. Lee, X. Xu, W. J. Chng, M. Watson, Y. W. Lim, C. I. Wong, P. Iau, N. Sukri, S. E. Lim, H. L. Yap, S. A. Buhari, P. Tan, J. Guo, B. Chuah, H. L. McLeod and B. C. Goh, *Pharmacogenet. Genomics*, 2009, **19**, 833–842.
156. C. Sotiriou and L. Pusztai, *N. Engl. J. Med.*, 2009, **360**, 790–800.
157. G. Ouyang, Y. Chen, L. Setkova and J. Pawliszyn, *J. Chrom. A*, 2005, **1097**, 9–16.
158. C. L. Silva, M. Passos and J. S. Camara, *Talanta*, 2012, **89**, 360–368.
159. D. Y. Wang, S. J. Done, D. R. McCready, S. Boerner, S. Kulkarni and W. L. Leong, *Breast Cancer Res.*, 2011, **13**, R92.
160. S. S. Tang and G. P. Gui, *Biomarkers Med.*, 2012, **6**, 567–585.
161. S. Srivastava, *Gastrointest. Cancer Res. : GCR*, 2007, **1**, S60–63.

III.4 Article III: Literature Review

In Silico Subtractive Genomics for Target Identification in Human Bacterial Pathogens

Debmalya Barh, Sandeep Tiwari, Neha Jain, Amjad Ali, Anderson Rodrigues Santos, Amarendra Narayan Misra, **Vasco Azevedo**, and Anil Kumar

Drug Development Research, 2011; 72 (2), 162-177
Impact Factor: 1.19 (2011)

In anti-bacterial drug and vaccine discovery process, target identification is the first step. While genomics approaches are concerned, subtractive and pan-genomics are well accepted and effective strategies in this regard. In this chapter, various aspects of subtractive genomics such as concept of essential non-host homologue genes, genome subtraction, subtraction based methodologies and tools, subtractive genomics based identified bacterial targets, application in reverse vaccinology, advantages and disadvantages of the methods, and future perspectives of subtractive genomics are discussed.

DDR

DRUG DEVELOPMENT RESEARCH 72:162–177 (2011)

Research Overview

In Silico Subtractive Genomics for Target Identification in Human Bacterial Pathogens

Debmalya Barh,[1,4]* Sandeep Tiwari,[2] Neha Jain,[2] Amjad Ali,[3] Anderson Rodrigues Santos,[3] Amarendra Narayan Misra,[4] Vasco Azevedo,[3] and Anil Kumar[2]

[1]*Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur WB-721172, India*

[2]*School of Biotechnology, Devi Ahilya University, Khandwa Rd., Indore 452001, India*

[3]*Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP 486, CEP 31270-901, Belo Horizonte, Minas Gerais, Brazil*

[4]*Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, 756020 Orissa, India*

Strategy, Management and Health Policy				
Enabling Technology, Genomics, Proteomics	Preclinical Research	Preclinical Development Toxicology, Formulation Drug Delivery, Pharmacokinetics	Clinical Development Phases I-III Regulatory, Quality, Manufacturing	Postmarketing Phase IV

ABSTRACT Target identification is the first step in the drug and vaccine discovery process; *in silico* subtractive genomics is widely used in this process. Using this approach, in recent years, a large number of targets have been identified for bacterial pathogens that are either drug resistant or for which no suitable vaccine is available; most such reports concern a specific pathogen. The *in silico* method reduces the time as well as the cost of target screening. Although a powerful technique that can be applied to a wide range of pathogens, there are many pitfalls in the analysis and interpretation of the data. We review this approach, including targets that have been identified with this technique and various other aspects, including advantages and disadvantages. We also discuss our own experiences using this technology. Drug Dev Res 72:162–177, 2011. © 2010 Wiley-Liss, Inc.

Key words: drug target; essential genes; subtractive genomics; bacterial pathogen

INTRODUCTION

Although high-throughput techniques and synthetic chemistry are an integral part of today's drug discovery process, accelerating the process manifold, the introduction of a new drug on the market still takes 10–15 years and therefore requires a huge investment [Plotkin, 2005]. Technological advancements, along with improved and innovative strategies, could reduce the cost and the time required to develop a new drug.

Most infectious diseases are caused by bacterial pathogens. An increase of 58% in the mortality rate due to such infectious diseases has been reported from 1980 to 1992 in the United States [Pinner et al., 1996]. According to the 2004 World Health Organization Report [www.who.int/whr/2004/annex/topic/en/annex_2_en.pdf], 16.4 million people died worldwide in that year from bacterial infectious diseases. Although several antibiotics are currently available for each bacterial pathogen, the emerging drug-resistant strains of such pathogens make them difficult to control,

*Correspondence to: Debmalya Barh, Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur WB-721172, India. E-mail: dr.barh@gmail.com

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/ddr.20413

perhaps due to decades-old uses of such drugs in human patients [Arias and Murray, 2009].

Drug target identification is the first step in the drug discovery process [Chan et al., 2010]. Because of the availability of both pathogen and host–genome sequences, it has become easier to identify drug targets at the genomic level for any given pathogen [Allsop, 1998; Stumm et al., 2002; Meinke et al., 2004; Owa, 2007]. In recent years, the strategies are shifting progressively from a generic approach to genomic and metabolomic approaches [Ishii et al., 2004; Lin and Qian, 2007] to identify novel drug targets that are required to design new defenses against antibiotic-resistant pathogens [McDevitt and Rosenberg, 2001; Mills, 2006; Fischbach and Walsh, 2009]. Tremendous advancements have been made in target identification and drug discovery since the human genome sequence became available [Lander et al., 2001; Venter et al., 2001].

Use of computational approaches, with integrated genomics, proteomics, transcriptomics, interactomics, signalomics, and metabolomics, are current trends in target discovery for most human diseases, especially for cancer, cardiovascular, neuroendocrine, and infectious diseases; they make the discovery process faster and more cost effective. Currently, genomics and more specifically *in silico* comparative, subtractive, and functional genomics are being widely used to identify novel drug and vaccine targets in order to develop effective antibacterial agents and vaccines against bacterial pathogens that are either resistant to existing antibacterial regimens or for which a suitable vaccine is not available [Ji, 2002; Pucci, 2006].

Computational metabolic flux modeling, along with systems approaches, have been found to be a great aid for understanding and manipulating microbial metabolism [Downs, 2006; Thykaer et al., 2009]. They can help in the identification of key essential or survival proteins (the targets) of the organism that can be inhibited by using appropriate lead molecule(s) identified by *in silico* virtual screening. Additionally, *in silico* comparative genomics-based subtraction analysis using host and pathogen genomes is a powerful approach for the identification of genus- or species-specific genes, or groups of genes that are responsible for a unique phenotype as well as the virulence factors of the pathogen [Huynen et al., 1997, 1998]. It is then necessary to determine whether these genes are essential survival genes of the pathogen and whether there are non-host homologues. Simultaneously, metabolic pathway subtraction is required to identify metabolic pathways that host and pathogen have in common and pathogen-specific pathways. Once the essential non-host homologue survival genes of a pathogen are identified, they need to be allocated to known pathways. If an essential non-host homologue survival gene is found crucial in any of the pathogen's metabolic pathways, it is considered a putative target. To identify a vaccine target, additional analyses, such as localization, antigenicity, and membrane topology, are required in order to design epitopes.

In this review we present an overview of *in silico* subtractive genomics approaches used to identify genomic targets in various human pathogenic bacteria, along with information from our own experiments using this approach. We also discuss various aspects, including advantages, disadvantages, and future prospectives for this approach.

STRATEGIES FOR SUBTRACTION-BASED TARGET DISCOVERY

Concept of Essential Non-Host Homologue Genes

Subtractive genomics-based target identification is based on essential genes and the non-host homologue. Essential genes are genes that are required for growth, adaptability and survival of an organism. Therefore, deficiency of any such gene should be lethal to the organism. Essential genes are likely to have a common function across all organisms [Mushegian and Koonin, 1996]. Often such essential genes are evolutionally conserved in different taxa [Itaya, 1995; Tatusov et al., 1997; Koonin et al., 1998; Jordan et al., 2002; Kobayashi et al., 2003]. Essential genes can be identified through random mutagenesis of bacterial genomes [Hood, 1999]. The Database of Essential Genes (DEG) [Zhang et al., 2004] is the main resource that lists experimentally validated essential genes in bacteria, fungi, plants, and animals. DEG can readily be used for target identification through comparative and subtractive genomics approaches. A non-host homologue (not present in the host but present in the pathogen) essential gene of a pathogen is considered a good target against the pathogen [Sakharkar et al., 2004]. Ideally, a target should fulfill four properties: (1) it must be an essential gene for survival or pathogenesis of the target organism; (2) druggability, i.e., having protein structure characteristics that make it amenable to bind to small inhibitor molecules; (3) functional and structural characterization, with established assays for screening small molecule inhibition; and (4) distinctness from current drug targets to avoid cross-resistance [Holman et al., 2009]. Both experimental and computational methods are available for essential gene-based target prediction; however, the computational methods are preferable as they require less time, labor, and are less expensive [Itaya, 1995; Kobayashi et al., 2003].

Genome Subtraction

Subtraction literally means "removed from below," more precisely, taking a smaller piece from a larger one. It is a mathematical approach to determine the difference between two amounts in the same category. Subtractive genomics is based on a comparative genomics approach; generally, we use two genomes and subtract the genomic data set of one from the other to obtain genus-, species-, and unique phenotype-specific genes. In target identification, the pathogen genome, within which target(s) have to be identified, is subtracted from the host genome, and subtracted genome sequences or genes (non-host homologues) are further analyzed to determine whether they are essential for pathogen survival. These essential and non-host homologues must be a critical component in vital physicochemical and metabolic pathways, so that a designed drug or a lead compound specific to such target(s) will only impact on the pathogen's system, without hampering host physiology or any aspect of host biology. Identified targets may be used to design and develop drugs, vaccines, or dual-purpose targets [Sakharkar et al., 2004; Dutta et al., 2006; Barh and Kumar, 2009; Barh et al., 2009]. In general, enzyme targets located in the cytoplasm are good candidates for drug development; exo-membrane (surface-exposed) and secreted protein targets, based on their antigenicity, can be used for peptide vaccine design. Exo-membrane enzyme or transporter targets are most suitable for dual purpose [Barh et al., 2009].

METHODS FOR IN SILICO SUBTRACTIVE GENOMICS FOR TARGET DISCOVERY

Subtractive genomics provides new opportunities for finding optimal targets among unexplored cellular functions, based on an understanding of related biological processes in bacterial pathogens and their hosts [Dong et al., 2009]. The *in silico* method follows a similar strategy of subtractive hybridization, suppressive subtractive hybridization, positional cloning, and comparative genomics that can be used for the identification of drug targets in the wet lab. This strategy was first applied to *Helicobacter pylori* [Huynen et al., 1997, 1998]. A differential genome display approach was used in this case; it relies on the fact that parasitic bacterial genomes are smaller and encode fewer proteins than a closely related free-living bacterial organisms. Hence, genes which are present in the parasitic bacterium, but absent in closely related free-living taxa, are responsible for adaptability and pathogenicity and therefore may be considered candidate targets. This strategy has evolved over time and has become much faster and more sensitive as a result of the availability of the complete genome sequence of several pathogenic bacteria, improved computational tools, and various databases. The efficiency of this method was further boosted manifold with the development and availability of DEG. This approach was successfully used for the first time to identify essential genes and targets in *Pseudomonas aeruginosa* by Saharker et al. [2004], using DEG. Since then, this approach has been widely applied with slight modifications to identify targets in several pathogenic bacteria, including *P. aeruginosa* [Sakharkar et al., 2004; Perumal et al., 2007], *H. pylori* [Dutta et al., 2006], *B. pseudomallei* [Chong et al., 2006], *A. hydrophila* [Sharma et al., 2008], *N. gonorrhoeae* [Barh and Kumar, 2009], *N. meningitides* [Sarangi et al., 2009], *M. tuberculosis* [Asif et al., 2009], *S. typhi* [Rathi et al., 2009], *M. leprae* [Shanmugam and Natarajan, 2010], and *M. pneumonia* [Gupta et al., 2010]. Table 1 lists bacterial pathogens affecting humans to which this strategy has been applied in order to identify targets in these pathogens.

Current Methodology

The NCBI Genome database (www.ncbi.nlm.nih.gov/genome), the Swiss-Prot protein database (http://us.expasy.org/sprot) [Bairoch and Apweiler, 1997], DEG (http://tubic.tju.edu.cn/deg), KEGG [Ogata et al., 1999], BLAST tools (http://blast.ncbi.nlm.nih.gov/Blast.cgi), VFDB [Chen et al., 2005], cellular localization prediction tools, such as CELLO [Yu et al., 2004], PSLpred [Bhasin et al., 2005], PSORTb [Gardy et al., 2005], and SOSUI-GramN [Imai et al., 2008], are integral parts of current subtractive genomics-based bacterial target identification strategies. In general, the host (human) and the pathogen (in which the target is to be identified) genomes and proteomes are collected from the NCBI genome server. The pathogen genome is then subjected to NCBI human BLAST to subtract the non-human homologous genes of the bacteria. Each identified non-human homologue gene and protein sequence of the pathogen is then subjected to BLASTx and BLASTp, using the bacterial BLAST option in DEG. A BLAST hit with significant cutoff values against any bacterial sequence listed in DEG gives an indication that the query sequence of the bacteria under study is a putative essential gene in the organism. Identified putative, essential non-human homologues genes (targets) are then mapped in metabolic pathways (pathogen unique and host–pathogen common) in which they are involved, using comparative pathway analysis for humans and the pathogen, available in the KEGG database. Essential non-human homologues that are crucial in pathways are identified and subsequently analyzed to determine their localization (cytoplasmic, membrane, exo-membrane, or secreted), using appropriate localization prediction tools, and enzymatic activity-related information is

TABLE 1. Human Bacterial Pathogens to Which In Silico Subtractive Genomics Strategy Has Been Applied to Identify Drug Targets*

Sl. No	Pathogens	Disease caused	Associated diseases	Prevalent countries	Pathogenesis and epidemiology	Available drugs/ antibiotics	Available vaccines	Challenges in vaccine/drug development
1	*M. tuberculosis*	Tuberculosis		Russia, Israel, China, Asia, Africa	An infectious disease that affects lungs and kills young and middle-aged adults faster than any other disease	Isoniazid, rifampicin, ethambutol, streptomycin	Bacille Calmette–Guérin (BCG)	The vaccine is efficient in preventing the disease, but the efficacy in adults is doubtful. Worldwide emergence of extensively drug-resistant tuberculosis is a serious challenge
2	*M. leprae*	Leprosy		Central Africa, Southeast Asia	An infectious disease that primarily affects the skin, mucous membranes, and peripheral nerves causing deformities. Estimated number of existing leprosy patients in the world is 12–15 million	Quinolones, refampicin, dapsone, ofloxacin	BCG	Multi-drug resistance of the strain hinders the use of antibiotics against the pathogen. Biology of the pathogen is poorly understood, hence effective drug discovery is not a priority
3	*H. pylori*	Gastric ulcer	Coronary artery, liver diseases, and MALT-type lymphoma	US, China, Korea, developing countries	Highly infectious and present in approximately 50% of world's population. Pathogen is of serious concern in developing countries. Infected individuals are at high risk of gastric cancer	Clarithromycin, rifabutin, furazolidone	*H. pylori* whole-cell (HWC) vaccine	No improved vaccine available but new vaccines are under development. Multi-drug-resistant species are predominant
4	*V. cholerae*	Endemic and epidemic cholera, secretory diarrhea		South and Central America, Asia	Highly infectious water-born disease that causes rapid loss of body fluids, leads to dehydration and shock. Without treatment, death can occur within hours. Severe cases require intravenous fluid replacement	Tetracycline, azithromycin, ciprofloxacin, erythromycin	Dukoral, Mutacol	No long-term effective vaccine available. Requires new vaccine against the pathogen. Tetracycline and Ciprofloxacin are used, but many resistant strains are reported
5	*N. gonorrhoeae*	Gonorrhoea	Conjunctivitis, pharyngitis, proctitis, prostatitis, and orchitis	US and in underdeveloped countries	Most common sexually transmitted diseases. Every year about sixty million new cases are reported. Severity causes infertility, PID, and ectopic pregnancy	Cefotaxime, cefoperazone, moxalactam, piperacillin, mezlocillin	Not available	The pathogen is highly adaptive and antibiotic resistant
6	*A. hydrophila*	Water-associated traumatic secondary wound infection	Septicaemia, cellulitis, pneumonitis, necrotizing fasciitis, and gastroenteritis	US	Infects through contaminated refrigerated animal products. Causes food poising. It can be fatal if untreated	Cefotaxime, cephalosporins	Not available	The pathogen is resistant to chlorination of water and to a variety of antibiotics

TABLE 1. Continued

Sl. No	Pathogens	Disease caused	Associated diseases	Prevalent countries	Pathogenesis and epidemiology	Available drugs/ antibiotics	Available vaccines	Challenges in vaccine/drug development
7	*B. pseudomallei*	Melioidosis, septicemia	Acute pulmonary infection, subacute and chronic diseases	Tropical Australia, Southeast Asia, East Asia and northern Australia, northeastern Thailand, Brazil	The pathogen is a potential bioterrorism agent; mortality from melioidosis septic shock remains high despite appropriate antimicrobial therapy	Ceftazidime, chloram-phenicol, doxycycline, trimethoprim-sulfame-thoxazole	Not available currently but in under-developing stage	Antibiotics are recommended, but death rate is >40% of treated patients due to drug-resistant strains
8	*P. aeruginosa*	Pneumonia, septicemia, urinary tract infection, gastroin-testinal, and skin and soft tissue infections,	Chronic lung infection of cystic fibrosis, and contact lens-associated pseudomonal keratitis	Europe, Germany, Bulgaria, Malta	Leading cause of nosocomial infections. An important pathogen among debilitated, burned, and immunocopromized individuals	Amikacin, aminoglycoside, piperacillin, fluoroquinolones	Whole-cell *P. aeruginosa* vaccine	Though there are many drugs available but the pathogen exhibits multi-drug resistance
9	*S. typhi*	Typhoid fever		Morocco, Algeria, Tunisia, Libya, Egypt, England	Is one of the most highly host-adapted pathogens; 150 to 300 deaths occur each year in the UK	Ciprofloxacin, trimethoprim	M-01ZH09	Multi-drug resistance is of great concern. Vaccine is in an experimental stage
10	*N. meningitides*	Meningitis	Meningococcemia	America, Asia, Africa	2500 to 3500 cases reported every year in US. Children aged <5 years are at greatest risk. Severity leads to shock and death	Ceftriaxone, rifampicin	MeNZBTM vaccine, Anti-MenB vaccine	Ceftriaxone is more effective and cheaper than rifampicin, but its acceptability by patients may limit its use as a first-line prophylactic agent
11	*C. perfringens*	Gangrene and gastrointestinal disease (food poisoning and necrotic enteritis)	Enterocolitis, dysenteria, and enterotoxemia	Underdeveloped and developing countries as well as in some parts of UK	It is the most prolific producer of toxins	Clindamycin, penicillin, metronidazole	Not available	Strain is resistant to penicillin, erythromycin, and chloramphenicol
12	*M. pneumoniae*	Pneumonia	Meningo-encephalitis	Denmark, US, Europe, Japan	Frequently causes community-acquired respiratory infections in children and adults. Severity of illness may vary from severe pneumonia to asymptomatic infection	Clarithromycin, quinolone, tetracyclines, macrolides, ketolides	Not available	No specific vaccine is available until now. Polysaccharides or whole attenuated cell-based vaccines are of limited scope. The biology and molecular pathogenesis is suntil unknown to a certain extent
13	*S. pneumoniae*	Pneumonia	Meningitis, otitis media, and sinusitis	US (Atlanta)	The disease is fatal if untreated. Incidence of disease is highest in children <2 years of age and in adults >65 years of age	Vancomycin, cefotaxime, meropenem, trimethoprim-sulfame-thoxazole, clindamycin	Heptavalent protein-polysaccharide conjugate vaccine, 23-valent capsular polysaccharide vaccines	Multi-drug-resistant strains of bacteria have complicated treatment approaches

*Also indicates the diseases caused by the pathogen, their prevalence, epidemiology, available drugs, and vaccines against the pathogen, and challenges for control of the infection and treatment.

collected from www.expacy.org. Both the localization and enzyme-related information may also be collected from Swiss-Port (if available). Cytoplasmic and membrane channel proteins are selected for drug targeting, whereas membrane, exo-membrane, and secreted proteins are used to design peptide vaccines. The overall approach is shown in Figure 1. Drug and vaccine targets for several human bacterial pathogens have been reported, using this innovative strategy. Table 2 presents a list of pathogens along with the applied cutoff values for BLAST.

Our Strategy

In a modified approach that we developed [Barh and Kumar, 2009; Barh and Misra, 2009], we first screen the essential genes of the pathogen using DEG and then identify the non-human homologues to reduce the number of BLASTs. We also use pathway subtraction instead of using all pathways present in host and pathogen. Our pathway analysis-based target identification is based on criteria such as: (1) the target must be an essential non-host homologue; (2) the target must be a core gene of the pathogen; (3) the pathogen's unique pathway related targets are more favorable and will be superior if the target is involved in multiple pathways; (4) pathways having multiple targets are superior to those having single targets; (5) in the case of enzyme targets in host–pathogen common pathways, it should not be of the same class of protein, and the EC. no. of the target should not match that of any protein product

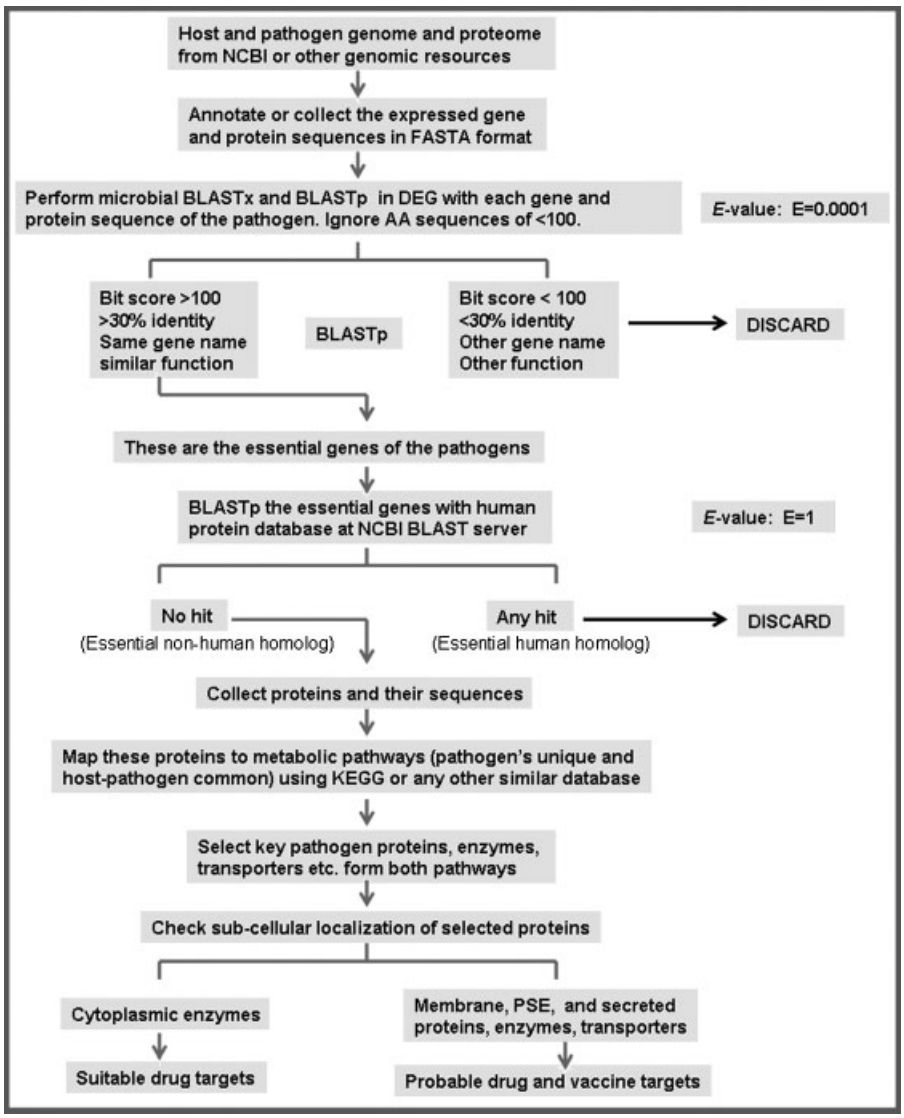


Fig. 1. Schematic representation of steps involved in *in silico* subtractive genomics-based target identification in bacterial pathogens. Identified targets can be used to develop drugs or vaccines, depending on their localization, exo-membrane topology, or secreted protein properties.

TABLE 2. Pathogens, Along With the Applied Cutoff Values for BLAST*

			BLASTp cutoff at amino acid (AA) level								
Sl. No.	Pathogen name	Genes in genome	Essential gene prediction (DEG)	Non-human homologue (NCBI)	No. of essential genes	No. of non-human homologues	No. of drug targets	No. of cytoplasmic targets	No. of membrane targets	Suggested targets	References
1	*B. pseudomallei*	5,855	E-value = 10^{-10}, 30% identity.	E-value = 10^{-3}, 30% identity	312	3,723	312	79.2%	20.8%	rpoE, OmpR	Chong et al. [2006]
2	*H. pylori*	1,590	E-value = 10^{-100}, bit score > 100, AA length > 100	E-value = 10^{-10}	178	40	40	30	10	rlpA, fecA fecA, dppA, nhaA, dppC, ftsX	Dutta et al. [2006]
3	*M. pneumoniae* (M129 strain)	693	E-value = 10^{-100}, bit score > 100, AA length > 100	E-value = 10^{-3}	220	375	112		12		Gupta et al. [2010]
4	*P. aeruginosa*	5,567	E-value = 10^{-3}	E-value = 10^{-10}	306	3,841				Genes Involved in transport of small molecules. Translation, post-translational modification and degradation	Sakharka et al. [2004]
5	*S. typhi*	4718	E-value = 10^{-4}, bit score > 100	E-value = 10^{-100}	300	149	149	138	11	ddl, trpB. motA, CheR, ppc	Rathi et al. [2009]
6	*M. leprae*	2,770	NA	E-value = 10^{-4}		179	62			Alr, rmlC, murC, murD, murE, murF, murG, murY	Shanmugam and Natarajan [2010]
7	*N. meningitides* (serogroup B)	2,001	E-value = $10^{-10,}$ 30% identity, bit-score > 100	E-value = 10^{-4}	362	1,413	35	26	9	Ppc, PilF, trpA, rpB, trpC, trpD, trpE	Sarangi et al. [2009]
8	*N. gonorrhoeae* (FA 10990)	2,002	E-value = $10^{-10,}$ 35% identity, bit-score > 100 AA length > 100	E-value = 10^{-10}	537	106	106	67	40	alf/tsr, ptsN, ddl, TbpA, afuB/fbpB ComL, cysW, PilF, pilV	Barh and Kumar [2009]
9	*A. hydrophila* (ATCC 7966)	4,287	E-value = $10^{-10,}$ bit-score > 100 AA length > 100	E-value = $10^{-10,}$ bit-score > 100	379	2,047	87			ddl, alr, Uroporphyrinogen-III synthase, glutathione S-transferase, biotin synthase	Sharma et al. [2008]
10	*Brugia malayi*	805		E = 10^{-25}	250						Holman et al. [2009]
11	*M. tuberculosis*	3,989	NA	NA	628	304	135			Genes related with Amino-acid biosynthesis	Asif et al. [2009]
12	*S. pneumoniae*	2,355		E-value cutoff < 0.005			161			Genes related with metabolism and cell wall biosynthesis	Singh et al. [2007]
13	*C. perfringens*	2,558	E-value = $10^{-10,}$ bit-score > 100 AA length > 100	E-value = $10^{-10,}$ bit-score > 100 AA length > 100	726	426				ABC transporter-ATP binding protein, FtsZ, RpoD, 50S ribosomal protein L13, and 30S subunit S5	Chhabra et al. [2010]

*Cutoff values for screening essential genes and non-human homologous genes are given separately. Genome size, number of essential genes, number of non-human homologues, number of cytoplasmic and membrane localized targets, and important targets are also presented for each pathogen.

of the host; and (6) pathogenic island-related or virulence proteins are considered superior targets.

TOOLS DEVELOPED FOR GENOME SUBTRACTION

Initial Approaches

The entire process of genome subtraction can be carried out using an *in silico* approach; to our knowledge, only two complementary *in silico* methods have been developed that allow genome subtraction. These are based on computed clusters of homologous proteins or on pairwise protein comparisons. First, all proteins of a sequence database, including those of complete genomes, are compared with each other using similarity search software, such as BLAST [Altschul et al., 1997, 1990] or FASTA [Pearson and Lipman, 1988]. Corresponding search outputs are then processed according to default constraints to extract significant hits. Finally, protein families are constructed using single transitive links. If proteins A and B are similar according to the constraints, and proteins B and C are also similar, proteins A, B, and C are then stored in the same cluster. Software tools and databases, such as CluSTr [Kriventseva et al., 2001], COG [Tatusov et al., 2001], Hobacgen [Perrière et al., 2000], ProtoMap [Yona et al., 1999], and Systers [Krause et al., 2002] provide access to such sets of homologous proteins. However, COG contains a tool called the "phylogenetic pattern search," which allows genome subtraction to select protein families. The second approach does not use fixed constraints. The user declines the similarity thresholds to decide whether a coding sequence is present or absent in a genome. The software Seebugs belongs to this category; it is based on a protein sequence comparison, using the FASTA program [Bruccoleri et al., 1998].

Current Approach-Based Tools

The target identification method involves a number of steps; therefore we need to develop bioinformatics tools that can perform the entire process on a single platform. Bruccoleri et al. [1998] developed a simple but efficient *in silico* tool that can predict putative targets based on subtraction of conserved sequences of essential genes in user-specified genomes. To develop an automated computational tool, FindTarget was built based on BLASTp comparative proteomes [Chetouani et al., 2001]. However, this tool cannot perform the entire process. In a further advancement, Singh et al. [2006] designed the T-iDT tool, which finds essential bacterial genes as well as non-human homologues, by using DEG and a human protein database. This tool can predict both the essential genes and potential targets in a pathogen genome at the same time. However, a pathway-based approach is not integrated into this tool. Recently, efforts have been made to integrate several parameters to enhance the efficacy of the prediction. The mGenomeSubtractor is one of such tools; it performs a rapid analysis of core, accessory, and essential genes, virulence factors, species-specific genes, and targets, using a mpiBLAST-based *in silico* subtractive genome hybridization method [Shao et al., 2010]. This tool can be accessed from http://bioinfo-mml.sjtu.edu.cn/mGS/. A list of available tools and databases useful for subtractive genomics-based bacterial target identification is given in Table 3.

Although in recent years several targets have been reported from various pathogenic bacteria, using genomic subtraction, no database has listed all such targets, except the Genomic Target Database (GTD) (www.iioab.webs.com/GTD.htm), which we started to develop in 2009 [Barh et al., 2009]. This database is a readily available resource that can be used to design mutagenesis studies to validate essential genes as well as the targets of pathogens listed in the database. However, currently, the database is not enriched with all targets available in the literature, as the number of reported pathogens and their targets is huge.

IDENTIFIED TARGETS

Target identification using subtractive genomics has generated a large number of targets from various pathogens. As this method is based on comparative genomics and DEG is most commonly used, several targets are found to be common in many bacteria; however, species or strain-specific, and novel targets have also been reported [Barh and Kumar, 2009; Sarangi et al., 2009]. Even though metabolic pathways related to both cytoplasmic and membrane-associated targets have been used to design both drugs and vaccines for many pathogens [Sakharkar et al., 2004; Sharma et al., 2008; Barh and Kumar, 2009], in some cases membrane-localized and secreted proteins were focused to identify potential vaccine targets [Dutta et al., 2006; Barh and Misra, 2009]. We have given more importance to targets related to pathways unique to bacteria [Barh and Kumar, 2009]; others have given equal importance to all targets [Sakharkar et al., 2004; Sharma et al., 2008]. Targets related to pathogens' unique pathways (e.g., D-alanine metabolism, two-component system, type II and III secretion systems, bacterial chemotaxis, and lipopolysaccharide and peptidoglycan biosynthesis) are ~60–70% in common, regardless of the genotype of the pathogen, and ~30–40% of the targets are genus or strain specific. Table 4 presents a list of targets from pathways unique to bacteria. The number of targets reported in the literature from host–pathogen common pathways is

TABLE 3. Databases and Tools Used in Subtractive Genomics-Based Bacterial-Target Identification†

	Utility	Website	References
Database			
NCBI bacterial genomes	Recourse of bacterial genomes	http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid = 2&type = 0&name = Complete%20Bacteria	
GOLD: Genomes	Recourse of genome projects	http://www.genomesonline.org/	Bernal et al. [2001]
Swiss-port	Proteome database	http://www.expasy.org/sprot/	Bairoch and Apweiler [1997]
Database of Essential Genes (DEG)	Screening of essential genes	http://tubic.tju.edu.cn/deg/	Zhang et al. [2004]
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Pathway comparison and subtraction	http://www.genome.jp/kegg/	Ogata et al. [1999]
Genomic Target Database (GTD)	List of bacterial targets based on subtractive genomics	www.iioab.webs.com/GTD.htm	Barh et al. [2009]
Virulence Factors of Pathogenic Bacteria Database (VFDB)	Resource of virulence factors of various medically significant bacterial pathogens	http://www.mgc.ac.cn/VFs/main.htm	Chen et al. [2005]
Tools			
CELLO	Subcellular localization prediction for bacteria and eukaryotes	http://cello.life.nctu.edu.tw/	Yu et al. [2004]
PSORTb	Subcellular localization prediction for gram-negative and gram-positive bacterial proteins	http://www.psort.org/psortb/	Gardy et al. [2005]
SOSUI-GramN	Subcellular localization prediction for gram-negative bacterial proteins	http://bp.nuap.nagoya-u.ac.jp/sosui/sosuigramn/sosuigramn_submit.html	Imai et al. [2008]
PSLpred	Subcellular localization prediction for gram-negative bacterial proteins	http://www.imtech.res.in/raghava/pslpred/	Bhasin et al. [2005]
NCBI human BLAST	Subtraction of non-human homologue genes	http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid = 9606	Altschul et al. [1990]
FindTarget	Subtractive genomics (link is not working)	http://bioweb.pasteur.fr/seqanal/findtarget	Chetouani et al. [2001]
T-iDT	Platform for identification of subtractive genomics based essential non-human homologue (link is not working)	http://www.milser.co.in/research.htm	Singh et al. [2006]
mGenomeSubtractor	in silico subtractive hybridization	http://bioinfo-mml.sjtu.edu.cn/mGS/	Shao et al. [2010]
SignalP*	Signal peptide prediction	http://www.cbs.dtu.dk/services/SignalP/	Bendtsen et al. [2004]
TMHMM*	Transmembrane domain prediction	http://www.cbs.dtu.dk/services/TMHMM/	Krogh et al. [2001]
LipoP*	Lipoprotein prediction	http://www.cbs.dtu.dk/services/LipoP/	Juncker et al. [2003]
SurfG*	Bacterial protein subcellular localization	http://genome.jouy.inra.fr/surfgplus/	Barinov et al. [2009]

†FindTarget and T-iDT web addresses mentioned in the references are currently not working. Tools used in reverse vaccinology are marked with an asterisk (*).

higher than the number of unique pathway targets. Some important targets from such common pathways are listed in Table 5.

ADVANTAGES OF IN SILICO SUBTRACTIVE GENOMICS FOR TARGET DISCOVERY

The importance of *in silico* subtractive genomics in drug-target identification is a function of its rapid and cost-effective screening of targets at the genome level. It also shortens the time required to develop immunomics-based antigens and thereby speeds up peptide vaccine design [Barh et al., 2010a,b]. Another major advantage is identification of putative essential genes in pathogens, which can be validated via mutagenesis studies [Sakharkar et al., 2004]. GTD has been developed with subtractive genomics-based targets to

TABLE 4. Selected Targets From Pathogen-Specific Metabolic Pathways*

	Pathways unique to bacteria	Genes	EC No.	Localization
1	Bacterial chemotaxis			
	Methyltransferase PilK	*pilK*	2.1.1.80	Cytoplasm
	Two-component sensor PilS	*pilS*	2.7.3.-	Membrane
	Chemotaxis-specific methylesterase		3.1.1.61	Cytoplasm
	Sensor histidine kinase		2.7.13.3	Cytoplasm
2	Polyketide sugar unit biosynthesis			
	Glucose 1-phosphate thymidylyltransfease	*rmlA*	2.7.7.24	Cytoplasm
	dTDP-D-Glucose 4,6 dehydratase	*rmlB*	4.2.1.46	Cytoplasm
	dTDP-4-dehydrorhamnose 3,5 epimerease	*rmlC*	5.1.3.13	Cytoplasm
	dTDP-4-dehydrorhamnose reductase	*rmlD*	1.1.1.133	Cytoplasm
3	Lipopolysaccharide biosynthesis			
	Probable glucosyltransferases		2.4.-	Cytoplasm
	3-deoxy-manno-octulosonate cytidylyltransferase	*kdsB*	2.7.7.38	Cell wall
	Putative 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase		3.1.3.45	Cytoplasm
	Tetraacyldisaccharide 4′-kinase	*lpxK*	2.7.1.130	Cell wall
	Lipid A-disaccharide synthase	*lpxB*	2.4.1.182	Cytoplasm
	Lipopolysaccharide core biosynthesis protein WaaP	*waaP*	2.7.-.-	Cytoplasm
	Poly(3-hydroxyalkanoic acid) synthase 1	*phaC1*	2.3.1.-	Cytoplasm
	UDP-3-*O*-[3-hydroxymyristoyl] glucosamine *N*-acyltransferase		2.3.1.-	Cytoplasm
	UDP-glucose:(heptosyl) LPS α 1,3-glucosyltransferase WaaG	*waaG*	2.4.1.-	Cytoplasm
	UDP-2,3-diacylglucosamine hydrolase		3.6.1.-	Cytoplasm
	UDP-3-*O*-acyl-*N*-acetylglucosamine deacetylase	*lpxC*	3.5.1.-	Cytoplasm
	UDP-*N*-acetylglucosamine acyltransferase	*lpxA*	2.3.1.129	Cytoplasm
	Putative sugar kinase/ADP heptose synthase	*rfaE*	2.7.1.-	Cytoplasm
	Lipopolysaccharide heptosyltransferase I	*rfaC*	2.4.-.-	Cytoplasm
	ADP-heptose–LPS heptosyltransferase II	*rfaF*	2.4.-.-	Cytoplasm
	ADP-L-glycero-D-mannoheptose 6-epimerase	*rfaD*	5.1.3.20	Cytoplasm
	2-dehydro-3-deoxyphosphooctonate aldolase (KDO 8-P-synthase)	*kdsA*	2.5.1.55	Cytoplasm
4	D-alanine metabolism			
	D-alanine-D-alanine ligase A	*Ddl*	6.3.2.4	Cell wall
	Biosynthetic alanine racemase	*Alr*	5.1.1.1	Cytoplasm
5	Carbon fixation in photosynthetic organisms			
	Fructose-1,6-bisphosphate aldolase	*alf/tsr*	4.1.2.13	Cytoplasm
6	Two-component system			
	Nitrite two-component system transcriptional response regulator	*narL*		Intracellular
	Two-component sensor PilS	*pilS*	2.7.3.-	Membrane
	Probable 2-(5′-triphosphoribosyl)-3′-dephosphocoenzyme-A synthase		2.7.8.25	Cytoplasm
	Serine protease MucD precursor	*mucD*	3.4.21.-	Cytoplasm
	Probable acyl-CoA thiolase		2.3.1.9	Cytoplasm
	Glutamine synthetase	*glnA*	6.3.1.2	Cytoplasm
	Citrate lyase β chain		4.1.3.6	Cytoplasm
	Putative nitrogen regulatory protein P-II	*glnB*		Cytoplasm
	Protein-PII uridylyltransferase	*glnD*	2.7.7.59	Cytoplasm
	β-lactamase precursor	*ampC*	3.5.2.6	Extracellular
	Anthranilate synthase component II	*trpG*	4.1.3.27	Membrane
	Anthranilate phosphoribosyltransferase	*trpD*	2.4.2.18	Cytoplasm
	Indole-3-glycerol-phosphate synthase	*trpC*	4.1.1.48	Cytoplasm
	Tryptophan synthase subunit β	*trpB*	4.2.1.20	Cytoplasm
	Tryptophan synthase α chain	*trpA*	4.2.1.20	Cytoplasm
	Potassium-transporting ATPase	*kdpA*	3.6.3.12	Membrane
	Probable methylesterase		3.1.1.61	Cytoplasm
	Alkaline phosphatase	*phoA*	3.1.3.1	Membrane
	Respiratory nitrate reductase α chain	*narG*	1.7.99.4	Cytoplasm
	Sensor histidine kinase		2.7.13.3	Cytoplasm
7	Type II secretion system			
	Two-component sensor PilS	*pilS*	2.7.3.-	Membrane
	Leader peptidase (prepilin peptidase)/*N*-methyltransferase	*pilD*	3.4.23.43	Membrane
	Methyltransferase PilK	*pilK*	2.1.1.80	Membrane
	Sensor histidine kinase		2.7.13.3	
	Type IV pilus assembly protein	*PilF*		Membrane

TABLE 4. Continued

	Pathways unique to bacteria	Genes	EC No.	Localization
	Putative type IV pilin protein	*PilV*		Fimbrium
8	Type III secretion system			
	Flagellum-specific ATP synthase	*fliI*	3.6.3.14	Cytoplasm
	ATP synthase F0, B subunit		3.6.3.14	Membrane
9	Flagellar assembly			
	ATP synthase F0, B subunit		3.6.3.14	Membrane
10	Phosphotransferase system (PTS)			
	Phosphotransferase system, fructose-specific IIBC component	*fruA*	2.7.1.69	Membrane
	Putative two-component system transcriptional response regulator	*pstN*		Cytoplasm
	Probable phosphotransferase system enzyme I		2.7.3.9	Cytoplasm
11	Biosynthesis of siderophore group nonribosomal peptides			
	Isochorismate synthase	*pchA*	5.4.4.2	Cytoplasm
	Isochorismate pyruvate lyase	*pchB*	4.1.99.-	Cytoplasm
12	1,2-Dichloroethane degradation			
	Quinoprotein alcohol dehydrogenase	*exaA*	1.1.99.8	Periplasm
	Probable aldehyde dehydrogenase	*calB*	1.2.1.3	Cytoplasm
13	Toluene and xylene degradation			
	Catechol 1,2-dioxygenase	*catA*	1.13.11.1	Cytoplasm
14	Peptidoglycan biosynthesis			
	UDP-*N*-acetyl glucosamine 1-carboxyvinyltransferase	*murA*	2.5.1.7	Cytoplasm
	UDP-*N*-acetyl muramyl tripeptide synthase	*murD*	6.3.2.9	Cytoplasm
	UDP-*N*-acetyl muramoyl alanyl-D-glutamyl-2,6-diamino pimelate–D-alanyl-D-alanyl ligase	*murF*	6.3.2.10	Cytoplasm

*Targets have been selected from *P. aeruginosa* [Sakharkar et al., 2004; Perumal et al., 2007], *H. pylori* [Dutta et al., 2006], *B. pseudomallei* [Chong et al., 2006], *A. hydrophila* [Sharma et al., 2008], *N. gonorrhoeae* [Barh and Kumar, 2009], *N. meningitides* [Sarangi et al., 2009], *M. tuberculosis* [Asif et al., 2009], *S. typhi* [Rathi et al., 2009], *M. leprae* [Shanmugam and Natarajan, 2010], and *M. pneumonia* [Gupta et al., 2010]. None of these targets are found in any single pathogen indicated here. The EC nos. and localization information of targets are also presented.

obtain a readily available resource of putative essential genes as well as drug targets in human bacterial pathogens [Barh et al., 2009]. Selective and essential genes of pathogens that are non-homologous to the host are considered putative targets. Such target sequences are not present in the host, making the identified target unique to the pathogen. Thus, inhibition of such targets with appropriate drug(s) should avoid cytotoxicity issues in the host and will reduce the cost of ADMET validation of newly designed drugs [Sakharkar et al., 2004, Barh and Kumar, 2009].

SCOPE IN REVERSE VACCINOLOGY

Reverse vaccinology (RV) is a computational approach that takes a path different from conventional approaches for the development of vaccines [Bambini and Rappuoli, 2009; Serruto and Rappuoli, 2006]. Rather than start from a set of proteins that have experimentally been proven antigenic, RV explores previously unconsidered possibilities. RV seeks candidate proteins to elicit immune responses in the entire genome of an organism against which a vaccine is required; however, special attention is given to proteins that are secreted or exposed on the cell wall of the organism [Rappuoli, 2000]. Subtractive genomics-based target discovery is applicable to both drug and vaccine targets. It is preferable that a vaccine candidate be a non-human homologue. In bacteria, it is known that exported proteins are the main forms of interaction with cells infected by such organisms; therefore they are potential candidates for vaccine targets [Sibbald and van Dij, 2009; Simeone et al., 2009; Stavrinides et al., 2008; Bhavsar et al., 2007]. Hence, a non-human homologue secreted, or exo-membrane, or exported protein will be a better option for developing vaccine following RV. Table 6 lists bacterial pathogens for which RV methods are used for developing vaccines.

RV-related tools (see Table 3, marked with an asterisk) are widely used by the scientific community to ensure viability and increase reliability. As an example, we can cite the software SignalP [Bendtsen et al., 2004] for protein motif identification, which indicates the existence of signal peptides; the software TMHMM [Krogh et al., 2001] indicates transmembrane motifs. RV makes use of large-scale software, analyzing all the proteins derived from the genome and combining results. An example of combined results is to determine whether a protein is secreted because of a signal peptide (SignalP), taking into account that there is only one possible transmembrane domain (TMHMM). Otherwise, even though there is a signal peptide, the protein remains anchored to the cell membrane, which

TABLE 5. Selected Targets From Metabolic Pathways That Host and Pathogen Have in Common*

	Host–pathogen shared pathways	Gene	EC No.	Localization
1	DNA replication, repair, and recombination			
	DNA polymerase III subunit ε		2.7.7.7	Cytoplasm
	Holliday junction DNA helicase motor protein	*ruvA*	3.6.1	Membrane
2	Cell cycle			
	Cell division protein	*MraZ*	NA	Cytoplasm
	Cell division membrane protein	*FtsW*	NA	Membrane
3	Pyrimidine metabolism			
	FAD-dependent thymidylate synthase	*thyX*	2.1.1.148	Cytoplasm
	Dihydroorotase		3.5.2.3	Cytoplasm
4	Purine metabolism			
	DNA-directed RNA polymerase subunit α		2.7.7.6	Cytoplasm
	DNA polymerase III, alpha subunit		2.7.7.7	Cytoplasm
	DNA polymerase III subunit β		2.7.7.6	Cytoplasm
5	Transcription and translation			
	Transcription anti-termination protein	*NusB*		Cytoplasm
	50S ribosomal protein L30	*rpmD*		Cytoplasm
	50S ribosomal protein L35	*rpmI*		Cytoplasm
	50S ribosomal protein L34	*rpmH*		Cytoplasm
	50S ribosomal protein L1	*rplA*		Cytoplasm
	50S ribosomal protein L28	*rpmB*		Cytoplasm
	elongation factor P	*efp*		Cytoplasm
	tRNA guanine-N 1-methyltransferase	*trmD*	2.1.1.31	Cytoplasm
6	Histidine metabolism			
	Imidazole glycerol-phosphate dehydratase	*hisB*	4.2.1.19	Cytoplasm
	Histidinol dehydrogenase	*hisD*	1.1.1.23	Cytoplasm
	ATP Phosphoribosyl transferase	*hisG*	2.4.2.17	Cytoplasm
	Imidazole glycerol phosphate synthase subunit	*hisH*	2.4.2.-	Cytoplasm
	Phosphoribosyl-AMP cyclohydrolase		3.5.4.19	Cytoplasm
7	Thiamin biosynthesis			
	Thiamine monophosphate kinase	*thiL*	2.7.4.16	Cytoplasm
	Phosphomethylpyrimidine kinase		2.7.4.7	Cytoplasm
	Cysteine desulfurase		2.8.1.7	Cytoplasm
8	Aminosugars metabolism			
	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase		2.5.1.7	Cytoplasm
	UDP-*N*-Acetylenolpyruvoylglucosamine reductase		1.1.1.158	Cytoplasm
9	Phenylalanine, tryptophan, porphyrin and chlorophyll metabolism			
	Glutamyl-tRNA reductase		1.2.1.-	Cytoplasm
	Chorismate mutase		4.2.1.51	Cytoplasm
	3-dehydroquinate synthase		4.6.1.3	Cytoplasm
	Shikimate dehydrogenase		1.1.1.25	Cytoplasm
	Phospho-2-dehydro-3-deoxyheptonate aldolase	*aroH*	2.5.1.54	Cytoplasm
10	Glycine, isoleucine, serine, threonine, lysine metabolism			
	Homoserine dehydrogenase	*thrA*	1.1.1.3	Cytoplasm
	Gycyl-tRNA synthetase subunit β	*glyS*	6.1.1.14	Cytoplasm
	Homoserine kinase	*thrB*	2.7.1.39	Cytoplasm
	Diaminopimelate epimerase		5.1.1.7	Cytoplasm
	Aspartate-semialdehyde dehydrogenase		1.2.1.11	Cytoplasm
11	Terpenoid backbone biosynthesis			
	4-hydroxy-3-methyl but-2-enyl diphosphate reductase	*ispH*	1.17.1.2	Cytoplasm
12	Riboflavin metabolism			
	Riboflavin synthase subunit β	*ribH*	2.5.1.9	Cytoplasm
	Riboflavin synthase subunit β		2.5.1.-	Cytoplasm
	GTP cyclohydrolase II		3.5.4.25	Cytoplasm
13	Biotin biosynthesis			
	Biotin synthase family transferase		2.8.1.6	Cytoplasm
	Biotin synthase	*bioB*	2.8.1.6	Cytoplasm
	Dethiobiotin synthetase	*bioD1*	6.3.3.3	Cytoplasm
14	Folate biosynthesis			
	Dihydropteroate synthase	*FolP*	2.5.1.15	Cytoplasm

TABLE 5. Continued

	Host–pathogen shared pathways	Gene	EC No.	Localization
	p-Aminobenzoate synthase component		2.6.1.85	Cytoplasm
15	Oxidative phosphorylation			
	F0F1 ATP synthase subunit A		3.6.1.34	Membrane
	F0F1 ATP synthase subunit B		3.6.3.14	Membrane
16	Environmental information processing and membrane transport			
	ABC transporter iron-uptake	*fbpB*		Membrane
17	Protein export			
	Preprotein translocase subunit SecA	*secA*		Membrane
	Preprotein translocase subunit SecY	*secY*		Membrane
	Preprotein translocase subunit SecD	*secD*		Membrane
18	Pyruvate, propanoate, taurine, and hypotaurine metabolism			
	Acetate kinase		2.7.2.1	Intracellular
	Phosphotransacetylase		2.3.1.8	Cytoplasm
19	Steroids and isoprene biosynthesis			
	1-deoxy-D-xylulose 5-phosphate reductoisomerase	*dxr*	1.1.1.267	Cytoplasm
	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase		4.6.1.12	Cytoplasm
20	Glutamate, D-glutamine, and D-glutamate metabolism			
	Glutathione synthetase		6.3.2.3	Cytoplasm
	UDP-*N*-acetylmuramate-L-alanine ligase		6.3.2.8	Cytoplasm
	Glutamate racemase		5.1.1.3	Cytoplasm

*Targets have been selected from *P. aeruginosa* [Sakharkar et al., 2004; Perumal et al., 2007], *H. pylori* [Dutta et al., 2006], *B. pseudomallei* [Chong et al., 2006], *A. hydrophila* [Sharma et al., 2008], *N. gonorrhoeae* [Barh and Kumar, 2009], *N. meningitides* [Sarangi et al., 2009], *M. tuberculosis* [Asif et al., 2009], *S. typhi* [Rathi et al., 2009], *M. leprae* [Shanmugam and Natarajan, 2010], and *M. pneumonia* [Gupta et al., 2010]. All these targets are not found in any single pathogen indicated here. EC nos. and localization information of targets are also presented.

TABLE 6. Bacterial Pathogens for Which Reverse Vaccinology Approaches Have Been Adopted to Develop Vaccines*

Bacteria	Disease	Vaccine approach	Vaccine development stage
B. anthracis	Anthrax	Reverse vaccinology, CGH microarray, microarray proteomics, immunoproteomics	Discovery/preclinical
C. pneumoniae	Pneumonia, meningitis, middle era infections	Reverse vaccinology, proteomics	Discovery/preclinical
H. pylori	Ulcer, atrophicgastritis, adenocarcinoma, lymphoma	Reverse vaccinology, immunoproteomics	Discovery/preclinical
M. tuberculosis	Tuberculosis	Reverse vaccinology	Discovery/preclinical
N. meningitidis Serogroup B	Bacterial meningitis, septicemia	Reverse vaccinology, microarray, proteomics	Phase II clinical trials
S. aureus	Variety of infections, including, pelvic syndrome, rapidly progressive pneumonia, ocular infections, septic thrombophlebitis	CGH microarray Immunoproteomics	Discovery/preclinical
S. pyogenes (GAS)	Many systemic invasive infections, including necrotizing fasciitis, myositis, pneumonia, sepsis, arthritis	Genome-wide analysis, proteomics	Discovery/preclinical
S. agalactiae (GBS)	Bacterial sepsis, pneumonia, meningitis	Reverse vaccinology, classical or comparative	Discovery/preclinical
S. pneumoniae	Bacterial pneumonia, sepsis, sinusitis, otitis media, bacterial meningitis	Classical or comparative, reverse vaccinology, proteomics	Discovery/preclinical

*Adapted from Bambini and Rappuoli [2009].

leads to the classification of a membrane protein. There is also the software based on Hidden Markov Models (HMM) to check whether a protein has classic signs of retention and the software LipoP [Juncker et al., 2003] to check whether it is a lipoprotein.

We can make a rational analysis in RV, increasing the speed and reliability of results. Electron microscopy can be used to measure the thickness of cell walls. Data on cell wall thickness is used as the cutoff in the TMHMM output. A recently developed tool, SurfG

plus, takes into account transmembrane domain positive prediction, and the estimated size of transmembrane domains is confronted with the estimated measure for the cell wall [Barinov et al., 2009]. In this way, it is possible to arrive at a more reliable estimate of the probability of a protein being characterized as an integrated membrane protein versus exposed on the surface. Besides developing a list of predicted proteins that could be exported, we can also make an analysis of possible B- and T-cell epitopes, in order to create an additional filter and minimize the list of targets that can be experimentally proven through this immunoinformatics approach [Serruto and Rappuoli, 2006].

DISADVANTAGES OF THE METHOD

Although the method has many advantages, there are certain concerns about the use of this technique. To perform subtraction, both the host and pathogen genomes are required; if one is not available, the analysis is difficult to perform. Similar to other *in silico* methods, targets derived from such analyses require experimental validation. In recent years, in almost all reports, DEG BLAST has been used for identification of essential genes of the pathogen based on gene or amino acid sequence similarities. The DEG is continuously enriched with mutagenesis-based new essential genes, also including new pathogens. An obvious concern is the consistency of the number of screened essential genes for a given pathogen with respect to time. We found that the number increases dramatically as a result of data enrichment of the DEG [Barh and Kumar, 2009]. Second, researchers, including ourselves, have not considered proteins with less than 100 amino acids [Dutta et al., 2006; Sharma et al., 2008; Barh and Kumar, 2009]. However, it has been found with DEG BLAST that many proteins listed in this essential gene database are <100 amino acids long. Therefore, when we exclude such small proteins, we may purge out some novel targets. This may not always be true, because it has been observed in mutagenesis studies that when there are insertion mutations in a nucleotide sequence of <300 bp, expression of nearby genes is altered, resulting in lethality, giving a false-positive result concerning the essentiality of the target gene. Also pathogen genes that are essential but non-homologous to any DEG-listed essential gene may be missed. Sakharkar et al. [2004] cautioned that because the method is based on BLAST results and does not consider specific growth conditions, care should be taken in interpreting the BLAST results. Otherwise, a conditional essential gene may be screened and selected. Hence a parallel method and tool independent of DEG should be developed. We found that if we increase the number of different species/strains within the same genus of the pathogen and use more than one host, the number of targets is considerably reduced (unpublished data). Hence, it is advisable to use multiple strains of a pathogen and all strain-specific hosts in the analysis to identify common targets for all strains as well as for a broad host range. Therefore, a pangenomics approach, including distant gene relationships, should be considered.

CONCLUSIONS AND PERSPECTIVES FOR THE FUTURE

In silico subtractive genomics is a rapid, powerful, and cost-effective approach for screening of drug and vaccine targets for any given pathogen, provided both the pathogen and host genomes are available. However, the identified targets require experimental validation. This approach requires multiple analyses at different stages that mostly use BLAST. Parameters of BLAST at different stages require optimization to standardize the method. Similarly, an efficient integrated platform needs to be developed to perform the entire analysis at the same time. A parallel method independent of DEG-based screening of essential genes is also required. Pangenomics-based conserved essential genes as the targets may be considered in such analyses. An *in silico* mutagenesis approach and other computational validation methods could be included in the analysis to improve the efficacy of the original method.

ACKNOWLEDGMENTS

D.B., S.T., and N.J. acknowledge the motivation and encouragement of all IIOAB members. S.T., N.J., A.N.M., and A.K. acknowledge facilities of DBT's Bioinformatics subcenters at their respective schools. A.A., A.R.S., and V.A. received research scholarships from CNPq.

AUTHOR CONTRIBUTIONS

D.B., S.T., and N.J. collected data and wrote the paper; A.A. and A.R.S. provided inputs on reverse vaccinology; A.N.M., A.K., and V.A. reviewed the article and provided technical guidance.

FINANCIAL DISCLOSURE

This work was carried out without any grant or other financial support, other than scholarships. There is no conflict of interest regarding this work.

REFERENCES

Allsop AE. 1998. Bacterial genome sequencing and drug discovery. Curr Opin Biotechnol 9:637–642.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Arias CA, Murray BE. 2009. Antibiotic-resistant bugs in the 21st century—a clinical super-challenge. N Engl J Med 360:439–443.

Asif SM, Asad A, Faizan A, Anjali MS, Arvind A, Neelesh K, Hirdesh K, Sanjay K. 2009. Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis. Bioinformation 4:245–248.

Bairoch A, Apweiler R. 1997. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25: 31–36.

Bambini S, Rappuoli R. 2009. The use of genomics in microbial vaccine development. Drug Discov Today 14:252–260.

Barh D, Kumar A. 2009. In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*. In Silico Biol 9:225–231.

Barh D, Misra AN. 2009. In silico identification of membrane associated candidate drug targets in *Neisseria gonorrhoeae*. Int J Integr Biol 6:65–67.

Barh D, Kumar A, Misra AN. 2009. Genomic Target Database (GTD): a database of potential targets in human pathogenic bacteria. Bioinformation 4:50–51.

Barh D, Misra AN, Kumar A, Azevedo V. 2010a. A novel strategy of epitope design in *Neisseria gonorrhoeae*. Bioinformation 5: 77–85.

Barh D, Misra AN, Kumar A. 2010b. In silico identification of dual ability of *N. gonorrhoeae* ddl for developing drug and vaccine against pathogenic *Neisseria* and other human pathogens. J Proteomics Bioinform 3:082–090.

Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M. 2009. Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. Proteomics 9:61–73.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795.

Bernal A, Ear U, Kyrpides N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res 29:126–167.

Bhasin M, Garg A, Raghava GPS. 2005. PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 21: 2522–2524.

Bhavsar AP, Guttman JA, Finlay BB. 2007. Manipulation of host-cell pathways by bacterial pathogens. Nature 449:827–834.

Bruccoleri RE, Dougherty TJ, Davison DB. 1998. Concordance analysis of microbial genomes. Nucleic Acids Res 26:4482–4486.

Chan JN, Nislow C, Emili A. 2010. Recent advances and method development for drug target identification. Trends Pharmacol Sci 31:82–88.

Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:325–328.

Chetouani F, Glaser P, Kunst F. 2001. FindTarget: software for subtractive genome analysis. Microbiology 147:2643–2649.

Chhabra V, Sharma P, Anant A, Deshmukh S, Kaushik H, Gopal K, Srivastava N, Sharma N, Garg LC. 2010. Identification and modeling of a drug target for *Clostridium perfringens* SM101. Bioinformation 4:278–289.

Chong CE, Lim BS, Nathan S, Mohamed R. 2006. In silico analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. In Silico Biol 6:341–346.

Dong QJ, Wang Q, Xin YN, Li N, Xuan SY. 2009. Comparative genomics of *Helicobacter pylori*. World J Gastroenterol 15: 3984–3991.

Downs DM. 2006. Understanding microbial metabolism. Annu Rev Microbiol 60:533–559.

Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D. 2006. In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. In Silico Biol 6:43–47.

Fischbach MA, Walsh CT. 2009. Antibiotics for emerging pathogens. Science 28:1089–1093.

Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FSL. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623.

Gupta SK, Singh S, Gupta MK, Pant KK, Seth PK. 2010. Identification of potential targets in *Mycoplasma pneumoniae* through subtractive genome analysis. J Antivir Antiretrovir 2: 038–041.

Holman AG, Davis PJ, Foster JM, Carlow CK, Kumar S. 2009. Computational prediction of essential genes in an unculturable endosymbiotic bacterium, *Wolbachia of Brugia malayi*. BMC Microbiol 9:243.

Hood DW. 1999. The utility of complete genome sequences in the study of pathogenic bacteria. Parasitology 118:S3–S9.

Huynen MA, Diaz-Lazcoz Y, Bork P. 1997. Differential genome display. Trends Genet 13:389–390.

Huynen M, Dandekar T, Bork P. 1998. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5.

Imai K, Asakawa N, Tsuji T, Akazawa F, Ino A, Sonoyama M, Mitaku S. 2008. SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. Bioinformation 2:417–421.

Ishii N, Robert M, Nakayama Y, Kanai A, Tomita M. 2004. Toward large-scale modeling of the microbial cell for computer simulation. J Biotechnol 113:281–294.

Itaya M. 1995. An estimation of minimal genome size required for life. FEBS Lett 362:257–260.

Ji Y. 2002. The role of genomics in the discovery of novel targets for antibiotic therapy. Pharmacogenomics 3:315–323.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12:962–968.

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12:1652–1662.

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al. 2003. Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA 100:4678–4683.

Koonin EV, Tatusov RL, Galperin MY. 1998. Beyond complete genomes: from sequence to structure and function. 8:355–363.

Krause A, Haas SA, Coward E, Vingron M. 2002. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. Nucleic Acids Res 30:299–300.

Kriventseva EV, Fleischmann W, Zdobnov EM Apweiler R. 2001. CluSTr: a database of clusters of SWISSPROT TrEMBL proteins. Nucleic Acids Res 29:33–36.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lin J, Qian J. 2007. Systems biology approach to integrative comparative genomics. Expert Rev Proteomics 4:107–119.

McDevitt D, Rosenberg M. 2001. Exploiting genomics to discover new antibiotics. Trends Microbiol 9:611–617.

Meinke A, Henics T, Nagy E. 2004. Bacterial genomes pave the way to novel vaccines. Curr Opin Microbiol 7:314–320.

Mills SD. 2006. When will the genomics investment pay off for antibacterial discovery? Biochem Pharmacol 30:1096–1102.

Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268–10273.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34.

Owa T. 2007. Drug target validation and identification of secondary drug target effects using DNA microarrays. Tanpakushitsu Kakusan Koso 52:1808–1809.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448.

Perrière G, Duret L, Gouy M. 2000. HOBACGEN: database system for comparative genomics in bacteria. Genome Res 10:379–385.

Perumal D, Lim CS, Sakharkar KR, Sakharkar MK. 2007. Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. In Silico Biol 7: 453–465.

Pinner RW, Teutsch SM, Simonsen L, Klug LA, Graber JM, Clarke MJ, Berkelman RL. 1996. Trends in infectious diseases mortality in the United States. JAMA 275:189–193.

Plotkin SA. 2005. Why certain vaccines have been delayed or not developed at all. Health Aff (Millwood) 24:631–634.

Pucci MJ. 2006. Use of genomics to select antibacterial targets. Biochem Pharmacol 71:1066–1072.

Rappuoli R. 2000. Reverse vaccinology. Curr Opin Microbiol 3: 445–450.

Rathi B, Aditya N, Sarangi AN, Trivedi N. 2009. Genome subtraction for novel target definition in *Salmonella typhi*. Bioinformation 4:143–150.

Sakharkar KR, Sakharkar MK, Chow VT. 2004. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. In Silico Biol 4: 355–360.

Sarangi AN, Aggarwal R, Rahman Q, Trivedi N. 2009. Subtractive genomics approach for in silico identification and characterization of novel drug targets in *Neisseria meningitidis* serogroup B. J Comput Sci Syst Biol 2:255–258.

Serruto D, Rappuoli R. 2006. Post-genomic vaccine development. FEBS Lett 580:2985–2992.

Shanmugam A, Natarajan J. 2010. Computational genome analyses of metabolic enzymes in *Mycobacterium leprae* for drug target identification. Bioinformation 4:392–395.

Shao Y, He X, Harrison EM, Tai C, Ou HY, Rajakumar K, Deng Z. 2010. mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes. Nucleic Acids Res 38:194–200.

Sharma V, Gupta P, Dixit A. 2008. In silico identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*. In Silico Biol 8:331–338.

Sibbald MJJB, van Dij JML. 2009. Secretome mapping in gram-positive pathogens. In: Wooldridge K, editor. Bacterial secreted protein: secretory mechanisms and role in pathogenesis. Norfolk, UK: Caister Academic Press. p 193–225.

Simeone R, Bottai D, Brosch R. 2009. ESX/type VII secretion systems and their role in host–pathogen interaction. Curr Opin Microbiol 12:4–10.

Singh NK, Selvam SM, Chakravarthy P. 2006. T-iDT: tool for identification of drug target in bacteria and validation by *Mycobacterium tuberculosis*. In Silico Biol 6:485–493.

Singh S, Malik BK, Sharma DK. 2007. Metabolic pathway analysis of *S. pneumoniae*: an in silico approach towards drug-design. J Bioinform Comput Biol 5:135–153.

Stavrinides J, McCann HC, Guttman DS. 2008. Host–pathogen interplay and the evolution of bacterial effectors. Cell Microbiol 10:285–292.

Stumm G, Russ A, Nehls M. 2002. Deductive genomics: a functional approach to identify innovative drug targets in the post-genome era. Am J Pharmacogenom 2:263–271.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28.

Thykaer J, Andersen MR, Baker SE. 2009. Essential pathway identification: from in silico analysis to potential antifungal targets in *Aspergillus fumigatus*. Med Mycol 47:S80–S87.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001.The sequence of the human genome. Science 291:1304–1351.

Yona G, Linial N, Linial M. 1999. ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins 37:360–378.

Yu CS, Lin CJ, Hwang JK. 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13:1402–1406.

Zhang R, Ou HY, Zhang CT. 2004. DEG: a database of essential genes. Nucleic Acids Res 32:D271–D272.

IV CHAPTERS/ RESEARCH ARTICLES

IV.1 Chapter I: Research Article

A novel *in silico* reverse-transcriptomics-based identification and blood-based validation of a panel of sub-type specific biomarkers in lung cancer

Debmalya Barh, Neha Jain, Sandeep Tiwari, John K Field, Elena Padin-Iruegas, Alvaro Ruibal, Rafael López, Michel Herranz, Antaripa Bhattacharya, Lucky Juneja, Cedric Viero, Artur Silva, Anderson Miyoshi, Anil Kumar, Kenneth Blum, **Vasco Azevedo**, Preetam Ghosh, Triantafillos Liloglou

BMC Genomics. 2013; 14 Suppl 6:S5. doi: 10.1186/1471-2164-14-S6-S5. Epub 2013 Oct 25.
[PMID: 24564251]
Impact Factor: 4.041(2013)

In this chapter, a novel *in silico* reverse-transcriptomics strategy is described to identify transcription factor (TF) biomarkers for lung cancer and its sub-types. Started with miRNA expression profile in lung cancers, we identified all targets of the miRNAs and then target enrichment and reverse annotation strategy (using top targets) was adopted to assign gene Ontology (GO) to each miRNA. Next, we developed and analyzed global PPI derived from cancer specific TF-TF interactions and from that network, cancer related cell cycle specific TF-TF interaction networks were identified. The miRNA-TF-miRNA or TF-miRNA-TF (miR-miR) interactions networks were developed and sub-type specific TF-miRNA-TF /TF-TF network were generated. A novel subtractive interactome, subtractive network, and GSEA analysis were performed to identify lung cancer sub-type specific potential TF markers. Out of several identified markers, we selected 7 TF markers for NSCLC for validation. We used stage-II and stage-IV NSCLC tissue samples for microarray analysis and blood samples for qPCR based validation. It was found that, upregulation of TFPD1, E2F6, IRF1, and HMGA1 + NO expression of SUV39H1, RBL1, and HNRPD in blood sample are characteristics of Adeno and Squamous cell lung carcinomas. E2F6 is a novel/newly identified marker for lung cancer. The miRNA-marker-miRNA interactions can give novel insight of the lung tumorigenesis. A modified strategy can be useful in early marker identification and in developing personalized medicine. The strategy can also be equally useful in identifying biomarkers in other complex diseases too.

BMC Genomics

RESEARCH Open Access

A novel *in silico* reverse-transcriptomics-based identification and blood-based validation of a panel of sub-type specific biomarkers in lung cancer

Debmalya Barh[1*], Neha Jain[1,2], Sandeep Tiwari[1], John K Field[3], Elena Padin-Iruegas[4], Alvaro Ruibal[5], Rafael López[4], Michel Herranz[6], Antaripa Bhattacharya[1], Lucky Juneja[1,2], Cedric Viero[1,7], Artur Silva[8], Anderson Miyoshi[9], Anil Kumar[2], Kenneth Blum[1,10], Vasco Azevedo[9], Preetam Ghosh[1,11], Triantafillos Liloglou[3]

From 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2012)
Campinas, Brazil. 14-17 October 2012

Abstract

Lung cancer accounts for the highest number of cancer-related deaths worldwide. Early diagnosis significantly increases the disease-free survival rate and a large amount of effort has been expended in screening trials and the development of early molecular diagnostics. However, a gold standard diagnostic strategy is not yet available. Here, based on miRNA expression profile in lung cancer and using a novel *in silico* reverse-transcriptomics approach, followed by analysis of the interactome; we have identified potential transcription factor (TF) markers that would facilitate diagnosis of subtype specific lung cancer. A subset of seven TF markers has been used in a microarray screen and was then validated by blood-based qPCR using stage-II and IV non-small cell lung carcinomas (NSCLC). Our results suggest that overexpression of HMGA1, E2F6, IRF1, and TFDP1 and downregulation or no expression of SUV39H1, RBL1, and HNRPD in blood is suitable for diagnosis of lung adenocarcinoma and squamous cell carcinoma sub-types of NSCLC. Here, E2F6 was, for the first time, found to be upregulated in NSCLC blood samples. The miRNA-TF-miRNA interaction based molecular mechanisms of these seven markers in NSCLC revealed that HMGA1 and TFDP1 play vital roles in lung cancer tumorigenesis. The strategy developed in this work is applicable to any other cancer or disease and can assist in the identification of potential biomarkers.

Introduction

Lung cancer is the leading cause among cancer related deaths worldwide, constituting 17% of new cancer cases and 23% of deaths from cancer. Although N. American and European countries show a slow decline in death rates due to lung cancer, deaths due to this form of cancer are increasing considerably in Asian and African countries [1]. Lung cancer is mainly divided into two subtypes, small cell lung cancer (SCLC), which accounts for 10-15% of all cases and non-small cell lung cancer (NSCLC, 85-90%). The latter group is further histologically subdivided into four categories; adenocarcinoma, squamous cell carcinoma, large cell carcinoma and 'others', for example cancers of neuroendocrine origin [2]. The overall 5-year survival rate for NSCLC ranges from 9% to 15% [3]. The high mortality from lung cancer is due a combination of lack of reliable early diagnostic tools [3,4] along with a poor arsenal of lung cancer regimens for stage I lung cancer, whose survival rate is also surprisingly low [5].

* Correspondence: dr.barh@gmail.com
[1]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB-721172, India
Full list of author information is available at the end of the article

Numerous studies have utilized different "-omics"-based approaches to identify molecular signatures in lung cancer with diagnostic or prognostic value while using minimally invasive processes. Some of these are as follows: 34 miRNA signatures [6], expression profiles of 11 miRNAs (miR-106a, miR-15b, miR-27b, miR-142-3p, miR-26b, miR-182, miR-126, let7g, let-7i and miR-30e-5p) from serum [7], 7 miRNA signatures [8], overexpression of six snoRNAs [9], and expression of 3 miRs (miR-205, miR-210 and miR-708) in sputum [10]. Additional signatures and markers have also been reported from the plasma proteome [11,12], the salivary proteome [13], the serum epigenome [14], sputum-based genomics [15], and blood-based gene expression studies [16]. However, none of these have progressed sufficiently to provide the necessary specificity and sensitivity required for clinical implementation.

microRNAs (miRNAs/miRs) are involved in a variety of biological processes, including cell cycle regulation, cell differentiation, development, metabolism, and aging [17]. They have also been shown to be aberrantly expressed in several cancers [18]. Lung cancer is no exception to this and miRNA signatures have been suggested to be useful in diagnosis, prognosis, and therapy [7,19-21]. miRNAs regulate posttranscriptional gene expression and a single miRNA can regulate up to 200 mRNAs including those for transcription factors (TFs) [22]. Because miRNA transcription is under the regulation of TFs, intriguing feed-back and feed-forward regulatory loops can be formed among TFs and miRNAs [17].

In this study we have developed a novel *in silico* reverse-transcriptomics strategy followed by interactome analysis to identify the sub-type specific diagnostic TF markers in lung cancer. The approach is novel as the sub-type specific TF markers were identified starting with experimentally validated miRNA profiles in lung cancer. We have also attempted to provide a molecular insight during the early events in lung cancer.

Materials and methods

Literature mining

Extensive literature and text mining was carried out to collect deregulated miRNAs in lung cancers (NSCLC and SCLC) using databases such as PubMed, Sirus, and Elsevier as well as search engines such as Google and Google Scholar. miR2Disease [23] was also used to gather lung cancer specific miRNAs information. Priority was given to reports that have used markers based on biopsy samples and patient's remote media (blood, serum, plasma, sputum, and bronchioalveolar lavage among others [24]). Selected miRNAs were then grouped into three categories: (1) NSCLC specific, (2) exclusively SCLC related, and (3) common in both the types. The up- and down-regulated miRNAs within each of these three groups were also noted.

GO assignment to miRNAs using reverse annotation strategy

No tool is currently available to classify or cluster miRNAs as per their GO (Gene Ontology) or functional annotation. We applied a reverse approach in which GO terms to a miRNA are assigned based on the functional annotation of the targets of the particular miRNA. In this approach, we first identified experimentally validated targets of each miRNA using miRNA target databases miRWalk [25], miRecords [26], miReg [17], and miRTarBase [27]. Next, targets for each miRNA were subjected to ToppGene Suite [28] for GSEA (Gene Set Enrichment Analysis) candidate gene prioritization. The top-ranked genes were used in DAVID v6.7 [29] analysis for functional annotation clustering and the assignment of GO terms to each miRNA which targets these genes. GO terms related to various aspects of cancer were considered. miRNAs and their corresponding targets that fall under these specific GO categories were selected, and the rest were ignored (Figure 1, Step-3).

miRNA-TF-miRNA or TF-miRNA-TF interactions

To date, there is no study reporting direct miRNA-miRNA interaction. However, it is well known that miRNAs can modulate post-transcriptional gene regulation as well as their own expression through feed-back and feed-forward loops that are mediated by various TFs. Therefore, there are miRNA-TF interactions. As TFs interact with other TFs and proteins, the known TF-TF networks can be complemented by integrating the relevant miRNA-TF interactions to make TF-miRNA-TF or TF-miRNA-TF-miRNA interactions. Such TF-miRNA-TF-miRNA interaction networks will indirectly represent the miRNA-miRNA interactions.

We thus created a cancer specific TF-TF interaction network using targets of miRNAs frequently deregulated in NSCLC, SCLC, or common to both of these types utilizing Osprey v1.0.1 [30] (Figure 1, Step-3). To achieve this, we selected all experimentally validated, highly ranked miRNA targets of NSCLC, SCLC, or common to both that were identified in the previous step and fed them into Osprey (Figure 1, Step-6). The protein-protein interaction (PPI) network for each cancer type generated by Osprey was first filtered sequentially with the "Transcription", "Cell cycle" and "Cell cycle biogenesis" GO filters in Osprey (Figure 1, Step-8). Therefore, the resultant TF-TF interaction network is cell cycle specific. The sequential filters were used because cell cycle deregulation is one of the major BPs (Biological Processes) that is affected during tumorigenesis.

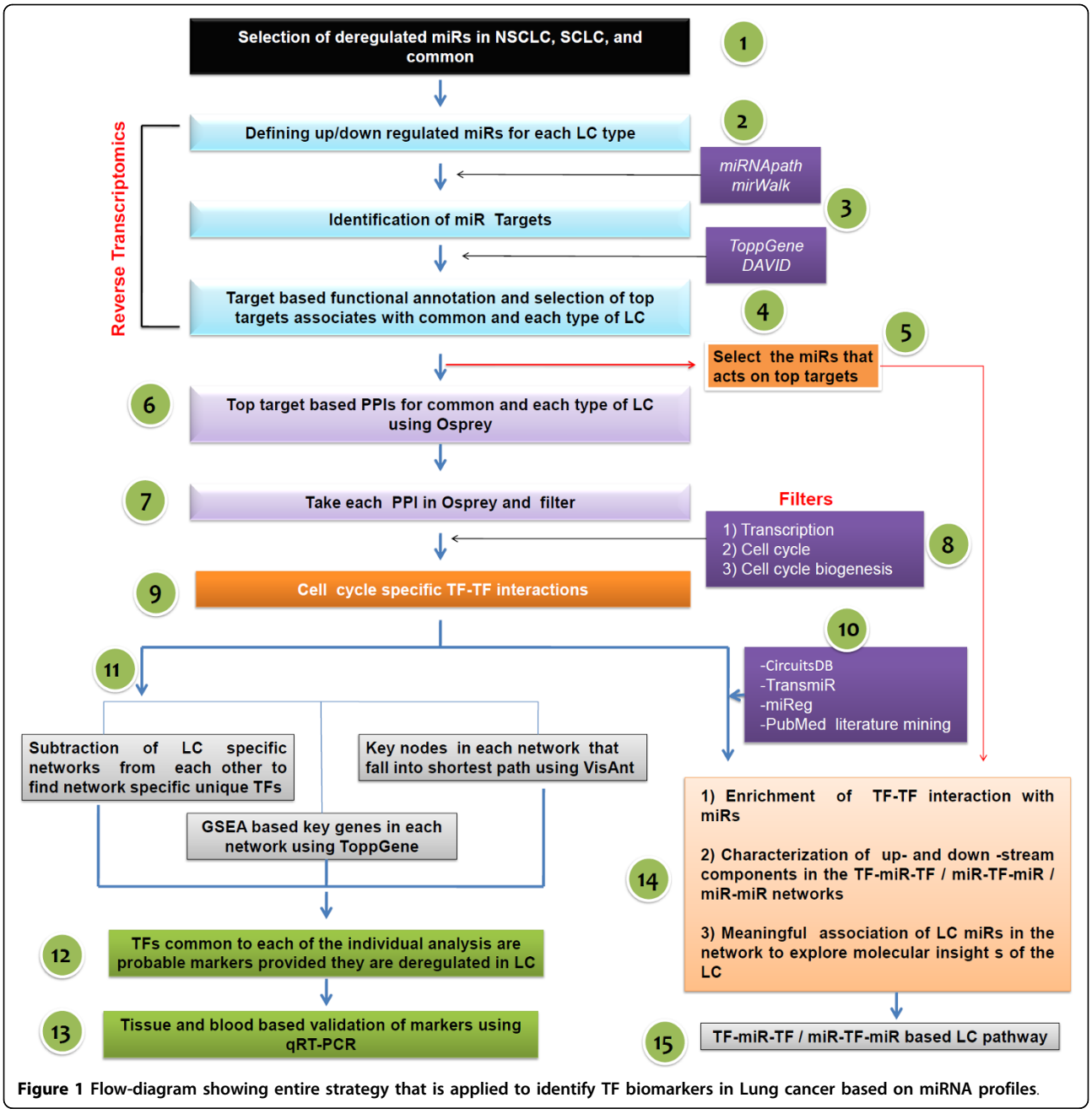


Figure 1 Flow-diagram showing entire strategy that is applied to identify TF biomarkers in Lung cancer based on miRNA profiles.

This cell cycle specific TF-TF network was further enriched by manually mapping the interacting miRNAs with data collected from the miReg [17], TransmiR [31], and CircuitsDB [32] databases and from literature mining to create a TF-miRNA-TF interaction map (Figure 1, Step-10). Because we have selected lung cancer related miRNAs (based on GO assignment in the previous step) and developed a network using their targets, this network represents the interaction of TFs involved in lung cancer tumorigenesis. Based on our earlier hypothesis, this interaction map also represents the miRNA-TF-miRNA or TF-miRNA-TF interaction map that is common to both NSCLC and SCLC. Similarly, NSCLC and SCLC specific miRNA-TF-miRNA or TF-miRNA-TF or miRNA-miRNA interaction maps were created using targets of NSCLC and SCLC unique miRNAs. Therefore, a total of three networks were generated (Figure 1, Steps-14-15).

Marker identification

The miRNA-TF-miRNA or TF-miRNA-TF interaction maps for NSCLC, SCLC, and common developed in the previous steps were analyzed by subtracting from each

other to identify the NSCLC, SCLC, and a common pathway that is specific unique TFs. Each network was further analyzed using the protein-protein interaction (PPI) analysis tool VisANT [33] to identify the key nodes and the shortest cancer specific pathways in each network. Key nodes in a PPI network are identified as having the highest number of interactions. Therefore, such key node proteins are often involved in multiple signaling pathways, and if a key node protein falls in a shortest path, the node might be treated as a marker of a disease provided that its expression is altered in that disease state. In the third strategy, we utilized GSEA identification of key genes in each network using ToppGene Suite [28]. When all of the data from each of these three analyses had been obtained, we identified the TFs common to each of the individual analyses (Figure 1, Steps-11-12). Therefore, these sets of common TFs were putative markers, and the TFs that were a part of NSCLC network could be treated as a NSCLC-specific marker.

Experimental validation of markers

Once we had selected the potential markers, we checked their expression levels initially in lung cancer tissue samples using microarrays and then further validated them using patient's blood samples and quantitative RT-PCR (qPCR) (Figure 1, Step-13).

Interrogation of data from expression microarray

The frozen tissue samples examined from 30 squamous cell carcinomas and 30 adenocarcinomas (each is a type of NSCLC) from the Liverpool Lung Project tissue bank. All samples were of pathological stage T2. RNA was extracted using the RNeasy kit (Qiagen). Five RNA pools from five adjacent normal lung tissues were also profiled for comparison purposes. The microarray experiments were performed by Almac (Belfast, UK). Total RNA was amplified using the NuGEN™ Ovation™ RNA Amplification System V2. First-strand synthesis of cDNA was performed using a unique first-strand DNA/RNA chimeric primer mix, resulting in cDNA/mRNA hybrid molecules. Following fragmentation of the mRNA component of the cDNA/mRNA molecules, second-strand synthesis was performed, and double-stranded cDNA was produced with a unique DNA/RNA heteroduplex at one end. In the final amplification step, RNA within the heteroduplex was degraded using RNaseH, and a replication of the resultant single-stranded cDNA was achieved using the DNA/RNA chimeric primer binding and DNA polymerase enzymatic activity. The amplified single-stranded cDNA was purified to allow accurate quantitation of the cDNA and to ensure optimal performance during the fragmentation and labeling process. The single-stranded cDNA was assessed using spectrophotometric methods in combination with the Agilent Bioanalyzer.

The appropriate amount of amplified single-stranded cDNA was fragmented and labeled using the FL-Ovation™ cDNA Biotin Module V2. The enzymatically and chemically fragmented product (50-100 nt) was labeled via the attachment of biotinylated nucleotides onto the 3'-end of the fragmented cDNA.

The resultant fragmented and labeled cDNA was added to the hybridization cocktail in accordance with the NuGEN™ guidelines for hybridization onto Affymetrix GeneChip® arrays. Following hybridization for 16-18 hours at 45°C in an Affymetrix GeneChip® Hybridization Oven 640, the array was washed and stained on the GeneChip® Fluidics Station 450 using the appropriate fluidics script and then inserted into the Affymetrix autoloader carousel and scanned using the GeneChip® Scanner 3000.

The Rosetta Error Model has been applied to the raw data to generate the processed data. The profile comparisons between cancerous lesions and normal RNA pools utilized Student's t-test. The Benjamini & Hochberg multiple test correction method was also employed.

Validation using quantitative RT-PCR (qPCR)

Blood samples, RNA isolation, and cDNA preparation

As our focus is NSCLC, blood samples from 8 metastatic lung adenocarcinoma, 8 metastatic squamous cell lung carcinoma patients, and 5 healthy volunteers (control) were used for the validation. Patient eligibility criteria were as follows: 18 years of age or older, in clinical stage II-IV based on the International TNM classification, performance status of 0 to 2, and no other malignances. All patients and volunteers have signed informed consent forms. Ten milliliters of EDTA blood sample was collected from the selected groups before chemotherapy treatment. Blood samples were centrifuged at 2000 g for 10 min and the serum phase was separated and frozen at -80°C. The Buffy Coat (white blood cells and circulating tumor cells) was collected and processed by lysis (Ammonium Chloride, TRIS, ddH_20) and then washed with PBS. The dry pellet was kept at -80°C until RNA isolation. RNA was purified by Quiamp RNA Blood Mini Kit (QIAGEN Inc., USA) according to the manufacturer´s instructions. cDNA was synthesized with random hexamer primers (Deoxynucleoside Triphosphate set, Roche, Germany) at 10 mM, $MgCl_2$, MuLV Reverse Transcriptase, PCR Buffer, RNAse Inhibitor, and random hexamers from Applied Biosystems USA. The resulting cDNA was stored at -20°C until further use.

Quantitative RT-PCR (qPCR)

qPCR was carried out using SYBR® Green Master Mix (Applied Byosistems, USA) and Applied Biosystem's 7500 real-time PCR system according to the manufacturer´s instructions. Primers for GAPDH were designed

with Vector NTI Advance™ 11 (Invitrogen) and primers for TFDP1, SUV39H1, RBL1, E2FG, IRF1, HMGA1, and HNRPD were designed using *qPrimerDepot* (http://primerdepot.nci.nih.gov/). To avoid the influence of genomic contamination, the amplicons spanned at least one intron. The primers used are listed in Additional file 1. qPCR was performed in a final volume of 20 μl with a SYBR PCR Master Mix, using 1 μl cDNA. Cycling conditions were 95°C for 10 min, followed by 40 cycles at 95°C for 15 s and 60°C for 1 min each to obtain the melting curve.

Relative gene expression levels were determined by the quantitative curve method. Quantitative normalization of the cDNA in each sample was performed using GAPDH gene expression as an internal control. Target gene mRNA levels were given as ratios to GAPDH mRNA levels. qPCR assays were performed in duplicate for each sample, and the mean value was used to calculate the mRNA expression levels.

Results

miRNA statistics in lung cancer

We selected 184 miRNAs for NSCLC and 62 for SCLC using literature mining and the miR2 Disease database. Among these 246 miRNAs, 41 were found to be involved in both of the lung cancers and therefore are common miRNAs involved in lung cancer regardless of the subtype (Figure 1, Step-1). In the common miRNA group, 13 and 11 miRNAs were found to be up- and downregulated, respectively; whereas 18 miRNAs showed differential expression, i.e., either upregulated in SCLC and downregulated in NSCLC or *vice versa* (Figure 1, Step-2) (Additional file 2). A total of 22 miRNAs were found to be unique to SCLC (16 upregulated and 6 downregulated) (Additional file 3). For NSCLC, the total number of unique miRNAs was 143, (89 upregulated and 43 downregulated) (Additional file 4).

Target-based functional annotation of miRNAs

Using miRWalK, miRBASE, miRecord, miRTarBASE, and miReg we identified several validated targets for each miRNA. Thereafter, as per our reverse transcriptomics strategy, targets for each miRNA were subjected to gene enrichment analysis using ToppGene Suite as described in Materials and Methods (Figure 1, Step-3). Top targets that are associated with common, NSCLC, and SCLC were identified. DAVID-based functional annotations of the top targets revealed that most of these targets are cell cycle related, so the miRNAs that have these targets are related to transcription, cell cycle regulation, cell biogenesis and organization, cell proliferation, and other biological processes related to tumorigenesis. The list of common miRNAs involved in lung cancer along with their corresponding GO terms is presented in Additional file 5. miRNAs involved uniquely in either NSCLC or SCLC and their corresponding GO terms were also defined (data not shown).

miRNA-miRNA interaction network in lung cancer

Interaction of common miRNAs

Based on the hypothesis that interactions of miRNA-TF-miRNA or TF-miRNA-TF-miRNA targets represent miRNA-miRNA interactions, we used gene enrichment based on the top targets of miRNAs common to NSCLC and SCLC in Osprey to create a protein-protein interaction map (Figure 1, Steps-6-7). In total, 638 targets corresponding to 40 common miRNAs generated a map having 1791 nodes in Osprey. Keeping in mind that miRNA genes are regulated by transcription factors (TF), miRNAs regulate TFs, and, as the gene enrichment analysis shows, most of the miRNAs regulate transcription, the network of 1791 nodes is filtered with the "Transcription factor" filter in Osprey and subsequently only 170 nodes are retained. This transcription network of 170 nodes is further filtered with "Cell cycle" and "Cell Organization and Biogenesis" filters, as per the enriched GO categories (Figure 1, Step-8), and finally the cell cycle specific total of 26 key TF nodes in common events, NSCLC, and SCLC are found (Figure 1, Step-9 and Figure 2).

Interactions of SCLC associated miRNAs

For SCLC, 634 nodes are used in total to create the interaction map in Osprey. The resultant map is sequentially filtered with "transcription factor", "Cell cycle", and "Cell organization and biogenesis" Filters and only 9 key nodes are obtained (Figure 1, Steps-6-9 and Figure 3).

Interactions of NSCLC linked miRNAs

Similar methods of network creation and filtering to those applied to identify key nodes in common and in SCLC (Figure 1, Steps-6-9) were adopted to generate a key interaction network in NSCLC. A total of 2421 nodes are filtered and finally 27 nodes are obtained (Figure 4).

SCLC network is a part of NSCLC

Next we subtracted the LC specific networks from each other to identify unique network specific TFs (Figure 1, Step-11). In the 27 nodes of the NSCLC network (Figure 4), all of the SCLC nodes (Figure 2) are found to be present (Figure 4, in red circle). Therefore, it is evident that there are additional pathways involved in NSCLC compared to SCLC and the SCLC network represents a subset of the NSCLC network.

Genes involved in common events in lung cancer

Next, we compared the common network (Figure 2) with the SCLC (Figure 3) and NSCLC and SCLC networks (Figure 4) by subtracting each from the other to identify key nodes that are common to (1) SCLC and NSCLC;

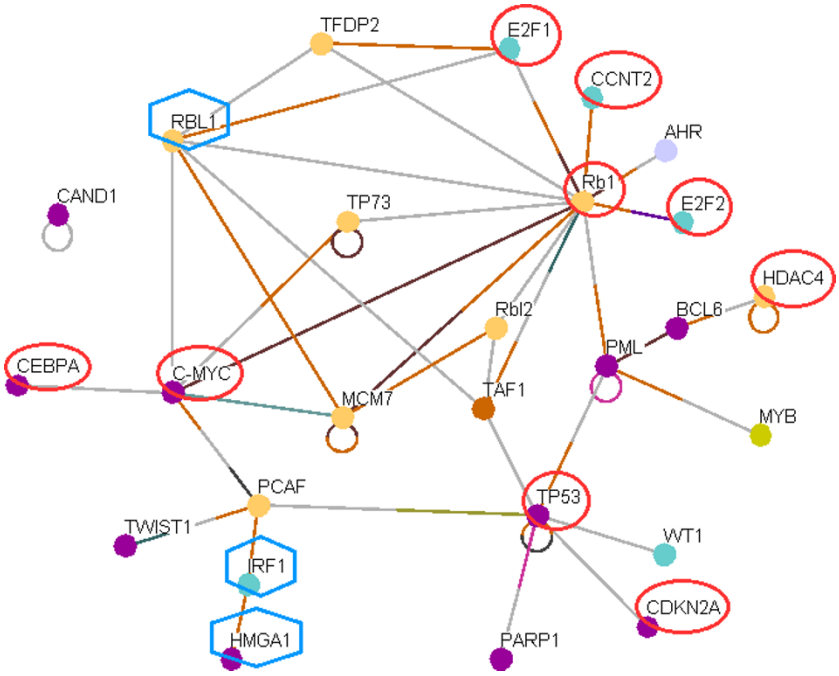


Figure 2 Cell cycle specific 26 key interacting TFs that are targets of miRNAs involved in common events in lung cancer as well as in NSCLC and SCLC. The network is created as described in the text. As per our hypothesis, this network also represents interactions of cell cycle regulating miRNAs associated with NSCLC, SCLC, and common events of lung cancer. TFs circled in red are shared by both NSCLC and SCLC. Molecules marked in hexagon are unique to common events. Other molecules in the map are shared by NSCLC and common events of lung cancers.

(2) general events, NSCLC, and SCLC; (3) NSCLC and general; (4) NSCLC specific; and (5) general events in lung cancers. The analysis revealed that nine genes (RB1, E2F1, E2F2, CCNT2, CMYC, CEBPA, TP53, CDKN2A, and HDAC4) that are key nodes in SCLC are common to both the (1) SCLC and NSCLC and (2) general events, NSCLC, and SCLC groups (Table 1, group-1-3). Therefore, all of the SCLC genes are involved in NSCLC and in general events in lung cancer. Fourteen unique genes (Table 1, group-4) are found to be involved in both NSCLC and general events. The comparison also shows that four genes (Table 1, group-5) are specific to NSCLC and three genes (Table 1, group-6) are unique to general events. Therefore, these gene sets can be used in combination and their expression signature may be useful as diagnostic markers for NSCLC.

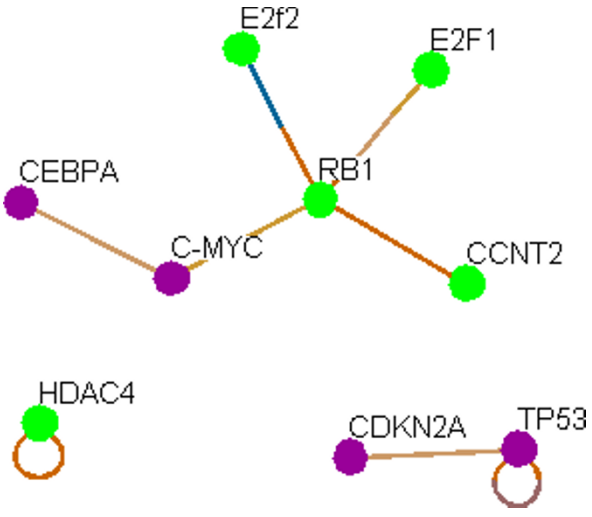


Figure 3 Cell cycle specific 9 key interacting TFs that are targets of miRNAs involved in SCLC. As per our hypothesis, this network represents interaction of cell cycle regulating miRNAs associated with SCLC. For detail, please see the text.

Validation of markers

We selected seven genes [4 unique genes (E2F6, TFDP1, SUV39H1, and HNRPD) for NSCLC and 3 genes (RBL1, IRF1, and HMGA1) for general events] for validation as diagnostic markers in lung cancer. Frozen NSCLC tissue-based microarray analysis revealed that E2F6, TFDP1, SUV39H1, and HMGA1 are significantly upregulated in both the adenocarcinoma and squamous cell carcinoma samples. The upregulation of RBL1 and downregulation of IRF1 in the microarray analysis was significant in squamous cell carcinoma but was statistically insignificant in adenocarcinoma (Additional file 6).

qPCR validation of markers based on blood samples showed expression patterns similar to the tissue based microarray analysis. TFPD1, E2F6, IRF1, and HMGA1 are

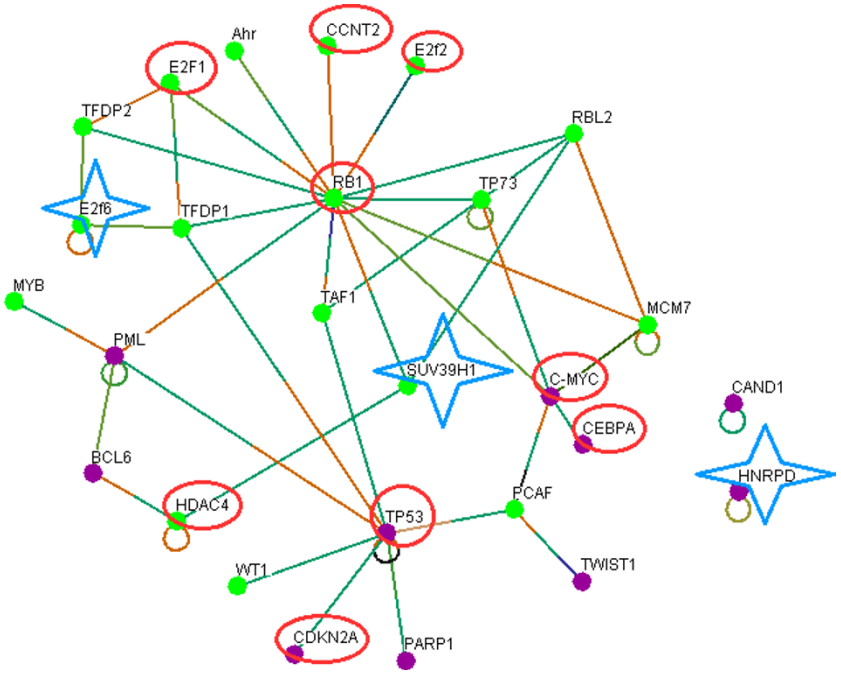


Figure 4 Interactions of TFs (as per our hypothesis miRNAs) associated with NSCLC and SCLC. TFs circled in red are shared by both NSCLC and SCLC. Molecules marked in star are unique to NSCLC. Other molecules in the map are shared by NSCLC and general events of lung cancers.

upregulated in all cancer samples. SUV39H1, RBL1, and HNRPD are downregulated or not expressed in all samples compared to the control (Figure 5). Therefore, combining the microarray and qPCR results, upregulation of E2F6, HMGA1, IRF1, and TFDP1 and downregulation or no expression of SUV39H1, RBL1, HNRPD can be used as diagnostic markers of NSCLC, and, in particular, adenocarcinoma and squamous cell carcinoma.

Discussion

In this work we have identified key transcription factors that can be useful biomarkers in diagnosis of lung cancer using an *in silico* reverse-transcriptomics approach. In this novel approach, starting with deregulated miRNAs in lung cancers we have identified transcription factors that can act as biomarkers, even for sub-type specific lung cancers. Out of several putative markers we identified, 7 NSCLC specific markers were validated. We found that E2F6, HMGA1, IRF1, and TFDP1 were upregulated and RBL1, SUV39H1, and HNRPD were downregulated or aberrantly expressed in adenocarcinoma and squamous cell carcinoma, which are the sub-types of NSCLC.

HMGA1 (High mobility group AT-hook 1) is an oncogene that is induced by Wnt/beta-catenin pathway and which positively regulates cell proliferation in gastric cancer [34]. By downregulating E-cadherin and upregulating expression of TWIST1, it enhances epithelial-mesenchymal transition and metastasis in colon cancer [35]. Upregulation of HMGA1 in glioblastoma positively correlates with malignancy, angiogenesis, and invasion [36]. In lung cancer, it is also overexpressed and increased nuclear expression correlates with poor survival in lung adenocarcinomas [37,38]. By upregulating PI3K and MMP2, it promotes cell migration and invasion [37,39] and by

Table 1 Identified putative markers in lung cancers using the *in silico* reverse transcriptomics approach

Group	LC Types	Gene sets
1	Unique to SCLC	RB1, E2F1, E2F2, CCNT2, CMYC, CEBPA, TP53, CDKN2A, HDAC4
2	Common to SCLC and NSCLC	RB1, E2F1, E2F2, CCNT2, CMYC, CEBPA, TP53, CDKN2A, HDAC4
3	Common to general, SCLC, and NSCLC	RB1, E2F1, E2F2, CCNT2, CMYC, CEBPA, TP53, CDKN2A, HDAC4
4	Common to NSCLC and general	TFDP2, AHR, CCND1, TP73, RBL2, TAF1, PML, BCL6, MYB, WT1, PARP1, PCAF, TWIST, MCM7
5	NSCLC specific	E2F6, TFDP1, SUV39H1, HNRPD
6	General/ common path specific	RBL1, IRF1, HMGA1

The markers can be used in combination to design panels for diagnosis of sub-type specific lung cancers.

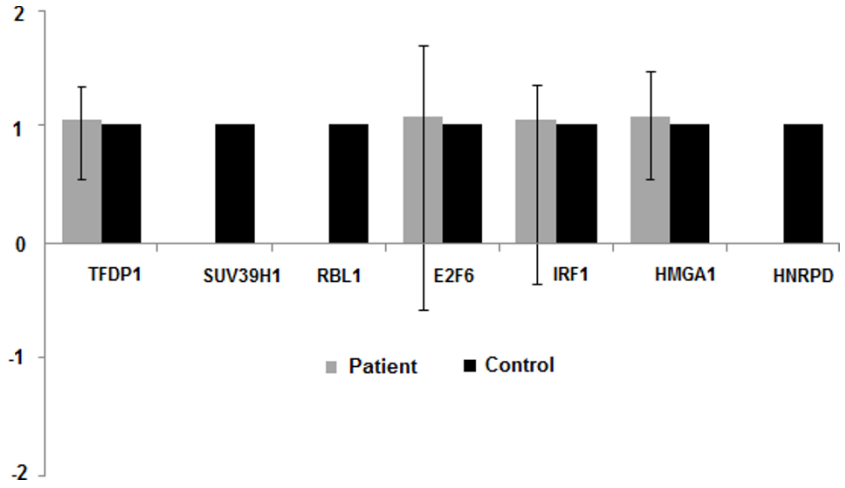


Figure 5 Blood based qPCR results for selected seven NSCLC specific markers. As compared to the control; HMGA1, TFPD1, E2F6, and IRF1 are upregulated and SUV39H1, RBL1, and HNRPD are downregulated or not expressed in all tested samples.

activating miR-222 oncomiR, it induces PPP2R2A mediated AKT signaling in NSCLC [40]. Therefore, upregulation of HMGA1 plays a significant role in tumor progression in NSCLC. In our study, we also observed that HMGA1 was upregulated in NSCLC supporting the previous findings.

TFDP1 (Transcription factor Dp-1) is a candidate oncogene that positively regulates S-phase entry and inhibits apoptosis in cooperation with E2F1 [41]. It is amplified and overexpressed in breast cancer [42] and upregulation of TFDP1 positively correlates with tumor size and progression of hepatocellular carcinomas [43] and increased cell viability in lung cancer [44]. In our observation, TFDP1 was overexpressed in all lung adenocarcinomas and squamous cell carcinomas, which supports the previous findings of Lu et al. (2000) in a SCLC cell line [45].

In our study, we observed **IRF1** (Interferon regulatory factor 1) was upregulated in all NSCLC samples tested, although it had been shown to be downregulated in lung cancer in a previous study [46]. IRF1 inhibits G1-S cell cycle progression through P53 and p21 mediated pathways [46] and may act as a tumor-suppressor gene. This finding is supported by the findings that it is downregulated in gastric [47] and recurrent breast cancers [48]. However, IRF1 may not always act as a tumor-suppressor, as there is a report that it is upregulated in skin squamous cell carcinoma [49]. Therefore, our observation of upregulated IRF1 in NSCLC samples requires further attention to explore the precise role of this TF in various cancers.

E2F6 (E2F transcription factor 6) inhibits entry into S phase of cells stimulated to exit G0 [50] and inhibits apoptosis through E2F1 [51]. It may therefore play a role in cell proliferation and cell survival. There is no report about this protein's expression pattern in any cancer. Here, we have, for the first time, observed that E2F6 was upregulated in all of our tested NSCLC samples. This finding supports E2F6's putative role in tumorigenesis and shows that it may be a novel marker for NSCLC.

SUV39H1 (Suppressor of variegation 3-9 homolog 1) is a histone methyltransferase that inhibits inflammatory responses by downregulating interleukin-6 production [52]. SUV39H1 inhibits the expression of CCND1 and may thereby negatively regulate cell proliferation [53]. However, its overexpression induces cell migration in breast and colon cancers [54] and negatively regulates apoptosis in a lung cancer model [55]. The expression level of SUV39H1 inversely correlates with stage, prognosis, and disease free survival in oral squamous cell carcinoma [56] and breast cancer [57]. Therefore, SUV39H1 may also have oncogenic properties. Although SUV39H1 was significantly upregulated in adenocarcinoma and squamous cell carcinoma tissue samples in our microarray analysis, supporting its positive role in tumorigenesis, it was found to be downregulated in blood samples in our qPCR validation. Therefore, SUV39H1 expression differs in lung cancer tissue and blood samples.

RBL1 (Retinoblastoma-like 1 (p107)) inhibits cell proliferation through G1 arrest [58] and positively regulates epidermal differentiation [59]. RBL1 is downregulated and inversely correlates with the histological grade of squamous cell carcinomas and adenocarcinomas [60]. Our qPCR validation shows downregulation in all squamous cell carcinoma and adenocarcinoma samples, which supports the previous findings and RBL1's function in tumors.

HNRPD/AUF1 is a RNA-binding protein that both positively and negatively regulates neoplastic gene regulatory networks in cancer depending on the type of neoplasm [61]. It binds to destabilize p21 mRNA and thereby inhibits its anti-apoptotic activity [62]. Although in our blood-based qPCR analysis AUF1 was downregulated in all NSCLC samples, it has been reported to be upregulated in HCC [63] and experimental murine lung cancer [64]. It has been patented to aid in the prediction of survival in lung cancer in a gene expression panel of biomarkers (US 20100267574).

miRNA-markerTFs correlation: The seven identified TFs that are aberrantly expressed in both the squamous cell carcinoma and adenocarcinoma were plotted for their interactions with miRNAs and other key TFs to obtain more insight into these markers in lung cancer pathogenesis (Figure 1, Steps-14-15). The miRNA-TF-Cancer relationships were gathered from the miReg [17], miR2Disease [23], miRWalk [25], miRecords [26], TransmiR [31], CircuitsDB [32], and miRDB [65] databases. The interaction map is represented in Figure 6.

The network clearly shows meaningful relationships between the TFs and miRNAs in lung cancer. The interactions show that the tumor suppressor miRNAs (miR-29a, miR-16, miR-125, and let-7) that could target the oncogene HMGA1 are downregulated. Upregulation of HMGA1 induces expression of oncogenic miR-122. Another two pro-oncogenic miRNAs that can also target HMGA1, miR-196a-2 and miR-155, are upregulated in lung cancers [66,67]. We observed that HMGA1 may inhibit the putative tumor-suppressor IRF1 (as per the interaction network) and that the miR-155 pro-oncomiR directly targeted IRF1. Therefore, in this network, HMGA1 is the key TF that positively regulates lung tumorigenesis through upregulation of miR-122 and perhaps by downregulation of IRF1. However, we found that IRF1 is upregulated in the samples so that the IRF1-HMGA1 interactions need further attention.

Tumor suppressor RBL1 is a target of the miR-17 oncomiR [68]. Furthermore, as per the interaction network, RBL1 is activated by TAF1 and cMYC, and regulates expression of E2F2, RB1, MCM7, and TFDP2.

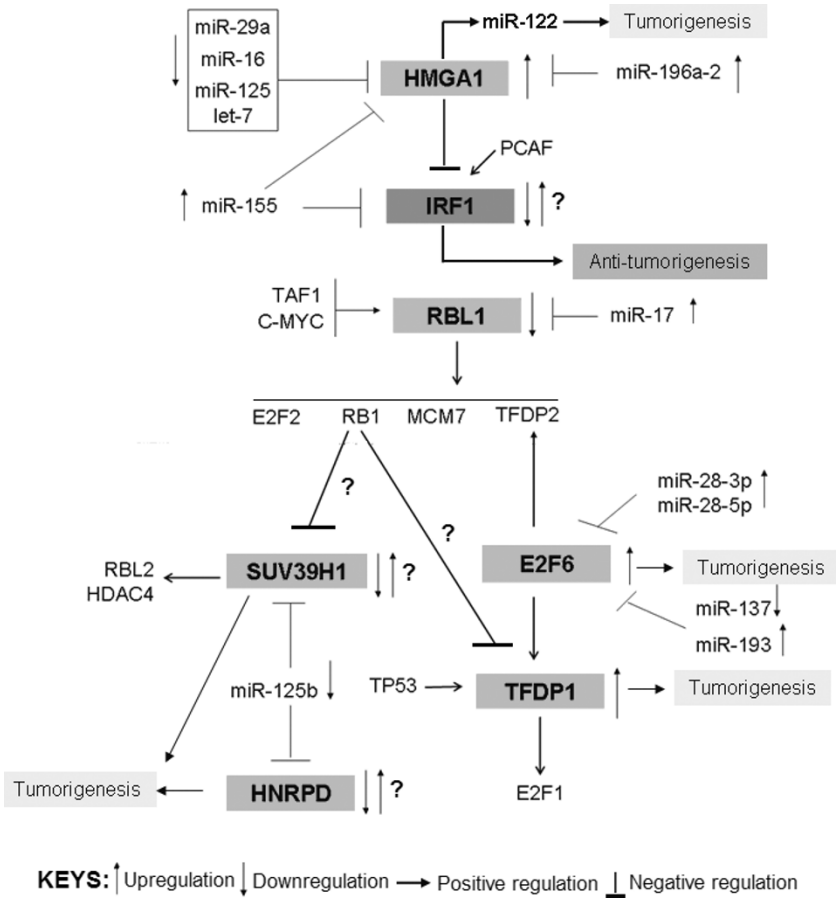


Figure 6 The correlations of identified seven TF markers and interacting miRNAs. The interactions provide better insights of molecular events and mechanisms during lung cancer tumorigenesis. For detail, please see the text.

It thereby regulates the cell cycle and cell proliferation. Therefore, RBL1 downregulation and upregulation of miR-17 provide a meaningful mechanism in lung cancer tumorigenesis [66,69].

The common pathway (of both NSCLC and SCLC) related genes HNRPD, E2F6, TFDP1, and SUV39H1 also showed the expected TF-miRNA relationship in the interaction map represented in Figure 6 based on the available experimental evidence. The literature shows that HNRPD and SUV39H1 may have positive roles in tumorigenesis [55,56,64]. Although in our blood-based qPCR, HNRPD and SUV39H1 are downregulated, they are reported to be upregulated in a mouse model of lung cancer [63], consistent with the tissue-based microarray analysis in our lung cancer samples. The involvement of HNRPD and SUV39H1 is further supported by reports that the tumor suppressor miR-125 is downregulated in both NSCLC and SCLC [70,71]. Furthermore, the tumor suppressor protein RB1 is downregulated in lung cancer [66] and may inhibit SUV39H1.

The other two markers, E2F6 and TFDP1, are upregulated in all of our blood samples. While two pro-oncogenic miRNAs, miR-28 and miR-193, are upregulated [40] the putative tumor-suppressor, miR-137, is downregulated in lung cancers [72,73]. All three of these miRNAs target E2F6 [74,75]. Furthermore, E2F6 putatively upregulates TFDP1 and is downregulated by RB1. It is also found from the interaction map that E2F6 inhibition by two upregulated pro-oncomiRs (miR-28 and miR-193) is not sufficient, as the E2F6 was found to be upregulated in lung cancer. Further, E2F6 has been reported to upregulate oncogene TFDP1 and to positively regulate cell proliferation and cell survival through E2F1 [41]. Additionally, downregulation of RB1 in lung cancer is not able to repress TFDP1 activity, and therefore, in lung cancer, tumorigenesis is mediated through upregulation of E2F6 and TFDP1. However, the role of SUV39H1 and HNRPD requires further exploration.

Conclusion

In this analysis, using an integrated reverse-transcriptomics-based bioinformatics approach, we have identified key transcription factors that may be useful in developing subtype specific biomarkers in lung cancer. Our proposed seven markers also have high potential to be used in lung cancer diagnostics for NSCLC subtypes. Of course, additional experimental validation in independent sets of patients is required to establish the diagnostic accuracy of this panel and we are currently conducting those experiments. The miRNA-TF-miRNA relationships with these seven miRNAs show meaningful associations with these TFs in lung cancer pathogenesis. The novel strategy developed in this research is powerful and can be applicable to identify molecular mechanisms and markers in other cancers as well.

Funding

This work was carried out without any grant. VA had funding from CNPq and FAPEMIG.

Additional material

Additional file 1: List of primers to amplify TFDP1, SUV39H1, RBL1, E2FG, IRF1, HMGA1, and HNRPD.

Additional file 2: Common miRNAs involved in both NSCLC and SCLC. The differentially expressed miRNAs are marked with blue.

Additional file 3: Small-cell-lung cancer (SCLC) specific 22 deregulated miRNAs (16 upregulated and 6 downregulated).

Additional file 4: Non-small-cell lung cancer (NSCLC) specific 143 deregulated miRNAs (89 upregulated and 43 downregulated). The miRNAs that are reported upregulated in one report but downregulated in other report or *vise versa* are highlighted in blue.

Additional file 5: Functional annotation of common miRNAs using the targets of these miRNAs and DAVID

Additional file 6: Microarray based expression analysis of NSCLC specific 6 identified markers [E2F6, TFDP1, and SUV39H1 for NSCLC and RBL1, IRF1, and HMGA1 for general events]. E2F6, TFDP1, SUV39H1, and HMGA1 are significantly upregulated in both the adenocarcinoma and squamous cell carcinoma samples. The upregulation of RBL1 and downregulation of IRF1 in the microarray analysis was significant in squamous cell carcinoma but was statistically insignificant in adenocarcinoma.

Conflict of interest

Authors declare no conflict of interest.

Authors' contributions

DB: Conceived the idea, designed the study, coordinated and leaded the entire project, and wrote the manuscript; **DB, NJ:** collected and analyzed primary data, **DB, NJ, ST, AB, LJ:** performed all *in silico* analyses; **JKF, TL:** performed microarray analysis; **EP, AR, RL, MH:** performed qPCR experiments; **PG**, **CV**, **AK, AS, AM, VA, KB:** cross verified all analyses. All authors have read and approved the manuscript.

Declarations

Publication for this article has been funded by NSF 1158608.
This article has been published as part of *BMC Genomics* Volume 14 Supplement 6, 2013: Proceedings of the International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2012). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S6.

Authors' details

[1]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB-721172, India. [2]School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India. [3]University of Liverpool, Department of Molecular and Clinical Cancer Medicine, 200 London Road, Liverpool L3 9TA, UK. [4]Medical Oncology Department, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain. [5]Nuclear Medicine Service, Complejo Hospitalario Universitario. Fundación Tejerina. Santiago de Compostela, A Coruña, Spain. [6]Molecular Oncology and Imaging Program, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain. [7]Institute of Molecular and Experimental Medicine, Cardiff University, Cardiff CF14 4XN, Wales, UK. [8]Instituto de Ciências Biológicas, Universidade Federal do Pará, Rua Augusto Corrêa, 01 - Guamá, Belém, PA, Brazil.

[9]Laboratorio de Genetica Celular e Molecular, Departmento de Biologia Geral, Instituto de Ciencias Biologics, Universidade Federal de Minas Gerais CP 486, CEP 31270-901 Belo Horizonte, Minas Gerais, Brazil. [10]Department of Psychiatry and Mcknight Brain Institute, College of Medicine, University of Florida, University Ave., Gainesville, FL 32601, USA. [11]Department of Computer Science and Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA.

Published: 25 October 2013

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D: **Global cancer statistics.** *CA Cancer J Clin* 2011, **61(2)**:69-90.
2. Petersen I: **The morphological and molecular diagnosis of lung cancer.** *Dtsch Arztebl Int* 2011, **108(31-32)**:525-531.
3. Wang T, Nelson RA, Bogardus A, Grannis FW Jr: **Five-year lung cancer survival: which advanced stage nonsmall cell lung cancer patients attain long-term survival?** *Cancer* 2010, **116(6)**:1518-1525.
4. Pastorino U: **Lung cancer screening.** *Br J Cancer* 2010, **102(12)**:1681-1686.
5. Granville CA, Dennis PA: **An overview of lung cancer genomics and proteomics.** *Am J Respir Cell Mol Biol* 2005, **32(3)**:169-176.
6. Bianchi F, Nicassio F, Marzi M, Belloni E, Dall'olio V, Bernard L, Pelosi G, Maisonneuve P, Veronesi G, Di Fiore PP: **A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer.** *EMBO Mol Med* 2011, **3(8)**:495-503.
7. Hennessey PT, Sanford T, Choudhary A, Mydlarz WW, Brown D, Adai AT, Ochs MF, Ahrendt SA, Mambo E, Califano JA: **Serum microRNA Biomarkers for Detection of Non-Small Cell Lung Cancer.** *PLoS One* 2012, **7(2)**:e32307.
8. Yu L, Todd NW, Xing L, Xie Y, Zhang H, Liu Z, Fang H, Zhang J, Katz RL, Jiang F: **Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers.** *Int J Cancer* 2010, **127(12)**:2870-2878.
9. Liao J, Yu L, Mei Y, Guarnera M, Shen J, Li R, Liu Z, Jiang F: **Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer.** *Mol Cancer* 2010, **9**:198.
10. Xing L, Todd NW, Yu L, Fang H, Jiang F: **Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers.** *Mod Pathol* 2010, **23(8)**:1157-1164.
11. Guergova-Kuras M, Kurucz I, Hempel W, Tardieu N, Kádas J, Malderez-Bloes C, Jullien A, Kieffer Y, Hincapie M, Guttman A, Csánky E, Dezso B, Karger BL, Takács L: **Discovery of lung cancer biomarkers by profiling the plasma proteome with monoclonal antibody libraries.** *Mol Cell Proteomics* 2011, **10(12)**:M111.010298.
12. Tong BC, Harpole DH Jr: **Molecular markers for incidence, prognosis, and response to therapy.** *Surg Oncol Clin N Am* 2012, **21(1)**:161-75.
13. Xiao H, Zhang L, Zhou H, Lee JM, Garon EB, Wong DT: **Proteomic analysis of human saliva from lung cancer patients using two-dimensional difference gel electrophoresis and mass spectrometry.** *Mol Cell Proteomics* 2012, **11(2)**:M111.012112.
14. Begum S, Brait M, Dasgupta S, Ostrow KL, Zahurak M, Carvalho AL, Califano JA, Goodman SN, Westra WH, Hoque MO, Sidransky D: **An epigenetic marker panel for detection of lung cancer using cell-free serum DNA.** *Clin Cancer Res* 2011, **17(13)**:4494-4503.
15. Jiang F, Todd NW, Li R, Zhang H, Fang H, Stass SA: **A panel of sputum-based genomic marker for early detection of lung cancer.** *Cancer Prev Res (Phila)* 2010, **3(12)**:1571-1578.
16. Showe MK, Vachani A, Kossenkov AV, Yousef M, Nichols C, Nikonova EV, Chang C, Kucharczuk J, Tran B, Wakeam E, Yie TA, Speicher D, Rom WN, Albelda S, Showe LC: **Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease.** *Cancer Res* 2009, **69(24)**:9202-9210.
17. Barh D, Bhat D, Viero C: **miReg: a resource for microRNA regulation.** *J Integr Bioinform* 2010, **7(1)**:144.
18. Blenkiron C, Miska EA: **miRNAs in cancer: approaches, aetiology, diagnostics and therapy.** *Hum Mol Genet* 2007, **16(1)**:R106-13.
19. Sempere LF, Liu X, Dmitrovsky E: **Tumor-suppressive microRNAs in Lung cancer: diagnostic and therapeutic opportunities.** *Scientific World Journal* 2009, **9**:626-628.
20. Barshack I, Lithwick-Yanai G, Afek A, Rosenblatt K, Tabibian-Keissar H, Zepeniuk M, Cohen L, Dan H, Zion O, Strenov Y, Polak-Charcon S, Perelman M: **MicroRNA expression differentiates between primary lung tumors and metastases to the lung.** *Pathol Res Pract* 2010, **206(8)**:578-584.
21. Hu Z, Chen X, Zhao Y, Tian T, Jin G, Shu Y, Chen Y, Xu L, Zen K, Zhang C, Shen H: **Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer.** *J Clin Oncol* 2010, **28(10)**:1721-1726.
22. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer.** *Nat Rev Cancer* 2006, **6(4)**:259-269.
23. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y: **miR2Disease: a manually curated database for microRNA deregulation in human disease.** *Nucleic Acids Res* 2009, **37(Database)**:D98-104.
24. Tsou JA, Galler JS, Siegmund KD, Laird PW, Turla S, Cozen W, Hagen JA, Koss MN, Laird-Offringa IA: **Identification of a panel of sensitive and specific DNA methylation markers for lung adenocarcinoma.** *Mol Cancer* 2007, **6**:70.
25. Dweep H, Sticht C, Pandey P, Gretz N: **miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes.** *J Biomed Inform* 2011, **44(5)**:839-47.
26. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T: **miRecords: an integrated resource for microRNA-target interactions.** *Nucleic Acids Res* 2009, **37(Database)**:D105-10.
27. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD: **miRTarBase: a database curates experimentally validated microRNA-target interactions.** *Nucleic Acids Res* 2011, **39(Database)**:D163-9.
28. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res* 2009, **37(Web Server)**:W305-11.
29. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources.** *Nature Protoc* 2009, **4(1)**:44-57.
30. Breitkreutz BJ, Stark C, Tyers M: **Osprey: a network visualization system.** *Genome Biol* 2003, **4(3)**:R22.
31. Wang J, Lu M, Qiu C, Cui Q: **TransmiR: a transcription factor-microRNA regulation database.** *Nucleic Acids Res* 2010, **38(Database)**:D119-22.
32. Friard O, Re A, Taverna D, De Bortoli M, Corá D: **CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse.** *BMC Bioinformatics* 2010, **11**:435.
33. Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C: **VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology.** *Nucleic Acids Res* 2009, **37(Web Server)**:W115-21.
34. Akaboshi S, Watanabe S, Hino Y, Sekita Y, Xi Y, Araki K, Yamamura K, Oshima M, Ito T, Baba H, Nakao M: **HMGA1 is induced by Wnt/beta-catenin pathway and maintains cell proliferation in gastric cancer.** *Am J Pathol* 2009, **175(4)**:1675-1685.
35. Belton A, Gabrovsky A, Bae YK, Reeves R, Iacobuzio-Donahue C, Huso DL, Resar LM: **HMGA1 induces intestinal polyposis in transgenic mice and drives tumor progression and stem cell properties in colon cancer cells.** *PLoS One* 2012, **7(1)**:e30034.
36. Pang B, Fan H, Zhang IY, Liu B, Feng B, Meng L, Zhang R, Sadeghi S, Guo H, Pang Q: **HMGA1 expression in human gliomas and its correlation with tumor proliferation, invasion and angiogenesis.** *J Neurooncol* 2012, **106(3)**:543-549.
37. Hillion J, Wood LJ, Mukherjee M, Bhattacharya R, Di Cello F, Kowalski J, Elbahloul O, Segal J, Poirier J, Rudin CM, Dhara S, Belton A, Joseph B, Zucker S, Resar LM: **Upregulation of MMP-2 by HMGA1 promotes transformation in undifferentiated, large-cell lung cancer.** *Mol Cancer Res* 2009, **7(11)**:1803-1812.
38. Sarhadi VK, Wikman H, Salmenkivi K, Kuosma E, Sioris T, Salo J, Karjalainen A, Knuutila S, Anttila S: **Increased expression of high mobility group A proteins in lung cancer.** *J Pathol* 2006, **209(2)**:206-212.
39. Scrima M, De Marco C, Fabiani F, Franco R, Pirozzi G, Rocco G, Ravo M, Weisz A, Zoppoli P, Ceccarelli M, Botti G, Malanga D, Viglietto G: **Signaling networks associated with AKT activation in non-small cell lung cancer (NSCLC): new insights on the role of phosphatydil-inositol-3 kinase.** *PLoS One* 2012, **7(2)**:e30427.
40. Zhang Y, Ma T, Yang S, Xia M, Xu J, An H, Yang Y, Li S: **High-mobility group A1 proteins enhance the expression of the oncogenic miR-222 in lung cancer cells.** *Mol Cell Biochem* 2011, **357(1-2)**:363-371.
41. Shan B, Farmer AA, Lee WH: **The molecular basis of E2F-1/DP-1-induced S-phase entry and apoptosis.** *Cell Growth Differ* 1996, **7(6)**:689-697.

42. Abba MC, Fabris VT, Hu Y, Kittrell FS, Cai WW, Donehower LA, Sahin A, Medina D, Aldaz CM: **Identification of novel amplification gene targets in mouse and human breast cancer at a syntenic cluster mapping to mouse ch8A1 and human ch13q34.** *Cancer Res* 2007, **67(9)**:4104-4112.
43. Yasui K, Okamoto H, Arii S, Inazawa J: **Association of over-expressed TFDP1 with progression of hepatocellular carcinomas.** *J Hum Genet* 2003, **48(12)**:609-613.
44. Castillo SD, Angulo B, Suarez-Gauthier A, Melchor L, Medina PP, Sanchez-Verde L, Torres-Lanzas J, Pita G, Benitez J, Sanchez-Cespedes M: **Gene amplification of the transcription factor DP1 and CTNND1 in human lung cancer.** *J Pathol* 2010, **222(1)**:89-98.
45. Lu K, Shih C, Teicher BA: **Expression of pRB, cyclin/cyclin-dependent kinases and E2F1/DP-1 in human tumor lines in cell culture and in xenograft tissues and response to cell cycle agents.** *Cancer Chemother Pharmacol* 2000, **46(4)**:293-304.
46. Usuda J, Saijo N, Fukuoka K, Fukumoto H, Kuh HJ, Nakamura T, Koh Y, Suzuki T, Koizumi F, Tamura T, Kato H, Nishio K: **Molecular determinants of UCN-01-induced growth inhibition in human lung cancer cells.** *Int J Cancer* 2000, **85(2)**:275-280.
47. Nozawa H, Oda E, Ueda S, Tamura G, Maesawa C, Muto T, Taniguchi T, Tanaka N: **Functionally inactivating point mutation in the tumor-suppressor IRF-1 gene identified in human gastric cancer.** *Int J Cancer* 1998, **77(4)**:522-7.
48. Cavalli LR, Riggins RB, Wang A, Clarke R, Haddad BR: **Frequent loss of heterozygosity at the interferon regulatory factor-1 gene locus in breast cancer.** *Breast Cancer Res Treat* 2010, **121(1)**:227-231.
49. Wenzel J, Tomiuk S, Zahn S, Küsters D, Vahsen A, Wiechert A, Mikus S, Birth M, Scheler M, von Bubnoff D, Baron JM, Merk HF, Mauch C, Krieg T, Bieber T, Bosio A, Hofmann K, Tüting T, Peters B: **Transcriptional profiling identifies an interferon-associated host immune response in invasive squamous cell carcinoma of the skin.** *Int J Cancer* 2008, **123(11)**:2605-15.
50. Gaubatz S, Wood JG, Livingston DM: **Unusual proliferation arrest and transcriptional control properties of a newly discovered E2F family member, E2F-6.** *Proc Natl Acad Sci USA* 1998, **95(16)**:9190-9195.
51. Yang WW, Shu B, Zhu Y, Yang HT: **E2F6 inhibits cobalt chloride-mimetic hypoxia-induced apoptosis through E2F1.** *Mol Biol Cell* 2008, **19(9)**:3691-3700.
52. Villeneuve LM, Reddy MA, Lanting LL, Wang M, Meng L, Natarajan R: **Epigenetic histone H3 lysine 9 methylation in metabolic memory and inflammatory phenotype of vascular smooth muscle cells in diabetes.** *Proc Natl Acad Sci USA* 2008, **105(26)**:9047-9052.
53. Yang YJ, Han JW, Youn HD, Cho EJ: **The tumor suppressor, parafibromin, mediates histone H3 K9 methylation for cyclin D1 repression.** *Nucleic Acids Res* 2010, **38(2)**:382-90.
54. Yokoyama Y, Hieda M, Nishioka Y, Matsumoto A, Higashi S, Kimura H, Yamamoto H, Mori M, Matsuura S, Matsuura N: **Cancer associated up-regulation of H3K9 trimethylation promotes cell motility in vitro and drives tumor formation in vivo.** *Cancer Sci* 2013, doi: 10.1111/cas.12166.
55. Watanabe H, Soejima K, Yasuda H, Kawada I, Nakachi I, Yoda S, Naoki K, Ishizaka A: **Deregulation of histone lysine methyltransferases contributes to oncogenic transformation of human bronchoepithelial cells.** *Cancer Cell Int* 2008, **8**:15.
56. Chen JH, Yeh KT, Yang YM, Chang JG, Lee HE, Hung SY: **High expressions of histone methylation- and phosphorylation-related proteins are associated with prognosis of oral squamous cell carcinoma in male population of Taiwan.** *Med Oncol* 2013, **30(2)**:513.
57. Patani N, Jiang WG, Newbold RF, Mokbel K: **Histone-modifier gene expression profiles are associated with pathological and clinical outcomes in human breast cancer.** *Anticancer Res* 2011, **31(12)**:4115-25.
58. Zhu L, van den Heuvel S, Helin K, Fattaey A, Ewen M, Livingston D, Dyson N, Harlow E: **Inhibition of cell proliferation by p107, a relative of the retinoblastoma protein.** *Genes Dev* 1993, **7(7A)**:1111-1125.
59. Paramio JM, Laín S, Segrelles C, Lane EB, Jorcano JL: **Differential expression and functionally co-operative roles for the retinoblastoma family of proteins in epidermal differentiation.** *Oncogene* 1998, **17(8)**:949-957.
60. Baldi A, Esposito V, De Luca A, Howard CM, Mazzarella G, Baldi F, Caputi M, Giordano A: **Differential expression of the retinoblastoma gene family members pRb/p105, p107, and pRb2/p130 in lung cancer.** *Clin Cancer Res* 1996, **2(7)**:1239-1245.
61. Zucconi BE, Wilson GM: **Modulation of neoplastic gene regulatory pathways by the RNA-binding factor AUF1.** *Front Biosci* 2011, **16**:2307-2325.
62. Shchors K, Yehiely F, Kular RK, Kotlo KU, Brewer G, Deiss LP: **Cell death inhibiting RNA (CDIR) derived from a 3'-untranslated region binds AUF1 and heat shock protein 27.** *J Biol Chem* 2002, **277(49)**:47061-47072.
63. Frau M, Tomasi ML, Simile MM, Demartis MI, Salis F, Latte G, Calvisi DF, Seddaiu MA, Daino L, Feo CF, Brozzetti S, Solinas G, Yamashita S, Ushijima T, Feo F, Pascale RM: **Role of transcriptional and posttranscriptional regulation of methionine adenosyltransferases in liver cancer progression.** *Hepatology* 2012, **56(1)**:165-175.
64. Blaxall BC, Dwyer-Nield LD, Bauer AK, Bohlmeyer TJ, Malkinson AM, Port JD: **Differential expression and localization of the mRNA binding proteins, AU-rich element mRNA binding protein (AUF1) and Hu antigen R (HuR), in neoplastic lung tissue.** *Mol Carcinog* 2000, **28(2)**:76-83.
65. Xiaowei W: **miRDB: a microRNA target prediction and functional annotation database with a wiki interface.** *RNA* 2008, **14(6)**:1012-1017.
66. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM: **A microRNA expression signature of human solid tumors defines cancer gene targets.** *Proc Natl Acad Sci USA* 2006, **103(7)**:2257-2261.
67. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu CG, Croce CM, Harris CC: **Unique microRNA molecular profiles in lung cancer diagnosis and prognosis.** *Cancer Cell* 2006, **9(3)**:189-198.
68. Trompeter HI, Abbad H, Iwaniuk KM, Hafner M, Renwick N, Tuschl T, Schira J, Müller HW, Wernet P: **MicroRNAs MiR-17, MiR-20a, and MiR-106b act in concert to modulate E2F activity on cell cycle arrest during neuronal lineage differentiation of USSC.** *PLoS One* 2011, **6(1)**:e16138.
69. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, Takahashi T: **A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation.** *Cancer Res* 2005, **65(21)**:9628-9632.
70. Miko E, Czimmerer Z, Csánky E, Boros G, Buslig J, Dezso B, Scholtz B: **Differentially expressed microRNAs in small cell lung cancer.** *Exp Lung Res* 2009, **35(8)**:646-664.
71. Raponi M, Dossey L, Jatkoe T, Wu X, Chen G, Fan H, Beer DG: **MicroRNA classifiers for predicting prognosis of squamous cell lung cancer.** *Cancer Res* 2009, **69(14)**:5776-5783.
72. Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, Cheng CL, Yu CJ, Lee YC, Chen HS, Su TJ, Chiang CC, Li HN, Hong QS, Su HY, Chen CC, Chen WJ, Liu CC, Chan WK, Chen WJ, Li KC, Chen JJ, Yang PC: **MicroRNA signature predicts survival and relapse in lung cancer.** *Cancer Cell* 2008, **13(1)**:48-57.
73. Dacic S, Kelly L, Shuai Y, Nikiforova MN: **miRNA expression profiling of lung adenocarcinomas: correlation with mutational status.** *Mod Pathol* 2010, **23(12)**:1577-1582.
74. Girardot M, Pecquet C, Boukour S, Knoops L, Ferrant A, Vainchenker W, Giraudier S, Constantinescu SN: **miR-28 is a thrombopoietin receptor targeting microRNA detected in a fraction of myeloproliferative neoplasm patient platelets.** *Blood* 2010, **116(3)**:437-445.
75. Kozaki K, Imoto I, Mogi S, Omura K, Inazawa J: **Exploration of tumor-suppressive microRNAs silenced by DNA hypermethylation in oral cancer.** *Cancer Res* 2008, **68(7)**:2094-2105.

doi:10.1186/1471-2164-14-S6-S5
Cite this article as: Barh *et al.*: A novel *in silico* reverse-transcriptomics-based identification and blood-based validation of a panel of sub-type specific biomarkers in lung cancer. *BMC Genomics* 2013 **14**(Suppl 6):S5.

IV.1.1 Conclusions from this research/ Chapter-1

i. We have developed an integrated *in silico* reverse-transcriptomics approach for sub-type specific biomarker identification in complex human diseases.
ii. The strategy is potential for identification of gene/protein biomarkers using miRNA expression profile.
iii. We applied the strategy in Lung cancer and several potential biomarkers are identified among which 7 markers (4-NSCLC and 3-common event in lung cancer) are validated.
iv. Upregulation of TFPD1, E2F6, IRF1, and HMGA1 + NO expression of SUV39H1, RBL1, and HNRPD in blood sample are characteristics of Adeno and Squamous cell lung carcinomas.
v. E2F6 is a novel/newly identified marker for lung cancer.
vi. miRNA-marker-miRNA interaction can give novel insight of the disease pathogenesis, therefore can be useful in developing disease management and therapeutic strategies.
vii. A modified strategy can be useful in early marker identification and in developing personalized medicine.
viii. The strategy can be used irrespective of any disease and sub-type specific markers may be identified.

IV.1.2 Media highlights of this research outcomes/ Chapter-1

Medical News Today (MTN)

Simple blood test for sub-type specific lung cancer diagnosis

Published: Tuesday 12 November 2013

Adapted Media Release

f 5 SHARE

Lung cancer is the leading cause of cancer specific death worldwide. 85% lung cancers are non-small cell lung cancers (NSCLC) while remaining are small cell lung cancers (SCLC). Although X-Ray and CT scan remain the main non-invasive diagnosis strategy, the painful lung tissue biopsy is essential for confirmation and staging of the disease.

An international team of researchers from India, UK, Spain, Brazil, and the USA led by Debmalya Barh from the Institute of Integrative Omics and Applied Biotechnology (IIOAB) in Nonakuri, Tamluk, Purba Medinipur, West Bengal, India have identified blood based novel diagnostic markers those can be useful in early detecting the lung cancer and even their sub-typing i.e. whether its SCLC or NSCLC.

The researchers have used a novel reverse-transcriptomics strategy to identify the biomarkers. The findings suggest that in a simple PCR based test if HMGA1, E2F6, IRF1, and TFDP1 are expressed and SUV39H1, RBL1, and HNRPD do not express in blood, the patient is having lung adenocarcinoma and/or squamous cell carcinoma sub-types of NSCLC.

Dr. Elena Padin-Iruegas form Medical Oncology Department, Complejo Hospitalario Universitario, Spain, who is associated with this research, says that "the identified markers can be made gold standard" and according Debmalya Barh who led the research "The novel reverse-transcriptomics-based integrated approach developed in this work can be applicable to identify early biomarkers not only for the lung cancer but can be applicable to other cancers and diseases too".

The study has been recently published in *BMC Genomics* 14 (Suppl 6), S5, 25 Oct, 2013.

http://www.medicalnewstoday.com/releases/268661.php

About Medical News Today *(www.medicalnewstoday.com): MTN is the healthcare internet publishing market leader for medical news. It is in the top 360 United States sites and top 120 United Kingdom sites and receives more than 12 million monthly visits, 10 million monthly unique visitors and 15 million monthly page views as reported by Quantcast. It contents are based on evidence-based, peer-reviewed studies, along with accurate, unbiased and informative content from governmental organisations (e.g. FDA, CDC, NIH, NHS), medical societies, royal colleges, professional associations, patients' groups, pharmaceutical and biotech companies, among others and targeted to an educated audience of both healthcare professionals and patients alike.*

Biomarkers for sub-type specific lung cancer identified

ENARADA, Bangalore, October 28, 2013:

More than 25% cancer related deaths are associated with Lung cancer globally. The high mortality is due lack of poor treatment regime and reliable non-invasive or minimal invasive early diagnostic markers. An international team of researchers from India, the UK, Spain, Brazil, and the USA led by Debmalya Barh from the Institute of Integrative Omics and Applied Biotechnology (IIOAB) in Nonakuri, Tamluk, Purba Medinipur, West Bengal, India has identified blood based novel diagnostic markers which can be useful in early detecting the lung cancer and even the subtypes such as small cell lung cancers (SCLC) and non-small cell lung cancers (NSCLC).

Using miRNA expression profile and a novel in silico reverse-transcriptomics based integrated bioinformatics approach, the researchers have first identified nine transcription factors (TFs) unique to SCLC, nine TFs common to SCLC and NSCLC, nine TFs common to general, SCLC, and NSCLC, fourteen TFs common to NSCLC and general, four TFs unique to NSCLC, and three TFs common to all types.

The researchers validated seven NSCLC specific TFs using stage-II and IV NSCLC patients' tissue microarray and blood samples based qPCR and concluded that "overexpression of HMGA1, E2F6, IRF1, and TFDP1 and downregulation or no expression of SUV39H1, RBL1, and HNRPD in blood is suitable for diagnosis of lung adenocarcinoma and squamous cell carcinoma sub-types of NSCLC".

"The novel reverse-transcriptomics-based integrated bioinformatics approach developed in this work can be applicable to identify early biomarkers not only for the lung cancer but can be applicable to other cancers and diseases also" says Debmalya Barh.

According to Dr. Elena Padin-Iruegas form Medical Oncology Department, Complejo Hospitalario Universitario, Santiago de Compostela, A Coruña, Spain, who is associated with this research, "we have just tested seven markers out of 48 and they have showed great potential. The other identified markers are under validation and our preliminary results are very satisfactory in early as-well-as sub-type specific detection of lung cancer".

http://enarada.com/biomarkers-for-sub-type-specific-lung-cancer-identified/

IV.2 Chapter II: Research Article

miRegulome: a knowledge-base of miRNA regulomics and analysis

Debmalya Barh, Bhanu Kamapantula, Neha Jain, Joseph Nalluri, Antaripa Bhattacharya, Lucky Juneja, Neha Barve, Sandeep Tiwari, Anderson Miyoshi, **Vasco Azevedo**, Kenneth Blum, Anil Kumar, Artur Silva, Preetam Ghosh

Scientific Reports. 2015 Aug 5; 5:12832. doi: 10.1038/srep12832
[PMID: 26243198]
Impact Factor: 5.2 (2015)

miRNAs targets mRNAs and regulate signaling pathways, biological processes, and patho-physiologies. Therefore, comprehensive understanding of miRNA regulatory networks is essential to develop miRNA based diagnostic and therapeutic strategies. This chapter describes about miRegulome, a novel knowledgebase that provides the entire regulatory modules of miRNAs such as upstream regulators, downstream targets, miRNA regulated pathways, functions, diseases etc. based on validated data manually curated from published literature. The basic, advanced / complex search, and intuitive schematic visualization interface for visualization of miRNA regulome provides a single window for a wide range of data exploration. Four novel analysis tools (Chemical-disease analysis, miRNA-disease analysis, Gene–disease analysis, and Disease-chemical/miRNA analysis) are plugged into miRegulome for exploration of new and novel biological events and for discovery of biomarkers and therapeutics with high precision. miRegulome is available at: http://bnet.egr.vcu.edu/miRegulome

SCIENTIFIC REPORTS

OPEN

miRegulome: a knowledge-base of miRNA regulomics and analysis

Debmalya Barh[1], Bhanu Kamapantula[2,*], Neha Jain[1,3,*], Joseph Nalluri[2,*], Antaripa Bhattacharya[1], Lucky Juneja[1,3,*], Neha Barve[1,3], Sandeep Tiwari[1,4], Anderson Miyoshi[4], Vasco Azevedo[4], Kenneth Blum[1,5], Anil Kumar[3], Artur Silva[6] & Preetam Ghosh[2]

Received: 08 September 2014
Accepted: 06 July 2015
Published: 05 August 2015

miRNAs regulate post transcriptional gene expression by targeting multiple mRNAs and hence can modulate multiple signalling pathways, biological processes, and patho-physiologies. Therefore, understanding of miRNA regulatory networks is essential in order to modulate the functions of a miRNA. The focus of several existing databases is to provide information on specific aspects of miRNA regulation. However, an integrated resource on the miRNA regulome is currently not available to facilitate the exploration and understanding of miRNA regulomics. *miRegulome* attempts to bridge this gap. The current version of *miRegulome* v1.0 provides details on the entire regulatory modules of miRNAs altered in response to chemical treatments and transcription factors, based on validated data manually curated from published literature. Modules of *miRegulome* (upstream regulators, downstream targets, miRNA regulated pathways, functions, diseases, etc) are hyperlinked to an appropriate external resource and are displayed visually to provide a comprehensive understanding. Four analysis tools are incorporated to identify relationships among different modules based on user specified datasets. *miRegulome* and its tools are helpful in understanding the biology of miRNAs and will also facilitate the discovery of biomarkers and therapeutics. With added features in upcoming releases, *miRegulome* will be an essential resource to the scientific community. Availability: http://bnet.egr.vcu.edu/miRegulome.

microRNAs (miRNAs) are small non-coding RNAs that inhibit post-transcriptional gene expression by complementary base pairing at the 3′-UTRs of target messenger RNAs (mRNAs)[1]. Transcription of a miRNA coding gene is under direct control of transcription factors (TFs). Expression of a miRNA is also regulated by environmental factors, xenobiotics, and drugs. These factors essentially regulate TFs and consequently regulate transcription of miRNAs[2]. A TF can positively or negatively regulate miRNA transcription. A transcribed miRNA, by virtue of its feed-back and feed-forward loop regulation mechanisms, may regulate its own transcription machinery or the expression of other genes and thereby impacts gene expression significantly[3]. A single miRNA may target nearly 200 mRNAs[4] and henceforth may regulate multiple signalling pathways and various essential biological processes (BPs) such as development[5], aging[6], immunity, and autoimmunity[7], etc.

Deregulations of miRNAs are well documented for their association with various patho-physiological conditions including different types of cancers[8], metabolic disorders[9], and neuronal diseases[10] among

[1]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB-721172, India. [2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA-23284, USA. [3]School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India. [4]Laboratorio de Genetica Celular eMolecular, Departmento de Biologia Geral, Instituto de Ciencias Biologics, Universidade Federal de Minas Gerais CP 486, CEP 31270-901 Belo Horizonte, Minas Gerais, Brazil. [5]Department of Psychiatry and McKnight Brain Institute, University of Florida, College of Medicine, Gainesville, Florida, USA. [6]Instituto de Ciências Biológicas, Universidade Federal do Pará, Rua Augusto Corrêa, 01 - Guamá, Belém, PA, Brazil. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to D.B. (email: dr.barh@gmail.com)

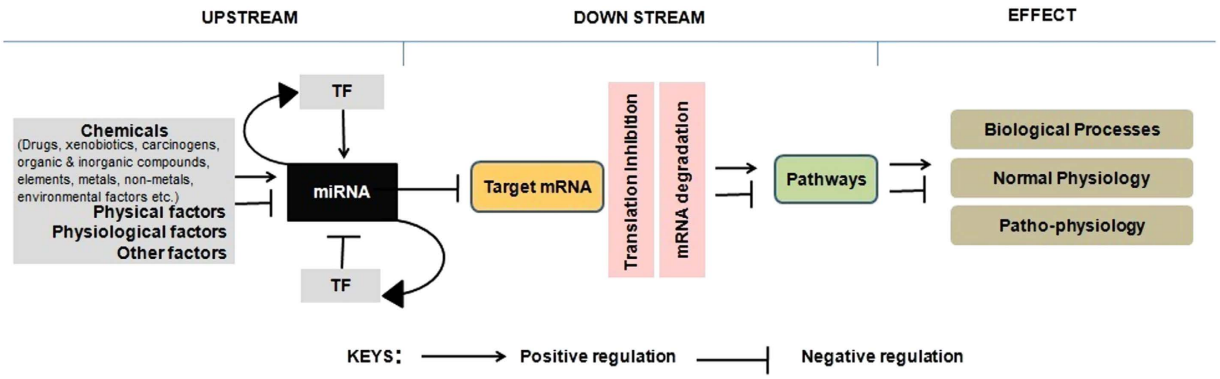


Figure 1. Schematic diagram of modules and their inter-relationships in a miRNA regulome.

others. Therefore, understanding miRNA regulation is highly important in bio-medical research. Exploration of the entire regulome of a miRNA is indispensable in understanding its biology and mechanisms through which it regulates gene expression under a given biological condition. This will also help in the development of diagnostic, prognostic, and therapeutic strategies[11–13]. The regulome of a miRNA essentially consists of modules such as upstream regulators, downstream targets, miRNA modulated pathways, and regulated BPs. The regulome also considers associated diseases when a miRNA is deregulated. A schematic of our proposed miRNA regulome is presented in Fig. 1.

Several existing miRNA-related databases individually provide information on specific aspects of a miRNA. For example, miRbase[14] maintains data on sequence repositories, mir2Disease[15] provides miRNA-disease relationships, TransmiR[16] maintains information on miRNAs and their upstream TFs, and miREnvironment[17] offers information on miRNA regulation in response to environmental factors. Various databases such as miRecords[18], miRWalk[19], mirDIP[20], miRTarBase[21], etc, have been developed to enlist predicted and experimentally validated targets of miRNAs. However, none of these databases provide the entire regulome of a miRNA or are helpful in understanding the biology or function of miRNAs by analysing a stand-alone database. Attempts have been made to understand the miRNA interactome at a systems level in *C. elegans* (TF-miRNA-TF interactions)[22] through computational simulation of miRNA regulated overall gene expression and cross-talk between miRNA targets[23] and by constructing regulatory models of miRNA-kinase-TF, miRNA-TF, and TF-TF[24]. Similarly, systems approaches for predicting TF-miRNA crosstalk in human protein interactome[25], demonstration of regulatory principles among miRNAs, TFs, and miRNA target genes[26], miRNA-mRNA and miRNA-miRNA interactions[27], and tissue-specific miRNA-TF regulatory networks[28] have also been attempted to explore the miRNA interactome. However, these works are mostly computational predictions and do not provide the entire regulome of a miRNA.

The web-based resource, *miReg*[29] provides basic correlations of various upstream regulators, downstream targets, BPs, and diseases of a miRNA based on experimentally validated data available in the PubMed literature. However, it is primitive in terms of data, completeness, functionality, and usability. Given the importance of miRNA in biomedical research, disease diagnosis, prognosis, and therapy, the huge inflow of new miRNA related information becomes a challenge. Therefore, a novel database that provides all the essential details of a miRNA regulome is necessary. Similarly, a state-of-the-art analysis platform to explore mechanisms behind various biological and patho-physiological processes that a miRNA regulates is also required.

miRegulome aims to address the need for such a novel database that represents the entirety of the miRNA regulome. In the current version of *miRegulome* (v1.0) we have incorporated all the downstream modules and TFs as well as the diverse group of chemicals as the upstream regulatory modules and their correlations. Several other aspects associated with the miRNA regulatory networks are also included (Fig. 1). The analysis tools for the current version of the miRNA regulome provide ranked association counts with Z-score statistical assessment for likely functions and disease associations of a set of input miRNAs. The list of resultant associated functions, diseases, and processes reported by our tools is similar to that provided by another tool that performs statistical enrichment analysis, TAM (http://210.73.221.6/tam).

Construction and Contents

In *miRegulome* v1.0, we have incorporated experimentally validated data for all the downstream modules (targets, modulated pathways, regulated BPs, and associated diseases) and chemical and TFs as the upstream modules of a miRNA regulome. The data are manually curated from published literature indexed in PubMed. In the current version of *miRegulome*, physical, physiological, and mechanical upstream factors are not included. Also, this version contains most of the modules of miRNAs for human, mouse, and rat. But for other species' miRNAs, mostly upstream modules (chemical or TF) are

incorporated since the validated downstream modules of these miRNAs are not yet available in published literature. The step-by-step data collection process and their sources is represented in Supplementary Fig. S1 and the home page of *miRegulome* v1.0 is shown in Supplementary Fig. S2.

Capturing miRNA and miRegulome modules

miRNAs and upstream chemical regulators. In *miRegulome* v1.0, first we have focused on miRNAs that are either up- or down-regulated in response to a chemical (drugs, xenobiotics, carcinogens, organic and inorganic compounds, elements, metal, non-metals, and environmental factors etc.). PubMed literature database was extensively searched manually with key word combinations (for example: chemical + miR/miRNA/microRNA + regulation) to identify publications describing experimentally validated chemical-miRNA relationships. Each selected article having chemical-miRNA relationships was then manually curated to capture the (i) chemical(s) (ii) miRNAs responding to the chemical(s), (iii) species of the miRNAs, (iv) expression of the miRNAs (up-/down-regulation) in response to the chemical(s), (v) experimental conditions, (vi) techniques used to detect the expression levels of miRNAs, and (vii) the corresponding PubMed ID (Supplementary Fig. S3).

Upstream TF regulators and downstream targets. Apart from the upstream chemical regulators, TFs that regulate the transcription of a miRNA gene are one of the main classes of upstream regulators of the miRNA. Since miRNA's basic function is to target specific mRNAs, such targets of miRNAs are the most important components in a miRNA regulome. In *miRegulome* v1.0, experimentally validated upstream TF regulators and the downstream target genes/mRNAs of each miRNA that have upstream chemical regulators, are manually curated from the PubMed indexed literature and incorporated into the database. Extensive manual search was carried out using specific key words such as target, regulator, transcription factor, "upregulates", and "downregulates" along with the name of the miRNA having upstream chemical regulators (identified in the previous step/search) to capture these relationships (Supplementary Fig. S4A,B).

Prioritized targets and miRNA functions. The availability of prioritized targets and target-based top functionalities of a miRNA are unique features of *miRegulome*. We manually performed the target prioritization based on the number of interactions of a target in a protein-protein interaction network using the ToppNet algorithm of ToppGene suite[30]. Since miRNA regulates the expression of a gene, we used 11 house keeping genes as described by Eisenberg and Levanon[31] in the training set and all experimentally validated targets of each miRNA (present in *miRegulome*) as the test set to perform the ToppNet analysis with its default parameters. Ranking or prioritized targets list were prepared and included in the *miRegulome* database based on the interaction counts of a target (higher count for higher rank). To assign the top functionalities of a miRNA, all the targets of a miRNA were subjected to ToppFun prediction analysis (with the default parameters) of the ToppGene suit[30]. In parallel, all the targets of that particular miRNA were analyzed using "Functional Annotation" module of DAVID[32] with its default p-value cut off = 0.1. The first 25 common top ranked predicted functions (BPs) derived from both the tools are listed under "Function of miRNA" module of *miRegulome* (Supplementary Fig. S4C,D).

miRNA involved pathways. The second unique feature of *miRegulome* consists of linking a miRNA to the pathways, it regulates. In order to do so, all the validated targets of each miRNA were subjected to DAVID[32] for enrichment into Kyoto Encyclopedia of Genes and Genomes (KEGG)[33] pathways. The ten enriched pathways are listed under "Pathways" tab in the second table with corresponding KEGG links. In order to give a broader picture of the pathways in which the miRNA is involved, a "More" option is given at the end of the pathways list, each of which is hyperlinked to the corresponding miRNA pathways listed in the miRNAPath database[34] (Supplementary Fig. S4E).

Disease module. The ultimate goal of miRNA research is to explore how a miRNA is associated with a patho-physiological process. Therefore, miRNA-disease association is an essential feature in a miRNA regulome. miRNA-disease relationships along with regulation of the miRNA (up- and down-regulation) in the disease condition were manually curated from PubMed listed published literature for those miRNAs that respond to chemical stimulus and were incorporated in *miRegulome* v1.0 under "Disease involvement" module. Similar to the "Pathways" module, a "More" option is given which is hyperlinked to that miRNA with all its associated diseases listed in the miR2Disease database[15] (Supplementary Fig. S4F).

Quality control. To ensure the accuracy and consistency of the data before recording it to the *miRegulome* database, quality control (QC) checks were performed thrice. Three individual team members manually cross-checked the curated information such as validated TF regulators, chemicals, targets, conditions, techniques, diseases etc. and their associations with the miRNA using the particular PubMed publication. For predicted data (prioritized targets, miRNA top functions, pathways etc.), the cross-checks were carried out by three different team members using the same tools and their fixed parameters. Inconsistency in the data, if observed during QC checks was manually corrected and incorporated into the database.

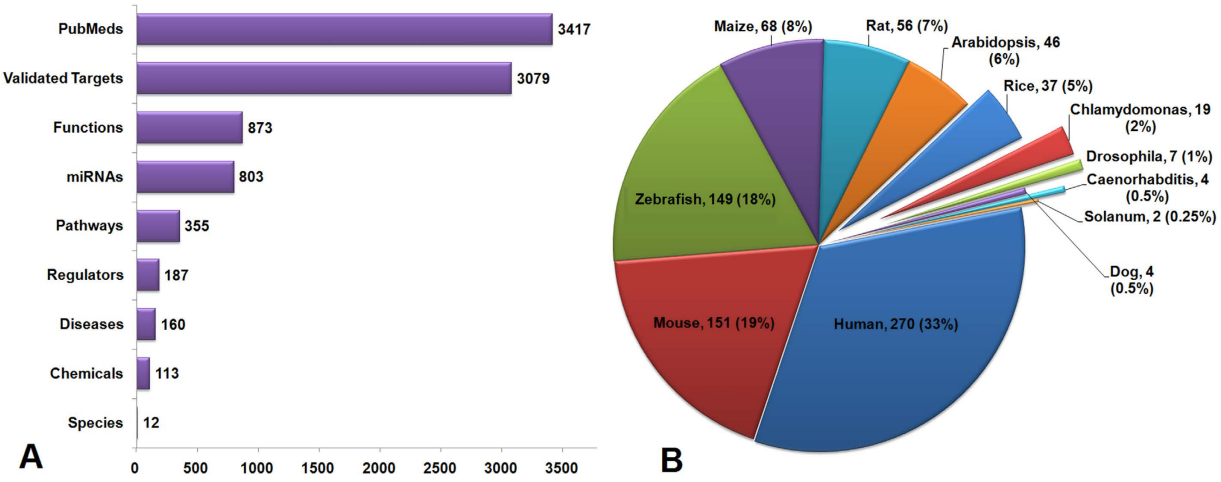


Figure 2. Distribution of *miRegulome* v1.0 contents. (**A**) Overall distribution of the database contents. (**B**) Species specific miRNA counts.

Database contents. *miRegulome* v1.0 contains experimentally validated information for 803 miRNAs from 12 species, 113 chemicals, 187 upstream TF regulators, 3079 targets, and 160 diseases manually curated from 3417 PubMed indexed articles. Predicted 873 functions and 355 pathways are currently available in this database. The distribution of these data is represented in Fig. 2. We aim to update the database with manually curated new data every six months through partial automation.

Database design. *miRegulome* v1.0's clean design helps users interact with the database in various ways. The regulome data has many-to-many relationships with its entities and this has been taken into consideration while designing the database. The database has been designed keeping the miRNA data as the central entity and all other database tables containing information about chemicals, functions, diseases, and genes are linked to the miRNA entry. Based on the current design, numerous specific combinations and associations among miRNAs, genes, conditions, techniques, diseases, and other related entities can be retrieved. This also allows retrieval of diverse nature of aggregated results based on dissimilar types of inputs the user may provide. The current design is beneficial in both - future approaches for analysis and incorporation of other tools and data into the existing system. The database design is presented in Supplementary Fig. S5.

Visualization of miRNA regulome. In order to represent the regulome of a miRNA, we developed an intuitive schematic visualization interface. Upon selection of a miRNA (under "miRNA Details" tab), the entire regulome of the selected miRNA will be visualized below the first table (Fig. 3). The schematic visualization provides the entire regulome of the miRNA with all its modules (chemicals, upstream activators and repressors, validated targets, enriched top targets, pathways, functions, and diseases) and their relationships with the miRNA. The relative impacts (activation, inhibition or association) of these modules on the miRNA are also graphically represented. JavaScript, HTML, and CSS are used to develop this complex interaction map in an intuitive way that is easy to interpret. This component is in addition to displaying the miRNA regulome information in a tabular format for better understanding of the miRNA regulome.

Search options and description. Users can search the database in two different ways. The option of "Search by Chemical" can be used to retrieve miRNAs that are associated with a specific chemical. Upon selection of a chemical that is listed in alphabetical order in a dropdown menu, a table representing the miRNA-chemical relationships along with the species of the miRNAs, experimental conditions, effect of the chemical on the miRNAs, techniques used to detect the effects, and the corresponding PubMed ID etc. can be obtained. The user will be provided with the detailed regulome of the particular miRNA in a second table by clicking on any miRNA name under the last column ("miRNA Details" tab) of the first table. This table contains the detailed regulome where each tab is having a specific module and relationships related to the miRNA. The modules are "Upstream Regulators", "Validated Targets", "Prioritized/Top Targets", "Function of miRNA", "Pathways", and "Disease Involvement". To provide detailed information, each module is hyperlinked with suitable external resources. For instance, miRNAs are linked to miRBase[14], chemicals to the Comparative Toxicogenomics Database (CTD)[35], genes to Entrez (http://www.ncbi.nlm.nih.gov/gene), functions/BPs to EBI-GO (http://www.ebi.ac.uk/GOA), pathways to KEGG[33] and miRNAPath[34], and diseases are linked to miR2Disease database[15] etc.

Using the "Advanced Search" option (Supplementary Fig. S6), the user can search information on all the twelve modules of the *miRegulome* v1.0 such as miRNA, regulators, targets, diseases etc. Importantly,

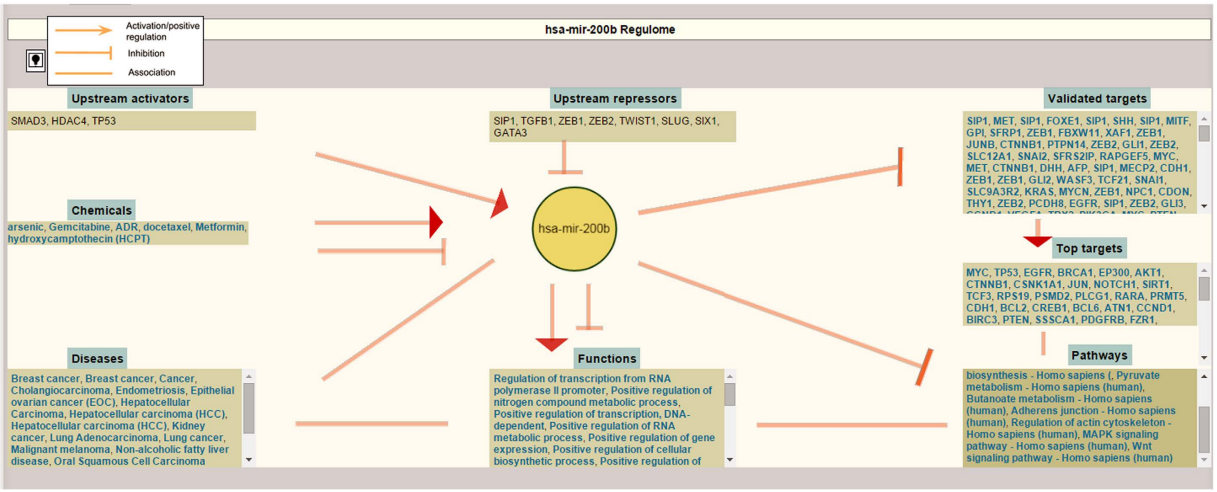


Figure 3. Visualization of miRNA regulome in *miRegulome* v1.0. The figure represents hsa-miR-200b regulome.

the user can obtain specific relationships as per their requirements using a combination of query modules instantly (see the example under section "*Exploration of new biological events*").

Analysis tools. *miRegulome* v1.0 is empowered with four unique tools to provide meaningful associations among chemical-disease, miRNA-disease, gene-disease, and disease-chemical-miRNA along with affected BPs based on user specific datasets. The results of these analyses correspond with a bipartite modelling approach which we developed to explore the associations among miRNAs and diseases available in the database. Maximum weighted matching algorithm is used to identify these associations[36]. In *miRegulome* v1.0, each association whether its chemical-miRNA, miRNA-disease, gene-miRNA etc. is manually curated from PubMed indexed literature and each of these relationships is tagged with specific PubMed ID from where the data are taken. These tools mine the database and give relevant associations to the user by querying the *miRegulome* database and counting the associations (direct or indirect) between entries. The output is returned as ranked association counts with Z-score statistical analysis rather than statistical enrichment measures. However, the results derived from these tools are quite similar to the ranked list of results returned by tools that use statistical enrichment and probability calculations (see the "Efficacy of *miRegulome* v1.0 tools" section). We calculated the Z-scores using the formula: $Z\text{- }score = (X - \mu)/\sigma$, where 'X' is the association count of the particular association i.e. the number of PMIDs citing the association, 'μ' is the mean of the association counts for the entire association type, and 'σ' is the standard deviation. Hence, a positive Z-score indicates that the count of the association is higher than the average of such associations, and a negative value indicates that it is below the average. A value of 0 would mean it is equal to the average.

Chemical-disease analysis. This analysis tool allows the user to explore the associations between a chemical to a disease via miRNAs. When a user selects a particular chemical, the tool retrieves all the miRNAs associated with the chemical. Thereafter, the tool retrieves all the diseases in which the miRNAs are associated. For example, if a user selects chemical '*C2*' in the *Chemical-Disease* analysis tool, miRNAs *M1* and *M3* are retrieved and subsequently, their associated diseases *D1*, *D2* and *D3* are retrieved. Finally, the tool ranks the diseases in which these miRNAs (which are associated with the chemical) are associated, counting the PubMed IDs (Supplementary Fig. S7A,B). The tool then displays the disease names, their association counts and their respective Z-scores for the counts. Using similar methodology, the BPs which are associated with the miRNAs are displayed according to the count of their associations as recorded in the database. It does not assert a direct link between the chemical to a disease or to the BPs via the miRNA, rather allows the user to explore and test their hypothesis for indirect associations between the chemical and the disease via the miRNA.

miRNA-disease analysis. In this analysis, when an input of one or more miRNAs is provided, the tool provides three tables for the user to get a comprehensive understanding of their results. The tool searches for all diseases associated with the provided miRNA(s) and the distinct miRNA-disease associations (based on PubMed IDs) (Supplementary Fig. S7B). Following which, it ranks the diseases based on their number of recorded (PubMed IDs) associations and displays them in the Table A (Supplementary Fig. S8, Top affected diseases for given miRNAs). The user can click on the '*Count of PMIDs*' and see the unique PubMed IDs supporting the results. The tool also displays the Z-scores for each disease along with its rank. Z-score is a standardized score for the count of each disease, indicating the resultant

disease's location in a distribution of other diseases, in relation to the mean and standard deviation of miRNA-disease counts. The Z-score of the disease tells the user, how many standard deviations it is from the mean of all miRNA-disease count distribution present in the database. To calculate this Z-score, we first calculated all the individual input miRNA and disease association counts and converted them to their respective Z-scores in Table C (Supplementary Fig. S8, Z-score for miRNA-disease associations). Thereafter, we added all the Z-scores associated to a single disease, thereby giving us the fair cumulative impact of all the input miRNAs with the single disease. This final cumulative Z-score is displayed for each disease. To further understand the relative impact of each miRNA (entered by the user) to the disease, we display Table C with each miRNA-disease edge with a Z-score. This value gives the user the individual miRNA-disease strength of association. It also displays which among the input miRNAs has the highest/lowest impact on the disease, thereby giving a more in-depth insight into the results displayed in Table A. Moreover, the '*Count of PMIDs*' gives the cumulative count of PubMed IDs citing the associations of the input miRNAs with the disease. However, it does not take into account the impact of each miRNA-disease association towards the count. For e.g. a certain disease '*D*' has '*Count of PMIDs*' as 15 with miRNAs *M1, M2, and M3* associated with it. Among them, the the count of PubMed IDs for *M1-D1, M2-D1* and *M3-D1* are 12, 2, and 1 respectively. Evidently, here M1-D1 will have a higher positive Z-score because of its high count (assuming the *mean* is 5) and M3-D1 will have a negative Z-score. Nevertheless, when we add these individual Z-scores together, we get a fair relative scoring of the disease with respect to the input miRNAs, capturing the cumulative effect of the group of miRNAs. This is especially helpful, when the '*Count of PMIDs*' for certain diseases are the same. In such cases, Z-score analysis would give a fine-grained ranking of the results. Table C explains the values obtained in Table A. This tool also searches the *miRegulome* database and identifies the most frequent BPs which are associated with the specified miRNA(s) and then displays them in Table B (Supplementary Fig. S8, Biological processes), ordered by rank, based on the count of associations present in this database.

Gene–disease analysis. When a list of genes is entered by the user in the input field, the tool searches for miRNAs associated with the set of genes and counts the number of gene-miRNA associations (i.e. PubMed IDs) recorded in the database. Thereafter, the tool searches and counts the existing relationships (i.e. PubMed IDs) between the observed miRNAs and diseases. Following which, the tool ranks the diseases based on their count of PubMed entries (Supplementary Fig. S9). Similarly, the tool also displays the list of BPs which are associated with the specified set of genes via miRNAs, and ranks them following the same principle of the miRNA-disease analysis tool. The tool does not assert a relationship between the entered genes and diseases but highlights the top diseases indirectly associated with the genes entered, via the miRNAs.

Disease-chemical/miRNA analysis. This tool works in the opposite way of the *Chemical – disease* analysis tool. It takes disease(s) as an input and searches the repository for specific miRNA(s)-disease associations. Thereafter, it retrieves the chemicals associated with the miRNAs. The tool displays these associations and ranks them based on the number of occurrences in the database (i.e. PubMed IDs). This gives the user an insight into possible role of chemicals in regulating miRNAs which are deregulated in the input disease(s).

Utility and Discussion

A single window for a wide range of data exploration. Information on validated upstream TF regulators can be obtained from TransmiR[16], validated targets from miRWalk[19], environmental factors acting on miRNAs from miREnvironment[17], effects of small molecules on miRNA expression from SM2miR[37], miRNA regulating pathways from miRNAPath[34], and disease related miRNAs from miR2Disease[15] databases. However, none of these databases provide additional information other than their respective specifics. Therefore, they are inadequate in terms of providing a comprehensive understanding of the miRNA regulome.

miRegulome is the first-of-its-kind integrated resource of miRNA regulomics having most of the modules of a miRNA regulome. Using *miRegulome* v1.0, in a single platform, the user can get almost all the information that is maintained by these databases along with several unique features such as prioritized targets and target based functional annotations of miRNAs among others. Therefore, it can be used in multiple ways to suit user needs.

Modules are hyperlinked to respective data resources so that if users are interested to explore additional information, it will ensure the comprehensive understanding of the data. Since, each miRNA regulating chemical is linked to the corresponding CTD webpage[35]; user can easily obtain the basic chemistry of the chemical and the details of gene interactions, associated diseases, other chemicals having comparable sets of interacting genes, BPs, pathways, etc. regulated by the chemical from the CTD (Supplementary Fig. S3). CTD, so far does not contain miRNA information for any chemical listed in the database. Therefore, the miRNA information of *miRegulome* for a chemical will be complementary to CTD thereby adding the entire range of regulatory network of the chemical. Similarly, basic information of a miRNA can be obtained by clicking on the name of the miRNA that is hyperlinked to miRBase[14] and miRBase contains several useful information and links for the miRNA including nomenclature, basic annotation, stem-loop and mature sequences, locus report, Entrez, and HUGO Gene Nomenclature

Committee (HGNC) (www.genenames.org) etc. From Entrez and HGNC, the user can get most of the resources which include associated published literature and even the clinically significant information on the miRNA. Since the targets and upstream TF regulators are hyperlinked to Entrez, user will also be able to get detailed information on these modules. Similarly, functions of miRNAs are hyperlinked to EBI-GO (http://www.ebi.ac.uk/GOA), pathways to KEGG[33] and miRNAPath[35], disease to miR2Disease database[15] etc. (Supplementary Fig. S4). Further, additional miRNA resources and tools have also been listed under the "Resources" page of *miRegulome* (Supplementary Fig. S2). Therefore, using *miRegulome*, users can explore most of the information and analysis related to a miRNA.

Visualization of regulome and data interpretation. To simplify the understanding of a miRNA regulome, *miRegulome* v1.0 is integrated with an intuitive and effective schematic visualization tool. The complex interactions and relationships of a miRNA with its various modules can be visualized, thereby providing a cursory overview of the miRNA biology. This visual schematic is displayed when the user clicks on a miRNA under "miRNA Details" available in the first table. Fig. 3 represents visualization of hsa-miR-200b. The visualization gives a glimpse of seven modules (chemicals, upstream activators and repressors, validated targets, prioritized targets, biological pathways, BPs/functions, and disease associations) of the miRNA and the manner in which they regulate the miRNA or are being regulated by the miRNA with the relative impacts such as activation or inhibition or association. Further analysis of the visualization may provide deeper understanding of the mechanism and the impact of each module component on the hsa-miR-200b, and hence the biology and patho-physiological significance of the miRNA. Considering an example of cancer, from Fig. 3 and tabular forms of the regulome description, it can be observed that hsa-miR-200b: (i) forms a feed-back loop with TGFB1, (ii) P53 activates and TGFB1 inhibits its expression, (iii) inhibits MAPK, WNT, and AKT signalling pathways, (iv) inhibits cell cycle and cell proliferation by targeting cell cycle regulators and oncogenes like CCND1, VEGFA, MYC, NOTCH1, MET, EGFR etc. (v) is involved in cancer associated pathways, (vi) is downregulated in response to arsenic carcinogen, (vii) is upregulated by chemotherapy drug Gemcitabine and downregulated in Docetaxel-resistant cancers, and (viii) is downregulated in several cancers (Fig. 4). Therefore, it may be implicated that hsa-miR-200b could be a tumor suppressor miRNA and may be a potential therapeutic for a wide range of cancers.

Exploration of new biological events. Since *miRegulome* gives most of the information associated with a miRNA, the user can explore the molecular mechanism and precise role of a miRNA behind a normal BP or a patho-physiological process. Identification of the particular role of a miRNA may consecutively help in developing miRNA-based diagnostics and therapeutic strategies.

User can search *miRegulome* v1.0 based on chemical using simple "Search by chemical" option or can use the "Advanced search" to get specific information as per the input search combinations. Using the "Search by Chemical", user can get the validated regulations of a list of miRNAs in response to a user defined chemical. Chemical responses to miRNAs for twelve species (Human, Mouse, Rat, Dog, Zebra fish, *Drosophila*, *C. elegans*, *Arabidopsis*, Maize, Rice, *Solanum*, and Chlamydomonas) can be explored using this feature. From the retrieved list of miRNAs, user can get the regulome having upstream regulators (chemicals and TFs), regulated BPs, pathways, and disease involvement of any selected miRNA (currently available for Human, Mouse, Rat) and therefore can analyze any event related to that miRNA. Current version of *miRegulome* does not provide other upstream regulators except chemicals for other species.

In the "Advanced search" option, each module of miRNA regulome along with species, regulations, technique and condition, totalling twelve options have been accommodated. Therefore, single or a combination of input keywords can be used to get highly selective information. For example, for a particular disease, we can easily get a list of all up- and down-regulated miRNAs separately, using a combination of two input fields: disease and regulation.

Similarly, we can explore complex associations among several modules and novel correlations using the "Advanced search" option. For example, it can be found that, hsa-mir-27b is down-regulated and hsa-mir-143 is up-regulated in obesity. Further, from *miRegulome* v1.0, it can also be established that hsa-mir-27b is involved in adipocytokine, insulin, and type-2 diabetes pathways and hsa-mir-143 acts in lipid metabolism pathway. These pathways are important events in obesity and therefore, deregulation of hsa-mir-27b and hsa-mir-143 may affect these pathways and eventually may lead to obesity and diabetes. Further, the database also provides correlation of obesity - mir-27b - Ribavirin and obesity - mir-143 - Benzo[a]pyrene. As per *miRegulome* v1.0, Benzo(a)pyrene and Ribavirin up-regulate mmu-mir-143 and hsa-mir-27b, respectively. It is reported that higher Body Mass Index (BMI) lowers bioavailability of Ribavirin and causes treatment failure in obese HCV patients[38]. On the other hand, Benzo[a]pyrene can induce obesity[39]. In summary, it can therefore be implicated that, (a) Benzo[a]pyrene upregulates mir-143 and affects lipid metabolism to induce obesity and (b) an aberrant expression of mir-27b may play a role in obesity-associated insulin resistance by modulating adipocytokines and Ribavirin resistance in obese patients. Similarly, it also suggests that these two miRNAs interlink obesity with diabetes at a new and deeper molecular level (Fig. 5). Therefore, *miRegulome* v1.0 may play an important role in exploring novel molecular mechanisms behind a disease as well as designing personalized medicine.

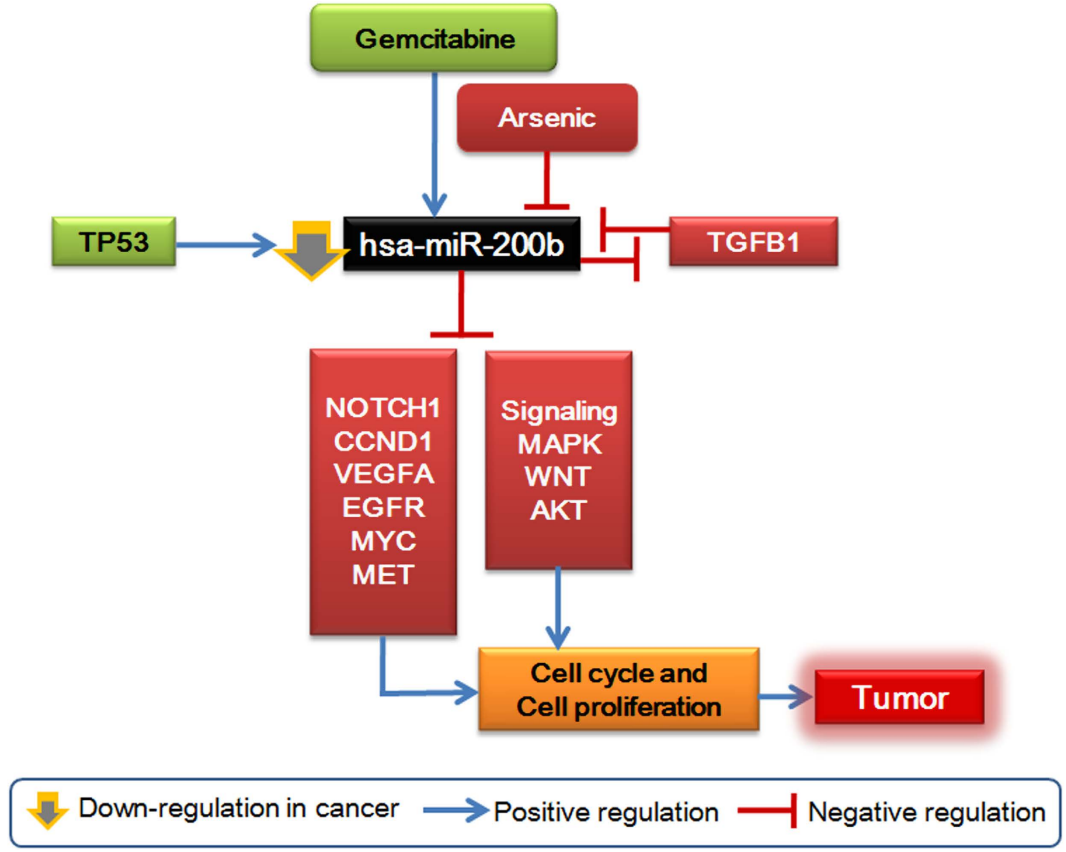


Figure 4. The relationships among various modules of hsa-mir-200b suggest that this miRNA is a probable tumor suppressor.

miRegulome v1.0 tools and analysis. The user-friendly tools of *miRegulome* v1.0 can help the users to understand and test their hypotheses by exploring relationships among miRNA - function - disease, gene - function - disease, chemical - function - disease, and disease associated chemicals and miRNAs.

A user chosen set of miRNAs or genes can be used to understand relevant diseases and BPs associated with the sets using *miRNA-Disease* and *Gene-Disease* tool, respectively. Similarly, single disease or a set of diseases can be used as input to find the common miRNAs, chemicals, and BPs associated with the set or individual disease using *Disease-Chemical/miRNA* tool. In *miRNA-Disease* tool, users can search for species-specific (Human, Mouse, Rat) miRNA analysis using the corresponding prefix (has-, mmu-, rno-) for the miRNA symbols, or can adopt a combined analysis including all three species. For the latter, the user should not use species-specific prefix, instead should use general miRNA symbols such as miR-21 etc. *Disease-Chemical/miRNA* tool can provide the associated miRNAs, chemicals, and affected BPs for a single or a combination of diseases.

Efficacy of miRegulome tools. To demonstrate the efficacy of *miRegulome* v1.0 tools, we used sets of miRNAs or genes that have been reported to show directly up- or down- regulated in a particular disease condition, and have been considered as biomarkers or signatures for the corresponding diseases. Ten such published datasets from ten PubMed literature (that are not incorporated so far in this version of *miRegulome*) were randomly selected and used for the analysis. The set of miRNAs or genes from each publication is fed into the corresponding tool and checked for the output results. Upon analysis, we look for the results if they match with the disease(s) mentioned in the corresponding publication for that particular set of miRNAs or genes. *miRegulome* v1.0 tools rank top 15 diseases associated with single or a set of miRNAs or gene(s) along with a count of association/count of PMIDs along with respective Z-scores based on the strategy as described in the "Analysis Tools" section. Out of the tested ten miRNA and 10 gene sets from twenty different publications, the *miRNA-Disease* and *Gene-Disease* association analysis tools are able to rank the diseases associated with the miRNA and gene sets as per the published literature within the top 15 diseases and mostly under the first five listed diseases.

Out of ten tested miRNA sets from ten different PubMed literature, in 7 cases the tool ranks the same disease associated with the miRNA set (as mentioned in the corresponding literature) within rank 5 and only 3 are between ranks 10 to 14 (Supplementary Table S1). As per the PubMed: 22213236[40], a set of human miRNAs comprising of miR-21, miR-31, mir-122, miR-221, miR-222, miR-145, miR-146a,

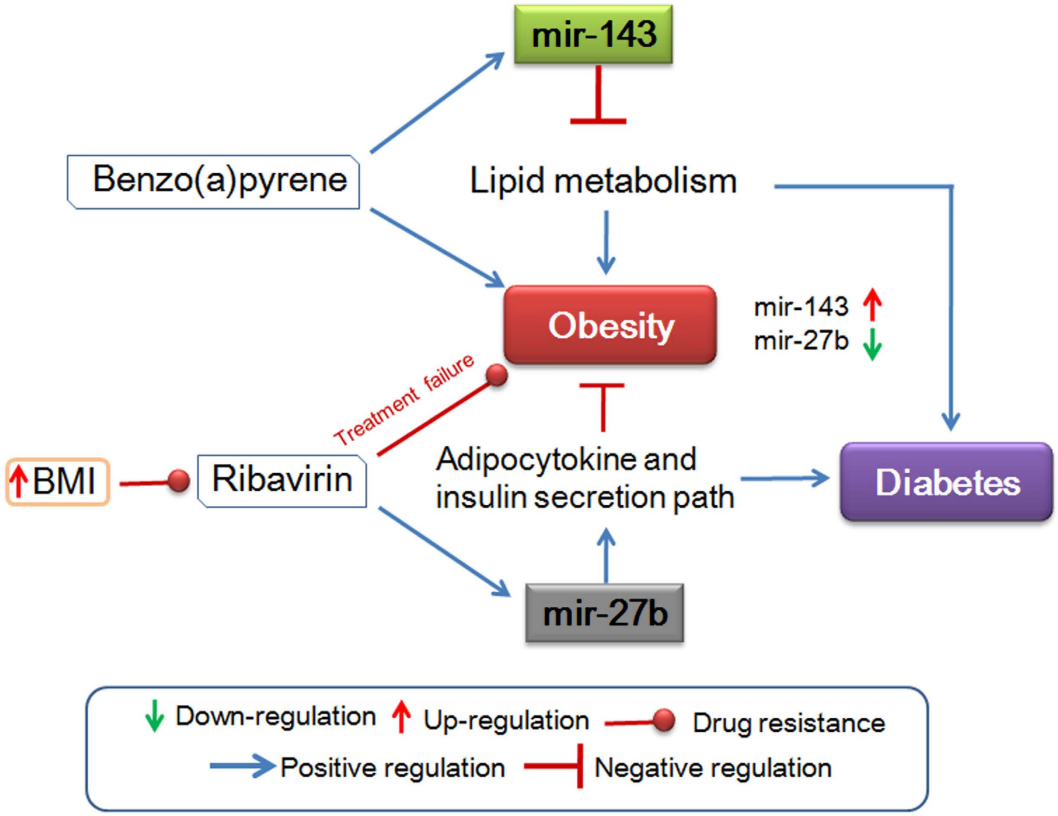


Figure 5. Emerging molecular mechanism and correlation between obesity and diabetes as identified by miRegulome data. For details, please see the text.

miR-200c, and miR-223 are deregulated in hepatocellular carcinomas. When this set of miRNA is used for analysis, the *miRNA-Disease* analysis tool gives hepatocellular carcinoma at rank-3 with an association score/count of PMIDs 50 (Z-score: 37.821). Another set of miRNA miR-1, miR-134, miR-186, miR-208, miR-223, and miR-499 are associated with acute myocardial infarction according to PubMed: 23641832[41]. When this set is used in *miRNA-Disease* analysis tool, the outcome ranks myocardial infarction at position 2 with an association score/count of PMIDs 6 (Z-score: 4.222). Additionally, the associated BPs ranked by the tool also correlate with the diseases; supporting the efficacy of the tool in associating diseases with the input miRNA sets (Supplementary Fig. S8).

The performance of miRegulome tools may be compared to tools that perform statistical enrichment analysis. We compared the results from *miRNA-Disease* analysis tool with the Tool for Annotations of human miRNAs (TAM) (http://210.73.221.6/tam)[42] that calculates the probability of a particular miRNA belonging to a cluster of miRNAs, using its expression values and gives enriched miRNA-associated disease and functions based on *p*-values. *miRegulome* v1.0 does not have expression values or information of cluster of miRNA families and thereby cannot do statistical significance analysis. However, we have used the Z-score statistical assessment of the results which gives a standardized scoring metric and credibility to the miRNA-disease associations. For comparison, we used the same miRNA sets (Supplementary Table S1), that are used to test efficacy of *miRNA-Disease* analysis tool and selected overrepresentation and set version 2 of TAM. As shown in Supplementary Table S2, the ranking of diseases for a particular set of miRNA based on *p*-values by TAM is comparable to the ranking by our *miRNA-Disease* analysis tool; although the absolute ranks differ. This variation could be due to more disease entries used by TAM analysis (as it uses the entire HMDD database[43]).

Supplementary Table S3 shows the results for the *Gene-Disease* association analysis tool. It is observed that, a 29-gene signature for lung cancer (PubMed: 19951989)[44] gives the same cancer at rank-5 with an association score 215. Two genes COL2A1 and ATP4B identified as markers in gastric cancer (PubMed: 23606240)[45] gives colorectal cancer as the rank-1 disease with an association score of 99. For the *Gene-Disease* tool, we observed that, in 50% cases the tool ranks the disease associated with the gene set within rank 5 and 100% within rank 10 (Supplementary Table S3). Similar to the *miRNA-Disease* analysis, the ranked BPs also correlate with the disease.

We have also tested the *Chemical-Disease* and *Disease-Chemical/miRNA* association tools and have received similar precision in results (data not shown). Therefore, these tools are useful in predicting or identifying disease associations or disease associated miRNAs/genes/chemicals for user specified data

sets e.g., (i) miRNA/gene biomarkers for a disease, (ii) disease susceptibility in response to a chemical based on miRNA profile, (iii) miRNA/gene signature of a disease, (iv) affected or regulated BPs by a group of miRNA/gene, and (v) developing therapeutic strategies.

Conclusions and Future development

miRegulome may be considered as the advanced version of *miReg*[29] in terms of data, functionality, usability, and completeness. *miRegulome* aims to provide the complete regulome of any miRNA listed in this database as derived from published literature. The current version of *miRegulome* v1.0 provides the complete regulome for chemically responsive 803 miRNAs from 12 species. However, the miRNAs from Human, Mouse, and Rat currently have most of the downstream modules of their regulome. For other species, as most of these modules are unavailable in published literature, they are partially available here. Once the data for these missing modules are available in the public domain, they will be incorporated in the upcoming versions of *miRegulome*.

Performance of *miRegulome* v1.0 tools depend on the available data present in the database. Since these tools work based on the association counts; the efficacy of their analysis can be further improved if more data are added to the database. For the next release, we aim to add new data and new upstream regulators (such as physical, physiological and mechanical factors) for the existing miRNAs. Similarly, miRNAs from new species will also be included to further enrich the database. New modules such as regulatory relationships between miRNAs and epigenetic modifications, miRNAs and disease related SNPs in their target genes, and variations in miRNA sequence and its associated phenotypes among others will be included. Interactions among diseases can be explored using miRNA-disease relation model of the current data. This can be achieved by representing the miRNA-disease information as a network. Further work would also aim to develop novel ways of interaction with the database and ascertain the nature of associations among different diseases, in which the effect of a certain disease on the occurrence or receding of other diseases can be evaluated. Tools for prediction of prognostic and early diagnostic markers will also be developed. In order to achieve that, high throughput genome-wide expression and association data will be incorporated in the next versions of *miRegulome* and the novel "reverse transcriptomics" approach by Barh *et al.*[46] will be implemented. Similarly, disease specific therapeutic miRNAs and small molecules targeting the disease causing precursor microRNAs also will be included. Further, the miRNAs may be classified based on -3p and -5p and the new version will be developed based on such strand specific miRNA information. We also aim to develop a better visualization and advanced integrated analysis system for next-level of interaction of *miRegulome* modules and interpretation of miRNA functions in various biological and patho-physiological processes and to understand if miRNA-miRNA direct interactions do also exist.

Availability and Requirements

miRegulome v1.0 can be accessed online at http://bnet.egr.vcu.edu/miRegulome and is free for academic research. Commercial use is not permitted. jQuery, JavaScript, and HTML have been used to design the user interface. PHP has been used for server-side scripting support. Google visualization library has been used to represent the data in an interactive form. MySQL v.5.1.63 is used as database to store regulome information. It runs on Apache 2.2.17 on Ubuntu 12.04. The database and its integrated tools can be best used using the latest versions of Google Chrome and Mozilla Firefox browsers.

References

1. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9,** 102–114 (2008).
2. Yokoi, T. & Nakajima, M. Toxicological implications of modulation of gene expression by microRNAs. *Toxicol Sci* **123,** 1–14 (2011).
3. Gurtan, A. M. & Sharp, P. A. The role of miRNAs in regulating gene expression networks. *J Mol Biol* **425,** 3582–3600 (2013).
4. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat Genet* **37,** 495–500 (2005).
5. Hossain, M. M., Salilew-Wondim, D., Schellander, K. & Tesfaye, D. The role of microRNAs in mammalian oocytes and embryos. *Anim Reprod Sci* **134,** 36–44 (2012).
6. Jung, H. J. & Suh, Y. MicroRNA in Aging: From Discovery to Biology. *Curr Genomics* **13,** 548–557 (2012).
7. Zhu, S., Pan, W. & Qian, Y. MicroRNA in immunity and autoimmunity. *J Mol Med (Berl)* **91,** 1039–1050 (2013).
8. Farazi, T. A., Hoell, J. I., Morozov, P. & Tuschl, T. MicroRNAs in human cancer. *Adv Exp Med Biol* **774,** 1–20 (2013).
9. Fernandez-Hernando, C., Ramirez, C. M., Goedeke, L. & Suarez, Y. MicroRNAs in metabolic disease. *Arterioscler Thromb Vasc Biol* **33,** 178–185 (2013).
10. Jin, X. F., Wu, N., Wang, L. & Li, J. Circulating microRNAs: a novel class of potential biomarkers for diagnosing and prognosing central nervous system diseases. *Cell Mol Neurobiol* **33,** 601–613 (2013).
11. Ishida, M. & Selaru, F. M. miRNA-Based Therapeutic Strategies. *Curr Anesthesiol Rep* **1,** 63–70 (2013).
12. Fabbri, M. MicroRNAs and cancer: towards a personalized medicine. *Curr Mol Med* **13,** 751–756 (2013).
13. Yuan, K., Orcholski, M., Tian, X., Liao, X. & de Jesus Perez, V. A. MicroRNAs: promising therapeutic targets for the treatment of pulmonary arterial hypertension. *Expert Opin Ther Targets* **17,** 557–564 (2013).
14. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34,** D140–D144 (2006).
15. Jiang, Q. *et al.* miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* **37,** D98–104 (2009).
16. Wang, J., Lu, M., Qiu, C. & Cui, Q. TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res* **38,** D119–D122 (2010).

17. Yang, Q., Qiu, C., Yang, J., Wu, Q. & Cui, Q. miREnvironment database: providing a bridge for microRNAs, environmental factors and phenotypes. *Bioinformatics* **27,** 3329–3330 (2011).
18. Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37,** D105–D110 (2009).
19. Dweep, H., Sticht, C., Pandey, P. & Gretz, N. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform* **44,** 839–847 (2011).
20. Shirdel, E. A., Xie, W., Mak, T. W. & Jurisica, I. NAViGaTing the micronome–using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs. *PLoS One* **6,** e17429 (2011).
21. Hsu, S. D. *et al.* miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* **42,** D78–D85 (2014).
22. Martinez, N. J. *et al.* A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev* **22,** 2535–2549 (2008).
23. Osella, M., Bosia, C., Cora, D. & Caselle, M. The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS Comput Biol* **7,** e1001101 (2011).
24. Naeem, H., Kuffner, R. & Zimmer, R. MIRTFnet: analysis of miRNA regulated transcription factors. *PLoS One* **6,** e22519 (2011).
25. Lin, C. C. *et al.* Crosstalk between transcription factors and microRNAs in human protein interaction network. *BMC Syst Biol* **6,** 18 (2012).
26. Nazarov, P. V. *et al.* Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function. *Nucleic Acids Res* **41,** 2817–2831 (2013).
27. Guo, L., Zhao, Y., Yang, S., Zhang, H. & Chen, F. Integrative analysis of miRNA-mRNA and miRNA-miRNA interactions. *Biomed Res Int* **2014,** 907420 (2014).
28. Guo, Z. *et al.* Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci Rep* **4,** 5150 (2014).
29. Barh, D., Bhat, D. & Viero, C. miReg: a resource for microRNA regulation. *J Integr Bioinform* **7,** (2010). doi: 10.2390/biecoll-jib-2010-144.
30. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37,** W305–W311 (2009).
31. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29,** 569–574 (2013).
32. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57 (2009).
33. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40,** D109–D114 (2012).
34. Chiromatzo, A. O. *et al.* miRNApath: a database of miRNAs, target genes and metabolic pathways. *Genet Mol Res* **6,** 859–865 (2007).
35. Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res* **41,** D1104–D1114 (2013).
36. Joseph Nalluri, B. K., Ghosh, Preetam, Barh, Debmalya, Jain, Neha, Juneja, Lucky, Barve, Neha. in BCB'13 Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. 672 (ACM New York, NY, USA, Washington DC, 2013).
37. Liu, X. *et al.* SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* **29,** 409–411 (2013).
38. Alsio, A. *et al.* Impact of obesity on the bioavailability of peginterferon-alpha2a and ribavirin and treatment outcome for chronic hepatitis C genotype 2 or 3. *PLoS One* **7,** e37521 (2012).
39. Irigaray, P., Newby, J. A., Lacomme, S. & Belpomme, D. Overweight/obesity and cancer genesis: more than a biological link. *Biomed Pharmacother* **61,** 665–678 (2007).
40. Karakatsanis, A. *et al.* Expression of microRNAs, miR-21, miR-31, miR-122, miR-145, miR-146a, miR-200c, miR-221, miR-222, and miR-223 in patients with hepatocellular carcinoma or intrahepatic cholangiocarcinoma and its prognostic significance. *Mol Carcinog* **52,** 297–303 (2013).
41. Li, C. *et al.* Serum microRNAs profile from genome-wide serves as a fingerprint for diagnosis of acute myocardial infarction and angina pectoris. *BMC Med Genomics* **6,** 16 (2013).
42. Lu, M., Shi, B., Wang, J., Cao, Q. & Cui, Q. TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics* **11,** 419 (2010).
43. Li, Y. *et al.* HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* **42,** D1070–D1074 (2014).
44. Showe, M. K. *et al.* Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. *Cancer Res* **69,** 9202–9210 (2009).
45. Yan, Z. *et al.* Highly accurate two-gene signature for gastric cancer. *Med Oncol* **30,** 584 (2013).
46. Barh, D. *et al.* A novel in silico reverse-transcriptomics-based identification and blood-based validation of a panel of sub-type specific biomarkers in lung cancer. *BMC Genomics* **14 Suppl 6**, S5 (2013).

Acknowledgement

We duly acknowledge all the external databases and resources that have been used for hyper linking various modules of *miRegulome* v1.0. *miRegulome* v1.0 is developed without any financial support.

Author Contributions

D.B. conceived, developed the prototype, formulate the design of the database, visualization, and analysis pipe line of the tools, coordinated and led the entire project; J.N. and B.K. performed all computational works and designed the database and tools; N.J., L.J., N.B. and A.B. curated all data, P.G. led the computational group, A.K., K.B., S.T., A.M., V.A. and A.S. cross-checked all data, data analysis, and tool performance. D.B. wrote the paper; All authors read and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Barh, D. *et al. miRegulome*: a knowledge-base of miRNA regulomics and analysis. *Sci. Rep.* **5**, 12832; doi: 10.1038/srep12832 (2015).

IV.2.1 Conclusions from this research/ Chapter-2

i. We have developed user-friendly miRegulome knowledgebase that provides complete regulome (upstream regulators, downstream targets, miRNA regulated pathways, functions, diseases etc.) of a miRNA where all modules are hyperlinked to an appropriate external resource to give overall idea about the miRNA
ii. All information in this database is based on validated data manually curated from published literature
iii. The database contains 803 miRNAs from 12 species (including Human, Mouse, Rat, and plants)
iv. It provides a comprehensive understanding of biology and function of a miRNA
v. Visualization of miRNA regulome and advance combined search provide a single window to explore wide range of data and data interpretation
vi. The database and its integrated four tools (Chemical-disease analysis, miRNA-disease analysis, Gene–disease analysis, and Disease-chemical/miRNA analysis) are helpful in understanding novel biology of miRNAs and their novel associations with diseases
vii. The database is essential resource for miRNA researchers and its four tools facilitates biomarkers and therapeutics discoveries with high precision

IV.2.2 Media highlights of this research outcomes/ Chapter-2

Medical News Today (MNT)

Novel association between obesity and diabetes through miRNA is identified

Published: Wednesday 28 October 2015

Adapted Media Release

In today's bio-medical research, microRNA (miRNAs) are becoming increasingly important for their various roles in human patho-physiologies. However, an integrated resource on how the miRNA regulates (miRNA regulome) such biological events are not available.

An international team of scientists led by Dr. Debmalya Barh from the Institute of Integrative Omics and Applied Biotechnology (IIOAB), Purba Medinipur, WB, India has developed the miRegulome- a unique knowledge-base of miRNA regulomics and analysis that provides a single window for a wide range of data exploration, visualization of miRNA regulome and data interpretation, exploration of new biological events, and biomarker discovery using its integrated highly efficient tools and search engines.

Using the miRegulome v1.0, these researchers are able to identify a novel relationship between obesity and diabetes. It is found that deregulation of mir-27b and mir-143 may affect lipid metabolism, adipocytokine, insulin, and type-2 diabetes pathways and eventually may lead to develop obesity and diabetes. Further, an aberrant expression of mir-27b may play a role in obesity-associated insulin resistance by modulating adipocytokines and Ribavirin resistance in obese patients thus linking obesity with diabetes at a new and deeper molecular level.

Dr. Preetam Ghosh form Dept. of Computer Science, Virginia Commonwealth University, USA, who headed the computing part, describes miRegulome as a powerful tool and according Dr. Barh it will be an essential resource in miRNOmics R&D in coming days.

The work was recently published in *Nature Scientific Reports*.

http://www.medicalnewstoday.com/releases/301718.php

About Medical News Today *(www.medicalnewstoday.com): MTN is the healthcare internet publishing market leader for medical news. It is in the top 360 United States sites and top 120 United Kingdom sites and receives more than 12 million monthly visits, 10 million monthly unique visitors and 15 million monthly page views as reported by Quantcast. It contents are based on evidence-based, peer-reviewed studies, along with accurate, unbiased and informative content from governmental organisations (e.g. FDA, CDC, NIH, NHS), medical societies, royal colleges, professional associations, patients' groups, pharmaceutical and biotech companies, among others and targeted to an educated audience of both healthcare professionals and patients alike.*

IV.3 Chapter III: Research Article

miRsig: a consensus-based network inference methodology to identify pan-cancer miRNA-miRNA interaction signatures.

Joseph J Nalluri, **Debmalya Barh**, **Vasco Azevedo**, Preetam Ghosh

Scientific Reports. 2017 Jan 3;7:39684. doi: 10.1038/srep39684.
[PMID: 28045122]
Impact Factor: 5.2 (2016)

In this chapter, a novel computational pipeline is developed based on consensus of six network inference algorithms on miRNA expression matrix and graph intersection analysis to predict the common and core miRNA-miRNA interaction signatures in multiple diseases especially in cancers towards identification of early deregulated miRNA networks and therefore the screening or early diagnostic biomarkers. The miRsig (miRNA Signature) online tool is also developed that performs the analysis and visualization of the disease-specific signature/core miRNA-miRNA interactions based on this novel computational pipeline. miRsig is available at: http://bnet.egr.vcu.edu/miRsig

SCIENTIFIC REPORTS

OPEN

miRsig: a consensus-based network inference methodology to identify pan-cancer miRNA-miRNA interaction signatures

Received: 29 April 2016
Accepted: 25 November 2016
Published: 03 January 2017

Joseph J. Nalluri[1], Debmalya Barh[2,3,4], Vasco Azevedo[3] & Preetam Ghosh[1]

Decoding the patterns of miRNA regulation in diseases are important to properly realize its potential in diagnostic, prog- nostic, and therapeutic applications. Only a handful of studies computationally predict possible miRNA-miRNA interactions; hence, such interactions require a thorough investigation to understand their role in disease progression. In this paper, we design a novel computational pipeline to predict the common signature/core sets of miRNA-miRNA interactions for different diseases using network inference algorithms on the miRNA-disease expression profiles; the individual predictions of these algorithms were then merged using a consensus-based approach to predict miRNA-miRNA associations. We next selected the miRNA-miRNA associations across particular diseases to generate the corresponding disease-specific miRNA-interaction networks. Next, graph intersection analysis was performed on these networks for multiple diseases to identify the common signature/core sets of miRNA interactions. We applied this pipeline to identify the common signature of miRNA-miRNA inter- actions for cancers. The identified signatures when validated using a manual literature search from *PubMed Central* and the *PhenomiR* database, show strong relevance with the respective cancers, providing an indirect proof of the high accuracy of our methodology. We developed *miRsig*, an online tool for analysis and visualization of the disease-specific signature/core miRNA-miRNA interactions, available at: http://bnet.egr.vcu.edu/miRsig.

MicroRNAs (miRNAs) are non-coding RNAs of ~22 nucleotides in length that inhibit gene expression at the post transcriptional level by binding to the 3′ UTR region of target mRNAs through complementary base pairing[1]. However, a couple of studies have instead reported an activation of target gene expression as well[2,3]. By virtue of this gene regulation mechanism, miRNAs play a critical role in several biological processes[4] and patho-physiological conditions, including cancers[5]. The role of miRNA regulations in diseases have been widely recorded[6], however the precise patterns through which a miRNA regulates a certain disease(s) are still elusive. For example, it is not yet clear *how* a miRNA's up/down regulation directly or indirectly affects a disease's progression or repression because of the many intermediate factors involved. Thus, predicting and identifying miRNA-disease associations has been a primary research area for several groups. Moreover, the multi-level interactions of miRNAs in cancer-like multi-factorial diseases are more complex due to the possibility of several types of interactions, such as, the classical miRNA-mRNA, miRNA-environmental factors, miRNA- transcription factors-miRNA[7], and our newly hypothesized direct miRNA-miRNA interactions without any intermediate linkers (e.g., transcription factors)[8]. However, till date, no experimental proof of direct miRNA-miRNA interactions exists except, a single study reported in mouse[9].

Although, the precise patterns or the reasons behind miRNAs' deregulation in cancers are not fully understood, it has been found that miRNAs tend to work together in groups[10], as evidenced in certain diseases[11]. Such co-ordinated

[1]Department of Computer Science, School of Engineering, Virginia Commonwealth University, Richmond, Virginia,USA. [2]Center for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Purba Medinipur, West Bengal, India. [3]Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Minas Gerais, Brazil. [4]Xcode Life Sciences, 3D Eldorado, 112 Nungambakkam High Road, Nungambakkam, Chennai, Tamil Nadu-600034, India. Correspondence and requests for materials should be addressed to J.J.N. (email: nallurijj@vcu.edu)

regulation, comprising mutual co-targeting and co-regulation, as well as miRNA regulation by other miRNAs are reported in many disease conditions, including various cancers[10]. To elucidate the miRNA-disease associations at the regulome level, we earlier developed the *miRegulome* database and corresponding analytic tools[12]. Furthermore, in cancers it has been observed that groups of miRNAs, known as *superfamilies*, express consistently across several cancers and may act as *drivers* of tumorigenesis, where few key miRNAs direct the global miRNA expression patterns[13]. Identification and existence of such groups or super-families of miRNAs obviously leads to the intuition, that the therapeutic suppression or expression of any one of the miRNAs in the family, would compensate for the other participants of the family[13]. Our central hypothesis in this paper is that, these miRNAs in such *superfamilies* may interact directly or indirectly, by forming a core miRNA-miRNA co-regulatory network and thereby acting as a signature component for prognosis, prediction, and early diagnosis of any disease including cancer.

Several computational efforts have been implemented to study and discover the disease-miRNA interaction networks based on functional enrichment analysis[14], social network analysis methods[15], similarity-based methods[16], and diffusion-based methods[17]. Some studies have integrated genomic and phenotype data sets to infer novel miRNA-disease associations[18]. A miRNA regulatory network was also constructed by integrating multidimensional high-throughput data and was used to identify the cancer-associated miRNAs[19]. Similarly, co-regulating miRNA clusters and prioritized candidate miRNAs across multiple types of diseases have been predicted. Using co-regulating functional modules, a miRNA-miRNA synergistic network was constructed to study the aspect of *homophily* among miRNAs associated with the same disease and subsequently disease-specific miRNAs were detected based on their network topological features. In this study, a miRNA-miRNA co-regulation network was constructed by selecting common miRNAs across various data sets related to the same disease, pairing them based on their sharing of common targets, and subsequently performing a GO enrichment analysis of their predicted targets. These miRNAs were qualified as co-regulating if they shared a significant amount of GO enrichment analyses of predicted targets[20]. Disease-specific miRNAs were also identified using the miRNA target-dysregulated network built on the assumption that causative miRNAs show abnormal regulation of their target genes[21]. Similarly, disease-specific miRNAs were also identified by integrating phenotype associations of diseases which had matching miRNA and mRNA expression profiles[22]. Network theoretic algorithms such as the biclique-based method[23], biclustering technique[24] and maximum weighted matching[25] among others have been deployed to discover and predict the patterns of miRNA regulation. Graph theoretical methods and network inference models have also been applied to analyze complex regulatory interactions and reconstruct the causative gene regulatory network and other biological networks[26–29].

In this work, we have used the miRNA expression data sets available at the *PhenomiR*[30] database to predict miRNA-miRNA core/signature interactions across several cancers using a combination of (i) six state-of-the-art network inference algorithms, (ii) a *wisdom of crowds*[31] based consensus approach[32] to generate disease-specific miRNA interaction networks with higher accuracy, and (iii) a simplified graph intersection analysis to identify the miRNA-miRNA core interactions across multiple diseases belonging to a particular disease class.

Methods

The methodology adopted in this paper is comprised of i) translating the miRNA-disease expression scores from the *PhenomiR* database into a miRNA expression matrix (Fig. 1, Step 1); ii) deploying six network inference algorithms on the expression matrix and deriving the miRNA-miRNA interaction scores from each algorithm (Fig. 1, Step 2); iii) performing a consensus-based approach, i.e. estimating an average score for every miRNA-miRNA interaction across its six predicted scores (Fig. 1, Step 3); iv) validating the resultant interactions using precision-recall analysis with a hypothetical true network generated using the PubMed IDs from *PhenomiR*; v) analyzing the miRNA-miRNA interaction networks for every disease and detection of the conserved miRNA-miRNA interactions across various groups of cancers and finally vi) validating the conserved miRNA-miRNA interactions in the identified group of cancers via manual literature search.

Data preparation and modeling. The data from the *PhenomiR* database is freely available and was used in this study. *PhenomiR 2.0* was downloaded for the purposes of this study. *PhenomiR 2.0* is a comprehensive data set containing 535 database entries across 345 articles recording miRNA expressions in diseases[30]. As shown in Fig. 2, the data from *PhenomiR* was converted into a disease-specific miRNA expression matrix (shown in Fig. 3). The miRNAs whose fold-change values were not available in *PhenomiR 2.0* data set were discarded from the study; this also includes some misformatted lines of data that were excluded from further processing as they were also missing the fold-change values. Here, the core idea is to consider a pair of miRNA and disease as a single miRNA-disease (*MD*) node, as seen in Fig. 3; note that, for ease of reference, we consider an M_iD_j pair as an *MD* node which conceptually designates a disease-specific miRNA. The same miRNA participating in multiple diseases will have different expression profiles in each of them and hence the disease specific miRNA terminology, i.e., *MD*, signifies a miRNA's expression profile in a particular disease. Thus, every unique miRNA-disease pair constitutes a unique *MD* type node. In this disease-specific miRNA expression matrix (Fig. 3-b), each row represents a study/experiment and each column represents an *MD*'s expression score in that study. The resultant expression matrix herein, has 4,343 unique nodes/columns (i.e., unique *MD*s in the network) for 267 samples (i.e., rows).

In the *PhenomiR* data set, some *MD*s have two fold-change values indicating minimum and maximum expression scores while other *MD*s only report a minimum fold-change expression score (for e.g., see Fig. 1, Step 1, *PhenomiR* data set, row 2). To assess these scenarios, we devised three different methodologies (described in the next section), generated separate expression matrices based on each methodology and performed the subsequent analysis on each of them.

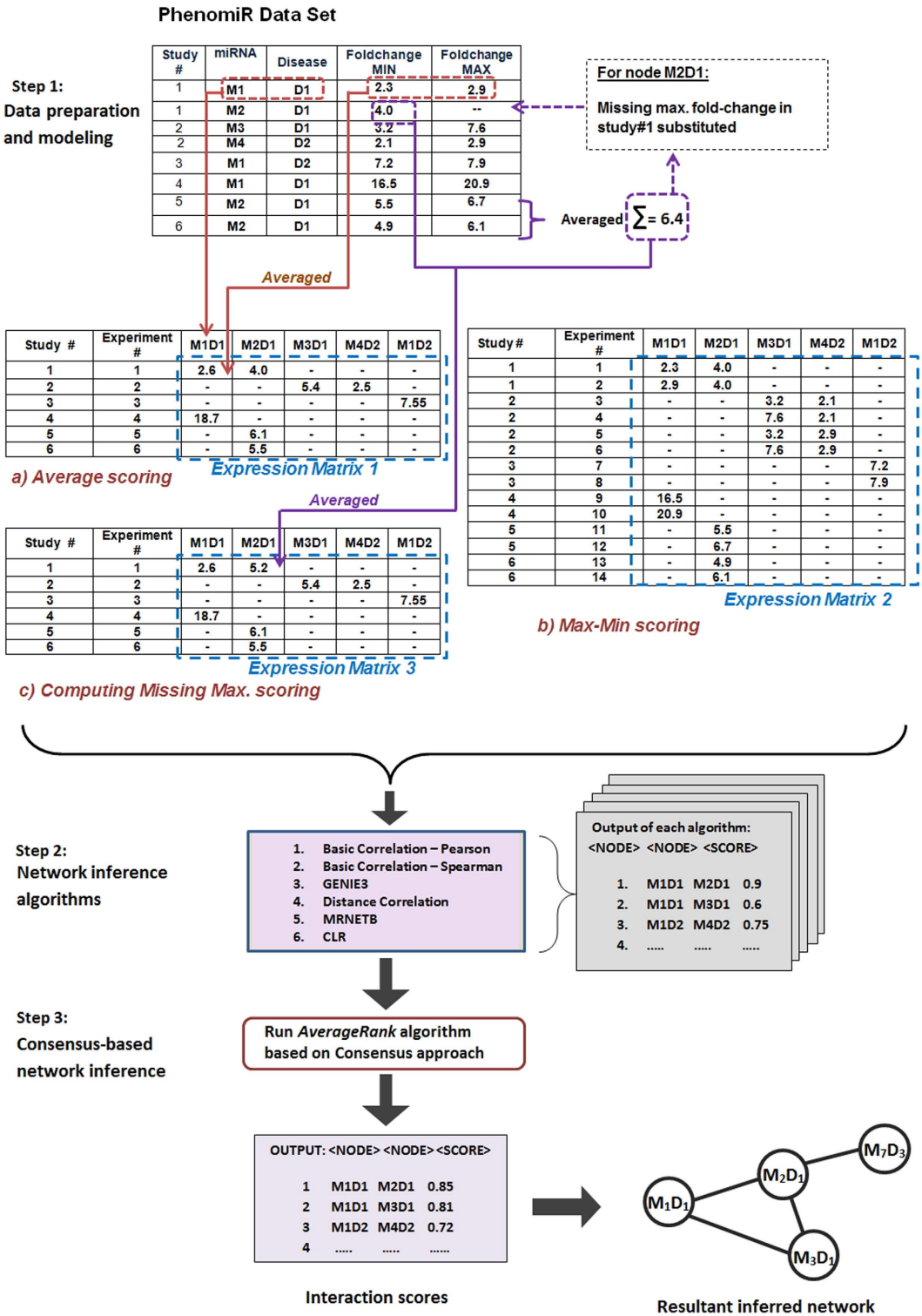


Figure 1. Overview of the methodology with M_i denoting miRNAs and D_j denoting the diseases. *Step 1* consists of translating the *PhenomiR* data set into three miRNA expression matrices (a, b and c) based on three approaches. In *Step 2*, each of these matrices are subjected to six network inference algorithms which produce the interaction scores across the different M_iD_j nodes. In *Step 3*, the six individual $M_iD_j - M_xD_y$ interaction scores are averaged into a final score designating its confidence.

Average scoring. Under *Average scoring* method, for the *MD*s having both minimum and maximum fold-change values per sample, their average was taken and considered as the final expression value in the expression matrix. As shown in Fig. 1, Step-1, the entry *M*1-*D*1 has two expression values - 2.3 and 2.9, i.e., minimum fold-change and maximum fold-change respectively, which were averaged to 2.6 in the *Expression Matrix 1* (see Fig. 1, Step 1-a).

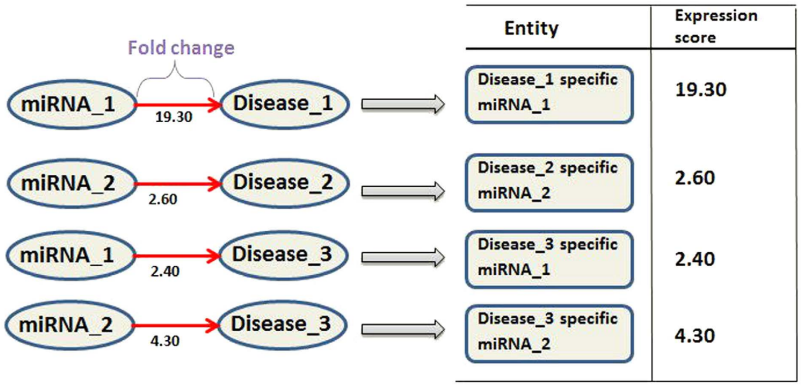


Figure 2. Schematic of the miRNA-disease regulation with fold-change values.

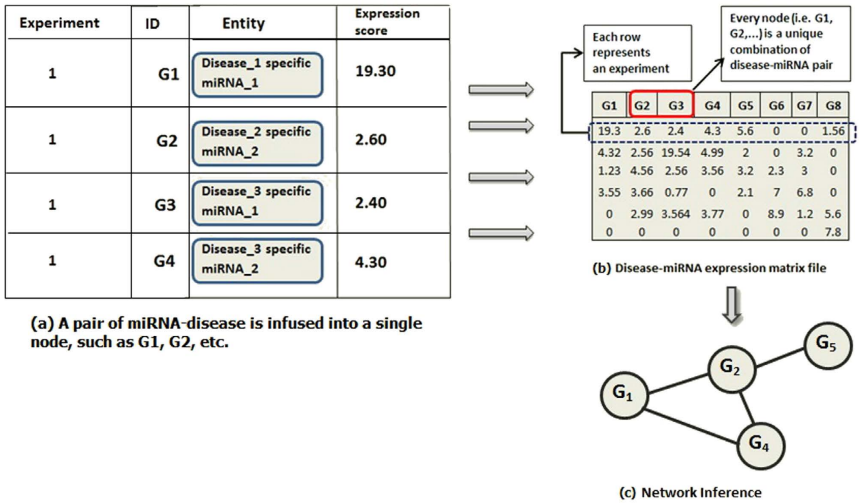


Figure 3. Schematic of the miRNA expression data set. [(**a**) and (**b**)] Data from *PhenomiR* is mapped into an miRNA expression matrix. (**c**) Network inference approach is applied to the matrix to derive the interaction network.

For *MD*s with only their minimum expression values reported, this single value was also considered to be its average expression value.

Retaining maximum and minimum expression values. The *Average scoring* method can lead to a potential loss of information as the individual maximum and minimum expression values (when available) were not retained. Hence we designed the following two methods to generate the expression matrix.

1. *Max-Min scoring*
   Under *Max-Min scoring* method, for the *MD*s having minimum and maximum fold-change expression values, (instead of taking their average) both these data points were considered as separate entries; thus, the same *MD* was considered twice in the expression matrix with the duplicate entry designating a new experiment. As displayed in Fig. 1, Step 1, the first row entry, *M*1-*D*1 in *Study-1* has two expression values; these values were individually considered as separate data points and included in the expression matrix accordingly along with their co-expressing miRNAs' expression values, providing us with *Expression Matrix 2* (see Fig. 1, Step 1-b).

2. *Computing Missing Max. scoring*
   Under *Computing Missing Max scoring* method, for the *MD*s which did not have a maximum fold-change expression value, we took an average of its maximum fold-change values across all its *other samples* and substituted this average score as it's maximum fold-change expression value. As shown in Fig. 1, Step 1, the entry *M*2-*D*1 on 2nd row does not have a maximum fold-change value. However, *M*2-*D*1 combination has maximum fold-change expression values of 6.7 and 3.1 from sample #5 and #6, respectively. Herein, we took an average of these two values, i.e. 6.4 and substituted it for the original missing value for *M*2-*D*1 in the 2nd row. This method overcomes the limitation posed due the non-availability of the expression value by giving

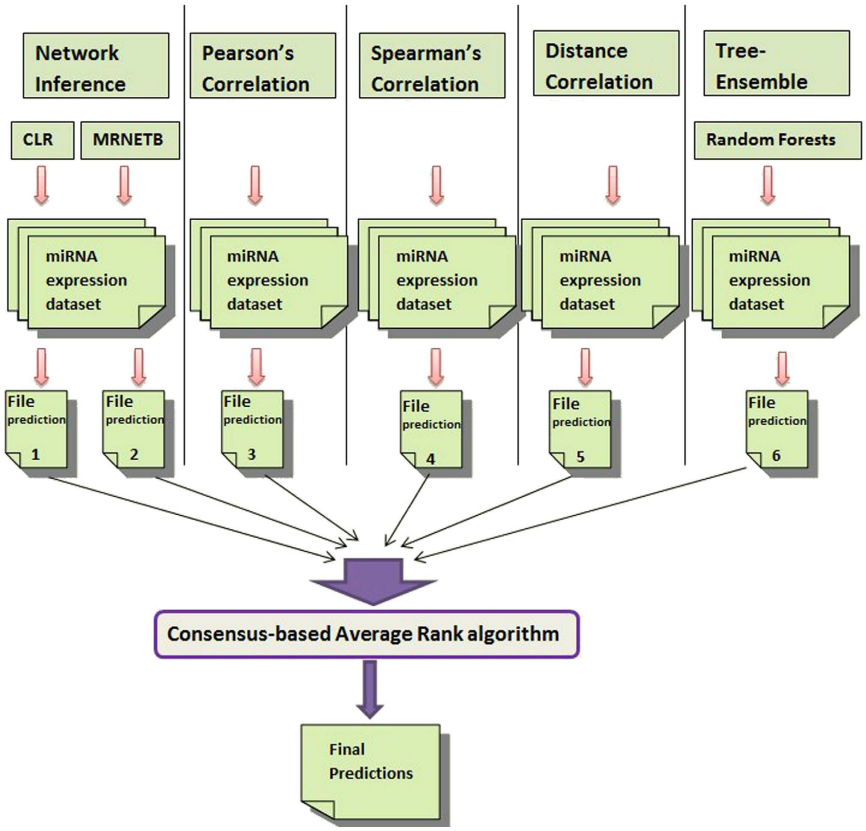


Figure 4. Workflow of the consensus-based miRNA network inference.

its closest approximation, based on the particular *MD*'s expression pattern across the sample spectrum. After applying this method, the *Average Scoring* method was performed on this matrix to obtain *Expression Matrix 3* (see Fig. 1, Step 1-c).

After the three expression matrices were derived, a reverse engineering methodology[32] was adopted to reconstruct the *MD-MD* regulatory network from these expression matrices (Fig. 3, *Network Inference*), by applying six widely used network inference algorithms along with a consensus-based ranking algorithm, which is explained in the next section.

Network inference algorithms. Each expression matrix has 4,343 nodes and therefore, there are potentially 4,343 × 4,343 (i.e. 18,861,649) *MD-MD* interactions in the network. Six different network inference algorithms were applied on the miRNA expression matrix, which gave prediction scores for every *MD-MD* interaction. We used the mutual information-based algorithm, Context Likelihood of Relatedness (CLR)[33], Maximum Relevance Minimum Redundancy Backward (MRNETB)[34], Basic Correlation methods (Pearson and Spearman), Distance Correlation (DC)[35], and regression-based Gene Network Inference with Ensemble of Trees (GENIE3)[36] algorithms for network inference. The details of the algorithms are given in Supplementary File S1. Note that, the Basic Correlation methods resulted in two different network inference algorithms based on the type of correlations implemented, i.e., one each for Pearson and Spearman correlations.

Consensus based network inference approach. Each of the six individual network inference algorithms produced a ranked list of prediction scores for every *MD-MD* interaction (see Fig. 1, Step-2). Thereafter, we used the *wisdom of crowds*[31] approach, which proposes that the aggregation of information from the community yields better results than the individual few. In this study, the consensus based approach aggregates the collective information (i.e. prediction scores) from the six individual network inference algorithms and computes a more accurate final score for *MD-MD* interactions. This rank is computed by taking an average of the predicted ranks of each interaction derived from the corresponding network inference algorithms. Figure 4 displays the workflow of this approach. This approach was earlier implemented to infer gene-regulatory networks and yielded highest accuracy compared to each of the individual network inference algorithms[32].

This consensus based network inference approach is executed in the *Average Rank*[32] algorithm which essentially computes the average score of a particular *MD-MD* interaction by taking the mean of its six predicted ranks. The ranking methodology used in this algorithm is based on the *Borda* count method. This method is used in elections during which voters rank candidates as per their preferences. The winning candidate is the one with the best average rank. Here, all the interactions are first ranked in descending order of their predicted scores (as

Borda rank (Norm. borda points)	Borda points	Rank	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6
1	3	1	I_4^*	I_2	I_2	I_2	I_2	I_3
0.6667	2	2	I_2	I_3	I_3	I_4^*	I_4^*	I_2
0.3334	1	3	I_1	I_4^*	I_1	I_3	I_3	I_4^*
0	0	4	I_3	I_1	I_4^*	I_1	I_1	I_1

Table 1. Ranked individual predictions of each algorithm for every interaction *I*. Borda points are allocated to each **Rank**. A relative ***Borda rank*** $\left(=\frac{Borda\ points\ for\ that\ rank}{maximum\ Borda\ points}\right)$ is computed for every **Rank**. *Borda* ranks for interaction I_4 (noted with *) are 1, 0.333, 0, 0.666, 0.666 and 0.334 by the six algorithms respectively.

Interaction	Averaging of *Borda* ranks	Final rank
I_2	$(0.66+1+1+1+1+0.66)/6$	0.88
I_3	$(0+0.66+0.66+0.33+0.33+1)/6$	0.49
I_4^*	$(1+0.33+0+0.66+0.66+0.33)/6$	0.49*
I_1	$(0.33+0+0.33+0+0+0)/6$	0.11

Table 2. Final ranks for each interaction; the final rank of interaction I_4 is 0.49.

Rank	Interaction	Score
1	Hepatocellular carcinoma:hsa-mir-183 ⇒ Hepatocellular carcinoma:hsa-mir-374a	0.9786
2	Hepatocellular carcinoma:hsa-mir-374a ⇒ Hepatocellular carcinoma:hsa-mir-182	0.9781
3	Breast cancer:hsa-let-7a-1 ⇒ Breast cancer:hsa-mir-30d	0.2985
4	Breast cancer:hsa-let-7a-1 ⇒ Breast cancer:hsa-mir-381	0.2426

Table 3. Format of the results based on the consensus approach.

seen in the column *Rank* in Table 1). Describing briefly, the *Borda* count method allocates points to each rank. The highest ranked interaction (meaning, 1) get the maximum *Borda* points (number of interactions - 1) and the lowest ranked interaction has 0 *Borda* points as demonstrated in the column *Borda points* in Table 1. In order to derive the final rank between 0 and 1, these points are thereafter normalized to derive a relative *Borda* rank. Thus, each rank has been translated to its new relative *Borda* rank. Note that, the *Borda* count ranking method is among the many other methods to perform *averaging* of the ranks in the consensus methodology.

The six network inference algorithms generate six different ranks for each interaction and the consensus algorithm next computes an average *Borda* rank for the interaction. Tables 1 and 2 display a scenario of ranking four *MD-MD* interactions I_1, I_2, I_3 and I_4 via a consensus-based approach as executed in *AverageRank* algorithm. Table 1 displays the ranked list of predictions for these interactions by all the six network inference algorithms based on their prediction scores. For example, in Table 1, *Algorithm 1* ranks *MD-MD* interactions in this order — I_4, I_2, I_1 and I_3 based on their prediction scores. The individual ranks for miRNA-miRNA interaction I_4 are 1, 3, 4, 2, 2 and 3 by the six algorithms respectively (noted with *), and their relative respective *Borda* ranks are 1, 0.333, 0, 0.666. 0.666 and 0.333. The final rank of interaction I_4 is the average of all the *Borda* ranks, i.e., 0.49, as demonstrated in Table 2 (noted with *). Similarly the final ranks of every other interaction is computed using the following formula,

$$Final\text{-}rank(I) = \frac{1}{K}\sum_{j=1}^{K} Borda\text{-}rank_j(I) \quad (1)$$

where, K is the number of algorithms (six, in our case). These results are displayed in Table 2.

An example of the final result listing of our *MD-MD* interactions is shown in Table 3 (also see Fig. 1 Step-3).

In these results, we noted all the different possibilities of interactions that can occur considering the miRNA-disease pair, i.e. *MD* as a node. There are essentially four types of interactions that can exist in this network. These are explained in Table 4. Among these types, *type 1* is a self-loop and not applicable for our purposes. For application purposes of our methodology, we focused on analyzing the set of interactions belonging to *type 3* which is further elaborated in the next section. Interactions of *type 2* and *type 4* will be studied in the future to analyze the relationship between diseases sharing a common miRNA (*type 2*) and the proximity between dissimilar miRNAs and dissimilar diseases (*type 4*) having high probabilities of interaction.

Disease-specific miRNA network construction. In this section, the results of the *type 3* interactions were selected for disease-specific analysis. There were 66 unique diseases in the final predicted list of interactions from the *Average Rank* algorithm; this list of diseases are provided in the Supplementary File S2. Under a specific disease D_x, all the miRNA-miRNA edges, i.e. $M_1D_x - M_2D_x$ edges were collected into a single D_x disease network; thereby giving us the **d**isease-specific **mi**RNA-miRNA **i**nteraction **n**etwork (*DMIN*) (Fig. 5, Step 1). *DMIN* is a network $G=(V, E)$, where $V=\{M_1D_x, M_2D_x, \ldots M_nD_x\}$ (i.e., set of miRNAs under disease name D_x)

Type #	Interaction type	Edge	Remark
1	$miRNAs_{same}$, $Diseases_{same}$	$M_1D_1 \rightarrow M_1D_1$	Self-loops, N/A
2	$miRNAs_{same}$, $Diseases_{different}$	$M_1D_1 \rightarrow M_1D_2$	Present in the result set
3	$miRNAs_{different}$, $Diseases_{same}$	$M_1D_1 \rightarrow M_2D_1$	Present and used for analysis
4	$miRNAs_{different}$, $Diseases_{different}$	$M_1D_1 \rightarrow M_2D_2$	Present in the result set

Table 4. Types of interactions in the network.

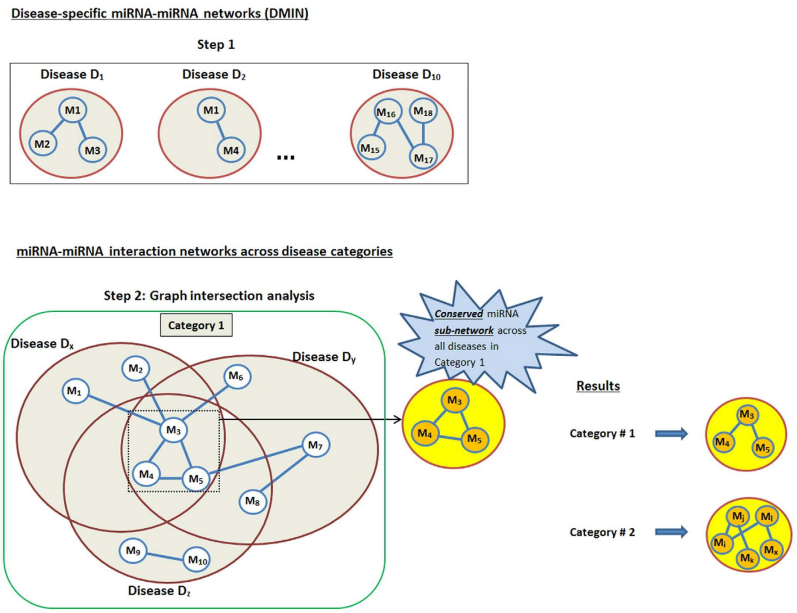


Figure 5. Overview of the disease analysis.

and E is the ordered set of edges, where edge $e = \{M_i, M_j\}$. We performed a similar network construction for every cancer-related disease, D_x. To pursue a more definitive and cancer-specific analysis, only cancer-related diseases were chosen and grouped into classes based on their tissue/organ specificity. We created four major classes: i) gastrointestinal cancers (esophageal, gastroesophageal, gastrointestinal, gastric, and colorectal cancer), ii) endocrine cancers (hepatocellular, pancreatic, and thyroid carcinoma follicular, and thyroid carcinoma papillary), iii) leukemia/blood cancers (hematological tumors, acute myeloid leukemia, chronic lymphatic leukemia, and acute myelogenous leukemia), and iv) nerve cancers (neuroblastoma, medulloblastoma, and glioblastoma).

Under a particular disease class, all the corresponding *DMIN*s were combined into a single network (Fig. 5, Step 2). Using graph intersection analysis, we mined the miRNA-miRNA interaction networks of all the cancers within the specific class to identify a conserved (signature/core) miRNA-miRNA interaction component. This identified miRNA-miRNA interaction component was present in all the diseases of that particular class. These findings are reported in the *pan-cancer miRNA signatures* section and the results are discussed in the *Discussion* section.

Results

Validation of interactions. After executing the *Consensus based network inference approach* on three input miRNA expression matrices derived from the three approaches mentioned in the *Data preparation and modeling* section (*Average scoring*, *Retaining Max-Min* and *Computing Missing Max.*), we obtained three sets of predicted miRNA-miRNA interactions. Each predicted interaction was validated by querying for PubMed IDs in the *PhenomiR* database which cited and reported the occurrence of miRNAs' association with the specific disease in a single PubMed ID. For e.g., for each predicted interaction, i.e. M_aD_x to M_bD_x, if a PubMed ID cited the occurrence of the association between the miRNAs (M_a, M_b) and the disease (D_x), the interaction was termed as true/validated (1); else the predicted interaction was termed as unknown/unverified (0). Based on this, labels were generated for every interaction in the resultant set forming the true network. We performed a precision-recall analysis to ascertain the accuracy of the consensus-based network inference method. The precision-recall values were calculated using the formula:

$$Precision = \frac{tp}{tp + fp} \qquad Recall = \frac{tp}{tp + fn} \tag{2}$$

where tp, fp, and fn are true-positives, false-positives and false-negatives respectively.

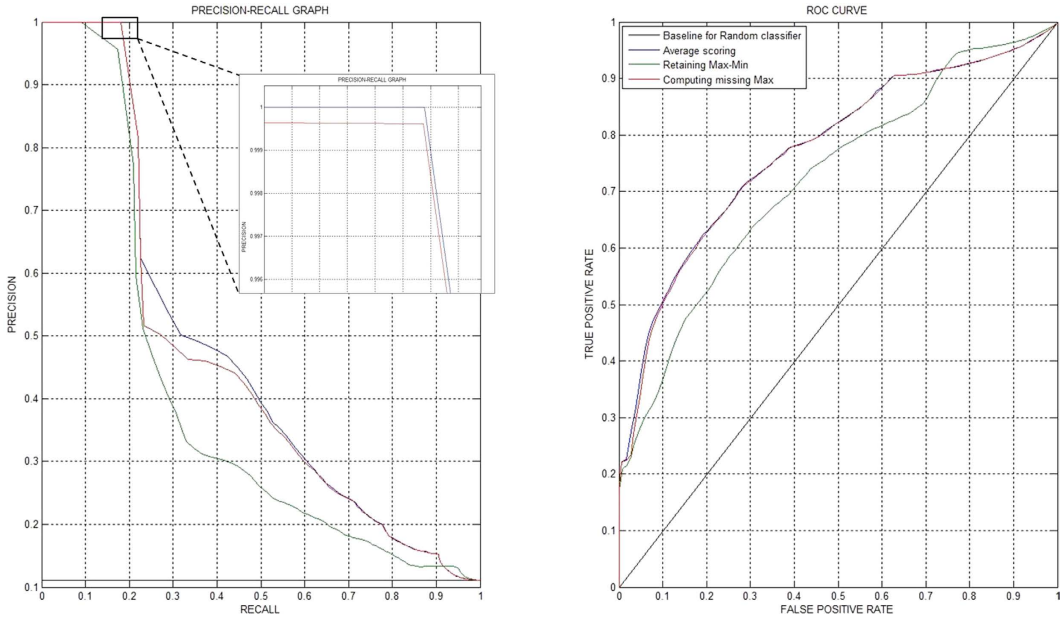


Figure 6. Precision-recall and ROC curves displaying the accuracy of the three methods. The figure demonstrates that the *Average scoring* (blue curve) method fared better than *Retaining Max-Min* (green curve) and *Computing Missing Max.* (red curve) methods. The inset image shows that the precision of *Average scoring* method slightly outperformed the *Computing Missing Max.* and was the best overall performer.

Figure 6 displays the results of the precision-recall analysis and the ROC curve for all the three approaches used. As demonstrated in the figure, the *Average scoring* method fared better than the other two methods; in fact the *Computing Missing Max.* method also performed well for low recall but gradually degraded for higher recall values. Based on this precision-recall curve, our proposed methodology displays a high precision (for up to a 30% recall) demonstrating its effectiveness in providing high confidence to the results. The ROC curve shows that both the *Average scoring* and *Computing Missing Max.* methods are comparable in predicting the true positives when compared to the number of false positives seen alongside.

Note that our true network generation method has some obvious limitations. While a true edge constituting the association of the two miRNAs with the same disease in the same PubMed ID is still acceptable (specifically because these edges were manually curated), the unverified edges may simply mean that a study has not yet been reported associating the miRNAs to the same disease. Hence, a high precision performance should be the best judge of our methodology whereas the recall curve can be somewhat circumstantial.

Pan-cancer miRNA signatures. After the *Validation of interactions*, in order to confidently detect miRNA signatures in the specified disease classes, only the top 10% interactions with the highest confidence scores were used in the construction of *DMIN* (Fig. 5, Step 1) and the subsequent graph intersection approach (Fig. 5, Step 2). Hence, all the considered miRNA-miRNA interactions had a confidence score of 0.9 and above. As reported in Fig. 7, under gastrointestinal cancers, we detected a signature component of three miRNAs (hsa-mir-30a, hsa-mir-181a-1, and hsa-mir-29c). For endocrine cancers, the signature component consisted of hsa-mir-221, hsa-mir-222, hsa-mir-155, hsa-mir-224, hsa-mir-181a-1, and hsa-mir-181b-1. For leukemia cancers, the signature component consisted of hsa-mir-29b-1, hsa-mir-106a, hsa-mir-20a, hsa-mir-126, and hsa-mir-130a. We observed two different signatures for nerve cancers. For subsequent validation of these cancer-specific signature set of miRNAs, we manually mined PubMed articles which corroborate our results, as reported in Fig. 7. We queried both the *PhenomiR* database and the *PubMed Central* database for these reported PubMed IDs; the results from these two sources are shown in different colors in Fig. 7. We also observed that, while hsa-mir-30 is common in gastrointestinal and nerve cancers; hsa-mir-181 is shared by gastrointestinal, endocrine and nerve cancers. The miRNA signature component of the category leukemia is found to possess a distinct group of miRNAs (Fig. 7). The role and involvement of these miRNAs in their associated diseases are further elaborated in the *Discussion* section.

The individual steps involved in the manual search process from *PubMed Central* are shown in Fig. 8. To summarize, we first searched *PubMed Central* with the list of core miRNAs and each disease for which they form a signature component. We next manually checked the 'search' results to confirm the associations (i.e., the pruning step for PMIDs). If not enough results were retrieved from this search, we entered each miRNA, disease pair individually for all the miRNAs forming the signature component in that disease; each of these results were then manually pruned and collated to give us the set of PMIDs corresponding to the core miRNAs for that disease. This process was repeated for all the other diseases of a particular disease class.

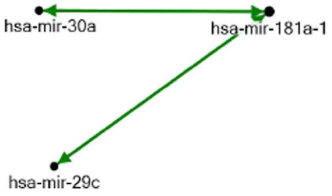
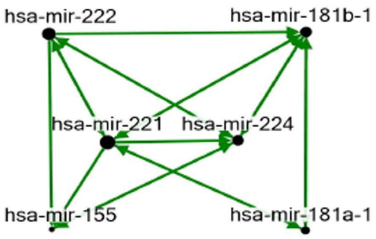
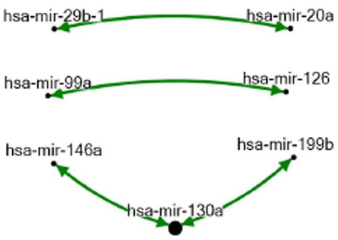
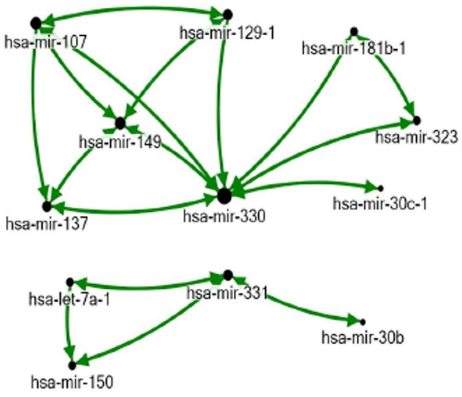
Category	Preserved miRNA-miRNA interaction	Critical miRNAs	PubMed IDs
Gastrointestinal Cancers		hsa-mir-30a hsa-mir-181a-1 hsa-mir-29c	Colorectal cancer: 18607389, 20480519, 22112324 Gastric cancer: 20022810, 21139804 Gastrointestinal cancer: 19030927, 24289824, 23950912 Gastroesophageal cancer: 19737949, 24289824 Esophageal cancer: 19737949, 20480519
Endocrine Cancers		hsa-mir-221, hsa-mir-222, hsa-mir-155, hsa-mir-224, hsa-mir-181a-1, hsa-mir-181b-1	HCC: 18223217, 21139804 Pancreatic cancer: 16192569, 16966691, 21139804, 24289824 Thyroid carcinoma, follicular: 18270258, 17205120 Thyroid carcinoma, papillary: 18270258, 17205120
Leukemia		hsa-mir-29b-1, hsa-mir-20a, hsa-mir-126, hsa-mir-130a, hsa-mir-99a, hsa-mir-146a, hsa-mir-199b	Hematological tumors: 16192569, 21139804 Leukemia, acute myeloid: 18337557, 21708028, 19602709 Leukemia, chronic lymphatic: 17934639, 20439436 Leukemia acute myelogenous: 18187662, 21708028, 19602709
Nerve Cancers		(A) hsa-mir-323, hsa-mir-129-1, hsa-mir-137, hsa-mir-330, hsa-mir-149, hsa-mir-107, hsa-mir-30c-1, hsa-mir-181b-1 (B) hsa-mir-30b, hsa-mir-331, hsa-mir-150, hsa-let-7a-1	(A)Medulloblastoma: 18973228, 24213470 Neuroblastoma: 17283129, 25238782 Glioblastoma somatic: 18577219, 17363563, 24213470 (B)Medulloblastoma: 18973228, 18756266, 25594007, 22623952 Neuroblastoma: 17283129, 24438171, 23220581 Glioblastoma somatic: 17363563, 26132860, 21139804, 26046581

Figure 7. Signature miRNA-miRNA interaction component identified in various cancer categories. The PubMed IDs citing the critical miRNAs with the disease from the *PhenomiR* database are in magenta while the PubMed IDs from the *PubMed Central* database are in blue.

***miRsig* - an online tool.** In order to aid researchers to identify disease-specific miRNA-miRNA interaction networks across several diseases, we developed the *miRsig* tool, available at http://bnet.egr.vcu.edu/miRsig. *miRsig* allows the user to visualize the miRNA-miRNA interaction network for each disease recorded in *PhenomiR* and also across multiple diseases. The results are based on the consensus-based network inference approach. *miRsig* also allows users to search for a common/core miRNA-miRNA interaction component in a user-specified selection of diseases (see Fig. 9). Users can create their own class/category of cancers by *selecting* more diseases, as shown in Fig. 9. The edges in the interaction have confidence scores as weights, from 0 (minimum) to 1 (maximum). Hence, the tool also allows the user to view only the higher/lower/specific confidence interactions by changing the *Maximum* and *Minimum* confidence score ranges. Currently, the total number of edges across the entire miRNA-miRNA interaction networks are more than 18 million. Hence, to avoid cluttering of the result set and to allow clear visibility and comprehension of the network, the *Minimum* score is set to 0.5, if not specified

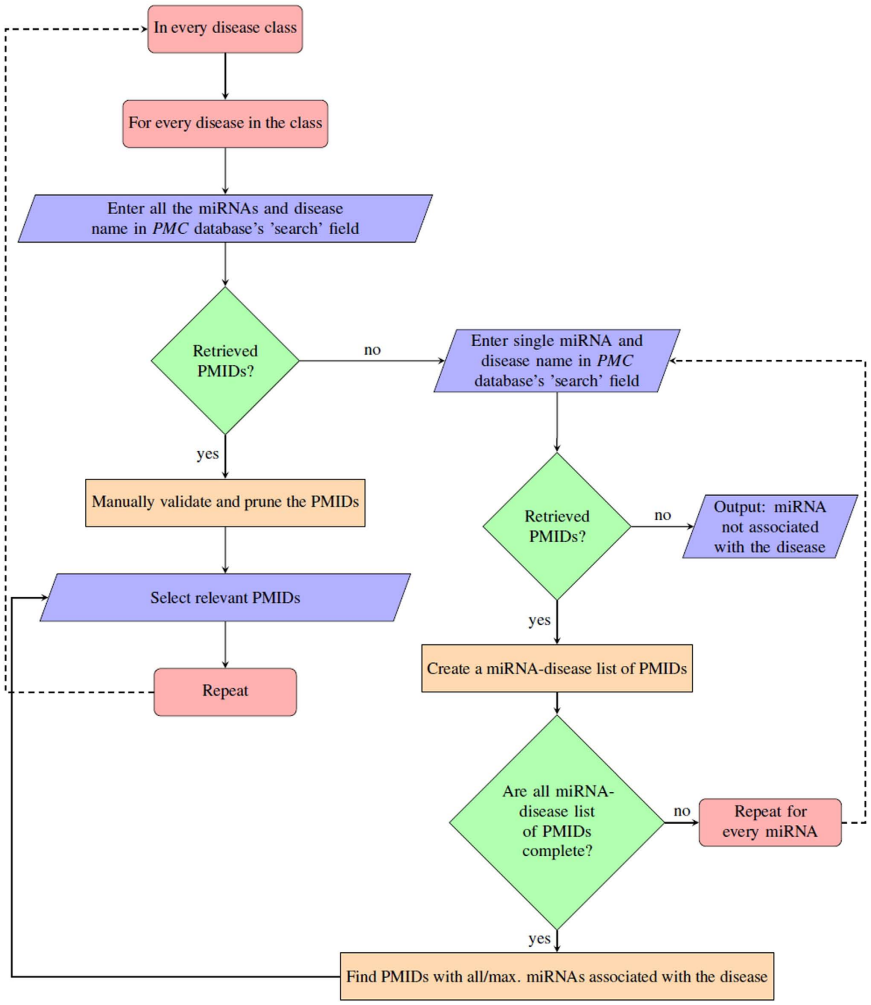


Figure 8. Flowchart of the workflow for manual literature search.

by the user. Users can also view and analyze the topological properties of miRNA clusters interacting in each or a set of diseases. The signature/core miRNA-miRNA interactions among esophageal, gastroesophageal, gastrointestinal, gastric, and colorectal cancers, as predicted and visualized is shown in Fig. 9. This network component consisting of three miRNAs (has-mir-30a, has-mir-181a-1, and has-mir-29c) is the signature component for all the aforementioned five cancers, and can be validated using simple literature search on *PubMed Central* database as demonstrated in Fig. 8.Users can also download the miRNA-interaction network in the format of an edge-list in a CSV file. This edge-list can be imported in various network analysis tools such as, *NodeXL, Cytoscape*, etc. for further study and analysis of the interaction network.

miRsig tool has been developed using MySQL as the back-end database and HTML, PHP, JavaScript, AJAX for front-end design. The interactive network visualization has been implemented using data visualization library, D3.js[37].

Discussion

miRNA-mRNA interactions have been substantially documented[38] and is a prime area of ongoing research. Similarly, miRNA- miRNA interactions through mutual co-expression[39], via transcription factor[40], and miRNA-disease associations[6] have also been reported. However, miRNA-miRNA interactions towards identification of a core miRNA-miRNA module that could potentially be a signature component for a particular disease have not been studied enough. Many studies have used computational approaches to study this aspect. A miRNA-miRNA co-regulation network in lung cancer was identified using a progressive data refining approach[20]. Similarly, miRNA expression profiling along with a genome-wide SNP approach was used to create a miRNA-miRNA synergistic network to study coronary artery disease[41]. miRNA-miRNA interactions were also identified in esophageal cancer using K-clique analysis on a bipartite network consisting of miRNAs and subpathways[42]. Additionally, miRNA-target interactions were integrated with miRNA and mRNA expressions to deduce miRNA-miRNA interactions in prostate cancer[43]. A network topological approach was also undertaken to identify disease miRNAs by constructing a miRNA-miRNA synergistic network consisting of co-regulating functional modules[44].

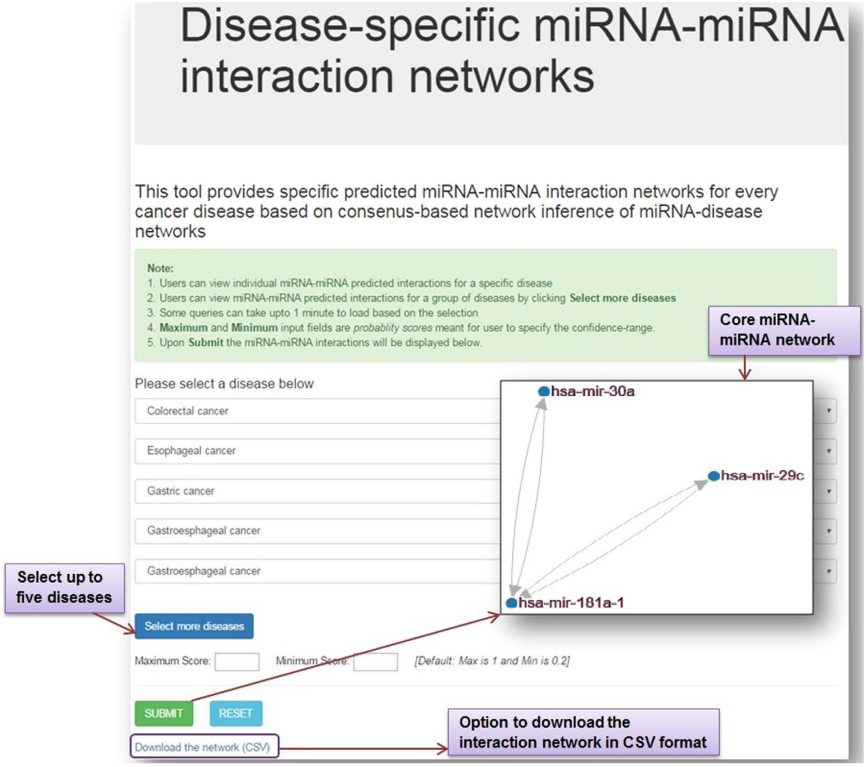


Figure 9. miRNA-miRNA interactions shown in *miRsig* for esophageal, gastroesophageal, gastrointestinal, gastric, and colorectal cancers.

In this work, we adopted a strategy that takes a miRNA expression profile and uses six different network inference algorithms (CLR[33], MRNETB[34], Basic Correlation (Pearson and Spearman), DC[35], GENIE3[36]), each varying in their inference strategies, integrated with a consensus approach and graph intersection to identify the conserved miRNA-miRNA interaction signature across a group of diseases (cancers, in this case). The identified signatures were validated via manual literature search and were found to be associated within the classes of the selected cancers, demonstrating the efficacy of the method. Under validation, we retrieved the PMIDs reporting the associations from the *PhenomiR* database and also performed a manual literature search in the *PubMed Central* database to separately corroborate our results, as displayed in Fig. 7.

Our results show that, the expression profile of hsa-mir-30a, hsa-mir-181a-1, and hsa-mir-29c could be a signature for gastrointestinal cancers that comprises of esophageal, gastroesophageal, gastrointestinal, gastric, and colorectal cancers (Fig. 7). These miRNAs are already reported to be associated with these cancers[45–48]. miRNAs (hsa-mir-30a, hsa-mir-29c, hsa-mir-181a-1) displayed the same trend of expression in a study of esophageal adenocarcinoma (EAC) and Barrett's esophagus (BE) and were differentially up-regulated in both the disease tissues. hsa-mir-181a and hsa-mir-29c showed higher expression levels in EAC to that of BE with high grade dysplasia[48]. Studies have also reported hsa-mir-181a, hsa-mir-30a and hsa-mir-29c being overexpressed in esophagela carcinoma (EC) and hsa-mir-29c to be underexpressed in EC[49,50] and therefore, this group of miRNAs may be considered for developing a pan-diagnostic tool for the aforementioned cancers.

We identified that hsa-mir-221, hsa-mir-222, hsa-mir-155, hsa-mir-224, hsa-mir-181a-1, and hsa-mir-181b-1 make the signature for endocrine cancers (hepatocellular, pancreatic, and thyroid cancers) (Fig. 7). Reports suggest that these miRNAs are predominantly associated with this group of cancers[51–54]. In another study analyzing molecular signatures for aggressive pancreatic cancer, all the miRNAs (hsa-mir-221, hsa-mir-222, hsa-mir-155, hsa-mir-224, hsa-mir-181a-1, and hsa-mir-181b-1) were significantly altered due to chronic exposure to conventional anti-cancer drugs[55]. A large-scale meta-analysis investigating candidate miRNA biomarkers for pancreatic ductal adenocarcinoma (PDAC) across eleven miRNA expression profiling studies, reported all the miRNAs to be up-regulated and having a consistent direction of change. miRNAs hsa-mir-221, hsa-mir-222, hsa-mir-155 were reported to be upregulated together in at least five of these studies with a consistent direction. Among them, miRNAs hsa-mir-221, hsa-mir-155 were identified as part of a meta-signature and biomarkers for PDAC[56]. Studies also report all these miRNAs to be associated with lung cancer[57]. Thus this set of miRNAs may be used/tested as a diagnostic tool for all the endocrine cancers considered here.

Seven miRNAs (hsa-mir-29b-1, hsa-mir-146a, hsa-mir-20a, hsa-mir-126, hsa-mir-99a, hsa-mir-199b and hsa-mir-130a) that are well documented for their association with various kinds of leukemia[54,58–63] are found to form the signature component of leukemia from our analysis (Fig. 7). miRNAs (hsa-mir-29b-1, hsa-mir-20a, hsa-mir-126, hsa-mir-146a, hsa-mir-199b) were differentially expressed in a blood stem cell study in which the blood stem cells were treated with plerixafor and granulocyte colony-stimulating factor. The miRNAs were

recorded to be expressed in this treated cell study analyzing acute lymphocytic leukemia conditions[64]. miRNAs (hsa-mir-126, hsa-mir-130a, hsa-mir-99a, hsa-mir-146a, hsa-mir-199b) have also been reported to express together in a myeloid cell study exploring transcription factor binding site motifs[65]. Therefore, this signature group of miRNAs can be potentially used as a screening or diagnostic tool for a range of different types of leukemia.

In case of neurone cancers (neuroblastoma, medulloblastoma, and glioblastoma) we detected two signatures: i) hsa-mir-323, hsa-mir-129-1, hsa-mir-137, hsa-mir-330, hsa-mir-149, hsa-mir-107, hsa-mir-30c-1, hsa-mir-181b-1 and ii) hsa-mir-30b, hsa-mir-331, hsa-mir-150, hsa-let-7a-1 (Fig. 7). Regarding the first signature network component, hsa-mir-137, hsa-mir-330, hsa-mir-149, hsa-mir-107, hsa-mir-181b were among the miRNAs whose experimentally validated targets (such as CTBP1, CDC42, CDK6, E2F1, VEGFA, AKT1, KAT2B) affect the pathways which play a crucial role in glioblastoma biology. Deregulations of hsa-mir-137, hsa-mir-330 and hsa-mir-149 lead to effects in the glioma de novo pathway, VEGF signaling pathway and Notch signaling pathway[66]. Among the miRNAs reported in the second signature component, hsa-mir-330 and hsa-mir-30b are among the top ten miRNAs having least coefficient of variation in the expression of benign kidney tumor and hsa-mir-150 is differentially expressed in metastatic clear cell renal cell carcinoma[67].

Comparing our results with other similar works has been challenging, primarily because there are not many studies that have reported direct miRNA-miRNA co-regulations across these disease classes. Similar studies[13,20,68] have used different disease and miRNA data sets which makes a one-to-one comparison challenging. In some previous works, miRNA-miRNA regulatory associations have been deduced based on the semantic similarities between the associated diseases[69] and based on the analysis of shared transcription factors, common targets, KEGG pathway analysis and corroboration from literature[20]. However, none of these methods allow for a network-level miRNA-miRNA analysis for a variety of diseases and hence cannot be used for comparison purposes to the predicted interaction networks in this paper.

Online analysis and visualization of results is an aid to the research community. Along these lines, several network analysis and visualization tools have been developed, such as *VisANT* for integrative online visual analysis of biological networks and pathways[70], *miRegulome* for miRNA regulome visualization and analysis[12] and *miRNet* for functional analysis of miRNAs within a high-performance network visual analytics system[71] among others. However, no tool is available so far which can perform an online visualization and analysis of signature miRNAs across multiple diseases. The *miRsig* tool developed here bridges this gap and provides an intuitive analysis and visualization of core/signature miRNA-miRNA interaction components for several diseases.

Conclusion

In this work, we have developed a novel consensus-based network analysis pipeline to identify disease-specific miRNA-miRNA interactions by combining the expression profiles of various miRNAs in specific diseases. This method can effectively identify the signature/core miRNA-miRNA interactions for a group of diseases; here tested on cancer. These signature miRNAs may have potential use for diagnostic, prognostic, or therapeutic applications for a group of related diseases such as cancers. The predicted miRNA-miRNA signature patterns were extensively validated by the PMIDs reported in the *PhenomiR* database as well as an independent manual literature search from *PubMed Central*. *miRsig* thus provides a powerful prediction and visualization tool for the identification of core/signature miRNA-miRNA interactions amongst a number of diseases. Our future work includes investigating the (i) $miRNA_{same}Disease_{different}$ category of interactions to study the dynamics of similar miRNAs across multiple diseases and also (ii) the $miRNA_{different}Disease_{different}$ category of interactions to understand the evolution of diseases based on the underlying miRNA expression patterns. As miRNAs may potentially serve as biomarkers for a wide variety of diseases, our proposed pipeline may motivate the study of several interesting questions both for particular diseases or across multiple diseases.

References

1. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature Reviews Genetics* **9,** 102–114 (2008).
2. Place, R. F., Li, L.-C., Pookot, D., Noonan, E. J. & Dahiya, R. Microrna-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences* **105,** 1608–1613 (2008).
3. Saraiya, A. A., Li, W. & Wang, C. C. Correction: Transition of a microrna from repressing to activating translation depending on the extent of base pairing with the target. *PloS one* **8** (2013).
4. Tüfekci, K. U., Meuwissen, R. L. J. & Genç, Ş. The role of microrna s in biological processes. *miRNomics: MicroRNA Biology and Computational Analysis* 15–31 (2014).
5. Blenkiron, C. & Miska, E. A. mirnas in cancer: approaches, aetiology, diagnostics and therapy. *Human molecular genetics* **16,** R106–R113 (2007).
6. Ardekani, M. A. & Moslemi Naeini, M. The role of microrna s in human diseases. *Avicenna journal of medical biotechnology* **2,** 161–180 (2011).
7. Ye, S. *et al.* Bioinformatics method to predict two regulation mechanism: Tf-mirna-mrna and lncrna-mirna-mrna in pancreatic cancer. *Cell biochemistry and biophysics* **70,** 1849–1858 (2014).
8. Barh, D., Malhotra, R., Ravi, B. & Sindhurani, P. Microrna let-7: an emerging next-generation cancer therapeutic. *Current Oncology* **17,** 70 (2010).
9. Tang, R. *et al.* Mouse mirna-709 directly regulates mirna-15a/16-1 biogenesis at the posttranscriptional level in the nucleus: evidence for a microrna hierarchy system. *Cell research* **22,** 504–515 (2012).
10. Shi, B., Zhu, M., Liu, S. & Zhang, M. Highly ordered architecture of microrna cluster. *BioMed research international* **2013** (2013).
11. Lu, M. *et al.* An analysis of human microrna and disease associations. *PloS one* **3,** e3420 (2008).
12. Barh, D. *et al.* miregulome: a knowledge-base of mirna regulomics and analysis. *Scientific reports* **5** (2015).
13. Hamilton, M. P. *et al.* Identification of a pan-cancer oncogenic microrna superfamily anchored by a central core seed motif. *Nature communications* **4** (2013).
14. Yuan, D. *et al.* Enrichment analysis identifies functional microrna-disease associations in humans. *PloS one* **10,** e0136285 (2015).
15. Zou, Q. *et al.* Prediction of microrna-disease associations based on social network analysis methods. *BioMed research international* **2015,** 810514 (2015).

16. Chen, H. & Zhang, Z. Similarity-based methods for potential human microrna-disease association prediction. *BMC medical genomics* **6,** 12 (2013).
17. Liao, B., Ding, S., Chen, H., Li, Z. & Cai, L. Identifying human microrna-disease associations by a new diffusion-based method. *Journal of bioinformatics and computational biology* **13,** 1550014 (2015).
18. Shi, H. *et al.* Integration of multiple genomic and phenotype data to infer novel mirna-disease associations. *PloS one* **11,** e0148521 (2016).
19. Yang, J.-H. & Qu, L.-H. Discovery of microrna regulatory networks by integrating multidimensional high-throughput data. In *MicroRNA Cancer Regulation* 251–266 (Springer, 2013).
20. Song, R., Catchpoole, D. R., Kennedy, P. J. & Li, J. Identification of lung cancer mirna-mirna co-regulation networks through a progressive data refining approach. *Journal of Theoretical Biology* (2015).
21. Xu, J. *et al.* Prioritizing candidate disease mirnas by topological features in the mirna target-dysregulated network: Case study of prostate cancer. *Molecular cancer therapeutics* **10,** 1857–1866 (2011).
22. Xu, C. *et al.* Prioritizing candidate disease mirnas by integrating phenotype associations of multiple diseases with matched mirna and mrna expression profiles. *Mol. BioSyst.* **10,** 2800–2809 (2014).
23. Yoon, S. & De Micheli, G. Prediction of regulatory modules comprising micrornas and target genes. *Bioinformatics* **21,** ii93–ii100 (2005).
24. Bandyopadhyay, S., Mitra, R., Maulik, U. & Zhang, M. Q. Development of the human cancer microrna network. *Silence* **1,** 1 (2010).
25. Nalluri, J. J. *et al.* Dismira: Prioritization of disease candidates in mirna-disease associations based on maximum weighted matching inference model and motif-based analysis. *BMC Genomics* **16,** S12 (2015).
26. Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData mining* **4,** 1 (2011).
27. Cho, H., Berger, B. & Peng, J. Reconstructing causal biological networks through active learning. *PloS one* **11,** e0150611 (2016).
28. Nagarajan, N. & Kingsford, C. Giraf: robust, computational identification of influenza reassortments via graph mining. *Nucleic acids research* gkq1232 (2010).
29. Pati, A., Vasquez-Robinet, C., Heath, L. S., Grene, R. & Murali, T. Xcisclique: analysis of regulatory bicliques. *BMC bioinformatics* **7,** 1 (2006).
30. Ruepp, A. *et al.* Phenomir: a knowledgebase for microrna expression in diseases and biological processes. *Genome biology* **11,** R6 (2010).
31. Surowiecki, J. *The wisdom of crowds* (Anchor, 2005).
32. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nature methods* **9,** 796–804 (2012).
33. Faith, J. J. *et al.* Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology* **5,** e8 (2007).
34. Meyer, P., Marbach, D., Roy, S. & Kellis, M. Information-theoretic inference of gene networks using backward elimination. In *Biocomp* 700–705 (2010).
35. Guo, X., Zhang, Y., Hu, W., Tan, H. & Wang, X. Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PloS one* **9,** e87446 (2014).
36. Irrthum, A., Wehenkel, L., Geurts, P. *et al.* Inferring regulatory networks from expression data using tree-based methods. *PloS one* **5,** e12776 (2010).
37. Bostock, M. Data-driven documents. http:https://d3js.org/. Accessed: 2016-04-18.
38. Afonso-Grunz, F. & Müller, S. Principles of mirna-mrna interactions: Beyond sequence complementarity. *Cellular and Molecular Life Sciences* **72,** 3127–3141 (2015).
39. Guo, L., Sun, B., Wu, Q., Yang, S. & Chen, F. mirna-mirna interaction implicates for potential mutual regulatory pattern. *Gene* **511,** 187–194 (2012).
40. Arora, S., Rana, R., Chhabra, A., Jaiswal, A. & Rani, V. mirna-transcription factor interactions: a combinatorial regulation of gene expression. *Molecular genetics and genomics* **288,** 77–87 (2013).
41. Hua, L., Xia, H., Zhou, P., Li, D. & Li, L. Combination of microrna expression profiling with genome-wide snp genotyping to construct a coronary artery disease-related mirna-mirna synergistic network. *Bioscience trends* **8,** 297–307 (2014).
42. Wu, B. *et al.* Dissection of mirna-mirna interaction in esophageal squamous cell carcinoma. *PloS one* **8,** e73191 (2013).
43. Alshalalfa, M. Microrna response elements-mediated mirna-mirna interactions in prostate cancer. *Advances in bioinformatics* **2012** (2012).
44. Xu, J. *et al.* Mirna-mirna synergistic network: construction via co-regulating functional modules and disease mirna topological features. *Nucleic acids research* **39,** 825–836 (2011).
45. Monzo, M. *et al.* Overlapping expression of microrna in human embryonic colon and colorectal cancer. *Cell research* **18,** 823–833 (2008).
46. Ueda, T. *et al.* Relation between microrna expression and progression and prognosis of gastric cancer: a microrna expression analysis. *The lancet oncology* **11,** 136–146 (2010).
47. Zhang, Y. *et al.* Profiling of 95 micrornas in pancreatic cancer cell lines and surgical specimens by real-time pcr analysis. *World journal of surgery* **33,** 698–709 (2009).
48. Yang, H. *et al.* Microrna expression signatures in barrett's esophagus and esophageal adenocarcinoma. *Clinical Cancer Research* **15,** 5744–5752 (2009).
49. Guo, Y. *et al.* Distinctive microrna profiles relating to patient survival in esophageal squamous cell carcinoma. *Cancer research* **68,** 26–33 (2008).
50. Zhou, S.-L. & Wang, L.-D. Circulating micrornas: novel biomarkers for esophageal cancer. *World J Gastroenterol* **16,** 2348–2354 (2010).
51. Jiang, J. *et al.* Association of microrna expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival. *Clinical Cancer Research* **14,** 419–427 (2008).
52. Roldo, C. *et al.* Microrna expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior. *Journal of Clinical Oncology* **24,** 4677–4684 (2006).
53. Nikiforova, M. N., Tseng, G. C., Steward, D., Diorio, D. & Nikiforov, Y. E. Microrna expression profiling of thyroid tumors: biological significance and diagnostic utility. *The Journal of Clinical Endocrinology & Metabolism* **93,** 1600–1608 (2008).
54. Jiang, J., Lee, E. J., Gusev, Y. & Schmittgen, T. D. Real-time expression profiling of microrna precursors in human cancer cell lines. *Nucleic acids research* **33,** 5394–5403 (2005).
55. Ali, S., Almhanna, K., Chen, W., Philip, P. A. & Sarkar, F. H. Differentially expressed mirnas in the plasma may provide a molecular signature for aggressive pancreatic cancer. *Am J Transl Res* **3,** 28–47 (2010).
56. Ma, M.-Z. *et al.* Candidate microrna biomarkers of pancreatic ductal adenocarcinoma: meta-analysis, experimental validation and clinical significance. *Journal of Experimental & Clinical Cancer Research* **32,** 1 (2013).
57. Leidinger, P., Keller, A. & Meese, E. Micrornas-important molecules in lung cancer research. *Frontiers in genetics* **2,** 104 (2012).
58. Zanette, D. *et al.* mirna expression profiles in chronic lymphocytic and acute lymphocytic leukemia. *Brazilian Journal of Medical and Biological Research* **40,** 1435–1440 (2007).
59. Garzon, R. *et al.* Microrna signatures associated with cytogenetics and prognosis in acute myeloid leukemia. *Blood* **111,** 3183–3189 (2008).

60. Garzon, R. *et al.* Distinctive microrna signature of acute myeloid leukemia bearing cytoplasmic mutated nucleophosmin. *Proceedings of the National Academy of Sciences* **105,** 3945–3950 (2008).
61. Colquhoun, J. A. S. *With the Kurram Field Force, 1878-79* (WH Allen & Company, 1881).
62. Contreras, J. R. *et al.* Microrna-146a modulates b-cell oncogenesis by regulating egr1. *Oncotarget* **6,** 11023 (2015).
63. Favreau, A. J., McGlauflin, R. E., Duarte, C. W. & Sathyanarayana, P. mir-199b, a novel tumor suppressor mirna in acute myeloid leukemia with prognostic implications. *Experimental hematology & oncology* **5,** 1 (2016).
64. Donahue, R. E. *et al.* Plerixafor (amd3100) and granulocyte colony-stimulating factor (g-csf) mobilize different cd34+ cell populations based on global gene and microrna expression signatures. *Blood* **114,** 2530–2541 (2009).
65. Jansen, B. J. *et al.* Microrna genes preferentially expressed in dendritic cells contain sites for conserved transcription factor binding motifs in their promoters. *BMC genomics* **12,** 1 (2011).
66. Singh, S. K., Vartanian, A., Burrell, K. & Zadeh, G. A microrna link to glioblastoma heterogeneity. *Cancers* **4,** 846–872 (2012).
67. Wu, X. *et al.* Identification of a 4-microrna signature for clear cell renal cell carcinoma metastasis and prognosis. *PloS one* **7,** e35661 (2012).
68. Bandyopadhyay, S. & Bhattacharyya, M. Analyzing mirna co-expression networks to explore tf-mirna regulation. *BMC bioinformatics* **10,** 163 (2009).
69. Wang, D., Wang, J., Lu, M., Song, F. & Cui, Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* **26,** 1644–1650 (2010).
70. Hu, Z. *et al.* Visant 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic acids research* **41,** W225–W231 (2013).
71. Lab, X. miRNet, network-based visual analysis of miRNAs, targets and functions. http://www.mirnet.ca/ (2015). [Online; Last accessed 4-April-2016].

Acknowledgements
We thank Daniel Marbach, Ph.D. for promptly responding to our queries regarding the network inference tools hosted on the GenePattern website.

Author Contributions
J.J.N. and P.G. conceived, conceptualized and implemented the computational methods. D.B. and V.A. provided biological insights and validated *miRsig* tool. J.J.N., D.B. and P.G. wrote the manuscript. All authors have read and approved the final manuscript.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Nalluri, J. J. *et al. miRsig*: a consensus-based network inference methodology to identify pan-cancer miRNA-miRNA interaction signatures. *Sci. Rep.* **7**, 39684; doi: 10.1038/srep39684 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

IV.3.1 Conclusions from this research/ Chapter-3

i. We have developed miRsig (miRNA Signature) online tool that analyses miRNA-miRNA interaction signatures
ii. A novel consensus-based network inference is developed from six network inference algorithms along with graph based approach is used in this analysis pipeline.
iii. The algorithm is applied to miRNA-disease expression profiles for identification of miRNA-miRNA interaction signatures
iv. Disease-specific miRNA-miRNA interaction networks as well as common conserved miRNA-miRNA interaction network signatures in multiple diseases can also be constructed, visualized, and validated using the methodology
v. We applied the miRsig methodology to identify pan-cancer miRNA signatures
vi. The identified pan-cancer signatures:
    a. Gastrointestinal cancers (esophageal, gastroesophageal, gastrointestinal, gastric, and colorectal): hsa-mir-30a, hsa-mir-181a-1, and hsa-mir-29c
    b. Endocrine cancers (hepatocellular, pancreatic, and thyroid cancers): hsa-mir-221, hsa-mir-222, hsa-mir-155, hsa-mir-224, hsa-mir-181a-1, and hsa-mir-181b-1
    c. Leukemias: hsa-mir-29b-1, hsa-mir-106a, hsa-mir-20a, hsa-mir-126, and hsa-mir-130a.
    d. Neurone cancers (neuroblastoma, medulloblastoma, and glioblastoma) two signatures:
        a) hsa-mir-323, hsa-mir-129-1, hsa-mir-137, hsa-mir-330, hsa-mir-149, hsa-mir-107, hsa-mir-30c-1, hsa-mir-181b-1 and
        b) hsa-mir-30b, hsa-mir-331, hsa-mir-150, hsa-let-7a-1
vii. Literature mining shows the identified signatures are 100% accurate proving the accuracy of the method.
viii. This novel computational method can help in identification of early deregulated miRNA signatures in diseases, therefore the screening or early diagnostic biomarkers can be discovered.

IV.4 Chapter IV: Research Article

Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis*, and *E. coli* targeted by Piper betel compounds.

Debmalya Barh, Krishnakant Gupta, Neha Jain, Gourav Khatri, Nidia Leo´n-Sicairos, Adrian Canizalez-Roman, Sandeep Tiwari, Ankit Verma, Sachin Rahangdale, Syed Shah Hassan, Anderson Rodrigues dos Santos, Amjad Ali, Luis Carlos Guimara˜es, Rommel Thiago Juca´ Ramos, Pratap Devarapalli, Neha Barve, Marriam Bakhtiar, Ranjith Kumavath, Preetam Ghosh, Anderson Miyoshi, Artur Silva, Anil Kumar, Amarendra Narayan Misra, Kenneth Blum, Jan Baumbach, **Vasco Azevedo**

Integrative Biology (Camb). 2013 Mar;5(3):495-509. doi: 10.1039/c2ib20206a.
[PMID: 23288366]
Impact Factor: 3.5 (2013)

Understanding protein–protein interactions (PPI) and host–pathogen interactions are important aspects to unveil molecular insights of bacterial pathogenesis and target identification. In this chapter we have discussed a novel integrated bioinformatics strategy combining PPI, host-pathogen interactions, and subtractive genomics to identify common conserved targets in *C. pseudotuberculosis (*Cp), *C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis,* and *E. coli.* The intra-species PPIs of for four Cp strains (Cp FRC41, Cp 316, Cp 3/99-5, and Cp P54B96) is identified first time. Inter-species bacterial PPI based conserved common host–pathogen interactions (HP-PPI) were determined and validated. Network analysis strategies and subtraction genomics approaches were applied to the HP-PPI networks to identify acetate kinase (Ack) as a common conserved target in *Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans, E. coli*, and all four *Cp* strains. Piper betel derived Piperdardine and Dehydropipernonaline are predicted to have superior effects compared to Penicillin and Ceftiofur on Ack as per our virtual screening. Piperdardine inhibits *E. coli* O157:H7 growth similar to penicillin and may also work on other studied pathogens in a similar way. Thus, Ack could be a broad spectrum target and Piperdardine is a potential targeting molecule that to be tested *in vitro* and *in vivo* against all these pathogens.

Integrative Biology

RSC Publishing

PAPER

Conserved host–pathogen PPIs† Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis*, *C. diphtheriae*, *M. tuberculosis*, *C. ulcerans*, *Y. pestis*, and *E. coli* targeted by *Piper betel* compounds

Debmalya Barh,‡*[ab] Krishnakant Gupta,[ac] Neha Jain,[a] Gourav Khatri,[ac] Nidia León-Sicairos,[d] Adrian Canizalez-Roman,[d] Sandeep Tiwari,[a] Ankit Verma,[ac] Sachin Rahangdale,[ac] Syed Shah Hassan,[e] Anderson Rodrigues dos Santos,[e] Amjad Ali,[e] Luis Carlos Guimarães,[e] Rommel Thiago Jucá Ramos,[f] Pratap Devarapalli,[g] Neha Barve,[ac] Marriam Bakhtiar,[e] Ranjith Kumavath,[g] Preetam Ghosh,[ah] Anderson Miyoshi,[e] Artur Silva,[f] Anil Kumar,[c] Amarendra Narayan Misra,[bi] Kenneth Blum,[ajkl] Jan Baumbach[m] and Vasco Azevedo‡[e]

Cite this: *Integr. Biol.*, 2013, **5**, 495

Received 27th August 2012,
Accepted 5th November 2012

DOI: 10.1039/c2ib20206a

www.rsc.org/ibiology

Although attempts have been made to unveil protein–protein and host–pathogen interactions based on molecular insights of important biological events and pathogenesis in various organisms, these efforts have not yet been reported in *Corynebacterium pseudotuberculosis* (*Cp*), the causative agent of Caseous Lymphadenitis (CLA). In this study, we used computational approaches to develop common conserved intra-species protein–protein interaction (PPI) networks first time for four *Cp* strains (*Cp FRC41*, *Cp 316*, *Cp 3/99-5*, and *Cp P54B96*) followed by development of a common conserved inter-species bacterial PPI using conserved proteins in multiple pathogens (*Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, *E. coli*, and all four *Cp* strains) and *E. Coli* based experimentally validated PPI data. Furthermore, the interacting proteins in the common conserved inter-species bacterial PPI were used to generate a conserved host–pathogen interaction (HP-PPI) network considering human, goat, sheep, bovine, and horse as hosts. The HP-PPI network was validated, and acetate kinase (Ack) was identified as a novel broad spectrum target. Ceftiofur, penicillin, and two natural compounds derived from *Piper betel* were predicted to inhibit Ack activity. One of these *Piper betel* compounds found to inhibit *E. coli O157:H7* growth similar to penicillin. The target specificity of these *betel* compounds, their effects on other studied pathogens, and other *in silico* results are currently being validated and the results are promising.

[a] *Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal-721172, India. E-mail: dr.barh@gmail.com; Fax: +91-944 955 0032; Tel: +91-944 955 0032*

[b] *Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, Orissa, India*

[c] *School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India*

[d] *Unidad de investigacion, Facultad de Medicina, Universidad Autónoma de Sinaloa. Cedros y Sauces, Fraccionamiento Fresnos, Culiacán Sinaloa 80246, México*

[e] *Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

[f] *Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, PA, Brazil*

[g] *Department of Genomic Science, School of Biological Sciences, Riverside Transit Campus, Central University of Kerala, Kasaragod, India*

[h] *Department of Computer Science and Center for the Study of Biological Complexity, Virginia Commonwealth University, 401 West Main Street, Room E4234, P.O. Box 843019, Richmond, Virginia 23284-3019, USA*

[i] *Center for Life Sciences, School of Natural Sciences, Central University of Jharkhand, Ranchi, Jharkhand State, India*

[j] *University of Florida, College of Medicine, Gainesville, Florida, USA*

[k] *Global Integrated Services Unit University of Vermont Center for Clinical & Translational Science, College of Medicine, Burlington, VT, USA*

[l] *Dominion Diagnostics LLC, North Kingstown, Rhode Island, USA*

[m] *Computational Biology Group, Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark*

† Electronic supplementary information (ESI) available: Supplementary Tables 1–8. See DOI: 10.1039/c2ib20206a

‡ DB and VA conceived the idea; DB designed the study, collected and analyzed primary data to finalize the protocol, coordinated and leaded the entire project, and wrote the manuscript. DB, KG, NJ, GK, ST, AV, and SR performed all *in silico* analyses; SSH, ARS, AA, LCG, and ATJR performed *Cp* genome annotation and cross checked all other analyses; PD, RK, MB, NB, and PG cross checked all analyses; NLS and ACR conducted microbial experiments with betel compounds; AK, KB, ANM, AM, PG, JB, and VA provided technical consultations and reviewed the manuscript. All authors have read and approved the final manuscript.

Insight, innovation, integration

Here, for the first time we represent the intra-species PPIs in *C. pseudotuberculosis* (*Cp*). Further, a novel method was used to develop common conserved inter-species bacterial PPIs for *C. pseudotuberculosis*, *Y. pestis*, *M. tuberculosis*, *C. diphtheriae*, *C. ulcerans*, *C. glutamicum*, and *E. coli* (pathogenic, nonpathogenic, closed and distant taxa) to identify the conserved common essential PPIs in these bacteria. This inter-species bacterial PPI was then used to make conserved common host–pathogen interactions. Using network analysis strategies and subtraction genomics approaches, from this conserved common host–pathogen interactions; Ack was identified as a key target for all these bacteria. Virtual screening shows Penicillin and Ceftiofur can inhibit Ack. However, *Piper betel* derived Piperdardine and Dehydropipernonaline are predicted to have similar or superior effects compared to Penicillin and Ceftiofur on Ack. Piperdardine inhibits *E. coli O157:H7* growth similar to penicillin and can also work on other pathogens in a similar way.

Introduction

Protein–protein interactions (PPIs) are crucial events in several biological processes. PPI-based decoding of the functionality of uncharacterized proteins can reveal unknown molecular mechanisms behind important biological events within a cell or at the system level.[1,2] Therefore, PPIs of an entire proteome or between a set of proteins in a pathogen and its corresponding host can be useful in identifying precise molecular mechanisms of host–pathogen interactions, thereby leading to the development of effective drug targets against the pathogen.[3–5] Initial computational approaches for the prediction of PPIs were based on the structural context of proteins. However, in the post-genomic era, the focus has shifted, and sequence information is now used.[6,7] The availability of genomic and proteomic data and the advent of yeast two-hybrid, affinity purification, mass spectrometry, and other high-throughput techniques have tremendously enriched the field. Recently, a number of computational approaches have also been developed to facilitate the prediction and study of these ubiquitous interactions. A number of *in silico* approaches were recently reviewed that highlight the use of genomic, structural, and biological contexts of proteins and genes in complete genomes for PPI predictions and determination of the functional relationship among them.[8] Using these approaches, the development of highly reliable PPIs in several organisms including yeast[9] and human[10] are close to completion. However, false-positive interactions are a concern.[11,12] Similarly, sequence-based computational methods including gene neighborhood,[13] phylogenetic profiles,[14] gene fusion,[15] co-evolution,[16] and domain interactions,[17] along with several newly developed methods, have been used to generate genome-/proteome-wide interactions in a number of organisms including, *M. tuberculosis*[18] and *E. coli*.[19] Genomic sequences are used as the primary data sources in these prediction techniques, which assume that evolutionary co-inherited gene pairs have a functional association.[20,21] Similarly, amino acid (AA) sequence-based PPIs identify interacting protein pairs that have specific AA residues due to their co-evolution or binding to one another.[22] Yeats *et al.* have catalogued the commonly occurring domains for PPIs.[23–25] However, in general, a PPI denotes the binding of proteins to other proteins.

Concurrently, *in silico* host–pathogen interactions have been reported in many organisms, including *Plasmodium*,[26,27] *M. tuberculosis*,[28] and *Streptococcus*.[29] Combined computational and yeast two-hybrid based approaches have been recently published for *B. anthracis*, *F. tularensis*, and *Y. pestis* PPIs.[5] Although, it gives only 20% positive interactions and therefore produces a high degree of false-negative interaction,[30] the yeast two-hybrid method and related high-throughput and computational interaction data have been analyzed to identify targets in many pathogens.

Although extensive studies have been conducted for host–pathogen interactions and target identification in *M. tuberculosis*[31–34] and *Corynebacterium diphtheriae*,[35–38] another member of the *Corynebacterium, Mycobacterium, Nocardia,* and *Rhodococcus* (CMNR) group of pathogens, *C. pseudotuberculosis*, remains uninvestigated with respect to both its PPI and host–pathogen interactions. *C. pseudotuberculosis* causes Caseous Lymphadenitis (CLA) or "cheesy gland" in small ruminants worldwide, which can result in a significant economic loss.

CLA is characterized by the formation of external or internal abscesses, chronic limb infections (lymphangitis) and lymphadenitis.[39,40] It also infects visceral organs such as the liver, spleen, kidneys and lungs.[41] Although the bacterium rarely infects humans, there are reports of human lymphadenitis, and clinical strains have been isolated.[42] Other important pathogens in the CMNR group, *M. tuberculosis* and *C. diphtheriae*, cause tuberculosis and diphtheria, respectively. According to the WHO, approximately 1.7 million people died from tuberculosis in 2009 and 50,000 died from diphtheria in 2004. *Yersinia pestis* causes plague and poses a threat for use in bioterrorism.[43] Most of its isolates are derived from *Y. pseudotuberculosis*,[44] and lymphadenitis or lymphadenopathy caused by *Cp* is one of the symptoms of a *Y. pestis*[45–47] and *M. tuberculosis*[48,49] infection.

Here, for the first time, using a combination of comparative, functional, and phylogenomics approaches, supported by published, experimentally validated data we report (a) a probable conserved PPIs in the *Cp* proteome. (b) Further, we created proteome-wide common conserved PPIs for a number of pathogenic and non-pathogenic bacteria (*C. pseudotuberculosis*, *C. diphtheriae, C. ulcerans, M. tuberculosis*, *Y. pestis*, and *E. coli*). (c) Thereafter, the proteins involved in this common conserved intra-species bacterial PPIs were used to generate host–pathogen interactions considering human, goat, sheep, and horse as hosts. This host–pathogen PPI was based on experimentally validated published host–pathogen interactions data. (d) By analyzing the host–pathogen interaction networks, we identified common conserved targets in these pathogens.

(e) Finally, we use the identified targets to develop broad spectrum drugs from an existing antibiotic regime and phytochemicals derived from *Piper betel*.

Materials and methods

Selection of highly identical conserved proteins in *Cp*, other bacteria, and hosts

Selection of conserved genes for intra-species *Cp* PPI. In this work, we aimed to develop PPIs based on sequences. Therefore, highly identical common conserved proteins of Cp were selected using comparative genomics/proteomics approaches using the BLAST tool.[50] As there is no report on *Cp* PPIs so far, first we approached to develop intra-species common conserved PPIs for four *Cp* strains (strain FRC41, 316, 3/99-5, and P54B96) that were isolated from four different hosts and recently sequenced. The strain *FRC41* (biovar *ovis*) was isolated from a human; strain *316* (biovar *equi*) was isolated from a horse; strain *3/99-5* (biovar *ovis*) was isolated from a sheep; and strain *P54B96* (biovar *ovis*) was isolated from an antelope. Highly conserved and common proteins of these four strains were selected using BLASTp cut off values: $E = 0.0001$ and $\geq 80\%$ identity. Such BLAST parameters were used to select identical sequences from different strains of a species.[51]

Selection of conserved genes for inter-species bacterial PPI. Next, the highly identical common conserved genes across a wide range of pathogenic and non-pathogenic bacteria from the same and distant taxa (*E. coli*, *Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, *C. glutamicum,* and all four *Cp* strains) were selected using the BLAST option available in the Prokaryotic Sequence homology Analysis (PSAT) Tool.[52] The PSAT tool was selected because it compares gene neighborhoods, gene clusters, homologs, and orthologs among multiple bacterial genomes in a single run. It also accounts information of gene context including weak alignment scores therefore provides better sensitivity compared to other available comparative analysis methods. To get the homolog list we used *Y. pestis* genome as reference and compared with *M. tuberculosis*, *C. diphtheriae, C. glutamicum, and E. coli.* The BLAST score thresholds were set to: $E = 0.01$, bit score ≥ 100, identity $\geq 35\%$ that was used in our previous report to identify homologs essential genes.[51] The common homolog genes in these bacteria were selected and further tested for their presence in *C. ulcerans* and pool of conserved common genes of four Cp strains, and other selected bacterial strains (Table S1, ESI†) using NCBI BLASTp with same parameters. Finally, the common conserved genes that are present in all these selected bacteria were collected and the common conserved *E. coli* K12 genes were used in further analysis as most of the required experimentally validated data are available for this species. The list of bacteria used in this analysis is represented in Table S1 (ESI†).

Selection of conserved genes in hosts. A range of hosts (human, goat, sheep, bovine, and horse) were selected based on the commonality of the pathogenesis from the selected pathogenic organisms. The conserved genes in these hosts were identified using the general NCBI BLASTp program (cut off values: $E = 0.01$, bit score ≥ 100, identity $\geq 35\%$). In all cases, the name of the protein or the functionality was matched during the selection.

Classification and functional annotations of common conserved bacterial proteins

The common conserved inter-species bacterial proteins were functionally classified as per the Clusters of Orthologous Groups classifications (COGs).[53] *E. coli* genes were subjected to the COGNITOR BLAST (using default parameters) to group the proteins under each COG functional classifications. Each class of COG consists of evolutionary conserved (at least 3 distant lineages) individual protein or groups of paralogs having similar cellular function under 18 classes. Therefore, the COG database and its classification are very useful in comparative, evolutionary, and phylogenetic analysis of new genome or gene to assign their biological functions.[54] Additionally, the proteins were annotated for their functionality using the NCBI and UniProt[55] databases. Pathogenicity islands (PAIs) encode various virulence factors including type III secretion system proteins of a bacterium that are required for infection. Hence, to check the virulence of the common bacterial proteins, each protein was tested with the help of the BLASTp option at the Pathogenicity Island Database (PAIDB) server.[56] The PIDB contains all reported PAIs from 497 pathogenic bacterial strains. The database also contains more than 310 predicted PAIs from 118 prokaryots. To map the pathway involvements of these conserved proteins, we used the KEGG pathway database.[57]

Generation of intra- and inter-species bacterial PPI, validation, and analysis

The bacterial PPIs were developed and analyzed using VisANT 3.0.[58,59] VisANT is an integrative platform for developing PPIs and network prediction, construction, editing, analysis, and visualization. It develops biological interactions based on data derived from 102 methods (computational and both high- and low- throughput experimental methods). The tool can integrate and mine KEGG[57] pathways in biological interactions and multi-scale analysis and visualization of multiple pathways can also be done.

Intra-species PPI of four Cp. The *Cp* genome is not available in VisANT. Therefore, a combination of genomic context-based methods including comparative and phylogenetic profiling,[14] gene or domain fusion,[15] and gene neighborhood methods[13] were used to develop the intra-species PPIs for the conserved *C. pseudotuberculosis* proteins of the selected four *Cp* strains. The resultant PPIs along with KEGG pathways were incorporated in the VisANT for network analysis and *in silico* validation of the intra-species *Cp* PPIs.

Inter-species bacterial PPI. The common conserved inter-species bacterial PPIs for *Y. pestis, E. coli, M. tuberculosis, C. glutamicum, C. diphtheriae,* and *C. ulcerans* and all four *Cp* strains were developed using VisANT. We used common conserved *E. coli K12* proteins to develop this PPI as multiple experimentally validated data for *E. coli* PPIs are available in VisANT. Additionally, the VisANT generated *E. coli* based conserved PPIs were evaluated using anti-tag co-immunoprecipitation-based

binding PPIs from *E. coli*.[60] Next, the COG-based classification was applied to construct interacting protein hubs (a group of proteins under a common COG). Further, KEGG pathways were incorporated into the PPI network and analyzed in VisANT for identification of correlations among the interacting individual proteins, hubs, connecting nodes, and pathways to determine if the selected common conserved proteins and their PPIs are involved in bacterial essential metabolic process as well as in pathogenesis. This is with the agreement that as we have taken common conserved proteins of multiple pathogenic and non-pathogenic bacteria from same and different taxa; the proteins and their resultant interspecies PPIs must be involved in bacterial essential metabolic as well as pathogenic pathways.

Host–pathogen protein–protein interactions (HP-PPIs)

Cp infects a broad range of hosts, commonly goat, sheep, and horse,[51] and in rare cases, human.[42] However, the other pathogens investigated in this analysis do affect humans. With the exception of the human host, the genomes of the other hosts (goat, sheep and horse) have not been fully characterized. It is presumed that the goat, sheep and horse genomes have protein products similar to those of human, as they are higher mammals.[51] Several symptoms are shared between a *Cp, Y. pestis,*[45–47] and *Mycobacterium*[48,49] infection. *Mycobacterium* also falls under the same bacterial group of *Cp* (the CMNR group of pathogens). Therefore, we used our identified common conserved proteins (that interact with each other and forms common conserved PPIs) in our previous analysis step (inter-species PPIs) to generate a common conserved host–pathogen interaction that will be common to all the selected pathogens and hosts.

Although several computational approaches based HP-PPIs have been reported over time for a number of pathogens,[26,28,61,62] instead of using computational methods, we made our HP-PPIs based on published experimentally validated host–pathogen protein–protein binding data. To achieve the HP-PPIs; yeast two-hybrid assay based *Y. pestis-human* PPIs,[5,63] liquid chromatography-tandem mass spectrometry based surface-affinity profiling data for *S. gallolyticus-human* PPIs,[64] and protein microarray based *streptococcu*-human PPIs[29] were extracted from corresponding published literatures. Although the yeast two-hybrid screens generate significant degree of false negatives interactions,[65] we had no other option to generate the host pathogen PPIs because of unavailability of any other high throughput experimental data.

In addition to these literature based data, 7180 experimentally validated host–pathogen protein binding interactions for 21 pathogens with the human proteins from the Patho-Systems Resource Integration Center (Patric) database[66] and 24 253 PPIs between 58 hosts and 416 pathogen species from HPIDB database[67] were downloaded to enrich our interaction data. While the Patric contains interactions of bacterial proteins with only human; the HPIDB provides PPIs data for multiple hosts (including human, mouse, rat, and bovine, chicken *etc.*).

Next, the identified common conserved bacterial proteins those interact with each other in intra-species bacterial PPIs were manually correlated with human interacting counterparts based on the collected experimentally validated host–pathogen interaction data. In some cases, the correlation was difficult as the interacting partner protein from the bacteria was from species that is not considered in our analysis. Therefore, we used comparative genomics BLAST to identify if the interacting bacterial partner is a homologue to any of our selected common conserved bacterial proteins and if there is a $>35\%$ identity, we considered the interaction for our purpose.

Towards validating and determining the significance of the HP-PPIs

To identify and evaluate the significance of the host–pathogen interactions involved in the host response to the pathogenesis and the key bacterial proteins involved in the pathogenesis, we performed two analyses of the HP-PPIs. First, we performed gene set enrichment and enriched functional clustering based on Gene Ontology using the well known tool: Database for Annotation Visualization and Integrated Discovery (DAVID Vs6.7)[68] for the host proteins in the HP-PPIs. Further, we used ToppGene[69] for candidate gene prioritization, identification of network key nodes, and centrality analysis of the interacting host proteins by mapping their involvement in host pathways affected due to infection. ToppGene is a platform for gene set enrichment, functional annotations, and protein interactions network based candidate gene prioritization. It also provides information about relative importance of a candidate gene in a PPI network. For ToppGene analysis, the training sets for the respective biological processes were collected from data available at the Molecular signature Database (MsigDB).[70] The key biological processes were selected that are modulated within the host such as TLR signaling and inflammatory pathways, immunity, cytoskeleton reorganization, phagocytosis, and apoptosis in response to infection of *Y. pestis, E. coli, M. tuberculosis* and several other pathogenic bacteria as described in manually curated PHIDIAS host–pathogen interactions database.[71] Finally, the interacting pathway-specific key host proteins were selected based on the ToppGene analysis.

The key bacterial proteins in the HP-PPIs that are involved in the pathogenesis were identified based on the functionality analysis. The functional annotation was done using the NCBI, UniProt,[55] and KEGG databases.[57] Additionally, the sub-cellular localization of the proteins were determined using CELLO[72] and ''Effective''[73] tools. While CELLO identifies extracellular, outer membrane, inner membrane, periplasmic, and cytoplasmic proteins; the ''Effective'' specifically predicts bacterial secreted proteins. The virulence was checked using PAIDB database.[56]

Identification of targets from the host–pathogen PPIs and virtual screening

From the host–pathogen interaction network, the interacting essential non-host homolog bacterial proteins were identified as probable targets based on the method and criteria as described by Barh *et al.*, 2011.[51] Briefly, the interacting essential bacterial proteins were selected based on Database of Essential Genes (DEG)[74] BLASTp (cut off values: $E = 0.01$, bit score ≥ 100,

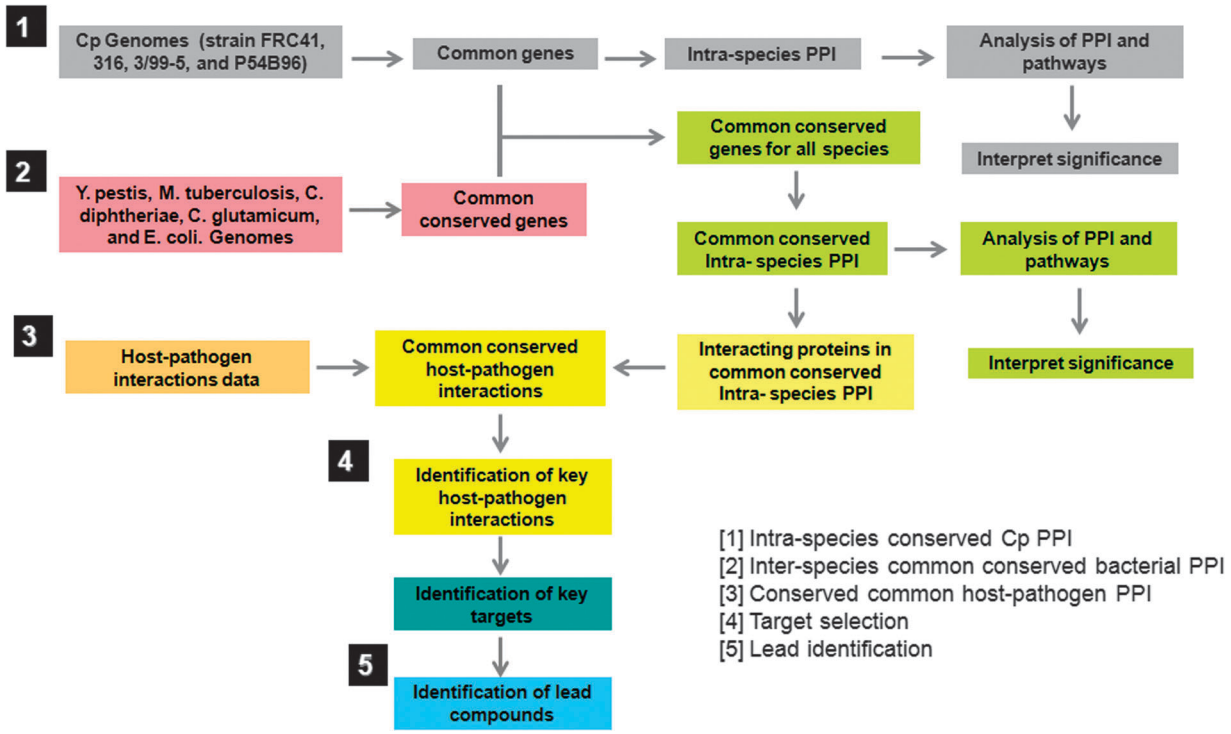


Fig. 1 Simple flow diagram of the overall strategy used to develop intra-species *Cp* PPI, inter-species bacterial PPI, host–pathogen interaction PPI, and identification of targets from the host–pathogen interactions.

identity $\geq$ 35%). Further, the non-host essential bacterial homologs were identified by subjecting essential proteins in NCBI BLASTp program against human, mouse, sheep, horse, and bovine proteomes. Finally, the bacterial essential non-host homolog core and PAI associated proteins having $\leq$100 KDa molecular weight and are involved in bacteria's multiple unique essential metabolic pathways were selected as putative targets.

The bacterial targets were modeled using the Phyre 2[75] and Swiss model servers[76] and validated using the SAVS server Vs.4. (http://services.mbi.ucla.edu/SAVES/). A ligand library was developed with 30 well known antibiotics used against the selected pathogens and effective drugs for *Cp*.[77] In India, a *Cp* infection is rare in areas where the cattle feed on betel vine leaves and stalks. Therefore, 120 compounds derived from betel vine were also used to enrich the ligand library and for testing these betel compounds on the identified targets. The catalytic pockets within the target proteins were determined using Molegro Virtual Docker.[78] The docking was performed using GOLD software[79] and the five best ligands based on their GOLD score. The overall strategy is represented in Fig. 1.

Growth inhibitory effect of *Piper betel* compounds: preliminary validation

The best lead compounds from *Piper betel* were tested for their individual growth inhibition efficacy against the pathogenic *E. coli* O157:H7. The bacteria were cultured in Mueller Hinton (MH) broth (Sigma-Aldrich Co. LLC) at 37 °C for 6 hours to reach the log phase. Then, cells were harvested by centrifugation and 10^7 CFU mL^{-1} cells were resuspended in tubes containing MH broth and 10, 100 μM or 1, 10, and 100 mM concentrations of the *Piper betel* compounds. Treatment with 100 μg ml^{-1} of ampicillin was used as control. Cultures were then incubated at 37 °C for 2 hours in a shaker. The number of colony-forming units (CFUs) was counted each 30 min interval by obtaining the CFU/ml from serial 10-fold dilutions prepared in MH agar (Sigma-Aldrich Co. LLC).

Results

Bacterial protein–protein interactions

Common conserved intra-strain PPI in Cp. We identified 1783 genes common to our 4 *Cp* selected strains. Using the computational approaches, we found 4186 conserved interactions common to these *Cp* strains. We found total 874 proteins are involved in these interactions. The number of predicted PPIs based on phylogenetic profile, domain fusion, and gene neighborhood methods are 2392, 2388, and 245, respectively. To analyze the pathways falling in these conserved interactions, we fed the PPIs and *Cp FRC41* metabolic pathways (obtained from KEGG) into VisANT. Upon analysis, we found that 68 pathways can be mapped in this intra-strain PPI of the Cp. These pathways include various metabolisms, two component systems, ABC transporters, and bacterial secretion systems among others that are important for bacterial survival and pathogenesis. Therefore, our selected conserved common proteins and the developed intra-strain PPI of *Cp* will be useful to explain the biology and pathogenesis of the bacteria if further analyzed.

Although this PPI of *Cp* is very preliminary of its kind, we are reporting it because there is no report so far on *Cp* PPI. As our main aims are to develop conserved common inter-species bacterial PPIs and use the same to develop conserved common host–pathogen interactions to finally identify conserved common broad spectrum target; we did not analyze the intra-strain PPI of *Cp* in detail.

Inter-species common conserved bacterial PPIs. To generate the common conserved inter-species bacterial PPIs, first we identified common conserved proteins in *Y. pestis CO92, E. coli K-12 DH10B, E. coli O157:H7, M. tuberculosi H37Rvs, C. diphtheriae,* and *C. glutamicum R* using PAST server. Seventy eight proteins were found to be conserved in all these species. Further, we checked if all these proteins are conserved in other virulent and non-virulent strains of various strains of these bacteria and *Cp* strains *i.e.* from closed and distance taxa. To achieve this we used amino acid sequences of these 78 *Y. pestis CO92* proteins and performed comparative BLASTp in NCBI server against proteomes of *E. coli str. K-12 substr. MG1655*, *C. glutamicum ATCC 13032 Kitasato*, *C. urealyticum DSM 7109*, *M. tuberculosis CDC1551*, *M. ulcerans Agy99*, and four of our *Cp* strains (*FRC41, 316, 3/99-5, and P54B96*). We found all these 75 proteins are conserved in all these selected species and strains (Table S2, ESI†).

As various experimental PPI data are available for *E. coli str. K-12*, we selected conserved 75 proteins of this species to make the common conserved inter-species PPIs using VisANT. In VisANT, these 75 proteins form a PPI network with 1674 interactions involving 666 interacting nodes where 1210, 755, and 281 interactions are based on the tandem affinity purification, inferred by authors, and anti tag co-immunoprecipitation methods, respectively. There are interactions based on computational and other experimental methods such as cross-linking studies among others. Twenty seven total pathways were mapped in this PPI (Table S3a, ESI†). However, while we did internal interactions among these 75 proteins, we found only 142 interactions involving 23 pathways (Table S3b, ESI†). These 75 interacting proteins fall under 14 COGs (Fig. 2) and with the exception of 3 proteins, all other proteins were found to be virulent as per the PAIDB – BLASTp analysis (Table S2, ESI†).

We selected pathogenic and non-pathogenic organisms from the same and distant taxa and their conserved genes to make the inter-species PPIs. Therefore, the resultant PPIs are common and conserved in all the bacterial species considered and the PPIs should involve pathways that are essential for bacterial survival as well as for pathogenesis. To check this, KEGG pathways were incorporated in the PPIs using VisANT's "expand pathways" option and the interactions along with the pathways were analyzed. The analysis showed that the interacting networks were well linked and fit with various pathways that are well known for their involvements in bacterial survival and virulence such as various metabolism, two-component

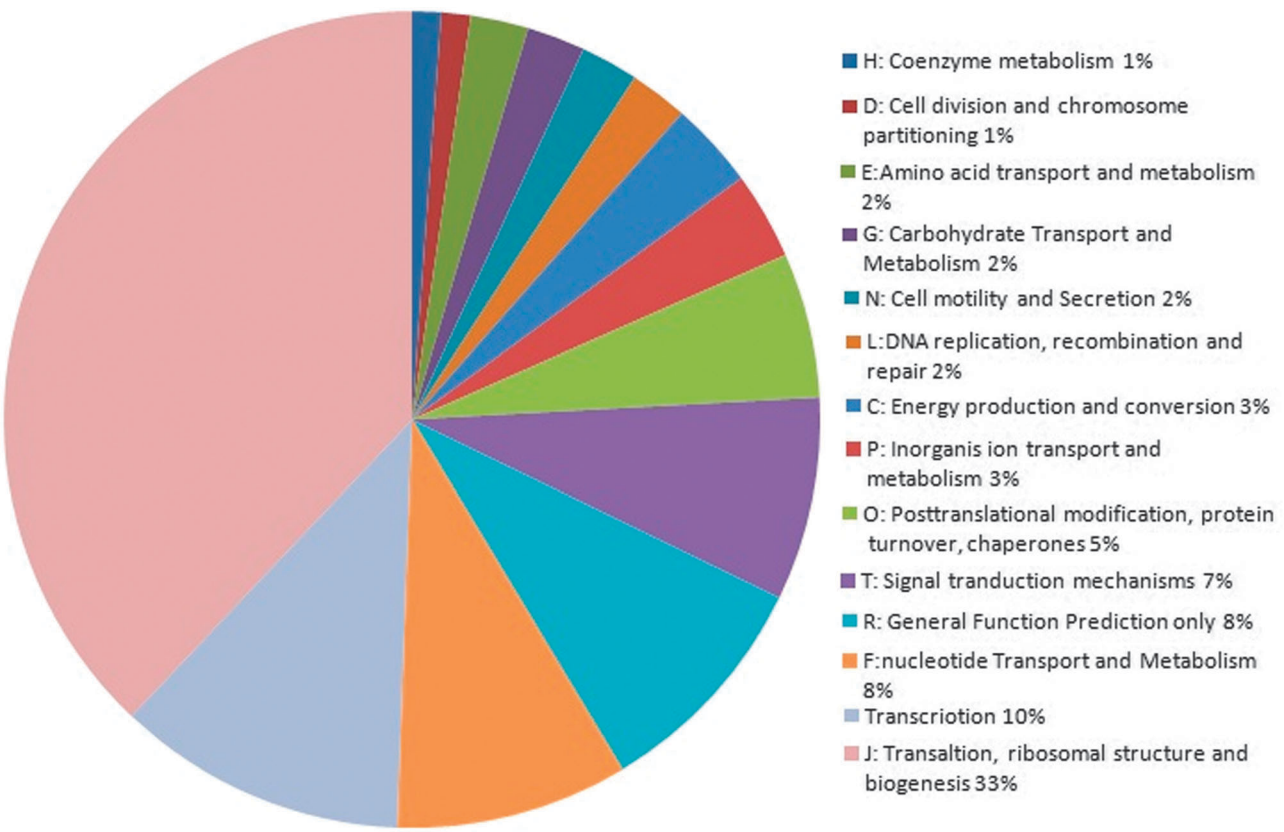
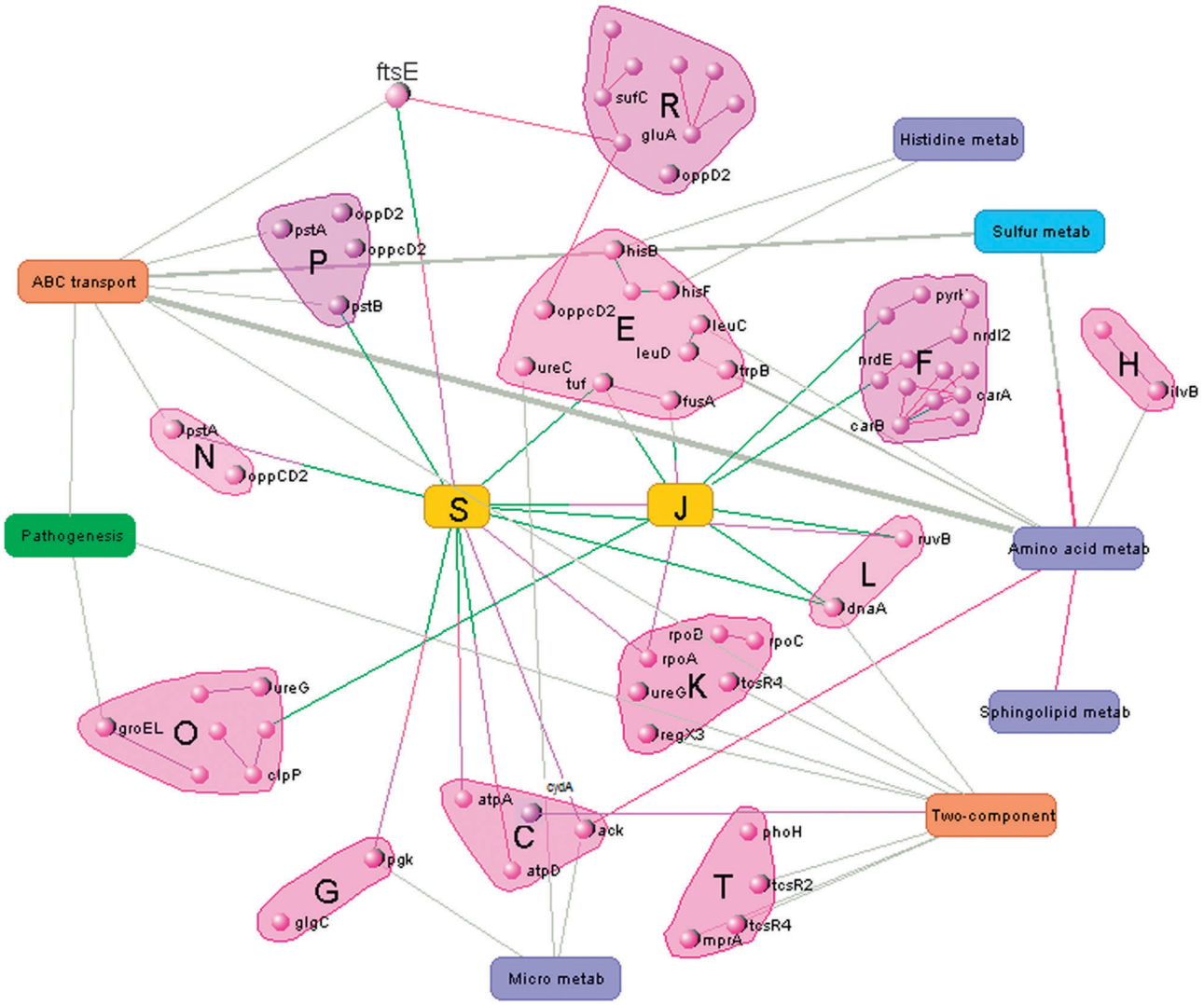


Fig. 2 Clusters of Orthologous Groups (COG) classifications of common conserved proteins of four *C. pseudotuberculosis* strains, *Y. pestis, M. tuberculosis, C. glutamicum, C. diphtheriae, C. ulcerans*, and *E. coli*.



C	Energy production and conversion	L	DNA replication, recombination and repair
E	Amino acid transport and metabolism	N	Cell motility and secretion
F	Nucleotide transport and metabolism	O	Posttranslational modification, protein turnover, chaperones
G	Carbohydrate transport and metabolism	P	Inorganic ion transport and metabolism
H	Coenzyme metabolism	R	General function prediction only
J	Translation, ribosomal structure and biogenesis	S	Function unknown
K	Transcription	T	Signal transduction mechanisms

Fig. 3 The conserved common PPIs with COG classifications of *Cp FRC41*, *Cp 316*, *Cp 3/99-5*, *Cp P54B96*, *Y. pestis*, *M. tuberculosis*, *C. gluticum*, *C. diptherae*, *C. ulcerans*, and *E. coli*. Important bacterial pathways involving these proteins and the relationship of these proteins and pathways are also shown. The relationships (edgs) between hubs and individual proteins are determined using VisANT.

system,[80] ABC transporter,[81,82] redox signaling,[83] and sphingolipid metabolism[84,85]-like pathways (Fig. 3), supporting the accuracy and significance of our PPIs.

Host–pathogen protein–protein interactions (HP-PPIs)

To make the HP-PPIs, we used the conserved bacterial proteins that interact with at least another protein of the bacteria in the inter-species bacterial PPI. Using the procedure described in the methods and such conserved interacting proteins, we identified 14 bacterial proteins that interact with 122 host proteins. Functional annotations of these bacterial proteins revealed that eight are cytoplasmic enzymes and five are membrane localized. All these 14 proteins were predicted to be involved in virulence as per the PAIDB and the DEG based analysis showed; all these proteins are encoded by essential genes. Further, the functional annotation of these 14 proteins revealed that, they are involved in bacterial various essential metabolic pathways as well as pathogenicity-related pathways

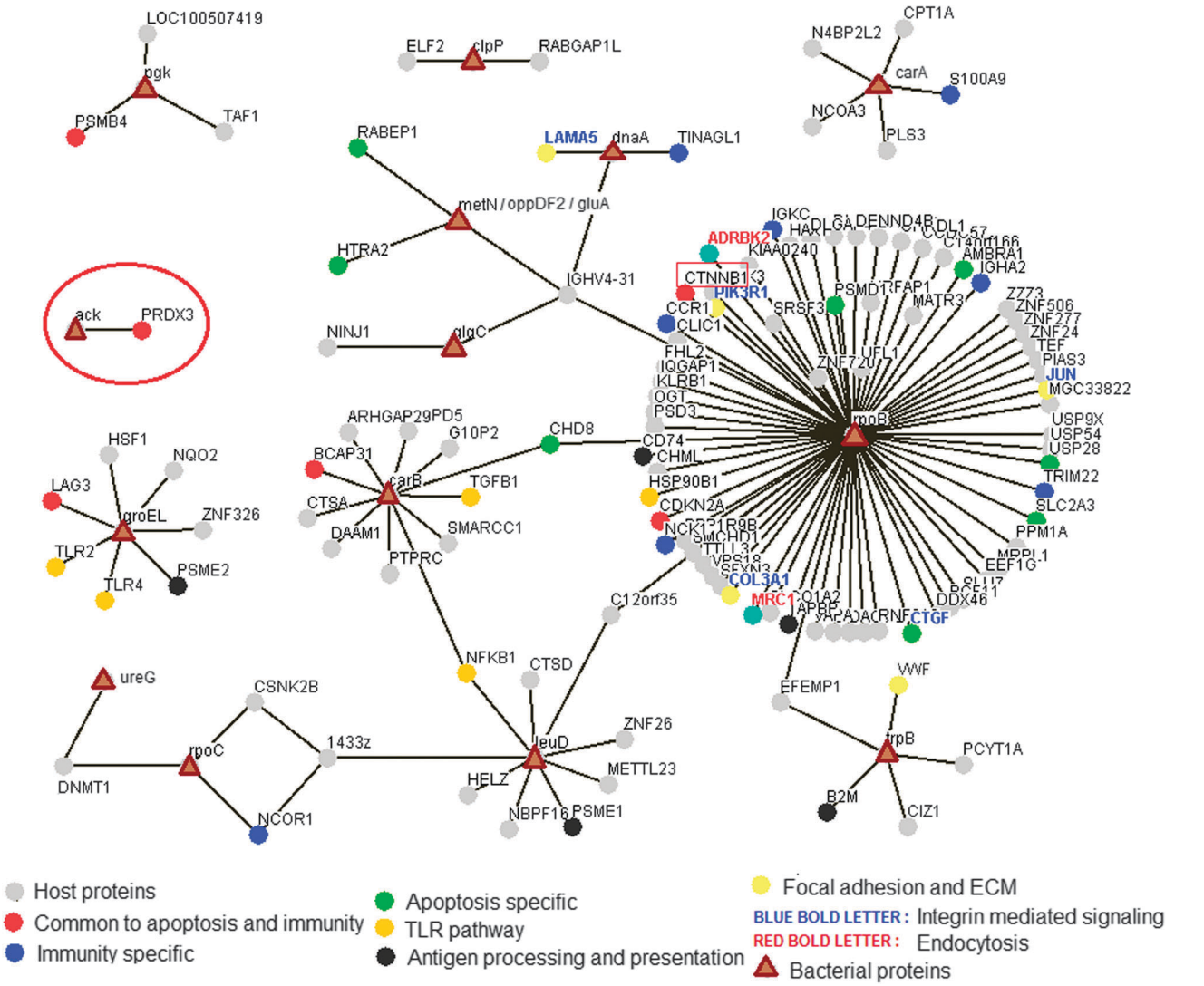


Fig. 4 Common conserved host–pathogen interaction network of multiple pathogens (four *C. pseudotuberculosis* strains, *Y. pestis, M. tuberculosis, C. glutamicum, C. diphtheriae, C. ulcerans*, and *E. coli*) and their usual hosts.

such as two-component systems (dnaA) and ABC transporters (gluA) (Table S4, ESI†). All 122 interacting host proteins were found in well-known bacterial infection associated host pathways such as integrin-mediated signaling, endocytosis, TLR signaling, immunity, apoptosis, inflammation, and redox signaling[71] (Table S5, ESI†). The ToppGene-based gene set enrichment analysis ranked CTNB1 and PIK3R1 at positions one and four, respectively. Both proteins interact with rpoB and are involved in immunity, apoptosis, and cell matrix adhesion (Table S6, ESI†). The bacterial proteins rpoB, carA, carB, leuD, groEL and their host interacting partners IGHV4-31, NFKB1, CHD8, and C12orf35, respectively were the key nodes in the host–pathogen protein–protein interaction network based on the degree of interactions and centrality analysis (Fig. 4).

Drug target and lead selection

From the host–pathogen protein–protein interaction network, we identified common conserved bacterial targets using subtractive genomics as described by Barh *et al.*, 2011.[51] The 14 identified genes were essential for the selected group of pathogens, and the cytoplasmic Acetate kinase (Ack) [EC = 2.7.2.1, Mass = 43.3 KDa] involved in the metabolism of taurine, hypotaurine, pyruvate, propanoate, and methane metabolism is the only non-host homolog satisfying most of the criteria of an ideal target for being (a) an essential non-host homolog enzyme for multiple organisms, (b) core gene of the organisms, (c) involvement in organisms' multiple unique and essential pathways, (d) PAI-related enzyme, and (e) less than 100 KDa molecular weight[51] (Table S4, ESI†). This common conserved target binds to host PRDX3 in yeast two hybrid assay (Fig. 4). PRDX3 is involved in the immune system, apoptosis, cell proliferation, and redox signaling-like pathways. Therefore, interaction of Ack-PRDX3 affects all these biological processes in the host, supporting a mechanism of bacterial infection.

Four active sites were found in the modeled Ack using the Molegro Virtual Docker (Table S7, ESI†). The GOLD fitness

score and MVD analysis of the docking showed that of the group of 30 selected antibiotics, ceftiofur and penicillin, commonly used to treat *Cp*, Diphtheria, Tuberculosis, and *Y. pestis* infections, were probably effective against Ack (Fig. 5 and Table S8, ESI†). Additionally, piperdardine and dehydropipernonaline derived from *Piper betel* were also predicted to be effective and possibly had a similar or superior inhibitory activity against the target as compared to penicillin and ceftiofur (Table S8, ESI†).

Piperdardine inhibits *E. coli O157:H7* growth

Viable cells were counted during the culture in MH media containing the compounds in order to investigate their growth-inhibiting effect on *E. coli* O157:H7. We observed that addition of 100.0 μM of piperdardine or their higher concentration dramatically decrease in the CFU counts, similar to bacteria treated with ampicillin (Fig. 6).

Discussion

PPIs derived information along with a molecular basis for host–pathogen interactions are important in finding effective targets against a pathogen. Computational or high-throughput approaches based on the development of genome- or proteome-wide PPI networks have been applied to various organisms,[9,10,18,19,26–29] allowing for the extraction of important information for specific biological processes. Predicted host–pathogen PPIs have been reported for *HIV*,[86,87] *Dengue virus*,[88] *Mycobacterium*, *apicomplexa*, *kinetoplastida*,[28] and *P. falciparum*.[26,27] Experimentally validated interactions and their implementations in drug or vaccine development against the various pathogens have also been reported for group-B *streptococcus*,[29] *Corynebacterium diphtheriae*,[36,89] *M. tuberculosis*,[31–33] *Yersinia pestis*,[90] and *Yersinia pseudotuberculosis*.[91] However, these experiments were conducted for a small fraction of pathogenic proteins. Recently, yeast-two hybrid-based proteome-wide

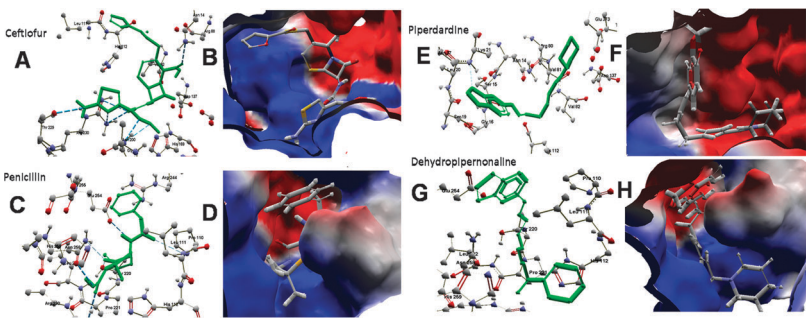


Fig. 5 Docking of Ack with ceftiofur (A–B), penicillin (C–D), piperdardine (E–F), and dehydropipernonaline (G–H).

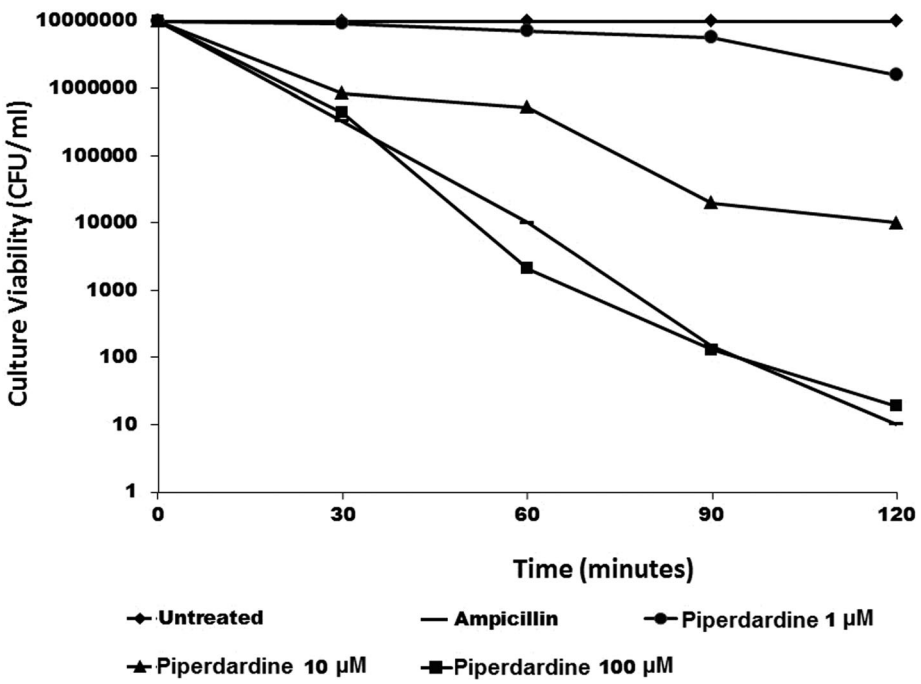


Fig. 6 Inhibitory effects of Piperdardine on the growth of *Escherichia coli* 0157:H7 as compared to ampicillin.

host–pathogen protein–protein binding interactions were reported for *B. anthracis*, *F. tularensis*, and *Y. pestis*,[5] and a number of novel interactions were documented for these pathogens.

In this report, for the first time we represent 4186 common conserved intra-species PPIs for four *Cp* strains (*Cp FRC41*, *Cp 316*, *Cp 3/99-5*, and *Cp P54B96)* using phylogenetic profile, domain fusion, and gene neighborhood methods. In *Cp*, we found 874 proteins are involved in these interactions. The recently reported experimental PPI data on *M. tuberculosis H37Rv*, another CMNR group of pathogens, revealed ~8000 novel interactions.[92] When we compared our intra-species PPIs of *Cp* with these *M. tuberculosis* data, we found half of the number of *M. tuberculosis* interactions in *Cp.* This difference may be due to the larger genome size of *M. tuberculosis* (4062 genes, almost double that of the *Cp* genome), the methods applied, and the phylogenetically conserved proteins in *Cp*. The sixty eight pathways mapped in the *Cp* PPI belong to both the bacterial essential metabolic and virulence pathways. Therefore, our developed *Cp* PPI will be significant in explaining biology and pathology of *Cp* upon further analysis.

While we developed the inter-species PPIs using common phylogenetically conserved proteins of different groups of organisms (pathogenic, non-pathogenic, and same and distance taxa), including four *Cp* strains*, Y. pestis, M. tuberculosis, C. glutamicum, C. diphtheriae, C. ulcerans*, and *E. coli*, we only observed 75 common interacting proteins that constituted a network of 142 interactions among each other and 1674 interactions involving 666 proteins to form the PPI network; however, important essential metabolic pathways and virulence related pathway can be mapped in these networks supporting the usefulness of the PPI in describing the common physiological process and virulence of these selected pathogens. It is also profound from these results that, the species-specific global PPIs exhibit a large number of interactions. However, number of interactions in conserved PPIs across a distantly related species of similar pathogenesis is reduced drastically, although essential and important pathogenesis-related proteins and pathways were found in the network.

Human-based host–pathogen interactions have been reported for a number of individual pathogens.[5,29,63,64] However, a common conserved HP-PPI for a number of pathogens and hosts have not been reported. Here, for the first time we used common conserved proteins from a broad spectrum of hosts (human, goat, sheep, and horse) to study the interactions. Additionally, for the first time, we have extended the strategy to generate conserved and common host–pathogen interactions for a group of pathogens using inter-species common conserved interacting proteins of *Cp, Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, and *E. coli* with a mode of pathogenesis common to these selected hosts. This strategy helped to gain insight into common conserved host–pathogen interactions across a wide range of organisms and to identify broad spectrum targets in a single analysis. The PAI-related proteins are thought to be involved in pathogenesis.[93] Our results support this finding, and we found that the 14 identified conserved pathogen proteins involved in host–pathogen interactions were located in PAIs. These proteins are also involved in essential metabolic and virulence pathways. Similarly, GSEA, candidate gene prioritization, key nodes, and centrality analysis of the interacting host proteins revealed that they are involved in most of the infection-related signaling pathways,[71] supporting the rationality of the developed host–pathogen interaction networks.

Based on the strategy of target identification,[51] Ack was selected as a broad spectrum target from the host–pathogen interaction network. Ack is essential to *E. coli*,[94] *M. genitalium*,[95] and *M. pulmonis*[96] and is predicted to be a target in *S. aureus*.[97] The HP-PPI showed that Ack interacted with Peroxiredoxin 3 (PRDX3) from the host. PRDX3 is a peroxidase and is involved in the NF-kappaB cascade, cell proliferation, apoptosis, and redox signaling. Redox-sensitive proteins in pathogens make them resistant to oxidative stress and antibiotics,[98] and manipulation of the redox state can be an important strategy for the management of Tuberculosis.[99] Ack, our identified target, is a kinase that interacts with the redox protein PRDX3 of the host. We hypothesized that the binding of Ack to PRDX3 modulates PRDX3 activity, thereby disrupting the redox signaling and immune system of the host. This interaction may help in SOD-mediated fibrocyte activation and scar or abscess formation[100] in lymphadenitis, the common symptom of *Cp* and *Y. Pestis* infections. It may also be a vital mechanism for drug resistance in these pathogens, disrupting the host redox system.

However, to interfere mitochondrial functions during pathogenesis, a bacterial protein needs to reach and bind to mitochondrial protein of the host.[101] Bacteria that possess type III and type IV secretion system like injection machinery can directly inject bacterial proteins into the host cell cytoplasm during infection process.[102,103] As per the ''Effective''[73] prediction, Ack of *M. tuberculosis H37Rv* is a type III secreted protein and according to *Couto et al.* (2012), Ack is probably secreted or localizes to bacterial surface during *M. mycoides* infection in cattle and plays a role in immunogenic responses in the host.[104] Therefore, it might be possible that Ack is injected into host cell through bacterial secretion system during infection and upon resealed into the host cytoplasm it interacts with mitochomdrial PRDX3. However, it should be proved experimentally and this is one of the future scopes of this research.

Virtual screening showed that ceftiofur and penicillin could be effective antibiotics against the selected pathogens considering the target Ack. The natural products piperdardine and dehydropipernonaline from *Piper betel* had shown a similar or superior effect on Ack as per our *in silico* analysis. Until now, no experimental data were available that tested the efficacy of compounds targeted to Ack, and validation is thereby necessary using conventional antibiotics and our identified *Piper betel* compounds. The leaf extract of *Piper betel* has proven to be useful as an antimicrobial,[105,106] antioxidant,[107] anti-inflammatory,[108] and immunomodulator.[109] However, the specific compounds in the plant that produce these properties are yet to be determined. In our preliminary validation, we observed that, 100.0 μM of piperdardine inhibits *E. coli* O157:H7 growth

similar to penicillin. Therefore, it is presumed that these compounds may also be effective against other pathogens considered in this work. We are currently testing the bactericidal effects of these betel compounds against *C. pseudotuberculosis*, *C. diphtheriae*, *M. tuberculosis*, *C. ulcerans*, and *Y. pestis* and their target specificity to Ack. The results are highly promising.

Conclusion

This study demonstrates intra-species PPI for *Cp* and illustrates the potential and importance of inter-species bacterial protein–protein and host–pathogen interactions in broad spectrum target identification. We report the conserved intra-species PPIs of *Cp* and a common conserved host pathogen-interaction network for *Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, *E. coli*, and four *Cp* strains. Ack was identified as a broad spectrum target for all these pathogens considering human, goat, sheep, and horse as hosts. Ceftiofur, penicillin and two natural compounds derived from *Piper betel*, piperdardine and dehydropipernonaline, were predicted to be effective against Ack activity. Validation shows piperdardine is a highly effective antibacterial agent. The *in silico* approaches used in this work were supposed to be effective in developing and analyzing inter-species global bacterial PPIs as well as host–pathogen interactions to identify drug targets.

Competing interests

The authors declare that they have no competing interests.

Financial disclosure

This work was carried out without any financial support or grant.

Note added after first publication

This article replaces the version published on 3rd January 2013, which contained an error in that the title was incomplete. The subtitle has been added to clarify this.

References

1 R. Sharan, I. Ulitsky and R. Shamir, Network-based prediction of protein function, *Mol. Syst. Biol.*, 2007, **3**, 88.
2 E. D. Levy and J. B. Pereira-Leal, Evolution and dynamics of protein interactions and networks, *Curr. Opin. Struct. Biol.*, 2008, **18**, 349–357.
3 F. Hormozdiari, R. Salari, V. Bafna and S. C. Sahinalp, Protein–protein interaction network evaluation for identifying potential drug targets, *J. Comput. Biol.*, 2010, **17**, 669–684.
4 Y. Y. Wang, J. C. Nacher and X. M. Zhao, Predicting drug targets based on protein domains, *Mol. BioSyst.*, 2012, **8**, 1528–1534.
5 M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali and B. W. Sobral, The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis, *PLoS One*, 2010, **5**, e12089.
6 J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, Predicting protein–protein interactions based only on sequences information, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(11), 4337–41.
7 X. W. Zhao, Z. Q. Ma and M. H. Yin, Predicting protein–protein interactions by combing various sequence-derived features into the general form of Chou's Pseudo amino acid composition, *Protein Pept. Lett.*, 2011, **19**(5), 492–500.
8 L. Skrabanek, H. K. Saini, G. D. Bader and A. J. Enright, Computational prediction of protein–protein interactions, *Mol. Biotechnol.*, 2008, **38**, 1–17.
9 H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill and M. Vidal, High-quality binary protein interaction map of the yeast interactome network, *Science*, 2008, **322**, 104–110.
10 U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach and E. E. Wanker, A human protein–protein interaction network: a resource for annotating the proteome, *Cell*, 2005, **122**, 957–968.
11 I. Ispolatov, A. Yuryev, I. Mazo and S. Maslov, Binding properties and evolution of homodimers in protein–protein interaction networks, *Nucleic Acids Res.*, 2005, **33**, 3629–3635.
12 M. P. Stumpf, T. Thorne, S. E. de, R. Stewart, H. J. An, M. Lappe and C. Wiuf, Estimating the size of the human interactome, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 6959–6964.
13 T. Dandekar, B. Snel, M. Huynen and P. Bork, Conservation of gene order: a fingerprint of proteins thatphysically interact, *Trends Biochem. Sci.*, 1998, **23**, 324–328.
14 M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 4285–4288.
15 A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, *Nature*, 1999, **402**, 86–90.
16 C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther and F. E. Cohen, Co-evolution of proteins with their interaction partners, *J. Mol. Biol.*, 2000, **299**, 283–293.
17 M. Singhal and H. Resat, A domain-based approach to predict protein–protein interactions, *BMC Bioinf.*, 2007, **8**, 199.
18 K. Raman and N. Chandra, Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance, *BMC Microbiol.*, 2008, **8**, 234.

19 M. Rashid, S. Ramasamy and G. P. Raghava, A simple approach for predicting protein–protein interactions, *Curr. Protein Pept. Sci.*, 2010, **11**(7), 589–600.
20 P. M. Bowers, S. J. Cokus, D. Eisenberg and T. O. Yeates, Use of logic relationships to decipher protein network organization, *Science*, 2004, **306**, 2246–2249.
21 D. Barker and M. Pagel, Predicting functional gene links from phylogenetic-statistical analyses of whole genomes, *PLoS Comput. Biol.*, 2005, **1**, e3.
22 N. Tuncbag, G. Kar, O. Keskin, A. Gursoy and R. Nussinov, A survey of available tools and web servers for analysis of protein–protein interactions and interfaces, *Briefings Bioinf.*, 2009, **10**, 217–232.
23 C. Yeats, J. Lees, P. Carter, I. Sillitoe and C. Orengo, The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences, *Nucleic Acids Res.*, 2011, **39**, W546–W550.
24 S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, C. E. de, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats and S. Y. Yong, InterPro in 2011: new developments in the family and domain prediction database, *Nucleic Acids Res.*, 2012, **40**, D306–D312.
25 J. Lees, C. Yeats, J. Perkins, I. Sillitoe, R. Rentzsch, B. H. Dessailly and C. Orengo, Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis, *Nucleic Acids Res.*, 2012, **40**, D465–D471.
26 M. D. Dyer, T. M. Murali and B. W. Sobral, Computational prediction of host–pathogen protein–protein interactions, *Bioinformatics*, 2007, **23**, i159–i166.
27 S. Wuchty, Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens, *PLoS One*, 2011, **6**, e26960.
28 F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow and A. Sali, Host pathogen protein interactions predicted by comparative modeling, *Protein Sci.*, 2007, **16**, 2585–2596.
29 I. Margarit, S. Bonacci, G. Pietrocola, S. Rindi, C. Ghezzo, M. Bombaci, V. Nardi-Dei, R. Grifantini, P. Speziale and G. Grandi, Capturing host–pathogen interactions by protein microarrays: identification of novel streptococcal proteins binding to human fibronectin, fibrinogen, and C4BP, *FASEB J.*, 2009, **23**, 3100–3112.
30 H. Huang, B. M. Jedynak and J. S. Bader, Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps, *PLoS Comput. Biol.*, 2007, **3**(11), e214.
31 S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell and P. M. Small, Variable host–pathogen compatibility in Mycobacterium tuberculosis, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 2869–2873.
32 D. P. Cifuentes, M. Ocampo, H. Curtidor, M. Vanegas, M. Forero, M. E. Patarroyo and M. A. Patarroyo, Mycobacterium tuberculosis Rv0679c protein sequences involved in host-cell infection: potential TB vaccine candidate antigen, *BMC Microbiol.*, 2010, **10**, 109.
33 K. Raman, A. G. Bhat and N. Chandra, A systems perspective of host–pathogen interactions: predicting disease outcome in tuberculosis, *Mol. BioSyst.*, 2010, **6**, 516–530.
34 Y. Wang, T. Cui, C. Zhang, M. Yang, Y. Huang, W. Li, L. Zhang, C. Gao, Y. He, Y. Li, F. Huang, J. Zeng, C. Huang, Q. Yang, Y. Tian, C. Zhao, H. Chen, H. Zhang and Z. G. He, Global protein–protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv, *J. Proteome Res.*, 2010, **9**, 6665–6677.
35 V. Kolodkina, T. Denisevich and L. Titov, Identification of Corynebacterium diphtheriae gene involved in adherence to epithelial cells, *Infect., Genet. Evol.*, 2011, **11**, 518–521.
36 L. Ott, M. Holler, R. G. Gerlach, M. Hensel, J. Rheinlaender, T. E. Schaffer and A. Burkovski, Corynebacterium diphtheriae invasion-associated protein (DIP1281) is involved in cell surface organization, adhesion and internalization in epithelial cells, *BMC Microbiol.*, 2010, **10**, 2.
37 L. Ott, M. Holler, J. Rheinlaender, T. E. Schaffer, M. Hensel and A. Burkovski, Strain-specific differences in pili formation and the interaction of Corynebacterium diphtheriae with host cells, *BMC Microbiol.*, 2010, **10**, 257.
38 E. Trost, J. Blom, S. S. de Castro, I. H. Huang, A. Al-Dilaimi, J. Schroder, S. Jaenicke, F. A. Dorella, F. S. Rocha, A. Miyoshi, V. Azevedo, M. P. Schneider, A. Silva, T. C. Camello, P. S. Sabbadini, C. S. Santos, L. S. Santos, R. Hirata, Jr., A. L. Mattos-Guaraldi, A. Efstratiou, M. P. Schmitt, H. Ton-That and A. Tauch, Pan-genomics of Corynebacterium diphtheriae: Insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis and pneumonia, *J. Bacteriol.*, 2012, **194**(12), 3199–3215.
39 L. H. Williamson, Caseous lymphadenitis in small ruminants, *Vet. Clin. North Am.: Food Anim Pract.*, 2001, **17**, 359–371, vii.
40 M. Aleman, S. J. Spier, W. D. Wilson and M. Doherr, Corynebacterium pseudotuberculosis infection in horses: 538 cases (1982–1993), *J. Am. Vet. Med. Assoc.*, 1996, **209**, 804–809.
41 R. G. Batey, Pathogenesis of caseous lymphadenitis in sheep and goats, *Aust. Vet. J.*, 1986, **63**, 269–272.
42 E. Trost, L. Ott, J. Schneider, J. Schroder, S. Jaenicke, A. Goesmann, P. Husemann, J. Stoye, F. A. Dorella, F. S. Rocha, S. C. Soares, V. D'Afonseca, A. Miyoshi, J. Ruiz, A. Silva, V. Azevedo, A. Burkovski, N. Guiso, O. F. Join-Lambert, S. Kayal and A. Tauch, The complete

genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence, *BMC Genomics*, 2010, **11**, 728.

43 G. J. Annas, Bioterror and ''bioart'' – a plague o' both your houses, *N. Engl. J. Med.*, 2006, **354**, 2715–2720.

44 M. Drancourt, Plague in the genomic area, *Clin. Microbiol. Infect.*, 2012, **18**, 224–230.

45 T. Karttunen, K. Nevasaari, O. Rasanen, P. J. Taskinen and M. Alavaikko, Immunoblastic lymphadenopathy with a high serum Yersinia enterocolitica titer. A case report, *Cancer*, 1983, **52**, 2281–2284.

46 S. J. Nesbitt, L. O. Neville, F. R. Scott and D. M. Flynn, Yersinia pseudotuberculosis in a 3 year old and rapid response to cefotaxime, *J. R. Soc. Med.*, 1994, **87**, 418–419.

47 J. E. Comer, D. E. Sturdevant, A. B. Carmody, K. Virtaneva, D. Gardner, D. Long, R. Rosenke, S. F. Porcella and B. J. Hinnebusch, Transcriptomic and innate immune responses to Yersinia pestis in the lymph node during bubonic plague, *Infect. Immun.*, 2010, **78**, 5086–5098.

48 P. R. Mohapatra and A. K. Janmeja, Tuberculous lymphadenitis, *J. Assoc. Physicians India*, 2009, **57**, 585–590.

49 J. Knox, G. Lane, J. S. Wong, P. G. Trevan and H. Karunajeewa, Diagnosis of Tuberculous Lymphadenitis Using Fine Needle Aspiration Biopsy, *Int. Med. J.*, 2012, DOI: 10.1111/j.1445-5994.2012.02748.x.

50 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.

51 D. Barh, N. Jain, S. Tiwari, B. P. Parida, V. D'Afonseca, L. Li, A. Ali, A. R. Santos, L. C. Guimaraes, S. S. de Castro, A. Miyoshi, A. Bhattacharjee, A. N. Misra, A. Silva, A. Kumar and V. Azevedo, A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens, *Chem. Biol. Drug Des.*, 2011, **78**, 73–84.

52 C. Fong, L. Rohmer, M. Radey, M. Wasnick and M. J. Brittnacher, PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes, *BMC Bioinf.*, 2008, **9**, 170.

53 R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin, The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 2001, **29**, 22–28.

54 M. Kaufmann, The Role of the COG Database in Comparative and Functional Genomics, *Curr. Bioinf.*, 2006, **1**, 291–300.

55 M. Magrane and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data. Database*, Oxford, 2011, bar009.

56 S. H. Yoon, Y. K. Park, S. Lee, D. Choi, T. K. Oh, C. G. Hur and J. F. Kim, Towards pathogenomics: a web-based resource for pathogenicity islands, *Nucleic Acids Res.*, 2007, **35**, D395–D400.

57 M. Kanehisa and S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 2000, **28**, 27–30.

58 Z. Hu, J. Mellor, J. Wu and C. DeLisi, VisANT: an online visualization and analysis tool for biological interaction data, *BMC Bioinf.*, 2004, **5**, 17.

59 Z. Hu, D. M. Ng, T. Yamada, C. Chen, S. Kawashima, J. Mellor, B. Linghu, M. Kanehisa, J. M. Stuart and C. DeLisi, VisANT 3.0: new modules for pathway visualization, editing, prediction and construction, *Nucleic Acids Res.*, 2007, W625–W632.

60 G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt and A. Emili, Interaction network containing conserved and essential protein complexes in Escherichia coli, *Nature*, 2005, **433**, 531–537.

61 N. Tyagi, O. Krishnadev and N. Srinivasan, Prediction of protein–protein interactions between Helicobacter pylori and a human host, *Mol BioSyst.*, 2009, **5**(12), 1630–1635.

62 O. Krishnadev and N. Srinivasan, Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria, *Int. J. Biol. Macromol.*, 2011, **48**(4), 613–619.

63 H. Yang, Y. Ke, J. Wang, Y. Tan, S. K. Myeni, D. Li, Q. Shi, Y. Yan, H. Chen, Z. Guo, Y. Yuan, X. Yang, R. Yang and Z. Du, Insight into bacterial virulence mechanisms against host immune response *via* the Yersinia pestis-human protein–protein interaction network, *Infect. Immun.*, 2011, **79**(11), 4413–4424.

64 A. Boleij, C. M. Laarakkers, J. Gloerich, D. W. Swinkels and H. Tjalsma, Surface-affinity profiling to identify host–pathogen interactions, *Infect. Immun.*, 2011, **79**(12), 4777–4783.

65 T. Stellberger, R. Häuser, A. Baiker, V. R. Pothineni, J. Haas and P. Uetz, Improving the yeast two-hybrid system with permutated fusions proteins: the Varicella Zoster Virus interactome, *Proteome Sci.*, 2010, **8**, 8.

66 J. J. Gillespie, A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, E. K. Nordberg, M. Scott, J. R. Schulman, E. E. Snyder, D. E. Sullivan, C. Wang, A. Warren, K. P. Williams, T. Xue, H. S. Yoo, C. Zhang, Y. Zhang, R. Will, R. W. Kenyon and B. W. Sobral, PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species, *Infect. Immun.*, 2011, **79**, 4286–4298.

67 R. Kumar and B. Nanduri, HPIDB – a unified resource for host–pathogen interactions, *BMC Bioinf.*, 2010, **11**(Suppl 6), S16.

68 D. W. Huang, B. T. Sherman and R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, 2009, **4**(1), 44–57.

69 B. E. A. B. J. A. Chen, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, *Nucleic Acids Res.*, 2009, 37.

70 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.

71 Z. Xiang, Y. Tian and Y. He, PHIDIAS: a pathogen-host interaction data integration and analysis system, *Genome Biol.*, 2007, **8**(7), R150.

72 C. S. Yu, C. J. Lin and J. K. Hwang, Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n*-peptide compositions, *Protein Sci.*, 2004, **13**, 1402–1406.

73 M. A. Jehl, R. Arnold and T. Rattei, Effective–a database of predicted secreted bacterial proteins, *Nucleic Acids Res.*, 2011, D591–D595.

74 R. Zhang, H. Y. Ou and C. T. Zhang, DEG: a database of essential genes, *Nucleic Acids Res.*, 2004, **1**, D271–D272.

75 L. A. Kelley and M. J. Sternberg, Protein structure prediction on the Web: a case study using the Phyre server, *Nat. Protoc.*, 2009, **4**, 363–371.

76 K. Arnold, L. Bordoli, J. Kopp and T. Schwede, The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling, *Bioinformatics*, 2006, **22**, 195–201.

77 S. J. Spier, Corynebacterium pseudotuberculosis infection in horses: An emerging disease associated with climate change?, *Equine vet. Educ*, 2008, **20**, 37–39.

78 R. Thomsen and M. H. Christensen, MolDock: a new technique for high-accuracy molecular docking, *J. Med. Chem.*, 2006, **49**, 3315–3321.

79 M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, Improved protein-ligand docking using GOLD, *Proteins*, 2003, **52**, 609–623.

80 T. Tobe, The roles of two-component systems in virulence of pathogenic Escherichia coli and Shigella spp, *Adv. Exp. Med. Biol.*, 2008, **631**, 189–99.

81 J. S. Klein and O. Lewinson, Bacterial ATP-driven transporters of transition metals: physiological roles, mechanisms of action and roles in bacterial virulence, *Metallomics*, 2011, **3**(11), 1098–1108.

82 V. G. Lewis, M. P. Ween and C. A. McDevitt, The role of ATP-binding cassette transporters in bacterial pathogenicity, *Protoplasma*, 2012, **249**(4), 919–942.

83 A. Trivedi, N. Singh, S. A. Bhat, P. Gupta and A. Kumar, Redox biology of tuberculosis pathogenesis, *Adv. Microbiol. Physiol.*, 2012, **60**, 263–324.

84 L. J. Heung, C. Luberto and M. Del Poeta, Role of sphingolipids in microbial pathogenesis, *Infect. Immun.*, 2006, **74**(1), 28–39.

85 D. An, C. Na, J. Bielawski, Y. A. Hannun and D. L. Kasper, Membrane sphingolipids as essential molecular signals for Bacteroides survival in the intestine, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(Suppl 1), 4666–4671.

86 P. Evans, W. Dampier, L. Ungar and A. Tozeren, Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs, *BMC Med. Genomics*, 2009, **2**, 27.

87 O. Tastan, Y. Qi, J. G. Carbonell and J. Klein-Seetharaman, Prediction of interactions between HIV-1 and human proteins by information integration, *Pac. Symp. Biocomput.*, 2009, 516–527.

88 J. M. Doolittle and S. M. Gomez, Mapping protein interactions between Dengue virus and its human and insect hosts, *PLoS Neglected Trop. Dis.*, 2011, **5**, e954.

89 V. Kolodkina, T. Denisevich and L. Titov, Identification of Corynebacterium diphtheriae gene involved in adherence to epithelial cells, *Infect. Genet. Evol.*, 2011, **11**, 518–521.

90 B. Li and R. Yang, Interaction between Yersinia pestis and the host immune system, *Infect. Immun.*, 2008, **76**, 1804–1811.

91 C. G. Zhang, A. D. Gonzales, M. W. Choi, B. A. Chromy, J. P. Fitch and S. L. McCutchen-Maloney, Subcellular proteomic analysis of host–pathogen interactions using human monocytes exposed to Yersinia pestis and Yersinia pseudotuberculosis, *Proteomics*, 2005, **5**, 1877–1888.

92 Y. Wang, T. Cui, C. Zhang, M. Yang, Y. Huang, W. Li, L. Zhang, C. Gao, Y. He, Y. Li, F. Huang, J. Zeng, C. Huang, Q. Yang, Y. Tian, C. Zhao, H. Chen, H. Zhang and Z. G. He, Global protein–protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv, *J. Proteome Res.*, 2010, **9**, 6665–6677.

93 H. Schmidt and M. Hensel, Pathogenicity islands in bacterial pathogenesis, *Clin. Microbiol. Rev.*, 2004, **17**, 14–56.

94 S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabasi, Z. N. Oltvai and A. L. Osterman, Experimental determination and system level analysis of essential genes in Escherichia coli MG1655, *J. Bacteriol.*, 2003, **185**, 5673–5684.

95 J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, III, H. O. Smith and J. C. Venter, Essential genes of a minimal bacterium, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 425–430.

96 C. T. French, P. Lao, A. E. Loraine, B. T. Matthews, H. Yu and K. Dybvig, Large-scale transposon mutagenesis of Mycoplasma pulmonis, *Mol. Microbiol.*, 2008, **69**, 67–76.

97 V. K. Morya, V. Dewaker, S. D. Mecarty and R. Singh, In silico Analysis Metabolic Pathways for Identification of Putative Drug Targets for Staphylococcus aureus, *J. Comput. Sci. Syst. Biol.*, 2010, **3**, 062–069.

98 P. R. Chen, P. Brugarolas and C. He, Redox signaling in human pathogens, *Antioxid. Redox Signaling*, 2011, **14**, 1107–1118.

99 A. Kumar, A. Farhana, L. Guidry, V. Saini, M. Hondalus and A. J. Steyn, Redox homeostasis in mycobacteria: the key to tuberculosis control?, *Expert Rev. Mol. Med.*, 2011, **13**, e39.

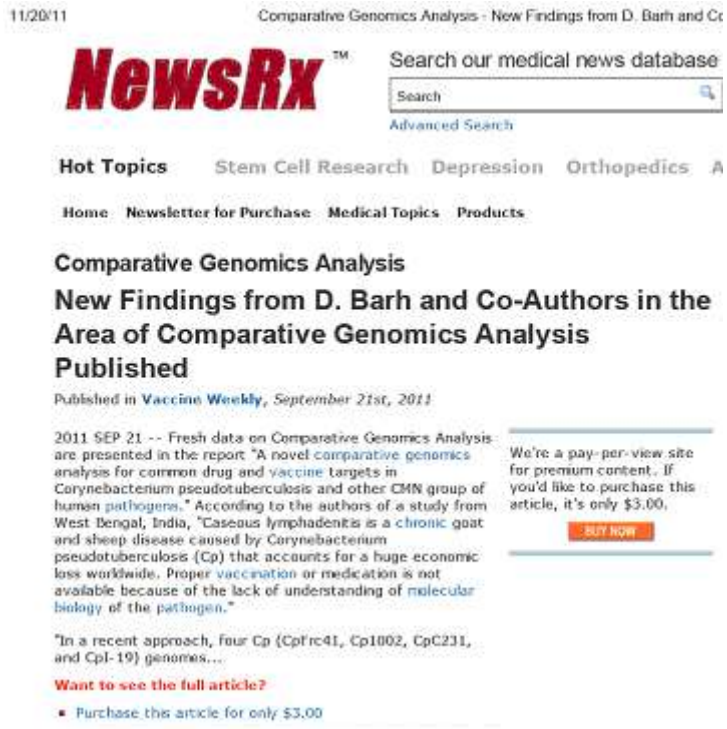
100 M. C. Vozenin-Brotons, V. Sivan, N. Gault, C. Renard, C. Geffrotin, S. Delanian, J. L. Lefaix and M. Martin, Antifibrotic action of Cu/Zn SOD is mediated by TGF-beta1 repression and phenotypic reversion of myofibroblasts, *Free Radical Biol. Med.*, 2001, **30**, 30–42.

101 V. Kozjak-Pavlovic, K. Ross and T. Rudel, Import of bacterial pathogenicity factors into mitochondria, *Curr. Opin. Microbiol.*, 2008, **11**(1), 9–14.
102 G. R. Cornelis, The type III secretion injectisome, *Nat. Rev. Microbiol.*, 2006, **4**(11), 811–25.
103 S. Backert and T. F. Meyer, Type IV secretion systems and their effectors in bacterial pathogenesis, *Curr. Opin. Microbiol.*, 2006, **9**(2), 207–217.
104 M. S. R. Couto, C. S. Klein, D. Voss-Rech and H. Terenzi, Extracellular Proteins of Mycoplasma synoviae, *ISRN Vet. Sci.*, 2012, **2012**, 6.
105 R. Nair and S. Chanda, Antimicrobial Activity of Terminalia catappa, Manilkara zapota and Piper betel Leaf Extract, *Indian J. Pharm. Sci.*, 2008, **70**, 390–393.
106 I. Ali, F. G. Khan, K. A. Suri, B. D. Gupta, N. K. Satti, P. Dutt, F. Afrin, G. N. Qazi and I. A. Khan, *In vitro* antifungal activity of hydroxychavicol isolated from Piper betle L, *Ann. Clin. Microbiol. Antimicrob.*, 2010, **9**, 7.
107 N. Dasgupta and B. De, Antioxidantactivity of PiperbetleL. leafextract *in vitro*, *Food Chem.*, 2004, **88**, 219–224.
108 S. Ganguly, S. Mula, S. Chattopadhyay and M. Chatterjee, An ethanol extract of Piper betle Linn. mediates its anti-inflammatory activity *via* down-regulation of nitric oxide, *J. Pharm. Pharmacol.*, 2007, **59**, 711–718.
109 D. G. Kanjwani, T. P. Marathe, S. V. Chiplunkar and S. S. Sathaye, Evaluation of immunomodulatory activity of methanolic extract of Piper betel, *Scand. J. Immunol.*, 2008, **67**, 589–593.

IV.4.1 Conclusions from this research/ Chapter-4

i. We have shown a novel integrated bioinformatics strategy combining PPI, host-pathogen interactions, and subtractive genomics to identify common conserved targets in *C. pseudotuberculosis (*Cp), *C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis,* and *E. coli.*
ii. The strategy have shown first time the intra-species PPI for Cp and the importance of inter-species bacterial PPI and HP-PPI in broad spectrum target identification.
iii. Ack was identified as a broad spectrum target for all these pathogens considering human, goat, sheep, and horse as hosts.
iv. Ceftiofur, penicillin and two natural compounds derived from Piper betel, piperdardine and dehydropipernonaline, were predicted to be effective against Ack activity.
v. Validation shows piperdardine is a highly effective broad spectrum antibacterial agent.
vi. This approach can be effective in developing and analysing interspecies global bacterial PPIs as well as host–pathogen interactions to identify drug targets in other pathogens too.

IV.4.2 Media highlights of this research outcomes/ Chapter-4



11/20/11 Comparative Genomics Analysis - New Findings from D. Barh and Co

NewsRx™

Search our medical news database

Search

Advanced Search

Hot Topics Stem Cell Research Depression Orthopedics A

Home Newsletter for Purchase Medical Topics Products

Comparative Genomics Analysis

New Findings from D. Barh and Co-Authors in the Area of Comparative Genomics Analysis Published

Published in Vaccine Weekly, September 21st, 2011

2011 SEP 21 -- Fresh data on Comparative Genomics Analysis are presented in the report "A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens." According to the authors of a study from West Bengal, India, "Caseous lymphadenitis is a chronic goat and sheep disease caused by Corynebacterium pseudotuberculosis (Cp) that accounts for a huge economic loss worldwide. Proper vaccination or medication is not available because of the lack of understanding of molecular biology of the pathogen."

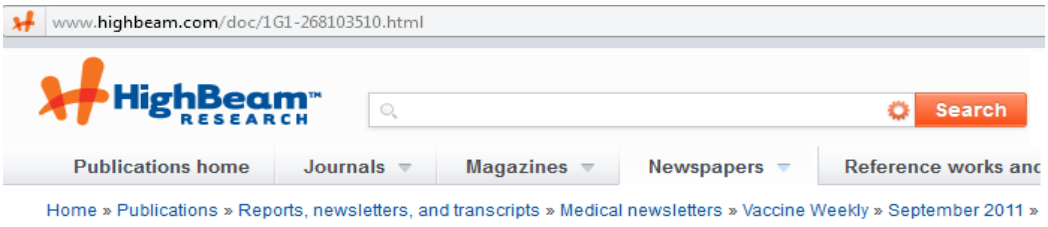
"In a recent approach, four Cp (CpFrc41, Cp1002, CpC231, and CpI-19) genomes...

Want to see the full article?

- Purchase this article for only $3.00

We're a pay-per-view site for premium content. If you'd like to purchase this article, it's only $3.00.

BUY NOW



www.highbeam.com/doc/1G1-268103510.html

HighBeam™ RESEARCH

Search

Publications home Journals Magazines Newspapers Reference works anc

Home » Publications » Reports, newsletters, and transcripts » Medical newsletters » Vaccine Weekly » September 2011 »

New Findings from D. Barh and Co-Authors in the Area of Comparative Genomics Analysis Published.

Newspaper

Vaccine Weekly

September 21, 2011 | Copyright

Fresh data on **Comparative Genomics Analysis** are presented in the report "**A novel comparative genomics analysis** for **common drug** and **vaccine targets** in **Corynebacterium pseudotuberculosis** and **other CMN group** of **human pathogens**." According to the authors of **a** study from West Bengal, India, "Caseous lymphadenitis is **a** chronic goat and sheep disease caused by **Corynebacterium pseudotuberculosis** (Cp) that accounts for **a** huge economic loss worldwide. Proper vaccination or medication is not available because of the lack of understanding of molecular biology of the pathogen."

"In **a** recent approach, four Cp (CpFrc41, Cp1002, CpC231, and CpI-19) genomes were sequenced to elucidate the ...

Medical News Today (MNT)

Discovery Of Novel Targets And Targeting Compounds For Fighting Deadly Infectious Diseases Including MtB

Published: Wednesday 13 February 2013 Adapted Media Release

An international team of researchers from India, Brazil, Mexico, the USA, and Denmark led by Debmalya Barh from the Institute of Integrative Omics and Applied Biotechnology (IIOAB) in Nonakuri, Tamluk, Purba Medinipur, West Bengal, India have claimed to have identified a novel drug target (Acetate kinase) common to pathogens causing hemorrhagic diarrohea; tuberculosis, plague diphtheria, and Caseous lymphadenitis and two novel targets (Undecaprenyl pyrophosphate phosphatase enzyme and Outer membrane protein ompU) that may help to develop both a drug and a vaccine against almost all pathogenic Vibrio spp including Vibrio cholerae.

Two studies have been published by the team, one in Integrative Biology on 3rd January, 2013 and the second in PLOS ONE on 30th January 30, 2013 describing these findings. The researchers have for the first time identified the targeting compounds isolated from Piper betel leaves. They have also demonstrated that among the several compounds one of the piper compounds, Piperdardine, is having higher efficacy than certain antibiotics such as Chloramphenicol, Penicillin, and Ampicillin. Dr. Nidia Leon-Sicairos form School of Medicine, Autonomous University of Sinaloa, Mexico, who headed the Mexican team associated with this research, describes these compounds as "golden compounds" and according to Prof. Vasco Azevedo, the Brazilian team led from UFMG "Identification of the common targets and targeting compounds for so many deadly pathogens in one shot is a big achievement and progress towards fighting against the diseases caused by these pathogens."

"Piper betel leaves are uses as a healthy mouth freshener and also having been used in Ayurvedic medicine since ages. But first time we are reporting the specific compounds from this plant having antibacterial properties against the deadly pathogens we studied." says Debmalya Barh. "We also have identified certain compounds in this plant that are effective against multi-drug resistant broad-spectrum deadly pathogens that we will disclose soon".

Betel leaves are an economic crop of IIOAB's surrounding locality which produces best quality leaves in India. Barh and his colleagues are now advancing the research to establish Piper betel as a source of next-generation antibiotics to fight against broad-spectrum deadly infectious diseases

http://www.medicalnewstoday.com/releases/256338.php

About Medical News Today *(www.medicalnewstoday.com): MTN is the healthcare internet publishing market leader for medical news. It is in the top 360 United States sites and top 120 United Kingdom sites and receives more than 12 million monthly visits, 10 million monthly unique visitors and 15 million monthly page views as reported by Quantcast. It contents are based on evidence-based, peer-reviewed studies, along with accurate, unbiased and informative content from governmental organisations (e.g. FDA, CDC, NIH, NHS), medical societies, royal colleges, professional associations, patients' groups, pharmaceutical and biotech companies, among others and targeted to an educated audience of both healthcare professionals and patients alike.*

IV.5 Chapter V: Research Article

Exoproteome and Secretome Derived Broad Spectrum Novel Drug and Vaccine Candidates in *Vibrio cholerae* Targeted by Piper betel Derived Compounds.

Debmalya Barh, Neha Barve, Krishnakant Gupta, Sudha Chandra, Neha Jain, Sandeep Tiwari, Nidia Leon-Sicairos, Adrian Canizalez-Roman, Anderson Rodrigues dos Santos, Syed Shah Hassan, Sıntia Almeid, Rommel Thiago Juca Ramos, Vinicius Augusto Carvalho de Abreu, Adriana Ribeiro Carneiro, Siomar de Castro Soares, Thiago Luiz de Paula Castro, Anderson Miyoshi5, Artur Silva, Anil Kumar, Amarendra Narayan Misra, Kenneth Blum, Eric R. Braverman, **Vasco Azevedo**

PLoS One. 2013;8(1):e52773. doi: 10.1371/journal.pone.0052773. Epub 2013 Jan 30 **[PMID: 23382822]**
Impact Factor: 4.4 (2013)

Targets from exoproteome and secretome are best candidates for developing next-generation antibacterial drugs and vaccines. In the omics era, peptide vaccines are the best choice. In this chapter, comparative proteomic strategy coupled with a modified reverse vaccinology approach are described to identify exoproteome and secretome derived novel broad spectrum drug and vaccine targets in 21 *Vibrio cholerae* serotypes. The strategy includes, subtractive proteomics, conventional and PPIs and host-pathogen interactions based target prioritization, antigenic B-cell derived T-cell epitope prediction, 3D modelling of drug and vaccine targets, epitope design, topology analysis of epitopes, validation of epitopes using IEDB, virtual screening of piper betel compounds against drug targets, and experimental validation of targeting compounds against *V. cholerae* O1 Inaba. ompU, uppP and yajC are found to be candidate targets for most of the pathogenic *Vibrio* strains where uppP and yajC are novel targets in *Vibrio*. uppP and ompU may be used to develop both drugs and vaccines against broad spectrum *Vibrio* serotypes. Seven Piper betel compounds found to target these targets in *in silico* and show anti- *Vibrio* effects *in vitro*.

OPEN ACCESS Freely available online

PLOS ONE

Exoproteome and Secretome Derived Broad Spectrum Novel Drug and Vaccine Candidates in *Vibrio cholerae* Targeted by *Piper betel* Derived Compounds

Debmalya Barh[1,2]*, Neha Barve[1,3], Krishnakant Gupta[1,3], Sudha Chandra[1,3], Neha Jain[1], Sandeep Tiwari[1], Nidia Leon-Sicairos[4], Adrian Canizalez-Roman[4], Anderson Rodrigues dos Santos[5], Syed Shah Hassan[5], Síntia Almeida[5], Rommel Thiago Jucá Ramos[6], Vinicius Augusto Carvalho de Abreu[5], Adriana Ribeiro Carneiro[6], Siomar de Castro Soares[5], Thiago Luiz de Paula Castro[5], Anderson Miyoshi[5], Artur Silva[6], Anil Kumar[3], Amarendra Narayan Misra[2,7], Kenneth Blum[1,8,9], Eric R. Braverman[10], Vasco Azevedo[5]

1 Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India, **2** Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, Orissa, India, **3** School of Biotechnology, Devi Ahilya University, Indore, India, **4** Unit for research, School of Medicine, Autonomous University of Sinaloa, Cedros y Sauces, Fracc. Fresnos, Culiacan, Mexico, **5** Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Minas Gerais, Brazil, **6** Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Para, Brazil, **7** Center for Life Sciences, School of Natural Sciences, Central University of Jharkhand, Ranchi, Jharkhand State, India, **8** College of Medicine, University of Florida, Gainesville, Florida, United States of America, **9** Center for Clinical and Translational Science, College of Medicine, University of Vermont, Burlington, Vermont, United States of America, **10** Weill-Cornell College of Medicine, Cornell University, New York, New York, United States of America

Abstract

Vibrio cholerae is the causal organism of the cholera epidemic, which is mostly prevalent in developing and underdeveloped countries. However, incidences of cholera in developed countries are also alarming. Because of the emergence of new drug-resistant strains, even though several generic drugs and vaccines have been developed over time, *Vibrio* infections remain a global health problem that appeals for the development of novel drugs and vaccines against the pathogen. Here, applying comparative proteomic and reverse vaccinology approaches to the exoproteome and secretome of the pathogen, we have identified three candidate targets (*ompU*, *uppP* and *yajC*) for most of the pathogenic *Vibrio* strains. Two targets (*uppP* and *yajC*) are novel to *Vibrio*, and two targets (*uppP* and *ompU*) can be used to develop both drugs and vaccines (dual targets) against broad spectrum *Vibrio* serotypes. Using our novel computational approach, we have identified three peptide vaccine candidates that have high potential to induce both B- and T-cell-mediated immune responses from our identified two dual targets. These two targets were modeled and subjected to virtual screening against natural compounds derived from *Piper betel*. Seven compounds were identified first time from *Piper betel* to be highly effective to render the function of these targets to identify them as emerging potential drugs against *Vibrio*. Our preliminary validation suggests that these identified peptide vaccines and *betel* compounds are highly effective against *Vibrio cholerae*. Currently we are exhaustively validating these targets, candidate peptide vaccines, and *betel* derived lead compounds against a number of *Vibrio* species.

Citation: Barh D, Barve N, Gupta K, Chandra S, Jain N, et al. (2013) Exoproteome and Secretome Derived Broad Spectrum Novel Drug and Vaccine Candidates in *Vibrio cholerae* Targeted by *Piper betel* Derived Compounds. PLoS ONE 8(1): e52773. doi:10.1371/journal.pone.0052773

Editor: Anil Kumar Tyagi, University of Delhi, India

Received September 18, 2012; **Accepted** November 21, 2012; **Published** January 30, 2013

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dr.barh@gmail.com

Introduction

Vibrio cholerae is a noninvasive gram-negative bacterium that causes water borne disease cholera, which is characterized by profuse watery diarrhea and vomiting [1]. The severity of the diarrhea and vomiting causes rapid dehydration and electrolyte imbalance that leads to death. The *V. cholerae O395* strain is a classical O1 serotype strain responsible for cholera epidemics in Asian countries, and the non-O1 sero-group *Vibrio cholerae O139* has been implicated as the causative agent of sporadic cases of gastro-enteritis and extra-intestinal infections [2,3]. Both of the strains have been reported to cause significant numbers of morbidities [4,5]. Although considerable research is ongoing to develop new drugs and vaccines and many antibiotics are already used to treat cholera, the infection remains frequently uncontrolled because of emerging antibiotic resistance of the pathogen [6,7,8]. Therefore, novel drugs and vaccines must be developed to tackle the *Vibrio* infection and transmission.

The identification of antigenic and virulence factors is paramount in developing antibiotics against a pathogen. In most cases, exomembrane (surface exposed) and secretary proteins exhibit antigenicity and virulence and are therefore suitable for targeting. Similarly, in the post-genomics era, computational approaches for the identification of genomic targets [9] and the use of reverse vaccinology [10] are becoming popular for rapid

identification of novel targets to develop both drugs and vaccines against any given pathogen.

The present study aims to identify broad spectrum and novel drug and vaccine targets for a number of *Vibrio* strains, including *V. cholerae* strains O395 and *O139*; to design peptide vaccines; and to identify lead natural compounds first time from *Piper betel*, a well-known plant with medicinal value, to make use of these targets.

Materials and Methods

Drug and Vaccine Target Prioritization Parameters

Target prioritization in pathogenic microorganisms is accomplished in various ways [11]. Among these important prioritization considerations are subcellular localization, non-host homolog essential genes, core pathogen genes, pathogenic island association, involvement in the pathogen's unique metabolic pathways, druggability, availability of 3D structural information, and low molecular weight of the target protein (≤110 kDa) [9,11–14]. The exoproteome and secretome are good source of targets for developing vaccines and drugs (dual targets) [9,10]. Therefore, we first screened the exoproteome and secretome of the pathogen for potential targets, followed by the application of other prioritization parameters to identify targets.

Screening of the Exoproteome and Secretome and Target Identification

We applied the classical reverse vaccinology strategy [10] and a modified method of subtractive proteomics [15,16] to identify candidate drug vaccine targets in *V. cholerae* strain *O395* and other *Vibrio* serotypes. In brief, the *Vibrio cholera O395* proteome, which consists of 3875 proteins, was screened using CELLO [17], PSLpred [18], PSORTb [19], SOSUI-GramN [20], and SurfG$^+$ [21] to identify the exomembrane and secreted proteins. Thereafter, the essential, non-human homolog *Vibrio* proteins (putative targets) from the pool of exoproteome and secretome were identified using the Database of Essential Genes (DEG) [22] and NCBI BLASTp [23], as described by Barh *et al.*, 2011 [15]. Selected non-human homolog essential *Vibrio* proteins were then checked for their pathway involvement, and the best targets were selected based on the involvements of these targets in the unique essential bacterial metabolic pathways and another twelve criteria as described by Barh *et al.*, 2011 [15] for target selection. The final list of identified targets was then checked for their presence in different *Vibrio* strains and related species using NCBI prokaryotic genome BLASTp.

Additional Evaluation of the Essentiality Parameters of Targets

The DEG-based essentiality of the identified targets was further validated using sequence-based computational approaches: (i) strand-bias; (ii) codon adaptation index (CAI); (iii) patterns of enzyme classes distributed, and (iv) clusters of orthologous groups (COG) of proteins. Essential genes are mostly located on leading strands and show strand bias [24]. We used Ori-Finder [25] to check the replication origin- and replication termini-based determination of strand-bias and the localization of the identified target genes in leading or lagging strands. CAI values are reported as one of the measures to evaluate essential genes, with a CAI >0.5 indicative of significant essentiality [26]. We used ACUA software [26] to calculate the CAI values of our identified targets. The distribution of enzyme classes of the targets was determined with BRENDA [27] and UniProtKB [28]. The targets were also examined for their bias toward COG functional subcategories for essentiality as per the findings of Lin *et al.*, 2010 [29].

Evaluation of Prioritization Parameters of Targets

We further checked the identified targets for their molecular weight if they are of ≤110 kDa using UniProtKB [28]. The druggability of the targets was determined using the DrugBank database [30]. The amino acid sequences of identified potential targets were aligned using BLASTp with a cutoff *E-value* = 0.01 against the DrugBank -listed targets for which specific compounds are available in the database. The availability of 3D structural information of targets was verified with PDB [31]. When structures were not available, a homology modeling or threading

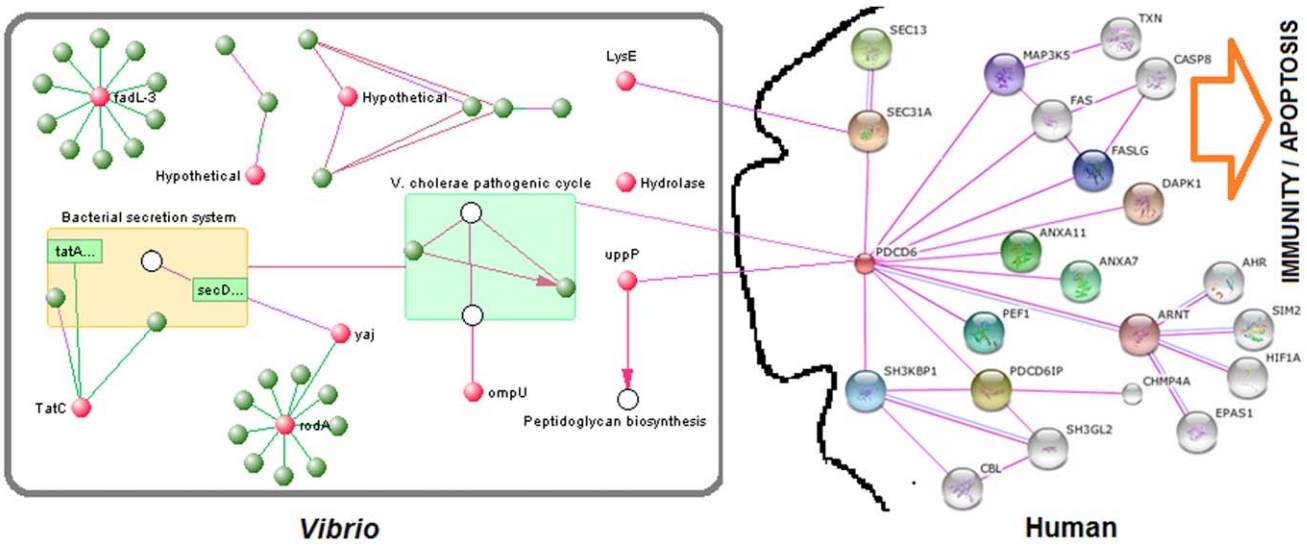


Figure 1. Protein-protein and host-pathogen interactions among ten preliminary identified *Vibrio* targets. The interactions demonstrate that the finally selected three targets (*ompU*, *yajC*, and *uppP*) are involved in *Vibrio* pathogenesis and modulate host response (immunity and apoptosis) by interacting with the host protein PDCD6.
doi:10.1371/journal.pone.0052773.g001

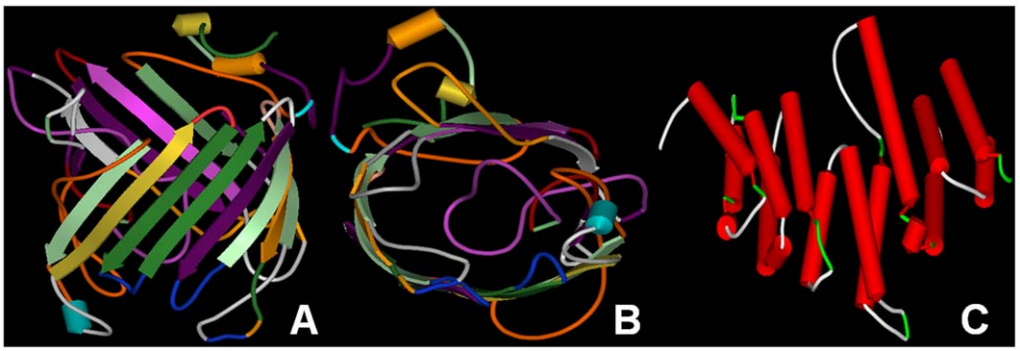


Figure 2. 3D models of *Vibrio* targets constructed using threading approaches. A) Front view of *ompU*, B) Side view of *ompU*, and C) Front view of *uppP*.
doi:10.1371/journal.pone.0052773.g002

approach was performed and verified with various 3D modeling parameters (see 3D modeling of targets section).

Additional Prioritization Parameters of Targets by PPIs and Hostpathogen Interactions

Target identification based on host-pathogen interactions has been implemented in many organisms, including *M. tuberculosis* [32]. Therefore, to verify the reliability of our identified targets, we searched for protein-protein interactions (PPIs) among the identified targets and also host-pathogen interactions. All *V. cholera O395* targets were selected to make PPI networks using VisAnt [33]. Further, KEGG pathways [34] were incorporated into the PPI networks and analyzed for their involvement in bacterial pathogenesis and essential pathways. To identify host-pathogen interactions, 20,000 experimentally-validated host-pathogen interactions for 24 pathogens were downloaded from the PathoSystems Resource Integration Center (PATRIC) database [35]. In PATRIC, *Vibrio*-specific host-pathogen interaction data are not available. Therefore, we used sequences from pathogens listed in PATRIC that are 90% homologous to our identified *Vibrio* targets to determine interactions and interacting human counterparts. The interacting human counterparts were also analyzed for their involvement in key biological processes and pathways involved in host response to infection, such as immunity and apoptosis, and examined whether they are key nodes in those pathways using the Search Tool for the Retrieval of Interacting Genes (STRING) [36] and the Database for Annotation, Visualization and Integrated Discovery (DAVID) [37]. Targets that are involved in bacterial pathogenesis or essential pathways and interact with key molecules in host response pathways are generally more effective targets.

Prediction of Antigenic B-cell Derived T-cell Epitopes

Once the targets are finalized, the novel strategy of epitope designing as described by Barh *et al.*, 2011 [16] was applied to design peptide vaccines from the vaccine targets. Briefly, the secreted and exomembrane proteins were checked for antigenicity using the VaxiJen v2.0 server (threshold = 0.4, ACC output) [38], and thereafter, their virulences were predicted using VirulentPred [39]. Proteins that were antigenic according to VirulentPred and showed an antigenicity score >0.5 in VaxiJen were selected. The exomembrane sequences of each virulent protein commonly derived from VaxiJen and VirulentPred analysis were determined by TMHMM v2.0 [40]. The BCPreds server [41] was used for B-cell epitope prediction (cutoff >0.8, 20-mer epitopes) and epitope sequences were matched with surface-exposed sequences of corresponding proteins. The surface-exposed B-cell epitope sequences were further checked for antigenicity using VaxiJen, and the best epitopes were selected for T-cell epitope prediction using ProPred [42] and ProPred I [43]. QSAR-based simulation analysis of each T-cell epitope was performed by MHCPred v.2 [44] and VaxiJen to detect half maximal (50%) inhibitory concentration (IC_{50}) and antigenicity, respectively. For a second level confirmation, the selected T-cell epitopes were further screened by T-epitope designer [45], and epitopes were selected that showed binding affinity to ≥80% of HLA molecules, including the A*0201, A*0204, and B*2705, DRB1*0101 and

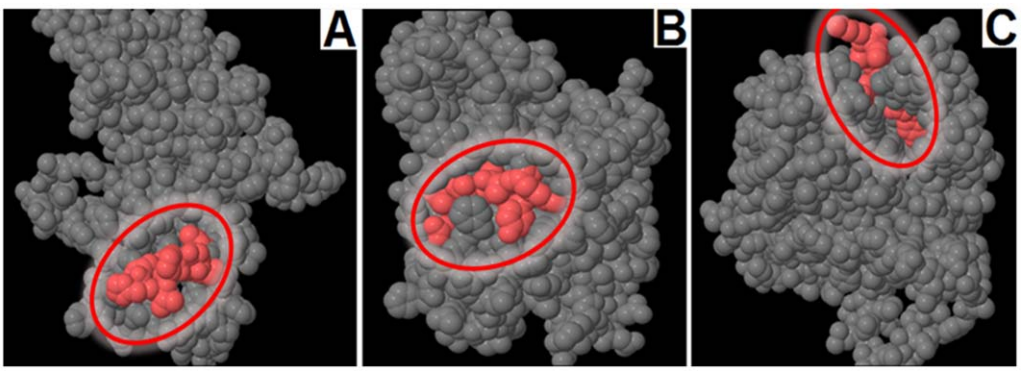


Figure 3. Pepitope analysis of identified T-cell epitopes for their exomembrane topology (colored in red) within the corresponding folded proteins. A) The ''VTSGEPVHS'' epitope of *uppP*, B) the ''VTETNAAKY'' epitope of *ompU*, and C) the ''YNNAETAKK'' epitope of *ompU*.
doi:10.1371/journal.pone.0052773.g003

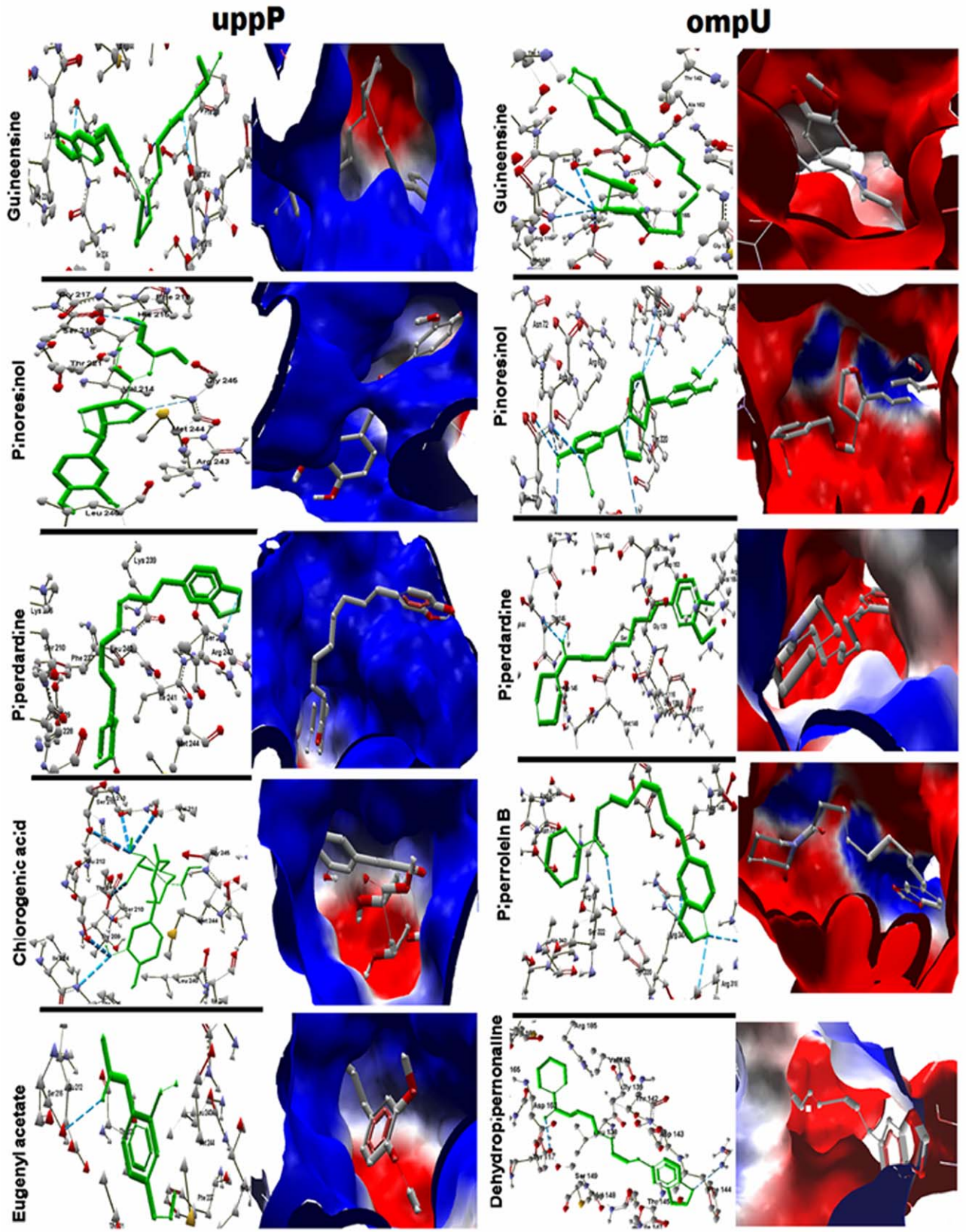


Figure 4. The best seven *Piper betel* compounds that may render activities of *Vibrio* targets *ompU* and *uppP*. GOLD fitness and Moldock scores were considered to select the compounds. Guineesine, Pinoresinol, and Piperdardine inhibit both targets. Dehydropipernonaline and Piperrolein B are effective on *ompU*. Chlorogenic acid and Eugenyl acetate are good ligands for *uppP*.
doi:10.1371/journal.pone.0052773.g004

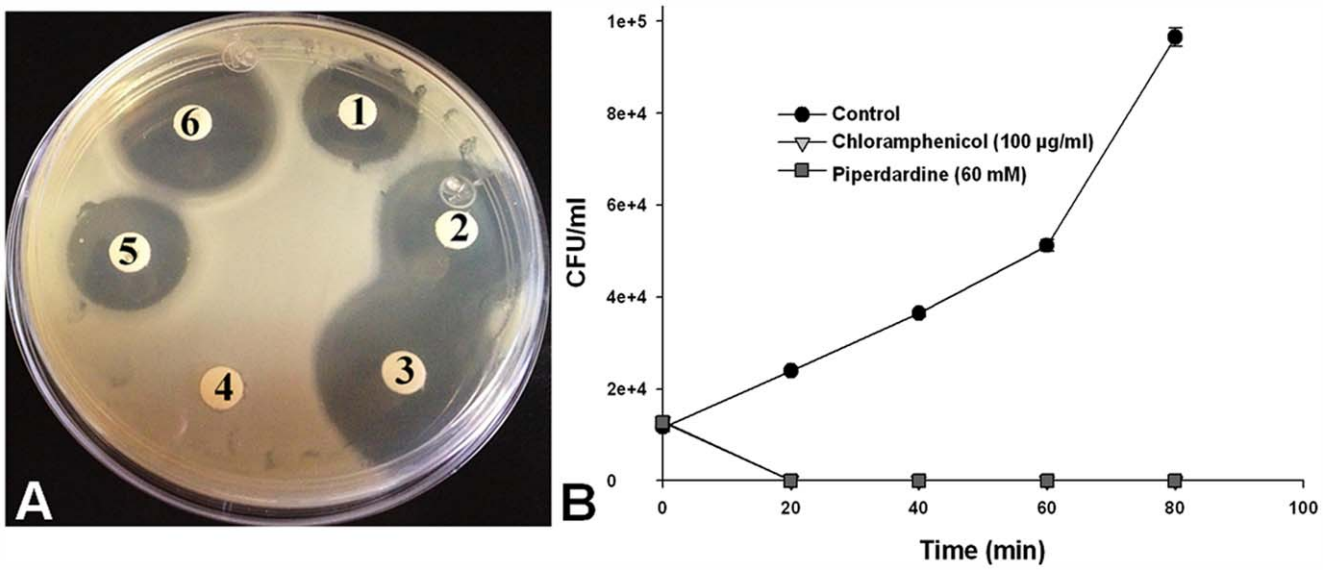


Figure 5. Anti-*Vibrio* activity of Piperdardine. A). Growth inhibition effects Piperdardine, Ampicillin, and Chloranphenicol on *V. Cholerae O1 Inaba* growth as per the disk diffusion method. 1) 100 mM, 2) 200 mM, and 3) 300 mM Piperdardine; 4) water; 5) Ampicillin (10 µg); and 6) Chloranphenicol (30 µg). The zones of inhibition (mm) around disks containing Piperdardine are concentration-dependent: 1) 19.3±0.03; 2) 26.23±0.1; 3) 28.65±0.16. Controls: 4) 0±0; 5) 18.51±0.16; and 6) 29.47±0.16. B). Effects on Piperdardine and Chloranphenicol on *V. Cholerae O1 Inaba* growth as per the Colony-forming units (CFU/ml) assay. As per the method described in the text, 60 mM of Piperdardine (squares) shows anti-*Vibrio* effect similar to 100 µg/ml of Chloramphenicol (triangles).
doi:10.1371/journal.pone.0052773.g005

DRB1*0401 alleles that are most common in the human population. Finally, epitopes that bound more than 13 MHC molecules in ProPred and ProPred-I with less than 100 nM IC_{50} for DRB1*0101 in MHCPred v2.0 and that bound ≥80% of HLA molecules in T-epitope designer were selected for fold-level topology analysis to select the best epitopes.

3D Modeling of Targets and Topology Analysis of Epitopes

For topology analysis of the identified epitopes and for virtual screening, the target proteins were modeled. The Phyre2 server [46] was used for homology modeling, and the threading approach was performed using the I-TASSER server [47]. The homology-based models were validated using the Structure Analysis and Verification Server (SAVS) Vs.4 (http://services.mbi.ucla.edu/SAVES/), and threading-based models were based on confidence scores (C-score range −5.0 to +2.0) and TM-scores of the resultant protein models. Further loop refinement of threading-based models was done by the ModLoop server [48], and finally, structure verification was performed by ERRAT plot version 2.0 [49], RAMPAGE [50], and the Dali server [51]. The localization and positioning of the epitopes within the folded proteins were analyzed using Pepitope server [52].

Ligand Library Preparation and Virtual Screening

Piper betel, one of the economic crops of West Bengal, India, is reported to have various medicinal and antimicrobial properties. However, no specific compound from this plant has so far been tested for antibacterial property. We collected 128 natural compounds of *Piper betel* from published literature to construct our ligand library. The library was also enriched with 35 well known antibiotics that are used to treat cholera with an aim to compare the efficacy of *betel* compounds with these antibiotics. The catalytic pockets of identified targets were determined using Molegro Virtual Docker (MVD) [53], CASTp [54], Pocket-Finder [55], and Active Site Prediction Server [56]. GOLD 4.1.2 software [57] was used for virtual screening. The best five *betel* derived ligands and antibiotics based on GOLD fitness scores and negative binding energy were selected and further validated using RMSD and MolDock scores in Molegro Virtual Docker 4.2.0 screening. The efficacy of top five *betel* compounds in respect to the top five antibiotics were determined based on GOLD fitness and Molegro Virtual Docker scores.

Preliminary Validation of Epitopes using IEDB and Betel Compounds against *V. cholerae O1 Inaba*

We preliminary validated the identified candidate peptide vaccines using the Immune Epitope Database (IEDB) [58]. One of the identified candidate *betel* compounds was also checked for its anti-*Vibrio* properties against *V. cholerae O1 Inaba*. The bacteria were maintained in Mueller-Hinton (MH) Broth, placed on a shaker incubator and grown at 37°C for 16–18 h, to reach the logarithmic phase. After that, bacterial cultures were adjusted to an absorbance of 0.1 at 600 nm (1×10^7 UFC/ml) to test the bactericidal activity of the candidate *betel* compound by two methods. **a)** *Disk diffusion method*: MH agar plates were prepared and spreaded with 1×10^7 UFC of bacterial cultures, and then sensi-disks (Ampicillin and Chloramphenicol) and disks impregnated with the *betel* compound (dissolved in water, at concentrations of 20, 40, 60, 80, 100, 200 and 300 mM); were placed on MH agar plates and were incubated at 37°C for 24 h. Finally, the zone inhibition was measured by using a Vernier caliper. To test the comparative efficacy of the candidate *betel* compound in respect to conventional anti-Vibrio antibiotics, we performed **b)** *Colony-forming units (CFU/ml) assay*: Here, 1×10^7 UFC/ml of bacterial suspension were resuspended in tubes containing MH broth, alone (control for bacterial growth) or incubated with 100 µg/ml of Chloramphenicol (control for bacterial inhibition), or with 20, 40, 60, 80, and 100 mM of candidate *betel* compound. After that, tubes were incubated at 37°C for 0, 20, 40, 60 and

80 min in shaking. Finally, the number of viable bacteria was counted each time by obtaining the CFU/ml from serial 10-fold dilutions prepared in MH broth and plating onto MH agar.

Results and Discussion

Genome Screening and Target Identification

We identified 513 membrane (160 from Ch-I and 353 from Ch-II) and 317 secreted (113 from Ch-I and 204 from Ch-II) proteins for a total of 830 proteins based on our exoproteome and secretome analysis of *V. cholerae* strain *O395*. The *V. cholerae* strain *O395* proteome consists of 3875 proteins; therefore, 13.2% and 8.18% of proteins of the entire *Vibrio* proteome constitutes exoproteome and secretome, respectively. DEG-based essential gene analysis revealed only 178 essential proteins (119 exomembrane and 59 secreted) out of the total 830. Only 10 essential proteins (7 exomembrane and 3 secreted) were found to be non-human homologs and therefore probable targets (**Table S1**).

As shown in **Table S2**, among these 10 proteins, 3 are hypothetical (*VC0395_0360*, *VC0395_A1375*, and *VC0395_A2856*). The antigenicity and virulence analysis showed that 9 out of these 10 proteins are antigenic and that 7 are virulent. All 10 proteins were further analyzed for their involvement in the pathogen's essential unique pathways using the KEGG pathway database. The three hypothetical proteins and *LysE* did not show any pathway involvement and therefore were removed from the analysis. *fadL-3* (Long chain fatty acid transport protein) does not show any vital role in any bacterial essential pathway and was therefore also eliminated. *rodA* (rod shape determining protein) is involved in the regulation of cell shape processes [59] and is essential for *Vibrio*; however, it is not an essential gene for *S. aureus* [60] and also did not provide any T-cell epitopes in further analysis.

Cell membrane-localized *TatC* (sec-independent translocase protein) was identified as an interesting target in *Vibrio*. *TatC* is a virulent protein and is involved in pathways such as membrane transport and the bacterial secretion system. *TatC* has been reported as a target in *M. leprae* [61] and *Klebsiella pneumonia MGH78578* [62]. However, according to the AEROPATH Target Database (http://aeropath.lifesci.dundee.ac.uk/), in *P. aeruginosa*, *TatC* is not an essential gene and it also did not generate any B-cell derived T-cell epitopes in further analysis.

The secreted protein *ompU/ VC0395_A0162* (Outer membrane protein *ompU*) was found to be an important target as it is involved in the *V. Cholerae* pathogenic cycle. *ompU* is involved in host cell invasion during *Vibrio* infections [63], and for pathogenic *Vibrio harveyi SF-1*, it is reported as a candidate subunit and DNA vaccine [64].

The second most important target is membrane-localized *yajC/ VC0395_A0472* (Preprotein translocase subunit *yajC*), which is involved in the bacterial secretion system, a vital pathway for bacterial survival. The *C. botulinum yajC* is reported as a putative target [65] and is also listed as a target in *M. leprae*, *M. tuberculosis*, and *Wolbachia endosymbiont* of *Brugia malayi* in the TDR Targets Database [66]. Our analysis also showed that both of these proteins from *Vibrio* are exomembrane/secreted, antigenic, and highly virulent and are therefore suitable for vaccine and drug design where *yajC* is a novel candidate target for *Vibrio* (**Table S2**).

Apart from these two vaccine targets, the third important target we identified is the membrane bound enzyme *uppP/ VC0395_A0054* (Undecaprenyl pyrophosphate phosphatase) because of its vital role in the bacterial-specific peptidoglycan biosynthesis pathway and its involvement in cell wall biosynthesis. *uppP* is reported as an antibiotic resistant gene [67] and is also a listed target for *M. Leprae* and *M. tuberculosis* in the TDR Targets Database [66]. However, we are reporting *uppP* for the first time as a target in *Vibrio*, therefore it is a novel target for this pathogen.

We used 21 *Vibro* species (both pathogenic and non-pathogenic) available in NCBI and when we searched these three targets (*ompU*, *yajC*, and *uppP*) for their presence among these *Vibrio* species using comparative BLASTp in NCBI server, we found that all three targets are present in 12 species, including the virulent strains *Vibrio anguillarum 775*, *Vibrio cholerae O1 biovar El Tor str. N16961*, *Vibrio splendidus LGP32*, *Vibrio cholerae O395*, and *Vibrio harveyi*, and the non-virulent strain *Vibrio fischeri ES114*. Therefore, all of these selected targets can be used for broad-spectrum drug and vaccine design for a number of *Vibrio* serotypes (**Table S2**).

Additional Evaluation of the Essentiality Parameters of Targets

The identified targets *ompU*, *yajC*, and *uppP* were further verified with additional parameters for essentiality in the pathogen genome. Essential enzymes are better targets [15], and most of the essential enzymes belong to the following enzyme classes: transferases, oxidoreductases, ligases, hydrolases, lyases, and isomerases [68]. Among the three targets, *uppP* (EC = 3.6.1.27) is a hydrolase and therefore meets the criteria to be an essential gene. The other two proteins (the secreted protein *ompU* and the preprotein translocase subunit *yajC*) are not enzymes; thus, additional analyses for essentiality were done using a combination of strand-bias, CAI, and COG-bias analysis. The strand-bias analysis showed that these three targets are located in the leading strand and that the codon adaptation indexes (CAI) are 0.63, 0.58, and 0.80, respectively, for *ompU*, *yajC*, and *uppP*, satisfying the cutoff value of >0.5 for being an essential gene. Previous reports have suggested that the essential genes of *M. ulcerans* belong to COG subcategories E, H, J. D, N, V and M [68]. Our identified targets *ompU*, *yajC*, and *uppP*, respectively, belong to M (cell envelope and membrane biogenesis), N (cell motility and secretion), and V (cellular processes and signaling) categories. Therefore, these three targets are essential as per the COG-bias analysis also.

Evaluation of the Prioritization Parameters of Targets

Proteins with molecular weight $\leq$110 kDa are proposed to be effective targets [68]. The molecular weights of *Yajc*, *uppP*, and *ompU* are 11.9 kDa, 29,3 kDa, and 37,7 kDa, respectively; therefore, these proteins are of the low molecular weight. This parameter is highly desirable for a target so that the target can be easily purified for further validation [69]. Targets are preferably druggable [70], and 3D structure is required for *in silico* drug discovery by modeling, virtual screening, and druggability analysis. The druggability of these three targets was first tested using a DrugBank search to determine if specific compounds are available against these targets. The results showed that only *ompU* is potentially druggable by small molecules such as N-(6,7,9,10,17,18,20,21-octahydrodibenzo[b,k] [1,4,7,10,13,16]hexaoxacyclooctadecin-2-yl) acetamide, Dodecane, and (Hydroxyethyloxy)Tri(Ethyloxy)Octane, N-Octyl-2-Hydroxyethyl Sulfoxide. However, the *E-values* were high. No molecule was found to target *Yajc* and *uppP* in DrugBank. The druggability analysis using DrugBank was negative, potentially because of the novel nature of these identified targets, the non-availability of their 3D structures in PDB, and no previous study on their druggability aspects. Therefore, in this study, we attempted to model these three targets and further tested for druggability using virtual screening.

PPIs and Host Pathogen Interactions

We used all 10 initially identified targets, including the hypothetical proteins, to make PPI networks of the targets in *V. cholera O395*. The phylogenetic analysis and domain fusion-based PPI networks show that with the exception of *VC0395_0360* (putative hydrolase/Hypothetical) and *LysE*, all targets interact with other *Vibrio* proteins. The KEGG-based analysis of the PPI networks reveals that the selected targets *ompU*, *yajC*, and *uppP* are involved in the *V. cholera* pathogenic cycle, bacterial secretion system, and peptidoglycan biosynthesis pathways, respectively. All of these pathways are unique to bacteria and involved in pathogenesis. Therefore, the PPIs-based analysis also supports our selected final three targets.

Our host-pathogen interaction analysis revealed that only *uppP* and *LysE* have host protein interacting counterparts. Our selected target *uppP*, which is involved in the peptidoglycan biosynthesis pathway, directly interacts with or binds to the PDCD6 (Programmed cell death 6) protein of the human host. The gene enrichment, pathway, and centrality analyses show that PDCD6 is a key molecule in host immunity and the apoptosis pathway (**Figure 1**). The other target, *LysE*, interacts with the host SEC31A (Protein transport protein SEC31A). SEC31A is also involved in immunity and apoptosis in the host but is not a key molecule in these pathways. *LysE* is also not involved in any bacterial pathogenesis pathway, and the exclusion of *LysE* from the final list of targets is therefore justified. Although the two selected targets *ompU* and *yajC* do not directly interact with any host protein, the network analysis showed that the pathways in which these two proteins are involved (the bacterial secretion system and the *V. cholerae* pathogenic cycle) are interlinked and that some proteins in the *V. cholerae* pathogenic cycle interact with PDCD6 (**Figure 1**). Therefore, these two selected targets (*ompU* and *yajC*) indirectly interact with PDCD6, leading us to the observation that all of our final selected targets (*ompU*, *yajC*, and *uppP*) interact with PDCD6 and modulate host response in terms of modulation of immunity and the apoptosis pathway in the host.

3D Modeling

We first attempted to model *ompU*, *yajC*, and *uppP* using the Phyre 2 server. However, the attempt failed because of unavailability of the proper template. We therefore developed threading-based 3D structure models of these proteins. We were able to model *ompU* and *uppP* using I-TASSER; however, we could not model the *yajC* protein using this approach. Models were validated using the RAMPAGE, ERRAT plot, and Dali servers. Models were found to satisfy all criteria (**Table S3, A-E**). The 3D models of *uppP* and *ompU* are represented in **Figure 2**.

Epitope Design

Antigenicity and cell-exposed sequences. A good epitope should be cell exposed and antigenic. Therefore, these three targets *ompU*, *yajC*, and *uppP* were first analyzed using VaxiJen and then by TMHMM. The antigenicity scores of these three proteins were found to be 0.766, 0.744 and 0.484, respectively, for *ompU*, *yajC*, and *uppP*; therefore, they are all highly antigenic (**Table S2, column-4**). The TMHMM-based exomembrane region for *ompU* is 1–350 amino acids and is therefore fully exposed to the outside of the membrane. The cell-exposed amino acid sequences of *uppP* are 30–84, 132–156 and 206–219, and for *yajC*, the sequence is 1–14 (**Table S4, column 6**).

Antigenic B-cell epitope-derived T-cell epitopes. Using the approach described above, we identified one B-cell epitope from *yajC*, two from *uppP*, and thirteen from *ompU* (**Table S4, column 2**). However, when we analyzed for the presence of T-cell epitopes within these B-cell epitopes according to our selected criteria, *yajC* did not produce any T-cell epitope. *ompU* generated two ("VTETNAAKY" and "YNNAETAKK") and *uppP* only one ("VTSGEPVHS") epitope satisfying all of our criteria (**Table S5**). The entire protein sequence of *uppP* is non-virulent, but this single epitope is highly virulent and antigenic. Therefore, the *uppP* protein is a candidate novel vaccine target for *Vibrio*. The Pepitope analysis also showed that all of the identified T-cell epitopes are of the exomembrane topology within their corresponding folded proteins (**Figure 3**).

Drug Target and Virtual Screening

Since *ompU* is a secreted and *uppP* is an exomembrane protein, they are also suitable drug targets. The *Piper betel* leaf is used in folk medicine for treatment of several situations [72], and the leaf extracts are experimentally shown to be useful as antimicrobial [73], anti-leishmanial [74], antimalarial [75], anti-filarial [76], anti-fungal [77], anti-allergic [78], immunomodulator [79], gastroprotective [80], antioxidant [81], and anti-inflammatory [82] agents. We performed literature mining and collected 128 active phytochemicals from *betel* leaf and used them to screen against these two targets. The docking was done against the best cavity according to the Molegro virtual docker (MVD), CASTp, Pocketfinder, and Active Site Prediction Server (**Table S6**). The docking results based on the GOLD fitness score and Moldock score show that Guineesine, Pinoresinol, and Piperdardine can bind and render the activities of both the targets with high specificity. Apart from these three common compounds, Dehydropipernonaline and Piperrolein B were found to be effective on *ompU* and Chlorogenic acid and Eugenyl acetate on *uppP* (**Figure 4, Table S7A**). Several other *betel* compounds such as Piperardine and Peridine are also found to be effective against these targets however their GOLD fitness and Moldock scores are less. It should also be noted from the docking results that the *Piper betel* compounds are superior to the conventional antibiotics that are prescribed for the treatment of cholera in inhibiting these two targets (**Table S7B**).

Validation of Epitopes and Betel Compounds

Among the identified three candidate peptide vaccines, we found *ompU* derived "VTETNAAKY" is 80% identical to an experimentally validated linear peptide vaccine derived from adhesin P1 of *Mycoplasma pneumoniae M129* [71]. However, we could not get any similar peptide in IEDB for other two identified epitopes ((*ompU* derived "YNNAETAKK" and *uppP* based "VTSGEPVHS"), perhaps due to unavailability of similar peptides in IEDB or because of their novelty as candidate vaccines.

Piperdardine was used in this preliminary validation. This *betel* compound is found to be highly effective against *V. cholerae O1 Inaba* and the effect is concentration-dependent (**Figure 5A**). While we tested Piperdardine for its efficacy in respect to Chloramphenicol using growth kinetics assay, we observed that 60 mM of Piperdardine was able to inhibit *V. cholerae O1 Inaba* growth similar to100 µg/ml of Chloramphenicol treatment (**Figure 5B**). Form these assays; it's also evident that the anti-*Vibrio* efficacy of Piperdardine is better than that of Chloramphenicol, although Piperdardine requires a higher concentration. In this study, we did not check the target specificity of Piperdardine in *V. cholerae O1 Inaba*. However, currently we are conducting in-depth validations and target specificities of all identified *betel* compounds against a number of *Vibrio* species.

Conclusion

In summary, in this analysis, we have identified *ompU*, *uppP*, and *yajC* from the *Vibrio cholerae* strain *O395* secretome and membrane proteome as novel targets that can be useful in designing broad-spectrum peptide vaccines or drugs against most of the virulent strains of the pathogen. YNNAETAKK and VTETNAAKY from *ompU* and VTSGEPVHS from *uppP* were found to be effective candidate peptide vaccines. *Piper betel*-derived Piperdardine, Pinoresinaol, and Guineensine can target both *ompU* and *uppP*, whereas Dehydropipernonaline and Piperrolein B are specific inhibitors of *ompU* and Eugenyl acetate and Chlorogenic acid are specific to *uppP*. Most of these compounds show better efficacy than the currently-used anti-*Vibrio* drugs in our *in silico* analysis. Our validation results first time demonstrate that Piperdardine exhibits anti-*Vibrio* effects in a dose dependent manner and 60 mM of Piperdardine is having similar anti-Vibrio effect as 100 µg/ml of Chloramphenicol has. We are currently validating all of our identified targets, candidate peptide vaccines, and *betel* derived lead compounds against most of the *Vibrio* strains and serotypes available.

Supporting Information

Table S1 Final statistics of membrane and secreted essential proteins. The proteome of the *Vibrio cholerae* strain *O395* was screened using CELLO, PSLpred, PSORTb, SOSUI-GramN, and SurfG$^+$ to identify the membrane proteome and secretome. The genome contains a total of 3998 genes encoding 3875 proteins. The essentialities of these membrane and secreted proteins were determined by DEG-based BLASTp. The cutoff values for bit score, *E-value*, and percentage of identity at the amino acid level, respectively, were ≥100, $E = 0.0001$, and ≥40%. A total of 178 essential proteins were identified in which 119 are membrane located and 59 are secreted. Essential non-host homologs of the pathogen were identified using NCBI Human BLASTp with default parameters. A total of 10 (7 membrane and 3 secreted) essential non-host homologs was found.
(DOC)

Table S2 Features of the identified 10 targets in *V. Cholerae*. Ten *V. cholerae O395* targets were selected based on subtraction proteomics. *VC0395_0360* and *VC0395_0374* are located in Chromosome-I (Ch-I), whereas the other eight targets are located in Chromosome-II (Ch-II). Column-1 and Column-3, respectively, represent locus tags and target names. The blue-colored (*ompU*, *uppP* and *yajC*) meet all conditions for good targets and may be used for broad-spectrum drug and vaccine designing. These three targets are also common to twelve *Vibrio* species. Column 4 represents the COG categories. Column 5 provides detailed annotation of the corresponding *Vibrio* target. Column 6 provides the information on Virulence based on VirulentPred. VaxiJen-based antigenicity of the target *Vibrio* protein is provided in Column 7. Column 8 provides PARTIC and other analysis-based host proteins that interact with the corresponding targets. Columns 9–29 represent *Vibrio* strains/species tested for having identical targets in their genome/proteome based on homology. X represents absence of the target and √ represents presence. The last column represents the BLAST results of corresponding *Vibrio* targets with the human genome/proteome, and all targets show non-homology.
(DOC)

Table S3 A) Template and structure selection for modeling. Because the homology-based approaches for 3D modeling failed, we performed modeling using a threading approach. The three target proteins were submitted to the I-TASSER server, and we observed that the C-score (−5, 2) and TM score were in acceptable ranges. (i) Template selection for modeling Column-1 (The rank of templates) represents the top ten threading templates used by I-TASSER. Ident1 (Column-3) is the percentage sequence identity of the templates in the threading-aligned region with the query sequence. The Ident2 (Column-4) is the percentage sequence identity of the entire template with the query sequence. Coverage (Column-5) represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of the query protein. Column-6 represents the normalized *Z-score* of the threading alignments. Alignment with a normalized *Z-score* >1 indicates a good alignment. (ii) Target protein structure selection B) Energy of the protein-modeled structures The modeled structures were subjected to energy minimization. We performed energy minimization in the Swiss PDB Viewer and then checked using RAMPAGE and ERRAT plot. The energies of these two proteins were as follows. C) RAMPAGE results To validate the stereochemical properties of the two targets' modeled proteins, we used the RAMPAGE server. The expected percentages for residues in the favored region, allowed region, and outliers region are 98%, 2% and 0%, respectively. Our results demonstrated that the parameters of our modeled proteins are close to these cutoff values, and the models are therefore acceptable. D) ERRAT plot results for *ompU* and *uppP*. To further examine the non-bonded interaction of atoms in the models of the two targets, we used the Erraplot server. This server provides the quality factor of the modeled structure. Good, high-resolution structures generally produce quality factor values of approximately 95% or higher. For lower resolutions (2.5 to 3A), the average overall quality factor is approximately 91%. The following ERRAT plot criteria clearly show that our modeled proteins are of high quality. E) Validation of structures using the Dali server To provide strong support of the modeled structure, we performed structure-structure alignment in the Dali server and examined the function. We observed *Z-scores* of 2 that were greater than the threshold for a good alignment for both of the modeled proteins. Therefore, the models are acceptable for further structure-based *in silico* analysis.
(DOC)

Table S4 Identification of B-cell epitopes. As described in the methods, the amino acid sequences of *yajC*, *uppP* and *ompU* were subjected to the BCPreds server for B-cell epitope identification. The BCPreds and VaxiJen scores and the transmembrane topology for the selected B-cell epitopes from each target are listed in this table.
(DOC)

Table S5 Selected B-cell epitope-derived T-cell epitopes and their properties. The method is adopted as described by Barh et al., 2010 [16]. The final selected epitopes are highlighted in red.
(DOC)

Table S6 Active residues of *ompU* and *uppP* in the best cavity. We predicted the active residues for the largest cavity from Molegro Virtual Docker (MVD), and we verified our predictions with Cast-P, Pocketfinder and Active site prediction server. All predictions were in good agreement with the predicted result of MVD. However, in *uppP*, we observe a Histidine residue that is well known for ligand specification.
(DOC)

Table S7 Virtual screening for *uppP* and *ompU*. The docking was performed as described in the methods. The top five

ligands were selected based on their GOLD fitness score, MolDock score and RMSD. A ligand with a GOLD fitness score >25 is considered to be a good ligand. Similarly, the standard RMSD ranges from 0 to 4. Apart from electrostatic and hydrophobic interactions, more than 2 H-bonds indicate the ligand stability in the docked position.
(DOC)

Acknowledgments

We thank Erika Acosta-Smith for her efficient technical support while we carrying out microbial experiments.

Author Contributions

Coordinated entire work: DB. Performed all in silico analysis: DB NB KG SC NJ ST. Cross-analyzed exoproteome, secretome, and core genome: ARS SA VACA SSH SCS TLPC RTJR ARC. Conducted microbial experiments: NLS ACR. Provided timely consultation and reviewed the manuscript: AM AS AK ANM KB ERB VA. Read and approved the final manuscript: ACR ARC ANM AM ARS AK AS DB ERB KB KG NB NJ NLS RTJR ST SA SCS SC SSH TLPC VA VACA. Conceived and designed the experiments: DB. Performed the experiments: DB NB KG SC NJ ST NLS SSH ACR. Analyzed the data: DB NB KG SC NJ ST ARS SA VACA SCS TLPC RTJR SSH ARC. Wrote the paper: DB.

References

1. Sack DA, Sack RB, Nair GB, Siddique AK (2004). Cholera. Lancet 363 (9404): 223–233.
2. Shimada T, Sakazaki R (1977) Additional serovars and inter-O antigenic relationships of V. cholerae. Jpn J Med Sci Biol. 30 (5): 275–277.
3. Faruque SM, Albert MJ, Mekalanos JJ (1998) Epidemiology, genetics, and ecology of toxigenic Vibrio cholerae. Microbiol Mol Biol Rev. 62 (4): 1301–1314.
4. Siddique AK, Zaman K, Baqui AH, Akram K, Mutsuddy P, et al. (1992) Cholera epidemics in Bangladesh: 1985–1991. J Diarrhoeal Dis Res. 10 (2): 79–86.
5. Siddique AK, Akram K, Zaman K, Mutsuddy P, Eusof A, et al. (1996) Vibrio cholerae O139: how great is the threat of a pandemic? Trop Med Int Health. 1 (3): 393–398.
6. Fazil MH, Singh DV (2011) Vibrio cholerae infection, novel drug targets and phage therapy. Future Microbiol. 6 (10): 1199–1208.
7. Mandal S, Mandal MD, Pal NK (2011) Cholera: a great global concern. Asian Pac J Trop Med. 4 (7): 573–580.
8. Tran HD, Alam M, Trung NV, Kinh NV, Nguyen HH, et al. (2012) Multi-drug resistant Vibrio cholerae O1 variant El Tor isolated in northern Vietnam between 2007 and 2010. J Med Microbiol. 61(3): 431–437.
9. Barh D, Tiwari S, Jain N, Ali A, Santos AR, et al. (2011), In silico subtractive genomics for target identification in human bacterial pathogens. Drug Dev Res. 72: 162–177.
10. Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, et al. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287 (5459): 1816–1820.
11. Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, et al. (2008) Genomic-scale prioritization of drug targets: the TDR Targets database. Nat Rev Drug Discov 7: 900–907.
12. Caffrey CR, Rohwer A, Oellien F, Marhöfer RJ, Braschi S, et al. (2009) A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni. PLoS One 4(2): e4413.
13. Crowther GJ, Shanmugam D, Carmona SJ, Doyle MA, Hertz-Fowler C, et al. (2010) Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. PLoS Negl Trop Dis 4: e804.
14. Abadio AK, Kioshima ES, Teixeira MM, Martins NF, Maigret B, et al. (2011) Comparative genomics allowed the identification of drug targets against human fungal pathogens. BMC Genomics 12: 75.
15. Barh D, Jain N, Tiwari S, Parida BP, D'Afonseca V, et al. (2011) A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens. Chem Biol Drug. 78 (1): 73–84.
16. Barh D, Misra AN, Kumar A, Azevedo V (2010) A novel strategy of epitope design in Neisseria gonorrhoeae. Bioinformation 5 (2): 77–85.
17. Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 13 (5): 1402–1406.
18. Bhasin M, Garg A, Raghava GP (2005) PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 21: 2522–2524.
19. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, et al. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21: 617–623.
20. Imai K, Asakawa N, Tsuji T, Akazawa F, Ino A, et al. (2008) SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. Bioinformation 2: 417–421.
21. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, et al. (2009) Prediction of surface exposed proteins in Streptococcuspyogenes, with a potential application to other Gram-positive bacteria. Proteomics 9: 61–73.
22. Zhang R, Lin Y (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 37: D455–D458.
23. Gish W, States JD (1993) Identification of protein coding regions by database similarity search. Nat Genet. 3: 266–272.
24. Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377–378.
25. Gao F, Zhang CT (2008) Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinformatics 9: 79.
26. Vetrivel U, Arunkumar V, Dorairaj S (2007) ACUA: a software tool for automated codon usage analysis. Bioinformation 2: 62–63.
27. Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information.Nucleic Acids Res. 30(1): 47–9.
28. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011:bar009.
29. Lin Y, Gao F, Zhang CT (2010) Functionality of essential genes drives gene strand-bias in bacterial genomes. Biochem Biophys Res Commun 396: 472–476.
30. Knox C, Law V, Jewison T, Liu P, Ly S, et al. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.Nucleic Acids Res. 39(Database issue): D1035–41.
31. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, et al. (2006) The RCSB PDB information portal for structural genomics. Nucleic Acids Res. 34: D302–5.
32. Lin MY, Ottenhoff TH (2008) Host-pathogen interactions in latent Mycobacterium tuberculosis infection: identification of new targets for tuberculosis intervention. Endocr Metab Immune Disord Drug Targets. 8 (1): 15–29.
33. Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, et al. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. 37: W115–21.
34. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28: 27–30.
35. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, et al. (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect. Immun. 79: 4286–4298.
36. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39: D561–568.
37. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4 (1): 44–57.
38. Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 8: 4.
39. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 9: 62.
40. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3): 567–580.
41. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. J Mol Recognit. 21(4): 243–255.
42. Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17(12): 1236–7.
43. Singh H, Raghava GP (2003) ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics 19(8): 1009–1014.
44. Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCPred: A server for quantitative prediction of peptide-MHC binding, Nucleic Acids Res. 31(13): 3621–3624.
45. Kangueane P, Sakharkar MK (2005) T-Epitope Designer: A HLA-peptide binding prediction server. Bioinformation 1(1): 21–4.
46. Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 4(3): 363–371.
47. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 5(4): 725–738.
48. Fiser A, Sali A (2003) ModLoop: automated modeling of loops in protein structures. Bioinformatics 19(18): 2500–2501.
49. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2(9): 1511–1519.
50. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, et al. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 50(3): 437–450.
51. Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res. 38: W545–9.

52. Mayrose I, Penn O, Erez E, Rubinstein ND, Shlomi T, et al. (2007) Pepitope: epitope mapping from affinity-selected peptides. Bioinformatics 23(23): 3244–3246.
53. Thomsen R, Christensen MH (2006) MolDock: A New Technique for High-Accuracy Molecular Docking. J Med Chem. 49(11): 3315–3321.
54. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, et al. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.Nucleic Acids Res. 34: W116–8.
55. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model. 15(6): 359–363.
56. Singh T, Biswas D, Jayaram B (2011) AADS–an automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. J Chem Inf Model. 51(10): 2515–2527.
57. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. Proteins 52(4): 609–623.
58. Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, et al. (2012) Immune epitope database analysis resource. Nucleic Acids Res. 40: W525–30.
59. Hunt A, Rawlins JP, Thomaides HB, Errington J (2006) Functional analysis of 11 putative essential genes in Bacillus subtilis. Microbiology 152(10): 2895–2907.
60. Chaudhuri RR, Allen AG, Owen PJ, Shalom G, Stone K, et al. (2009) Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). BMC Genomics 10: 291.
61. Shanmugam A, Natarajan J (2010) Computational genome analyses of metabolic enzymes in Mycobacterium leprae for drug target identification. Bioinformation 4(9): 392–395.
62. Georrge JN, Umrania V (2011) In silico identification of putative drug targets in Klebsiella pneumonia MGH78578. Indian Journal of Biotechnology 10: 432–439.
63. Duperthuy M, Schmitt P, Garzón E, Caro A, Rosa RD, et al. (2011) Use of *ompU* porins for attachment and invasion of Crassostrea gigas immune cells by the oyster pathogen Vibrio splendidus. Proc Natl Acad Sci U S A 108(7): 2993–2998.
64. Sperandio V, Bailey C, Girón JA, DiRita VJ, Silveira WD, et al. (1996) Cloning and characterization of the gene encoding the *ompU* outer membrane protein of Vibrio cholerae.Infect Immun. 64(12): 5406–5409.
65. Reddy GK, Rao KN, Prasad PR (2011) Identification of drug and vaccine targets in Clostridium botulinum-A by the approach in-silico subtractive genomics. International Journal of Pharmaceutical Studies and Research 2: 48–54.
66. Magariños MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, et al. (2012) TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res. 40: D1118–27.
67. Liu B, Pop M (2009) ARDB-Antibiotic Resistance Genes Database. Nucleic Acids Res. 37: D443–7.
68. Butt AM, Nasrullah I, Tahir S, Tong Y (2012) Comparative Genomics Analysis of Mycobacterium ulcerans for the Identification of Putative Essential Genes and Therapeutic Candidates. PLoS One. 7(8): e43080.
69. Duffield M, Cooper I, McAlister E, Bayliss M, Ford D, et al. (2010) Predicting conserved essential genes in bacteria: in silico identification of putative drug targets. Mol Biosyst 6: 2482–2489.
70. Keller TH, Pichota A, Yin Z (2006) A practical view of 'druggability'. Curr Opin Chem Biol 10: 357–361.
71. Dallo SF, Su CJ, Horton JR, Baseman JB (1998) Identification of P1 gene domain containing epitope(s) mediating Mycoplasma pneumoniae cytoadherence. J Exp Med. 167(2): 718–723.
72. Valentão P, Gonçalves RF, Belo C, de Pinho PG, Andrade PB, et al. (2010) Improving the knowledge on Piper betle: targeted metabolite analysis and effect on acetylcholinesterase. J Sep Sci. 33(20): 3168–3176.
73. Ali I, Khan FG, Suri KA, Gupta BD, Satti NK, et al. (2010) In vitro antifungal activity of hydroxychavicol isolated from Piper betle L. Ann Clin Microbiol Antimicrob 9: 7.
74. Sarkar A, Sen R, Saha P, Ganguly S, Mandal G, et al. (2008) An ethanolic extract of leaves of Piper betle (Paan) Linn mediates its antileishmanial activity via apoptosis. Parasitol Res. 102(6): 1249–1255.
75. Al-Adhroey AH, Nor ZM, Al-Mekhlafi HM, Amran AA, Mahmud R (2010) Antimalarial activity of methanolic leaf extract of Piper betle L. Molecules. 16(1): 107–118.
76. Singh M, Shakya S, Soni VK, Dangi A, Kumar N, et al. (2009) The n-hexane and chloroform fractions of Piper betle L. trigger different arms of immune responses in BALB/c mice and exhibit antifilarial activity against human lymphatic filarid Brugia malayi. Int Immunopharmacol.b9(6): 716–28.
77. Trakranrungsie N, Chatchawanchonteera A, Khunkitti W (2008) Ethnoveterinary study for antidermatophytic activity of Piper betle, Alpinia galanga and Allium ascalonicum extracts in vitro. Res Vet Sci. 84(1): 80–84.
78. Wirotesangthong M, Inagaki N, Tanaka H, Thanakijcharoenpath W, Nagai H (2008) Inhibitory effects of Piper betle on production of allergic mediators by bone marrow-derived mast cells and lung epithelial cells. Int Immunopharmacol. 8(3): 453–457.
79. Kanjwani DG, Marathe TP, Chiplunkar SV, Sathaye SS (2008) Evaluation of immunomodulatory activity of methanolic extract of Piper *betel*. Scand J Immunol. 67(6): 589–593.
80. Majumdar B, Ray Chaudhuri SG, Ray A, Bandyopadhyay SK (2003) Effect of ethanol extract of Piper betle Linn leaf on healing of NSAID-induced experimental ulcer–a novel role of free radical scavenging action. Indian J Exp Biol. 41(4): 311–315.
81. Dasgupta N, De B (2004) Antioxidant activity of Piper betle L. leaf extract in vitro. Food Chem. 88: 219–224.
82. Ganguly S, Mula S, Chattopadhyay S, Chatterjee M (2007) An ethanol extract of Piper betle Linn. mediates its anti-inflammatory activity via down-regulation of nitric oxide. J Pharm Pharmacol. 59(5): 711–718.

IV.5.1 Conclusions from this research/ Chapter-5

i. We have shown a novel bioinformatics approach to identify exoproteome and secretome derived common conserved drug and vaccine targets in 21 *Vibrio cholerae* serotypes.
ii. An integrative strategy was applied that includes subtractive proteomics, conventional and PPIs and host-pathogen interactions based target prioritization, antigenic B-cell derived T-cell epitope prediction, 3D modelling of drug and vaccine targets, epitope design, topology analysis of epitopes, validation of epitopes using IEDB, virtual screening of piper betel compounds against drug targets, and experimental validation of targeting compounds against *V. cholerae* O1 Inaba.
iii. ompU, uppP and yajC identified as candidate targets for most of the pathogenic *Vibrio* strains
iv. uppP and yajC novel targets in *Vibrio*.
v. uppP and ompU may be used as dual targets i.e. both drugs and vaccines can be developed against these proteins
vi. Seven Piper betel compounds found to target these targets in *in silico* and show anti-*Vibrio* effects *in vitro*.
vii. The strategy can equally be applied to any other pathogenic bacteria and their serotypes to identify common conserved broad spectrum drug and vaccine targets.

IV.5.2 Media highlights of this research outcomes/ Chapter-5

DownToEarth
FORTNIGHTLY ON POLITICS OF DEVELOPMENT, ENVIRONMENT AND HEALTH
25 years

Betel benefit

Smita Pandey Friday 15 March 2013

Compounds in paan leaf can kill several pathogenic bacteria

Photo: Vaibhav Raghunandan

PAAN needs no introduction. From being a delightful end-of-the-meal treat to being used in religious ceremonies, the heart-shaped leaf is ubiquitous in Indian traditions. And with good reason—it is not just a mouth freshener but also has known medicinal properties. It forms an important component of ayurvedic medicine and has been used since ages to treat several ailments, like inflammation, headache, constipation and respiratory problems.

Now, for the first time, an international team of researchers from India, Mexico, Brazil, Denmark and the US has found that certain compounds in the dainty Piper betel leaf have antibiotic properties too. Two studies, published by the team in Integrative Biology on January 3 and in PLoS ONE on January 30, show how these compounds can be used to produce potent antibiotics against a broad spectrum of bacteria. The researchers have also identified novel targets in these bacteria that the betel leaf compounds can act on.

"The medicinal properties of Piper betel have been known since vedic times. But we have, for the first time, shown that specific compounds in paan leaf have broad-spectrum antibiotic properties and are effective even against several multi-drug resistant deadly pathogens," says Debmalaya Barh of the Institute of Integrative Omics and Applied Biotechnology in Nonakuri, West Bengal, who led the research.

For their studies, the researchers used an integrative-omics approach consisting of substractive proteomics, comparative genomics, immunomics, reverse vaccinology, and in silico and in vitro drug discovery strategies to identify enzymes or proteins essential for the survival of the bacteria that cause cholera, diphtheria, tuberculosis, plague, hemorrhagic diarrohea and Caseous lymphadenitis, a disease in cattle. These enzymes and proteins could, therefore, be treated as effective drug targets.

Once the targets were identified, the researchers looked at several betel compounds to find out the ones that can inhibit the activity of these targets and kill the bacteria.

They found piperdardine, one of the several antimicrobial compounds found in paan leaf, is very effective against all the pathogens and has a higher efficacy than currently used antibiotics like Chloramphenicol and Ampicillin. "This is the golden compound from Piper betel," says one of the researchers, Nidia Leon-Sicairos of the School of Medicine at Autonomous University of Sinaloa in Mexico.

The researchers are now working further to establish Piper betel as a source of next-generation antibiotics for deadly infectious diseases. "Identification of the common targets and targeting compounds for so many deadly pathogens in one shot is a big achievement towards fighting against a whole list of diseases," says Vasco Azevedo of the Brazilian team from Universidade Federal de Minas Gerais, Brazil.

http://www.downtoearth.org.in/news/betel-benefit-40450

Pesquisador da UFMG comprova a eficácia de trepadeira indiana contra doenças

Trepadeira é usada há anos pela medicina ayurvédica na Índia, mas só agora um grupo internacional de cientistas, inclusive brasileiros, comprova seus efeitos medicinais

postado em 03/06/2013 09:29 / atualizado em 03/06/2013 09:29

Ludymilla Sá /Estado de Minas

A tuberculose é uma doença que não para de crescer no mundo. A afirmação é do professor Vasco Azevedo, do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais (ICB/UFMG), residente do Instituto de Estudos Avançados Transdisciplinares (Ieat) e um dos pesquisadores que comprovaram a eficácia de uma planta, originalmente, indiana no combate à enfermidade e outras infecções causadas por bactérias.

Muito usada na Índia, a folha da trepadeira Betel, cientificamente conhecida por Piper betle, já tinha suas funções terapêuticas conhecidas pela medicina ayurvédica, ciência milenar indiana cujos princípios deram origem às medicinas chinesa, árabe, romana e grega. A planta era indicada no tratamento de inflamações, dores de cabeça e problemas respiratórios e o povo indiano, especificamente o do Norte do país, sempre usou suas folhas como forma de também auxiliar na digestão e limpeza bucal.

Mas só agora ela teve a sua eficácia comprovada cientificamente por uma equipe internacional de pesquisadores da Índia, Brasil, México, EUA e Dinamarca, liderada pelo indiano Debmalya Barh, do Instituto de Omics Integrativas e Biotecnologia Aplicada (IIOAB), em Nonakuri, na Índia. "Realizamos testes in vitro em duas bactérias, a Corynebacterium pseudotuberculosis, que causa uma falsa tuberculose em caprinos e ovinos, e a Vibrio cholerae, causadora do cólera. E pela primeira vez, destrinchamos os compostos isolados das folhas da Betel", conta Vasco Azevedo, líder da equipe brasileira.

Professor do ICB/UFMG, Vasco Azevedo, ao lado do pesquisador indiano Sandeep Tiwari, que faz doutorado em bioinformática em BH: grupo multidisciplinar para avaliar a planta

Segundo o biólogo, os pesquisadores observaram que a Betel teve efeito superior a certos antibióticos como penicilina, ampicilina e cloranfenicol. Desta forma, ela pode ser usada na fabricação de novos medicamentos que combaterão, além da tuberculose, as bactérias causadoras de diarreia, cólera e peste bubônica. "A identificação das metas comuns e compostos de segmentação de tantos patógenos mortais é uma grande conquista e considerado um progresso para o combate às doenças causadas por esses agentes patogênicos", diz Vasco.

Em razão disso, a doutora Nidia Leon Sicairos, da Faculdade de Medicina da Universidade Autônoma de Sinaloa, no México, que liderou a equipe mexicana, definiu os compostos da planta como "compostos de ouro". Comprovada a eficácia da Betel, os pesquisadores precisam agora testar os compostos da planta in vivo. E torcem pelo interesse de grandes laboratórios para dar continuidade ao trabalho e, consequentemente, desenvolver remédios.

"Se as empresas quiserem nos contratar, ótimo. Se não quiserem, também ótimo porque fizemos o nosso dever social, afinal esse é o objetivo das pesquisas. Tornamos público nosso conhecimento, mas o grande problema no Brasil é que não achamos empresas dispostas a investir num projeto desde o início. O país precisa investir nas pesquisas para dar um salto tecnológico", critica o professor da UFMG, destacando que o trabalho foi publicado na Europa antes mesmo de ser patenteado, para tentar despertar o interesse de empresários do ramo de medicamentos.

"Sem patentear, você perde a primazia, mas foi a forma que encontramos para tentar despertar o interesse para a causa. Publicamos para ver se há interesse, se eles querem trabalhar com a gente ou querem nos contratar para trabalhar com eles, mas os empresários no Brasil querem sempre pagar menos. Na Europa, por exemplo, você tem 400 pesquisadores, hipoteticamente, trabalhando, enquanto no nosso país, não fazem pesquisa de desenvolvimento", desabafa.

Vasco espera que os grandes laboratórios se interessem pela pesquisa, levando em consideração o aumento significativo dos casos de tuberculose no país. "A doença já foi considerada um mal de países em desenvolvimento mas, com a epidemia de HIV, ressurgiu com força total em nações consideradas de primeiro mundo. Ficamos quase 40 anos sem a fabricação de um novo medicamento para a tuberculose porque não era interessante para o país rico, não era doença de país rico. Mas com a Aids, começou-se a desenvolver novas drogas. Esperamos ganhar com isso."

A tuberculose

De acordo com estimativas da Organização Mundial de Saúde (OMS), 8,7 milhões de pessoas foram infectadas com a doença e 1,4 milhão morreram de tuberculose em 2011. As estatísticas no Brasil também não são animadoras. Segundo o Ministério da Saúde, o Brasil ocupa o 15º lugar entre os 22 países responsáveis por 80% do total de casos da enfermidade no mundo. No país, são notificados, anualmente, 85 mil casos novos. Além disso, são registrados cerca de 6 mil mortes por ano em decorrência da doença.

http://www.em.com.br/app/noticia/tecnologia/2013/06/03/interna_tecnologia,398342/pesquisador-da-ufmg-comprova-a-eficacia-de-trepadeira-indiana-contra-doencas.shtml

GEN
Genetic Engineering & Biotechnology News

Send to printer »

Insight & Intelligence™ : May 13, 2014

Infectious Disease Vaccines in the Omics Era

A shift from conventional vaccines to more rationally designed approaches is a hallmark of the new revolution.

Richard A. Stein, M.D., Ph.D.

Infectious diseases have shaped humanity more than any other single factor in history. Despite the substantial morbidity and mortality associated with pathogens, eradicating infectious diseases has been challenging, both in terms of generating vaccines and in ensuring that immunization campaigns can be affordable, effective, and feasible on a global scale. The only human infectious disease that was eradicated worldwide as a result of vaccination initiatives was smallpox, and with three countries—Afghanistan, Nigeria, and Pakistan—remaining endemic in early 2014, poliomyelitis is expected to become the next target for global eradication.

Reverse Vaccinology

Advances in sequencing technologies and the increasing availability of microbial genomes catalyzed one of the great strides in vaccinology. This consisted in the shift from conventional vaccines, which are based on inactivated or killed microorganisms or subunit vaccines, to more rationally designed approaches. One of these advances consisted in the development of reverse vaccinology, a genome-based approach in which scanning the entire genome of a pathogen allowed, without the need to grow the microorganism, the identification of genes encoding proteins with certain desirable characteristics, which were then expressed and tested experimentally for their ability to confer protection in vivo.

“Several modifications have been developed to the reverse vaccinology strategy that was originally pioneered

by Dr. Rino Rappuoli," said Debmalya Barh, Ph.D., principal scientist at the Institute of Integrative Omics and Applied Biotechnology (IIOAB), India. "Peptide vaccines based on immunogenic pan- and subtractive-proteomic strategies mining the exoproteome/surfectome and secretome emerged as some of the most promising strategies. We have designed and are testing such peptide vaccines against gonorrhea, tuberculosis, and the *Corynebacterium-Mycobacterium-Nocardia* group of pathogens."

However, rationally designed vaccines open multiple challenges. "Most rationally designed vaccines do not work as expected during their initial developmental stages and fail to produce immune protection, even after multiple dosages," Dr. Barh commented.

Dr. Barh and his colleague Vasco Azevedo, D.V.M, Ph.D., professor at Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Brazil, have made significant advances toward designing drugs and vaccines for cholera, an infectious disease that remains a global public health concern. In the case of cholera, multiple serotypes of the pathogen are responsible for epidemics. The oral whole-cell, killed cholera vaccine, which is based on cholera toxin, is only effective against two serogroups, O1 and O139, and no broad-spectrum vaccine has been available so far.

"These two serogroups are not the only pathogenic ones, and several other *Vibrio* serogroups cause the epidemic in various geographical regions," said Dr. Barh. Therefore, there is a need to develop multivalent broad-spectrum vaccines. "The main constraint in designing next-generation vaccines for cholera is the identification of proper vaccine candidates and their clinical evaluation, and this is due to the large genome size, the genomic variations among the serotypes, the lack of proper animal models, and the lack of international and collaborative efforts," he explained.

In Vivo-Induced Antigen Technology

"When we approached the vaccine, we wanted to have information available about all the genes that are encoded in the genome, to identify the proteins that are expressed and elicit an immune response during infection," said Harry L.T. Mobley, Ph.D., professor and chair in the department of microbiology and immunology, University of Michigan Medical School. Several major projects in Dr. Mobley's lab focus on developing a vaccine against uropathogenic *Escherichia coli*, the primary cause of urinary tract infections.

In one of the recent experimental strategies that they used to identify vaccine candidates, Dr. Mobley and colleagues removed surface-exposed peptides from the bacterial cell surface by limited proteolytic digestion, and identified the fragments by using mass spectrometry. “This identified not only the proteins that are expressed and are on the bacterial cell surface, but also revealed the specific parts of the proteins that are exposed,” he said. The fragments that are exposed have a higher likelihood of being important in eliciting an immune response, and they can subsequently be tested in vivo for their ability to confer protection. This approach provides a promising strategy to identify vaccine candidates for other pathogens as well, if an animal model of pathogenesis is available and results from omics studies can be integrated with the in vivo studies.

In another strategy recently used in the Mobley lab, a genomic expression library, generated from a human pathogenic strain, was screened against sera collected from mice chronically infected with the strain and adsorbed against bacteria cultured under in vitro conditions. This immunoscreening approach, known as in vivo-induced antigen technology, identified, in the first stage, 93 genes from a pool of 40,000 clones. The in vivo expression of several representative genes identified from this pool was examined by quantitative PCR, and mutants that targeted these genes for deletion revealed, in an animal model, that one of them is an important contributor to virulence, emerging as a promising vaccine candidate.

“Our approach is to analyze the cell surface of parasites at different stages of the life cycle, and try to dissect their contribution to pathogenesis to develop more rational vaccines,” said Igor C. Almeida, Ph.D., professor of biological sciences at the University of Texas at El Paso (UTEP). Dr. Almeida and colleagues are focusing on *Trypanosoma cruzi*, the causative agent of Chagas disease, a condition that has become increasingly more relevant in the United States and many other countries worldwide particularly as a result of the very intensive migration of people from endemic to nonendemic countries. This parasitic infection is particularly challenging when it occurs during pregnancy.

“It is estimated that about 8–10 million people are chronically infected just in Latin America, and this does not include other countries, such as the U.S., Canada, and several European countries,” Dr. Almeida pointed out. “Overall, less than 1% of the patients are currently treated in the world.”

The challenge in designing a vaccine is that the only way to learn about the proteins expressed on the surface

of the trypomastigote, which is the infective form of the parasite and lives in the mammalian host, is by proteomic analysis. “To learn about the antigens that are expressed on the parasite surface and can be therapeutically targeted, proteomic analysis has to be performed not only in different parasite stages but also in different strains, and this is what our lab has been doing in recent years,” Dr. Almeida said. An additional hurdle in generating vaccines against protozoan parasites is that sometimes, even though they work in the experimental settings, this may not always be true in the field, due to the fact that field isolates do not always express the specific antigens that the vaccine was developed against.

“This is happening for other parasites as well, and it is the reason why so far we do not have a single vaccine against any parasites, despite the fact that they cause diseases affecting over 1 billion people worldwide,” he added.

Research in Dr. Almeida’s lab has focused on the analysis of surface antigens mainly at the trypomastigote and intracellular amastigote forms, in an attempt to find molecules that are conserved and can be recognized by patients from different geographic locations. “Unfortunately, the parasite has a very complex cell surface, and this makes it challenging to generate a vaccine,” he said. The strategy that Dr. Almeida and colleagues used involved a proteomic analysis to identify and characterize, at each of the parasite stages, conserved antigens that are vaccine candidates.

“As a proof of concept, using this approach, my group in collaboration with the group of Dr. Rosa Maldonado at UTEP were able to develop a vaccine based on a protein that is found in the *T. cruzi* secretome,” Dr. Almeida said. After fractionating culture supernatants, a proteomic analysis of the trypomastigote stage of the parasite secretome identified thousands of peptides, from which and a more narrow number of candidates was selected. Subsequent immunoinformatics studies to predict T-cell and antibody epitopes revealed a peptide of interest, which, after being synthesized and attached to a carrier protein, was able to prime the immune system. “In a mouse model, this peptide was 90% effective in controlling the infection,” he commented.

Dr. Almeida’s group has also been developing a vaccine based on sugars that are unique to *T. cruzi*. “The parasite surface is covered by a thick coat of sugar-containing molecules or glycoconjugates,” he said. “Although highly immunogenic to humans, these glycoconjugates have not been exploited as vaccine targets, mainly because of the technical difficulties in structurally characterizing and synthesizing carbohydrate epitopes

or glycotopes." Using transgenic mice that do not express a particular highly immunogenic sugar expressed by the parasite, Dr. Almeida's group has recently developed a fully protective vaccine that is now undergoing trials in nonhuman primates.

One of the many disciplines that benefited from the omics revolution, infectious diseases have entered a new era, characterized by the transition to rationally designed vaccines as one of its defining features. Human pathogens are expected to continue to inflict significant morbidity and mortality worldwide, both as a result of acute and chronic infectious diseases and in terms of the chronic medical conditions, such as cancer, which were causally linked to certain pathogens. The transition to novel vaccination strategies has far-reaching implications that extend to all groups of human pathogens and promise to fill a longtime gap in preventive medicine and global public health.

V. GENERAL CONCLUSIONS

In this research we have developed novel various bioinformatics strategies to identify biomarkers in cancers and targets in bacterial pathogens.

The *in silico* reverse-transcriptomics approach (Chapter-I) can be useful in identifying subtype specific and early diagnostic/ screening markers in lung cancer. A modified strategy can be helpful in drug target identification towards precision medicine. The miRegulome and its integrated tools (Chapter-II) can further boost the reverse-transcriptomics strategy and biomarkers and therapeutics discoveries. It also helps in exploring novel patho-physiological roles of miRNAs. The consensus-based network inference oriented miRsig (Chapter-III) can identify common miRNA-miRNA interaction network signatures in multiple diseases therefore, early deregulated common miRNA signatures in such diseases can also be identified using this method and thus screening or early diagnostic biomarkers can developed. Although, all these three strategies have been applied to identify cancer biomarkers in this research; they can also be equally applicable to other complex human diseases.

Towards identification of common, conserved, and broad spectrum targets in multiple pathogens including *M. tuberculosis*, we have shown a novel integrated bioinformatics strategy in Chapter-IV that combines intra- and inter-species PPI, host-pathogen PPI, and subtractive genomics. We have also shown a modified exoproteome and secretome base reverse vaccinology strategy to identify common conserved drug and vaccine targets in *Vibrio cholerae* serotypes (Chapter-V). Both the strategies can be applicable to identify globally conserved novel targets in any set of pathogens or serotypes of any pathogenic bacteria. Further, first time we have shown Piper betel derived photychemicals have broad spectrum antibacterial properties that may be effective against *C. pseudotuberculosis (*Cp), *C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, E. coli,* and *Vibrio cholerae*. The efficacies of the betel compounds are better than some third generation antibiotics indicating them potential in developing next-generation antibiotics.

VI. FUTURE PERSPECTIVES

Lung cancer is the leading cause of all cancer related deaths with recently estimated 1.6 million deaths worldwide [JEMAL ET AL., 2011]. Similar to lung cancer, pulmonary tuberculosis caused by M. tuberculosis is one of the major causes of death amongst infectious diseases and according to WHO 2013 report, it is estimated that 9 million people are infected and 1.5 million are died from tuberculosis in 2012 [GLAZIOU ET AL., 2015]. Several reports have documented co-existence of tuberculosis and lung cancer [SKOWRONSKI ET AL., 2015; YU ET AL., 2008; LIANG ET AL., 2009; WU ET AL., YU ET AL., 2011] and pulmonary tuberculosis is a risk factor for developing lung cancer [YU ET AL., 2008; LIANG ET AL., 2009; WU ET AL., YU ET AL., 2011]. However, it is not yet fully established at molecular level, how the Mycobacterium increases susceptibility to lung cancer.

Several miRNAs have been implemented to be associated with lung cancer having causative roles and diagnostic potentials [BARH ET AL., 2013]. Similarly, a number of miRNAs are found deregulated in pulmonary tuberculosis patients [LATORRE ET AL., 2015; XU ET AL., 2015; ZHENG ET AL., 2015]. Therefore, there could be common miRNA signature pretending to development of both the pulmonary tuberculosis and lung cancer.

The future scope of this research is to develop bioinformatics strategies to:

i. Identify the common miRNA signature associated with both pulmonary tuberculosis and lung cancer
ii. Understand the role this miRNA signature in *M. tuberculosis* infection and lung carcinogenesis.
iii. Explore the possibility of use of this miRNA signature for developing screening, early diagnosis, and prognosis tools for pulmonary tuberculosis and lung cancer.
iv. Identify the human genetic predisposition factors and factors in *Mycobacterium* that increases susceptibility or risk in developing lung cancer.
v. If such factors are identified, it can also be useful tool towards risk prediction and also for therapy or managing both the diseases.

VII BIBLIOGRAPHY

Almeida, S., Tiwari, S., Mariano, D., Souza, F., Jamal, S. B., Coimbra, N., …… Azevedo, V. (2016). The genome anatomy of *Corynebacterium pseudotuberculosis* VD57 a highly virulent strain causing Caseous lymphadenitis. *Stand Genomic Sci. 11*:29. doi: 10.1186/s40793-016-0149-7.

Ali, A., Naz, A., Soares, S. C., Bakhtiar, M., Tiwari, S., Hassan, S. S., . . . Azevedo, V. (2015). Pan-Genome Analysis of Human Gastric PathogenH. pylori: Comparative Genomics and Pathogenomics Approaches to Identify Regions Associated with Pathogenicity and Prediction of Potential Core Therapeutic Targets. *BioMed Research International, 2015*, 1-17. doi:10.1155/2015/139580

Bakhtiar, S. M., Ali, A., Baig, S. M., Barh, D., Miyoshi, A., & Azevedo, V. (2014). Mini Review Identifying human disease genes: advances in molecular genetics and computational approaches. *Genetics and Molecular Research, 13*(3), 5073-5087. doi:10.4238/2014.July.4.23

Barh, D., Agte, V., Dhawan, D., Agte, V., & Padh, H. (2012). Cancer Biomarkers for Diagnosis, Prognosis and Therapy. 18-68. doi:10.1002/9781119967309.ch2

Chen, Y. T., Tang, H. J., Chao, C. M., & Lai, C. C. Clinical manifestations of non-O1 *Vibrio cholerae* infections. *PLoS One. 10*(1):e0116904. doi: 10.1371/journal.pone.0116904

Dix, A., Vlaic, S., Guthke, R., & Linde, J. (2016). Use of systems biology to decipher host–pathogen interaction networks and predict biomarkers. *Clinical Microbiology and Infection, 22*(7), 600-606. doi:10.1016/j.cmi.2016.04.014

El-Mosalamy, H., Salman, T.,M., Ashmawey, A. M., & Osama, N. (2012). Role of chronic *E. coli* infection in the process of bladder cancer- an experimental study. *Infect Agent Cancer*. *7*(1):19. doi: 10.1186/1750-9378-7-19

El-Sharif, A., Afifi, S., El-Dahshan, R., Rafeh, N., & Eissa, S. (2012). Characterization of Mycobacterium tuberculosis isolated from cancer patients with suspected tuberculosis infection in Egypt: identification, prevalence, risk factors and resistance pattern. *Clinical Microbiology and Infection, 18*(11), E438-E445. doi:10.1111/j.1469-0691.2012.03974.x

Glaziou, P., Sismanidis, C., Floyd, K., & Raviglione, M. (2014). Global epidemiology of tuberculosis. *Cold Spring Harb Perspect Med, 5*(2), a017798. doi:10.1101/cshperspect.a017798

Gomes, D. L., Martins, C. A., Faria, L. M., Santos, L. S., Santos, C. S., …….. Mattos-Guaraldi, A. L. (2009). *Corynebacterium diphtheriae* as an emerging pathogen in nephrostomy catheter-related infection: evaluation of traits associated with bacterial virulence. *J Med Microbiol. 58*(Pt 11):1419-27. doi: 10.1099/jmm.0.012161-0

Granville, C. A., & Dennis, P. A. (2005). An overview of lung cancer genomics and proteomics. *Am J Respir Cell Mol Biol, 32*(3), 169-176. doi:10.1165/rcmb.F290

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA Cancer J Clin, 61*(2), 69-90. doi:10.3322/caac.20107

Kim, H., Watkinson, J., & Anastassiou, D. (2011). Biomarker Discovery Using Statistically Significant Gene Sets. *Journal of Computational Biology, 18*(10), 1329-1338. doi:10.1089/cmb.2010.0085

Kulasingam, V., & Diamandis, E. P. (2008). Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature Clinical Practice Oncology, 5*(10), 588-599. doi:10.1038/ncponc1187

Latorre, I., Leidinger, P., Backes, C., Dominguez, J., de Souza-Galvao, M. L., Maldonado, J., . . . Meyerhans, A. (2015). A novel whole-blood miRNA signature for a rapid diagnosis of

pulmonary tuberculosis. *Eur Respir J, 45*(4), 1173-1176. doi:10.1183/09031936.00221514

Liang, H. Y., Li, X. L., Yu, X. S., Guan, P., Yin, Z. H., He, Q. C., & Zhou, B. S. (2009). Facts and fiction of the relationship between preexisting tuberculosis and lung cancer risk: a systematic review. *Int J Cancer, 125*(12), 2936-2944. doi:10.1002/ijc.24636

Mandal, S., Mandal, M. D., & Pal, N. K. (2011). Cholera: a great global concern. *Asian Pac J Trop Med, 4*(7), 573-580. doi:10.1016/S1995-7645(11)60149-1

Mattos-Guaraldi, A. L., Formiga, L. C., Camello, T. C., Pereira, G. A., Hirata, R. Jr., & Halpern, M. (2001). *Corynebacterium diphtheriae* threats in cancer patients. *Rev Argent Microbiol. 33*(2):96-100

McDermott, J. E., Wang, J., Mitchell, H., Webb-Robertson, B. J., Hafen, R., Ramey, J., & Rodland, K. D. (2013). Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opin Med Diagn, 7*(1), 37-51. doi:10.1517/17530059.2012.718329

Moore, L. S., Leslie, A., Meltzer, M., Sandison, A., Efstratiou, A., & Sriskandan, S. (2015). *Corynebacterium ulcerans* cutaneous diphtheria. *Lancet Infect Dis. 15*(9):1100-7. doi: 10.1016/S1473-3099(15)00225-X

Nath, G., Gulati, A. K., & Shukla, V. K. (2010). Role of bacteria in carcinogenesis, with special reference to carcinoma of the gallbladder. *World J Gastroenterol, 16*(43), 5395-5404.

Plump, A. S., & Lum, P. Y. (2009). Genomics and Cardiovascular Drug Development. *Journal of the American College of Cardiology, 53*(13), 1089-1100. doi:10.1016/j.jacc.2008.11.050

Radusky, L. G., Hassan, S., Lanzarotti, E., Tiwari, S., Jamal, S., Ali, J., . . . Azevedo, V. A. C. (2015). An integrated structural proteomics approach along the druggable genome of Corynebacterium pseudotuberculosis species for putative druggable targets. *BMC Genomics, 16*(Suppl 5), S9. doi:10.1186/1471-2164-16-s5-s9

Rogers, M. B. (2011). Mycoplasma and cancer: in search of the link. *Oncotarget, 2*(4), 271-273. doi:10.18632/oncotarget.264

Shen, D., Chen, R., Ye, L., Luo, Y., & Tang, Y. W. (2010). Robinsoniella peoriensis Bacteremia in a Patient with Pancreatic Cancer. *Journal of Clinical Microbiology, 48*(9), 3448-3450. doi:10.1128/jcm.00477-10

Skowronski, M., Iwanik, K., Halicka, A., & Barinow-Wojewodzki, A. (2015). Squamous cell lung cancer in a male with pulmonary tuberculosis. *Pneumonol Alergol Pol, 83*(4), 298-302. doi:10.5603/PiAP.2015.0049

Ursu, A., Sen, A., & Ruffin, M. (2015). Impact of Cervical Cancer Screening Guidelines on Screening for Chlamydia. *The Annals of Family Medicine, 13*(4), 361-363. doi:10.1370/afm.1811

von Mering, C., Cohen Freue, G. V., Meredith, A., Smith, D., Bergman, A., Sasaki, M., . . . McMaster, W. R. (2013). Computational Biomarker Pipeline from Discovery to Clinical Implementation: Plasma Proteomic Biomarkers for Cardiac Transplantation. *PLoS Computational Biology, 9*(4), e1002963. doi:10.1371/journal.pcbi.1002963

Wang, T., Nelson, R. A., Bogardus, A., & Grannis, F. W., Jr. (2010). Five-year lung cancer survival: which advanced stage nonsmall cell lung cancer patients attain long-term survival? *Cancer, 116*(6), 1518-1525. doi:10.1002/cncr.24871

World Health Organization. Global Status Report on non-communicable diseases 2010.

Wu, C. Y., Hu, H. Y., Pu, C. Y., Huang, N., Shen, H. C., Li, C. P., & Chou, Y. J. (2011). Pulmonary tuberculosis increases the risk of lung cancer: a population-based cohort study. *Cancer, 117*(3), 618-624. doi:10.1002/cncr.25616

Xu, Z., Zhou, A., Ni, J., Zhang, Q., Wang, Y., Lu, J., . . . Yao, Y. (2015). Differential expression of miRNAs and their relation to active tuberculosis. *Tuberculosis (Edinb), 95*(4), 395-403. doi:10.1016/j.tube.2015.02.043

Yeruham, I., Elad, D., Van-Ham, M., Shpigel, N. Y., & Perl, S.(1997). *Corynebacterium pseudotuberculosis* infection in Israeli cattle: clinical and epidemiological studies. *Vet Rec.* 140(16):423-7

Yu, Y. H., Liao, C. C., Hsu, W. H., Chen, H. J., Liao, W. C., Muo, C. H., . . . Chen, C. Y. (2011). Increased lung cancer risk among patients with pulmonary tuberculosis: a population cohort study. *J Thorac Oncol, 6*(1), 32-37. doi:10.1097/JTO.0b013e3181fb4fcc

Yu, Y. Y., Pinsky, P. F., Caporaso, N. E., Chatterjee, N., Baumgarten, M., Langenberg, P., . . . Engels, E. A. (2008). Lung cancer risk following detection of pulmonary scarring by chest radiography in the prostate, lung, colorectal, and ovarian cancer screening trial. *Arch Intern Med, 168*(21), 2326-2332; discussion 2332. doi:10.1001/archinte.168.21.2326

Zhang, A., Sun, H., Yan, G., Wang, P., & Wang, X. (2015). Metabolomics for Biomarker Discovery: Moving to the Clinic. *BioMed Research International, 2015*, 1-6. doi:10.1155/2015/354671

Zheng, L., Leung, E., Lee, N., Lui, G., To, K. F., Chan, R. C., & Ip, M. (2015). Differential MicroRNA Expression in Human Macrophages with Mycobacterium tuberculosis Infection of Beijing/W and Non-Beijing/W Strain Types. *PLoS One, 10*(6), e0126018. doi:10.1371/journal.pone.0126018.

Zińczuk, J., Wojskowicz , P., Kiśluk, J., Fil, D., Kemona, A., & Dadan, J. (2015). Mesenteric lymphadenitis caused by *Yersinia enterocolitica*. *Prz Gastroenterol*. *10*(2):118-21. doi: 10.5114/pg.2014.47504

VIII. APPENDIX

A. Published Research & Review Articles

A.a: Journal articles with Prof. Vasco Azevedo

A.a.1 Bacterial Genomics

1. Folador EL, de Carvalho PV, Silva WM, Ferreira RS, Silva A, Gromiha M, Ghosh P, **Barh D**, Azevedo V, Röttger R. [2016]**. In silico identification of essential proteins in Corynebacterium pseudotuberculosis based on protein-protein interaction networks**. *BMC Syst Biol. 10(1):103.* [PMID: 27814699] **[IF=3.15]**
2. Oliveira A, Teixeira P, Azevedo M, Jamal SB, Tiwari S, Almeida S, Silva A, **Barh D**, Dorneles EM, Haas DJ, Heinemann MB, Ghosh P, Lage AP, Figueiredo H, Ferreira RS, Azevedo V. [2016]. **Corynebacterium pseudotuberculosis may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenic inference from sequence and structural analysis**. *BMC Microbiol. 2016 Jun 2;16:100. doi: 10.1186/s12866-016-0717-4.* [PMID:27251711] **[IF=2.5]**
3. Mariano DC, Sousa Tde J, Pereira FL, Aburjaile F, **Barh D**, Rocha F, Pinto AC, Hassan SS, Saraiva TD, Dorella FA, de Carvalho AF, Leal CA, Figueiredo HC, Silva A, Ramos RT, Azevedo VA. **Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002**. *BMC Genomics. 2016 Apr 30;17:315. doi: 10.1186/s12864-016-2673-7.* [PMID: 27129708] **[IF=3.8]**
4. Radusky LG, Hassan SS, Lanzarotti E, Tiwari S, Jamal SB, Ali J, Ali A, Ferreira RS, **Barh D**, Silva A, Turjanski AG, Azevedo V. [2015]. **An Integrated Structural Proteomics Approach Along The Druggable Genome of Corynebacterium Pseudotuberculosis Species For Putative Druggable Targets**. *BMC Genomics*, 16 (Sippl 5): S9 [PubMed: 26041381] **[IF=4.04]**
5. Guimarães LS, de Castro S, Trost E, Blom J, Ramos RTJ, Silva A, **Barh D**, Azevedo V. [2015]. **Genome informatics and vaccine targets in Corynebacterium urealyticum using two whole genomes, comparative genomics, and reverse vaccinology**. *BMC Genomics*. 16(Suppl 5):S7. [PubMed: 26041051] **[IF=4.04]**
6. Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, Hanan F, Ramos R, Pereira U, **Barh D**, Figueiredo HC, Ussery DW, Miyoshi A, Silva A, Azevedo V. [2015]. **Pan-Genome Analysis of Human Gastric Pathogen H. pylori: Comparative Genomics and Pathogenomics Approaches to Identify Regions Associated with Pathogenicity and Prediction of Potential Core Therapeutic Targets**. *Biomed Res Int.*. 2015:139580 [PubMed: 25705648] **[IF=2.7]**
7. Hassan SS, Tiwari S, Guimarães LC, Jamal SB, Folador E, Sharma NB, de Castro Soares S, Almeida S, Ali A, Islam A, Póvoa FD, de Abreu VA, Jain N, Bhattacharya A, Juneja L, Miyoshi A, Silva A, **Barh D**, Turjanski A, Azevedo V, Ferreira RS. [2014]. **Proteome scale comparative modeling for conserved drug and vaccine targets identification in Corynebacterium pseudotuberculosis.** *BMC Genomics; 15(7):S3.* [PubMed: 25573232]. **[IF=4.04]**
8. Tiwari S, da Costa MP, Almeida S, Hassan SS, Jamal SB, Oliveira A, Folador EL, Rocha F, de Abreu VA, Dorella F, Hirata R, de Oliveira DM, da Silva Teixeira MF, Silva A, **Barh D**, Azevedo V. [2014]. **C. pseudotuberculosis Phop confers virulence and may be targeted by natural compounds.** *Integr Biol (Camb). 6(11):1088-99*. [PMID: 25212181] **[IF=4.45]**
9. Pereira UP, Soares SC, Blom J, Leal CA, Ramos RT, Guimarães LC, Oliveira LC, Almeida SS, Hassan SS, Santos AR, Miyoshi A, Silva A, Tauch A, **Barh D**, Azevedo V, Figueiredo HC. [2013]. **In silico prediction of conserved vaccine targets in Streptococcus agalactiae strains isolated from fish, cattle, and human samples**. *Genetics and Molecular Research* 12(3):2902 – 2912 [PubMed: 24065646**] [IF=1.18]**
10. Silva WM, Seyffert N, Ciprandi A, Santos AV, Castro TLP, Pacheco LGC, **Barh D**, Loir YL, Pimenta AMC, Miyoshi A, Silva A, Azevedo V. [2013]. **Differential Exoproteome Analysis of Two Corynebacterium pseudotuberculosis Biovar Ovis Strains Isolated from Sheep (1002) and Goat (C231).** *Curr Microbiol*. *67(4):460-5* [PMID: 23699973] **[IF: 1.8]**
11. Silva WM, Seyffert N, Santos AV, Castro TL, Pacheco LG, Santos AR, Ciprandi A, Dorella FA, Andrade HM, **Barh D**, Pimenta AM, Silva A, Miyoshi A, Azevedo V. [2013]. **Identification of 11 new exoproteins in Corynebacterium pseudotuberculosis by comparative analysis of the exoproteome**. *Microb Pathog*. *61-62:37-42.* [PMID: 23684727] **[IF: 2.1]**
12. **Barh D**, Gupta K, Jain N, Khatri G, León-Sicairos N, Canizalez-Roman A, Tiwari S, Verma A, Rahangdale S, Shah Hassan S, Rodrigues Dos Santos A, Ali A, Carlos Guimarães L, Thiago Jucá Ramos R, Devarapalli P, Barve N, Bakhtiar M, Kumavath R, Ghosh P, Miyoshi A, Silva A, Kumar A, Narayan Misra A, Blum K, Baumbach J, Azevedo V. [2013]. **Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli targeted by Piper betel compounds.** *Integrative Biology*, *5: 495-509.* [PubMed: 23288366] **[IF: 4.5]**
13. **Barh D**, Barve N, Gupta K, Chandra S, Jain N, Tiwari S, Leon-Sicairos N, Canizalez-Roman A, Rodrigues Dos Santos A, Hassan SS, Almeida S, Thiago Jucá Ramos R, Augusto Carvalho de Abreu V, Ribeiro Carneiro A, de Castro Soares S, Luiz de Paula Castro T, Miyoshi A, Silva A, Kumar A, Narayan Misra A, Blum K, Braverman ER, Azevedo V. [2013]. **Exoproteome and Secretome Derived Broad Spectrum Novel Drug and Vaccine Candidates in Vibrio cholerae Targeted by Piper betel Derived Compounds**. *PLOS ONE, 8(1): e52773.* [PubMed: 23382822] [**IF: 4.09**]
14. Santos AR, Carneiro A, Gala-García A, Pinto A, **Barh D**, Barbosa E, Aburjaile F, Dorella F, Rocha F, Guimarães L, Zurita-Turk M, Ramos R, Almeida S, Soares S, Pereira U, Abreu VC, Silva A, Miyoshi A, Azevedo V. [2012]. **The Corynebacterium pseudotuberculosis in silico predicted pan-exoproteome.** *BMC Genomics; 13 (Suppl 5):S6.* [Pubmed: 23095951] **[IF: 4.07]**
15. Carneiro AR, Ramos RT, Barbosa HP, Schneider MP, **Barh D**, Azevedo V, Silva A. [2012]. **Quality of prokaryote genomes assembly: Indispensable issues of factors affecting prokaryote genome assembly quality.** *Gene*. *505(2):365-7* [PubMed: 22721771] [**IF: 2.26**]
16. Hollmann A, Saviello M, Delfederico L, Saraiva TDL, **Barh D**, Chandra S, Gupta K, Jain N, Zambare V, Kumar A, Misra AN, Christopher L, Azevedo V, Semorile L, Miyoshi A. [2012]. **Tight controlled expression and secretion of Lactobacillus brevis SlpA in Lactococcus lactis**. *Biotechnol Lett; 34(7):1275-81*.[PubMed: 22391736] [**IF:** 1.76]
17. Santos A, Ali A, Barbosa E, Silva A, Miyoshi A, **Barh D**, Azevedo V. [2011]. **The reverse vaccinology- a contextual overview**. *The IIOAB Journal; 2 (4): 8-15.* **[IC value: 4.55]**
18. **Barh D**, **Jain N,** Tiwari S**, Parida BP**, D'Afonseca V, Li L, Ali A, Santos AR, Guimarães LC, de Castro Soares S, Miyoshi A, Bhattacharjee A, Misra AN, Silva A, Kumar A, Azevedo V. [2011]. **A novel comparative genomics analysis for common drug and**

vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens. *Chemical Biology & Drug Design; 78(1):73-84.* [PubMed: 21443692] **[IF: 2.48]**

19. **Barh D**, Tiwari S**, Jain N**, Ali A, Santos AR, Misra AN, Azevedo V, Kumar A. [2011**]. In silico subtractive genomics for target identification in human bacterial pathogens**. *Drug Development Research; 72, 1-16*. [IF: 1.19]
20. Ali A, Soares SC, Barbosa E, Santos AR, **Barh D**, Bakhtiar SM, Hassan SS, Ussery DW, Silva A, Miyoshi A, Azevedo V. [2013]. **Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus *Corynebacterium***. *J Bacteriol Parasitol 4: 167. doi:10.4172/2155-9597.1000167*
21. **Barh D**, Misra AN, Kumar A, Azevedo V. [2010]. **A novel strategy of epitope design in Neisseria gonorrhoeae**. *Bioinformation; 5*

A.a.2 Bioinformatics Software

22. Nalluri JJ, **Barh D**, Azevedo V, Ghosh P. [2017]. **miRsig: a consensus-based network inference methodology to identify pan-cancer miRNA-miRNA interaction signatures**. *Sci Rep. 7:39684. doi: 10.1038/srep39684*. [PMID: 28045122] [**IF: 5.2**]
23. Mariano DC, Pereira FL, Aguiar EL, Oliveira LC, Benevides L, Guimarães LC, Folador EL, Sousa TJ, Ghosh P, Barh D, Figueiredo HC, Silva A, Ramos RT, Azevedo VA. [2016]. **SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology**. *BMC Bioinformatics. 17(Suppl 18):456. doi: 10.1186/s12859-016-1344-7.* [PMID: 28105921]
24. **Barh D**, Kamapantula B, Jain N, Nalluri J, Bhattacharya A, Juneja L, Barve N, Tiwari S, Miyoshi A, Azevedo V, Blum K, Kumar A, Silva A, Ghosh P [2015]. **miRegulome: a knowledge-base of miRNA regulomics and analysis**. *Nature, Scientific Reports,* 5:12832 [PubMed: 26243198] [**IF: 5.7**]
25. Mariano DCB, Pereira FL, Ghosh P, **Barh D**, Figueiredo HCP, Silva A, Ramos RTJ, Azevedo V. [2015]. **MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes**. *Bioinformation 11(6): 276-279.* [PubMed: 26229287]
26. Nalluri JJ, Kamapantula BK, **Barh D**, Jain N, Bhattacharya A, de Almeida SS, Ramos RTJ, Silva A, Azevedo V, Ghosh P. [2015]. **DISMIRA: Prioritization of disease candidates in miRNA-disease associations based on maximum weighted matching inference model and motif-based analysis**. *BMC Genomics.* 16(Suppl 5):S12. [PubMed: 26040329] **[IF=4.04]**
27. Folador EL, Hassan SS, Lemke N, **Barh D**, Silva A, Ferreira RS, Azevedo V. [2014]. **An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage.** *Integr Biol (Camb). 6(11):1080-7.* [PMID: 25209055] **[IF=4.45]**
28. Bakhtiar SM, Ali A, Baig SM, **Barh D**, Miyoshi A, Azevedo V. [2014]. **Identifying human disease genes: advances in molecular genetics and computational approaches.** *Genet Mol Res. 13(3):5073-87.* [PMID: 25061732] **[IF=1.18]**
29. Santos AR, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, Abdelzaher A, Ghosh P, Tiwari S, Barve N, Jain N, **Barh D**, Silva A, Miyoshi A, Azevedo V. [2013]. **PANNOTATOR: an automated tool for annotation of pan-genomes.** *Genetics and molecular research* 12(3):2982-2989 [PubMed: 24065654] **[IF=1.18]**
30. Ramos RTJ, Carneiro AR, Caracciolo PH, Azevedo V, Schneider MPC**, Barh D**, Silva A. [2013]. **Graphical Contig Analyzer for All Sequencing Platforms (G4ALL): a new stand-alone tool for finishing and draft generation of bacterial genomes.** *Bioinformation;* 9(11):599-604. [PubMed: 23888102] [IF=1.15]
31. Ramos RTJ, Carneiro AR, Azevedo V, Schneider MP, **Barh D**, Silva A. [2012]. **Simplifier: a web tool to eliminate redundant NGS contigs.** *Bioinformation 8(20): 996-999.* [PubMed: 23275695] [IF: 1.15]
32. Sá P, Pinto A, Ramos R, Coimbra N, Baraúna R, Azevedo V, **Barh D**, Silva A. [2012]. **FunSys: Software for functional analysis of prokaryotic transcriptome and proteome.** *Bioinformation 8(11): 529-531.*[PubMed: 22829724] [IF: 1.15]

A.a.3. Bacterial genomes

33. Tiwari S, Jamal SB, Oliveira LC, Clermont D, Bizet C, Mariano D, de Carvalho PV, Souza F, Pereira FL, de Castro Soares S, Guimarães LC, Dorella F, Carvalho A, Leal C, Barh D, Figueiredo H, Hassan SS, Azevedo V, Silva A. [2016]. **Whole-Genome Sequence of Corynebacterium auriscanis Strain CIP 106629 Isolated from a Dog with Bilateral Otitis from the United Kingdom**. *Genome Announc. 2016 Aug 11;4(4). pii: e00683-16. doi: 10.1128/genomeA.00683-16*. [PMID: 27516502]
34. Almeida S, Tiwari S, Mariano D, Souza F, Jamal SB, Coimbra N, Raittz RT, Dorella FA, Carvalho AF, Pereira FL, Soares Sde C, Leal CA, **Barh D**, Ghosh P, Figueiredo H, Moura-Costa LF, Portela RW, Meyer R, Silva A, Azevedo V. **The genome anatomy of Corynebacterium pseudotuberculosis VD57 a highly virulent strain causing Caseous lymphadenitis**. *Stand Genomic Sci. 2016 Apr 8;11:29. doi: 10.1186/s40793-016-0149-7*. [PMID: 27066196] **[IF=1.5]**
35. Hassan SS, Guimarães LC, Pereira Ude P, Islam A, Ali A, Bakhtiar SM, Ribeiro D, Rodrigues Dos Santos A, Soares Sde C, Dorella F, Pinto AC, Schneider MP, Barbosa MS, Almeida S, Abreu V, Aburjaile F, Carneiro AR, Cerdeira LT, Fiaux K, Barbosa E, Diniz C, Rocha FS, Ramos RT, Jain N, Tiwari S, **Barh D**, Miyoshi A, Müller B, Silva A, Azevedo V. [2012] **Complete genome sequence of Corynebacterium pseudotuberculosis biovar ovis strain P54B96 isolated from antelope in South Africa obtained by rapid next generation sequencing technology**. *Stand Genomic Sci.* 7(2):189-99. [PubMed: 23408795] **[IF: 3.6]**
36. Hassan SS, Schneider MP, Ramos RT, Carneiro AR, Ranieri A, Guimarães LC, Ali A, Bakhtiar SM, Pereira Ude P, Dos Santos AR, Soares Sde C, Dorella F, Pinto AC, Ribeiro D, Barbosa MS, Almeida S, Abreu V, Aburjaile F, Fiaux K, Barbosa E, Diniz C, Rocha FS, Saxena R, Tiwari S, Zambare V, Ghosh P, Pacheco LG, Dowson CG, Kumar A, **Barh D**, Miyoshi A, Azevedo V, Silva A. [2012]. **Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp162, Isolated from Camel.** *Journal of Bacteriology; 194(20):5718-19.* [PubMed: 23012291] **[IF: 3.92]**
37. Cerdeira LT, Carneiro AR, Bol E, Barbosa MS, Coimbra N, Ramos RTJ, Almeida SS, Santos AR, Soares SC, Pinto AC, Ali A, Barbosa E, Dorella FA, Rocha FS, Guimaraes LG, Figueira F, Ghosh P**,** Zambare V**,** Barve N**,** Tiwari S, **Barh D**, Miyoshi A, Schneider MPC, Azevedo V, Silva A. [2011]. **Complete Genome Sequence of *Corynebacterium pseudotuberculosis strain - CIP 52.97*- isolated from Horse in Kenya.** *Journal of Bacteriology, 193(24):7025-6.* [PubMed: 22123771] **[IF: 3.92]**
38. Cerdeira LT, Pinto AC, Cruz Schneider MP, Silva S, Santos AR, Vieira Barbosa EG, Ali A, Barbosa MS, Carneiro AR, Jucá Ramos RT, Santos O, **Barh D**, Barve N**,** Zambare V, Estevão S, Guimarães LC, de Castro S, Dorella FA, Rocha FS, Augusto V, Tauch A,

Trost E, Miyoshi A, Azevedo V, Silva A. [2011]. **Whole genome sequence of Corynebacterium pseudotuberculosis PAT10 of strain isolated from sheep in Patagonia, Argentine**. *Journal of Bacteriology*, *193(22):6420-1*. [PubMed: 22038974] **[IF: 3.72]**

A.a.4 Other articles

39. **Barh D**, Ivanova ME, Azevedo V. [2016]**. Are We Ready for Real-Time Applications of Clinical NGS?** *Next Generat Sequenc & Applic. 2:122.doi:10.4172/2469-9853.1000122*
40. Sousa CS, Barros BA, **Barh D**, Ghosh P, Azevedo V, Barros EG, Moreira MA. [2016]. **In silico characterization of 1,2-diacylglycerol cholinephosphotransferase and lysophospha-tidylcholine acyltransferase genes in Glycine max L. Merrill**. *Genet Mol Res. 15(3). doi: 10.4238/gmr.15038974.* [PMID: 27706605] **[IF=1.8]**
41. Kumavath RN, **Barh D**, Azevedo V, Kumar AP. [2017]. **Potential pharmacological applications of enzymes associated with bacterial metabolism of aromatic compounds**. *Journal of Microbiology and Antimicrobials. 9 (1), 1-13*
42. R Kumavath, M Azad, P Devarapalli, S Tiwari, S Kar, **D Barh**, V Azevedo, Kumar AP**. Novel aromatase inhibitors selection using induced fit docking and extra precision methods: Potential clinical use in ER-alpha-positive breast cancer**. *Bioinformation 12 (6), 324-331*
43. JJ Nalluri, **D Barh**, V Azevedo, P Ghosh. [2016]. **Towards a Comprehensive Understanding of miRNA Regulome and miRNA Interaction Networks**. *J Pharmacogenomics Pharmacoproteomics. 7 (160), 2153-0645.1000160*
44. Kumavath RN, Ramana ChV, Sasikala Ch, **Barh D**, Kumar AP, Azevedo V. [2015]. **Isolation and characterization of L-tryptophan ammonia lyase from Rubrivivax benzoatilyticus strain JA2**. *Curr Protein Pept Sci. DOI:10.2174/1389203716666150505235929* [PubMed: 25961404] **[IF=2.33]**
45. Foladora EL, de Oliveira AF, Jamala SB, Silva A, Ferreira SR, **Barh D**, Ghosh P, Azevedo V. [2015]. ***In silico* protein-protein interactions: avoiding data and method biases over sensitivity and specificity**. *Curr Protein Pept Sci. DOI:10.2174/1389203716666150505235437* [PubMed: 25961403] **[IF=2.33]**
46. Devarapalli P, Kumavath RN, **Barh D**, Azevedo V. [2014]. The **conserved mitochondrial gene distribution in relatives of Turritopsis nutricula, an immortal jellyfish.** *Bioinformation.*10(9):586-91. [PMID: 25352727]
47. **Barh D**, Jain N, Tiwari S, Field JK, Padin-Iruegas E, Ruibal A, López R, Herranz M, Bhattacharya A, Juneja L, Viero C, Silva A, Miyoshi A, Kumar A, Blum K, Azevedo V, Ghosh P, Liloglou T. [2013]. **A novel in silico reverse-transcriptomics-based identification and blood-based validation of a panel of sub-type specific biomarkers in lung cancer**. *BMC Genomics*, 14 (Suppl 6)**, S5** [PubMed: 24564251] **[IF: 4.07]**

A.b: Journal articles with other co-authors

A.b.1 Bioinformatics / Software with other co-authors

48. Talukder AK, Ravishankar S, Sasmal K, Gandham S, Prabhukumar J, Achutharao PH, **Barh D**, Blasi F. [2015]. **XomAnnotate: Analysis of heterogeneous and complex exome- a step towards translational medicine**. *PLOS ONE*; 10(4):e0123569. [PubMed: 25905921] **[IF=4.04]**
49. **Barh D,** Jain N. **A novel omics strategy to identify biomarkers for early diagnosis and classification of lung cancer.** *Journal of Thoracic Oncology, 2012, 7 (11), S5, S471* **[IF: 5.28]**
50. **Barh D**, Gupta K, Khatri G, Rahangdale S, Verma A. [2012]. **An Integrative Omics Strategy for Identification of Skin Cancer Biomarkers**. *Eur J Cancer, 48 (S6), 79.* **[IF: 5.62]**
51. **Barh D**, Tiwari S, Bhat D. [2012]. **Next-generation markers and targets in breast cancer: an integrative omics approach.** *Mol Cancer Ther. 10: (11), S1.* **[IF: 5.22]**
52. **Barh D**, Bhat D. [2011]. **Integrative-omics based next-generation molecular markers and targets in colorectal cancer**. *Annals of Oncology, 21(Suppl 8):viii70*: **[IF: 6.8]**
53. **Barh D**, Bhat D, Viero C. [2010]. **miReg: a resource for microRNA regulation**. *Journal of Integrative Bioinformatics; 7(1); 144.* [PubMed: 20693604
54. **Barh D**, Sindhurani P, Bhattacharjee A. [2010]. **NR2E1 inhibits cell proliferation in lung squamous cell carcinomas by regulating entry to cell cycle**. *Journal of Thoracic Oncology; 5 (5), S46*: **[IF: 5.2]**
55. **Barh D**, Sindhurani P. [2010]. **BLU is a classical tumor suppressor gene acts in DNA repair Pathway**. *Journal of Oncology Pharmacy Practice; 16(2), S34*. [PubMed: 20351168]
56. **Barh D**, Kumar A, Misra AN. [2010]. **In silico identification of dual ability of *N. gonorrhoeae* ddl for developing drugs and vaccines against pathogenic Neisseria and other human pathogens**. *Journal of Proteomics & Bioinformatics; 3(3), 082-090.*
57. **Barh D**, Kumar A, Misra AN. [2009]. **Genomic Target Database (GTD): A database of potential targets in human pathogenic bacteria**. *Bioinformation; 4 (1), 50-51.* [PubMed: 20011153] [IF: 1.15]

A.b.2 Other articles with other co-authors

58. Zolnikova IV, Strelnikov VV, Skvortsova NA, Tanas AS, **Barh D**, Rogatina EV, Egorova IV, Levina DV, Demenkova ON, Prikaziuk EG, Ivanova ME. [2016]. **Stargardt disease-associated mutation spectrum of a Russian Federation cohort**. *Eur J Med Genet. 60(2):140-147. doi: 10.1016/j.ejmg.2016.12.002.* [PMID: 27939946] **[IF=1.5]**
59. Barh D. [2016]. **NTEGRATING FITNESS GENOMICS IN PRECISION MEDICNE**. *Precision Med. 1 (1): 7-9*
60. Gupta KK, Rahangdale S, Barh D. [2016]. **AN INTEGRATIVE BIOINFORMATICS APPROACH FOR IDENTIFICATION OF BIOMARKERS IN MYOCARDIAL INFARCTION**. *Precision Med in Cardiol. 1(1): 8-12*
61. Barh D. [2016]. **PRECISION MARKERS IN LUNG CANCER: AN INDIAN SCENARIO OF MUTATIONS**. *Precision Med. In Onclogy. (1): 24-26.*
62. T McLaughlin, M Febo, RD Badgaiyan, **Barh D**, K Dushaj, ER Braverman, K Blum. [2016]. **KB220Z™ a Pro-Dopamine Regulator Associated with the Protracted, Alleviation of Terrifying Lucid Dreams. Can We Infer Neuroplasticity-induced Changes in the Reward Circuit**. *J Reward Defic Syndr Addict Sci. 2 (1), 3-13*
63. Blum K, Oscar-Berman M, Waite RL, Braverman ER, Kreuk F, Li M, Dushaj K, Madigan MA, Hauser M, Simpatico T, **Barh D**. [2014]. **A Multi-Locus Approach to Treating Fibromyalgia by Boosting Dopaminergic Activity in the Meso-Limbic System of the Brain.** *J Genet Syndr Gene Ther. 5(1):213.* [PMID: 24883230]
64. K Blum, M Oscar-Berman, SH Blum, MA Madigan, RL Waite, M Thomas, **Barh D**. [2014]. **Can Genetic Testing Coupled with Enhanced Dopaminergic Activation Reduce Recidivism Rates in the Workers Compensation Legacy Cases?** *J Alcohol Drug Depend*. 2:3. DOI: 0.4172/2329-6488.1000161
65. Blum K, Oscar-Berman M, Demetrovics Z, **Barh D**, Gold MS. [2014]. **Genetic Addiction Risk Score (GARS): Molecular Neurogenetic Evidence for Predisposition to Reward Deficiency Syndrome (RDS)**. *Mol Neurobiol*. 50(3):765-96. [PMID: 24878765] **[IF=6.5]**
66. Blum K, Oscar-Berman M, Downs W, Braverman ER, Kreuk F, Dushaj K, Truesdell C, Li M, Giordano J, Borsten J, Simpatico T, **Barh D**, Madigan MA, Jones S, Schoenthaler S. [2014]. **Hypothesizing that Putative Dopaminergic, Melatonin, Benzodiazepine Reward Circuitry Receptor(s) Activator Provides Sleep Induction Benefits**. *Journal of Sleep Disorders & Therapy*, 3(153). DOI:10.4172/2167-0277.1000153
67. RL Waite, M Oscar-Berman, E RBraverman, D Barh, K Blum. [2014]. **Quantitative Electroencephalography Analysis (qEEG) of Neuro-Electro-Adaptive Therapy 12T™[NEAT12] Up-Regulates Cortical Potentials in an Alcoholic during Protracted Abstinence**: Putative Anti-Craving Implications. *J Addict Res Ther*. 5:1. DIO: 0.4172/2155-6105.1000171
68. Blum K , Han D, Oscar-Berman M, Reinl G, DiNubile N, Madigan M, Bajaj A, Downs B, Giordano J, Westcott W, Smith L, Braverman E, Dushaj K, Hauser M, Simpatico T, McLaughlin T, Borsten J, **Barh D.** [2013]. **Iatrogenic opioid dependence is endemic and legal: Genetic addiction risk score (GARS) with electrotherapy a paradigm shift in pain treatment programs.** *Health, 5, 16-34.* doi: 10.4236/health.2013.511A1004
69. Blum K, Thompson B, Oscar-Berman M, Giordano J, Braverman E, Femino J, **Barh D**, Downs W, Smpatico T, Schoenthaler S. **[2013]. Genospirituality: Our Beliefs, Our Genomes, and Addictions.** *J Addict Res Ther. 5(4). pii: 162.* [PubMed: 24971227]
70. Hill E, Han D, Dumouchel P, Dehak N, Quatieri T, Moehs C, Oscar-Berman M, Giordano J, Simpatico T, **Barh D**, Blum K. [2013]. **Long term suboxone™ emotional reactivity as measured by automatic detection in speech.** *PLOS ONE*. 9;8(7):e69043. [PubMed: 23874860] **[IF=4.09]**
71. Downs BW, Oscar-Berman M, Waite RL, Madigan MA, Giordano J, Beley T, Jones S, Simpatico T, Hauser M, Borsten J, Marcelo F, Braverman ER, Lohmann R, Dushaj K, Helman M, **Barh D**, Schoenthaler ST, Han D, Blum K. [2013]. **Have We Hatched the Addiction Egg: Reward Deficiency Syndrome Solution System™.** *J Genet Syndr Gene Ther*; 4: 136. [PMID: 24077767]
72. Kushner S, Han D, Oscar-Berman M, Downs WB, Madigan MA, Giordano J, Beley T, Jones S, **Barh D**, Simpatico T, Dushaj K, Lohmann R, Braverman ER, Schoenthaler S, Ellison D, Blum K. [2013]. **Declinol, a Complex Containing Kudzu, Bitter Herbs (Gentian, Tangerine Peel) and Bupleurum, Significantly Reduced Alcohol Use Disorders Identification Test (AUDIT) Scores in Moderate to Heavy Drinkers: A Pilot Study**. *J Addict Res Ther*, 4: 153. [PubMed: 24273684]
73. Mclaughlin T, Oscar-Berman M, Simpatico T, Giordano J, Jones S, **Barh D**, Downs, Waite RL, Madigan M, Dushaj K, Lohmann R, Braverman ER, Han D, Blum K. [2013]. **Hypothesizing repetitive paraphilia behavior of a medication refractive Tourette's syndrome patient having rapid clinical attenuation with KB220Z-nutrigenomic amino-acid therapy (NAAT).** *Journal of Behavioral Addictions*, 2 (2): 117-124, DOI: 10.1556/JBA.2.2013.2.8
74. Blum K, Oscar-Berman M, Femino J, Waite RL, Benya L, Giordano J, Borsten J, Downs WB, Braverman ER, Loehmann R, Dushaj K, Han D, Simpatico T, Hauser M, **Barh D**, McLaughlin T. [2013]. **Withdrawal from Buprenorphine/Naloxone and Maintenance with a Natural Dopaminergic Agonist: A Cautionary Note.** *J Addict Res Ther 4: 146.* [PMID: 24273683]
75. Campbell HB, Oscar-Berman M, Giordano J, Beley TG, **Barh D**, Downs BW, Blum K. [2013]. **Common Phenotype in Patients with Both Food and Substance Dependence: Case Reports**. *J Genet Syndr Gene Ther 4: 122.* [PubMed: 23543232]
76. Blum K, Oscar-Berman, **Barh D**, Giordano J, Gold MS. [2013]. **Dopamine Genetics and Function in Food and Substance Abuse**. *J Genet Syndr Gene Ther 4: 121.* [PubMed: 23543775]
77. Blum K, Oscar-Berman M, Dinubile N, Giordano J, Braverman ER, Truesdell CE, **Barh D**, Badgaiyan R. [2013].**Coupling Genetic Addiction Risk Score (GARS) with Electrotherapy: Fighting Iatrogenic Opioid Dependence.** *J Addict Res Ther*. 4(163):1000163. [PubMed: 24616834]
78. Blum K, Han D, Hauser M, Downs BW, Giordano J, Borsten J, Winchell E, Simpatico T, Madigan MA, **Barh D**. [2013]. **Neurogenetic Impairments of Brain Reward Circuitry Links to Reward Deficiency Syndrome (RDS) as evidenced by Genetic Addiction Risk Score (GARS): A case study**. *The IIOAB Journal. 4 (1): 4-9.*
79. **Barh D** and Vedamurthy AB. [2013]. **Therapeutic potential of Let-7, miR-125, miR-205, and miR-296 in breast cancer: An update.** *The IIOAB Journal. 4 (1): 25-26*
80. Miller M, Chen A, Stokes A, Silverman S, Bowirrat A, Manka M, Manka D, Miller D, Perrine K, Chen T, Bailey J, Downs BW, Waite R, Madigan M, Braverman E, Damle U, Kerner M, Giordano J, Morse S, Oscar-Berman M, **Barh D**, Blum K. [2012]. **Early Intervention of Intravenous KB220IV- Neuroadaptagen Amino-Acid Therapy (NAAT)™ Improves Behavioral Outcomes in a Residential Addiction Treatment Program: A Pilot Study**. *J Psychoactive Drugs.* 44(5):398-409. [PubMed: 23457891] **[IF: 1.72]**

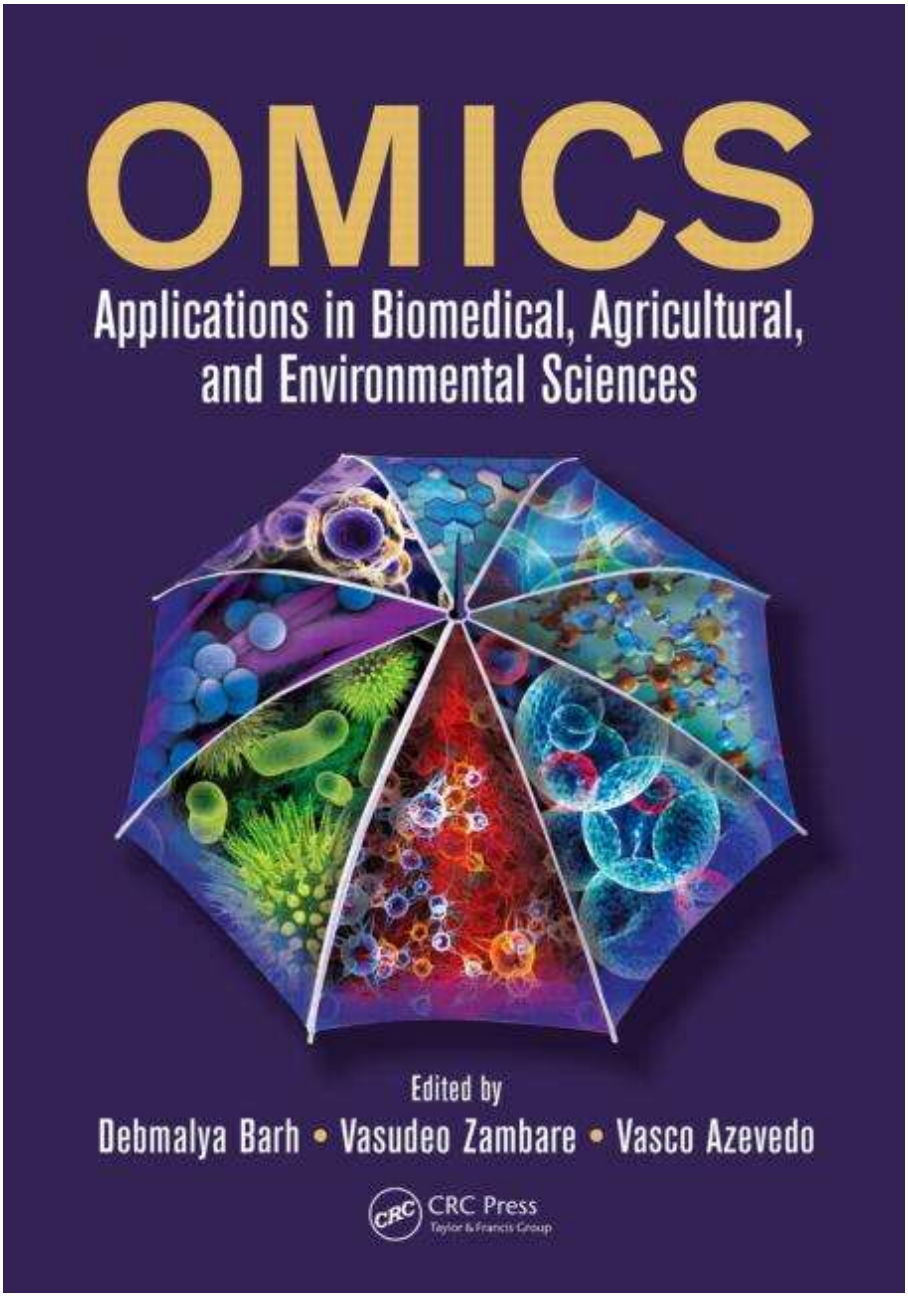
81. Blum K, Oscar-Berman M, Stuller E, Miller D, Giordano J, Morse S, McCormick L, Downs WB, Waite RL, **Barh D**, Neal D, Braverman ER, Lohmann R, Borsten J, Hauser M, Han D, Liu Y, Helman M, Simpatico T. [2013]. **Neurogenetics and Nutrigenomics of Neuro-Nutrient Therapy for Reward Deficiency Syndrome (RDS): Clinical Ramifications as a Function of Molecular Neurobiological Mechanisms.** *J Addict Res Ther.* 2012 Nov 27;3(5):139. [PubMed: 23926462]
82. Blum K, Giordano J, Borsten J, Downs BW, Hauser M, Simpatico T, Lohmann R, Braverman ER, **Barh D**. [2012]. **Translational Research to Uncover Diagnostic & Therapeutic Gene Targets Emerging in a Genomic Era: from Bench to Bedside**. *J Genet Disor Dis Inf 2012, 1:1*. doi.org/10.4172/jgddi.1000e103
83. Blum K, Giordano J, Oscar-Berman M, Bowirrat A, Simpatico T, **Barh D**. [2012]. **Diagnosis and Healing In Veterans Suspected of Suffering from Post-Traumatic Stress Disorder (PTSD) Using Reward Gene Testing and Reward Circuitry Natural Dopaminergic Activation.** *J Genet Syndr Gene Ther*. 3(3):1000116. [PubMed: 23264885]
84. Blum K, Chen ALC, Giordano J, Borsten J, Chen TJH, Hauser M, Simpatico T, Femino J, Braverman ER, **Barh D**. [2012]. **The Addictive Brain: All Roads Lead to Dopamine.** *J Psychoactive Drugs, 44 (2), 134–143.* [PubMed: 22880541] [IF: 1.72]
85. Blum K, Oscar-Berman M, Bowirrat A, Giordano J, Madigan M, Braverman ER, **Barh D**, Hauser M, Borsten J, Simpatico T.. [2012]. **Neuropsychiatric Genetics of Happiness, Friendships, and Politics: Hypothesizing Homophily ("Birds of a Feather Flock Together") as a Function of Reward Gene Polymorphisms.** *J Genet Syndr Gene Ther 3:112.* doi:10.4172/2157-7412.1000112 [PubMed: 23336089]
86. Chen ALC, Blum K, Chen TJH, Giordano J, Downs BW, Han D, **Barh D**, Braverman ER. [2012]. **Correlation of the Taq1 Dopamine D2 Receptor Gene and Percent Body Fat in Obese and Screened Control Subjects: A Preliminary Report**. *Food & Function; 3(1):40-8.* [PubMed: 22051885] **[IF:** 2.79]
87. Blum K, Chen TJH, Bailey J, Bowirrat A, Femino J, Chen ALC, Madigan M, Simpatico T, Morse S, Giordano J, Damle U, Kerner M, Braverman ER, Fornari F, Downs BW, Rector C, **Barh D**, Berman MC. **Can the chronic administration of the combination of buprenorphine and naloxone block dopaminergic activity causing anti-reward and relapse potential**? *Molecular Neurobiology, 44(3):250-68.* [PubMed: 21948099] **[IF: 6.068]**
88. Blum K, Giordano J, Morse S, Anderson A, Carbaja J, Waite R, Downs B, Downs J, Madigan M, **Barh D**, Braverman E. [2011]. **Hypothesizing Synergy between Acupuncture/ Auriculotherapy and Natural Activation of Mesolimbic Dopaminergic Pathways: Putative Natural Treatment Modalities for the Reduction of Drug Hunger and Relapse**. *IIOAB Letters. 1: 8-20.* DOI: 10.5195/iioablett.2011.9
89. Blum K, Chen ALC, Oscar-Berman M, Chen TJH, Lubar J, White N, Lubar J, Bowirrat A, Braverman E, Schoolfield J, Waite RL, Downs BW, Madigan M, Comings DE, Davis C, Kerner MM, Knopf J, Palomo T, Giordano JJ, Morse SA, Fornari F, **Barh D**, Femino J, Bailey JA. [2011]. **Generational Association Studies of Dopaminergic Genes in Reward Deficiency Syndrome (RDS) Subjects: Selecting Appropriate Phenotypes for Reward Dependence Behaviors**. *Int J Env Res Public Health; 8(12):4425-4459.* [PubMed: 22408582] **[IF: 2.49]**
90. Blum K, Bagchi D, Bowirrat A, Downs WB, Waite RL, Madigan M, Downs JM, Giordano J, Morse S, Braverman ER, Polanin M, **Barh D**, Fornari F, Simpatico T. [2011]. **Nutrigenomics of Neuradaptogen Amino-Acid-Therapy and Neurometabolic Optimizers: Overcoming carbohydrate bingeing and overeating through neurometabolic mechanisms**. *Functional Foods in Health and Disease; 9:310-378.*
91. Morse S, Giordano J, Perrine K, Downs BW, Waite RL, Madigan M, Bailey J, Braverman ER, Damle U, Knopf J, Simpatico T, Moeller MD, **Barh D**, Blum K. (2011). **Audio Therapy Significantly Attenuates Aberrant Mood in Residential Patient Addiction Treatment: Putative Activation of Dopaminergic Pathways in the Meso-Limbic Reward Circuitry of Humans**. *J Addict Res Ther; S3:001.* doi:10.4172/2155-6105.S3- 001
92. Muhammad SA, Ali A, Naz A, Hassan A, Riaz N, Saeed-ul-Hassan S, Andleeb S, **Barh D**. **A New Broad-Spectrum Peptide Antibiotic Produced by Bacillus brevis Strain MH9 Isolated from Margalla Hills of Islamabad, Pakistan**. *Int J Pept Res Ther. 22: 271. doi:10.1007/s10989-015-9508-2* **[IF: 1.9]**
93. Kumavath R, Prasad S, Devarapalli P, **Barh D**. [2014]. **In silico identification of novel candidate drug targets in Haemophilus influenzae Rd KW20**. *Int J of Genet and Genomics, 2(4): 62-67.* doi: 10.11648/j.ijgg.20140204.13
94. Jha UC, Chaturvedi SK, Bohra A, Basu PS, Khan MS, **Barh D**. [2014]. **Abiotic stresses, constraints, and improvement strategies in chickpea**. *Plant Breeding*, DOI: 10.1111/pbr.12150. **[IF=1.8]**
95. Nalluri J, Kamapantula B, **Barh D**, Ghosh P, Jain N, Juneja L, Barve N. [2013]. **Determining miRNA-disease associations using Bipartite Graph Modelling**. *ACM 978-1-4503-2434-2/13/09*. BCB '13, September 22 - 25, 2013, Washington, DC
96. **Barh D.** [2012]. **A normal 46 XX karyotype does not always represent female phenotype**. *The IIOAB Journal 3 (3), 49-50*
97. Hemaiswarya S, Raja R, Carvalho IS, Ravikumar R, Zambare V**, Barh D**. [2012]. **An Indian scenario on renewable and sustainable energy sources with emphasis on algae.** *Appl Microbiol Biotechnol; 96(5):1125-35.* [PubMed: 23070650] **[IF: 3.42]**
98. Zambare V. Zambare A, **Barh D**, Christopher LP. [2012]. **Optimization of enzymatic hydrolysis of prairie cordgrass for improved ethanol production.** *J. Renewable Sustainable Energy 4, 033118* **[IF:** 1.14]
99. **Barh D**, Malhotra R, Ravi B, Sindhurani P. [2010]. **microRNA Let-7: an emerging next-generation cancer therapeutic**. *Current Oncology; 17 (1), 70-80.* [PubMed: 20179807] **[IF: 2.4]**
100. Shanthi PA, Singh P, Barh D, Venkatachalaiah G. [2010]. **Comparative karyology based systematics of *Euphlyctis hexadactylus* and *Euphlyctis cyanophlyctis***. *International Journal of Integrative Biology; 9 (1), 6-9.*
101. **Barh D**, Misra AN. [2009]. **Epitope Design from Transporter Targets in *N. gonorrhoeae***. *Journal of Proteomics & Bioinformatics; 2, 475-480.*
102. **Barh D**, Kumar A, Chatterjee S, Liloglou T. [2009]. **Molecular features, markers, drug targets, and targeted therapeutics in cardiac myxoma**. *Current Cancer Drug Targets; 9 (6), 705-716.* [PubMed: 19754355] **[IF: 4.3]**
103. **Barh D**, Misra AN. [2009]. **In silico Identification of membrane associated candidate drug targets in *Neisseria gonorrhoeae***. *International Journal of Integrative Biology; 6(2), 65-67.*
104. **Barh D**. [2009]. **Targeted therapy in unusual cancers with special reference to cardiac myxoma and male breast cancer**. *New Biotechnology; Vol. 25, S11.* **[IF: 2.8]** ***(Conference Abstract)***
105. **Barh D**. [2009]. **In silico critical disease pathway mapping using public domain data and tools for development of personalized medicine**. *Advanced Biotech; June, 35-37.*
106. **Barh D** and Kumar A. [2009]. ***In silico* identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae**. *In Silico Biol.; Vol. 9, 0019.* [PubMed: 20109152] [Cited in Nature Reviews Microbiology, 9(7): 2011]

107. **Barh D**, Parida S. [2009]. **Cardiac myxoma: molecular markers, critical disease pathways, drug targets, and putative targeting miRs**. *Cancer Therapy; 7, 77-96*.
108. **Barh D**. [2009]. **Biomarkers, critical disease pathways, drug targets, and alternative medicine in male breast cancer**. *Current Drug Targets; 10, 1-8*. [PubMed: 19149530] **[IF: 3.2]**
109. **Barh D**, Das. K. [2008]. **Targeting critical disease pathways in male breast cancer: a pharmacogenomics approach**. *Cancer Therapy, 6: 193-212*.
110. **Barh D**. [2008]. **Let-7 replacement therapy: applicability in cancer**. *Cancer Therapy; 6, 939-984.*
111. **Barh D**, Parida S. Parida BP, Viswanathan G. [2008]. **Let-7, miR-125, miR-205, and miR-296 are prospective therapeutic agents in breast cancer molecular medicine**. *Gene Therapy and Molecular Biology; 12, 189-206.* [IF: 0.83]
112. **Barh D**, Viswanathan G. [2008]. **Syzigium Cumini Inhibits Growth and Induces Apoptosis in Cervical Cancer Cell Lines: A Primary Study**. *ecancermedicalscience*, 2: 83 DOI: 10.3332/eCMS.2008.83 **[IF: 1.2]**
113. **Barh D**, Srivastava HC, Mazumdar, BC. [2008]. **Self Fruit Extract and Vitamin-C Improves Tomato Seed Germination**. *Journal of Applied Sciences Research, 4(2): 156-165*.
114. **Barh D.** [2008]. **Dietary Phytochemicals: a Promise to Chemoprevention.** *Advanced Biotech, 21-23*
115. **Barh D**, Mazumdar BC. [2008]. **Comparative nutritive values of palm saps before and after their partial fermentation and effective use of wild date (Phoenix sylvestris Roxb.) sap in treatment of anemia**. *Research Journal of Medicine and Medical Sciences, 3 (2), 173-176*.
116. **Barh D**, Mazumdar BC [2007]. **Role of Calotropis procera** flowers for use in vase. The Lalbagh, 30 (1-2): 79-80.
117. **Barh D**, Mukhopadhyaya P, Mazumdar BC. [2005]. Analytical studies on sap and fruits of Palmyra and wild date grown in West Bengal. Indian Agric, 49 (1-2): 111-115.

VIII. APPENDIX

B. Published Book

B.a: Book with Prof. Vasco Azevedo



OMICS: Applications in Biomedical, Agricultural, and Environmental Sciences
Editors: Debmalya Barh, Vasudeo Zambare, Vasco Azevedo
Publication Date: March 26, 2013 by CRC Press
Reference: 713 Pages, 97 B/W Illustrations
ISBN: 9781466562813
Publisher: CRC Press, Taylor & Francis Group, USA

Amazon Best Sellers Rank: #4,136,008 in Books (As on 20 January, 2017)

Features

- Covers almost all current and emerging omics applications in one volume
- Explores omics in the biomedical, environmental, and agricultural fields
- Discusses epigenomics, pharmacogenomics glycomics, spliceomics, metagenomics, toxicogenomics, environomics, and other cutting-edge areas
- Draws on the work of active, expert researchers around the world
- Includes extensive references to aid in further research
- Contains an eight-page color insert and more than 95 black-and-white illustrations

Summary

With the advent of new technologies and acquired knowledge, the number of fields in omics and their applications in diverse areas are rapidly increasing in the postgenomics era. Such emerging fields—including pharmacogenomics, toxicogenomics, regulomics, spliceomics, metagenomics, and environomics—present budding solutions to combat global challenges in biomedicine, agriculture, and the environment. **OMICS: Applications in Biomedical, Agricultural, and Environmental Sciences** provides valuable insights into the applications of modern omics technologies to real-world problems in the life sciences. Filling a gap in the literature, it offers a broad, multidisciplinary view of current and emerging applications of omics in a single volume.

Written by highly experienced active researchers, each chapter describes a particular area of omics and the associated technologies and applications. Topics covered include:

- Proteomics, epigenomics, and pharmacogenomics
- Toxicogenomics and the assessment of environmental pollutants
- Applications of plant metabolomics
- Nutrigenomics and its therapeutic applications
- Microalgal omics and omics approaches in biofuel production
- Next-generation sequencing and omics technology for transgenic plant analysis
- Omics approaches in crop improvement
- Engineering dark-operative chlorophyll synthesis
- Computational regulomics
- Omics techniques for the analysis of RNA splicing
- New fields, including metagenomics, glycomics, and miRNA
- Breast cancer biomarkers for early detection
- Environomics strategies for environmental sustainability

This timely book explores a wide range of omics application areas in the biomedical, agricultural, and environmental sciences. Throughout, it highlights working solutions as well as open problems and future challenges. Demonstrating the diversity of omics, it introduces readers to state-of-the-art developments and trends in omics-driven research.

Book inauguration at ICB/UFMG by Prof. Vasco Azevedo

NDC NÚCLEO DE DIVULGAÇÃO CIENTÍFICA

CEDECOM - Centro de Comunicação da UFMG

Busca

Home | Conheça o NDC | Projetos | Agenda | Notícias | Imagens | Vídeos | Áudios | Fale conosco | Sobre divulgação científica

Notícias

Professor do ICB lança livro que descreve universo das ciências `ômicas'

20 de agosto de 2013 . Comentários (0)

Utilizado para descrever diversos campos emergentes da biologia, o sufixo omic abrange áreas que abrem vastas possibilidades de conhecimento e podem levar a caminhos para o enfrentamento de demandas globais em biomedicina, agricultura e meio ambiente, comenta o professor Vasco Azevedo [em foto de Isabella Lucas/UFMG], do Departamento de Biologia Geral do Instituto de Ciências Biológicas (ICB).

Editor de livro sobre o tema lançado este ano em parceria com pesquisadores indianos, Azevedo afirma que o principal desafio das chamadas ciências ômicas é "minerar e organizar dados no dilúvio de informação que cresce a cada hora, para que se possa interpretá-los".

Ele lembra que a palavra "genômica" nasceu em meados da década de 1980, junto com as primeiras tecnologias de sequenciamento de genomas. "A princípio, achava-se que bastava a genômica estrutural, ou seja, era só sequenciar o DNA. Depois é que se percebeu a importância de outras duas: a funcional, que faz o estudo do RNA e das proteínas, e a comparativa", lembra.

Logo começaram a aparecer as primeiras derivações para as ômicas, com a transcriptômica, ou seja, sequenciamento de RNA, e a proteômica, sequenciamento de proteína. Na mesma linha, surgiram áreas de estudo como farmacogenômica, epigenômica, toxicogenômica, metabolômica e interatômica.

No livro Omics: applications in biomedical, agricultural, and environmental sciences, Vasco e os coeditores indianos propõem o termo unknowwn-omics/unknowwnomics para fazer referência aos novos campos de pesquisa que, acreditam, em futuro próximo serão incorporados sob essa terminologia.

Segundo ele, com a descoberta de novas tecnologias e conhecimentos, o número de campos em omics e as suas aplicações tem aumentado rapidamente na era pós-genômica. A farmacogenômica, por exemplo, revela que determinados medicamentos – para câncer ou para doenças autoimunes – e até vacinas podem agir em um indivíduo e não em outro, já que o efeito depende da resposta imune, e seria preciso conhecer a genômica do paciente para garantir resultado satisfatório.

"Mas isso teria um preço enorme, seria economicamente inviável em grande escala", adverte. Já a interatômica lida com a ligação que ocorre, por exemplo, entre o genoma humano e o das bactérias que habitam esse organismo.

"É comum que falhas nessa interface gerem problemas, como doença inflamatória de intestino. Há estudos que sugerem que esse tipo de interação pode estar ligado a situações como diabetes e obesidade", comenta o professor, citando também outros campos, como toxicogenômica e avaliação de poluentes ambientais, nutrigenômica e suas aplicações terapêuticas, e biomarcadores de câncer de mama para detecção precoce.

As ômicas também têm ampla aplicação na melhoria de plantas e produtos e no controle de pragas, e as grandes commodities como soja, milho e arroz já foram sequenciadas. Vasco ressalta que, através da genômica, é possível inserir genes que confiram às plantas resistência a solos ácidos ou arenosos ou desenvolver espécies capazes de transferir certos genes para que culturas de ecossistemas áridos se tornem menos dependentes de água.

No entanto, a ciência ainda não conhece o real efeito no organismo humano de uma planta modificada. "É uma tecnologia extremamente poderosa, que precisa ser testada e validada, pois podem ocorrer acidentes graves, até ambientais", observa o pesquisador, que defende a existência de uma espécie de selo, emitido pelas universidades, para garantir a segurança de produtos transgênicos.

"Sou a favor da ideia de transgênico, mas cada transgênico é um, cada evento é diferente do outro, por isso cada um precisa de certificação. E as empresas devem investir recursos para obterem um selo de certificação expedido pelas instituições públicas de pesquisa", reitera.

Vazio coberto

Em parceria com os pesquisadores indianos Debmalya Barh e Vasudeo Zambare, Vasco Azevedo lançou o primeiro livro que traz uma visão geral da diversidade das ômicas e suas aplicações em vários campos da biologia.

Os artigos foram escritos por 78 docentes e pós-doutorandos de países como Dinamarca, Austrália, Estados Unidos, Índia, Portugal, Paquistão e Brasil. Cada capítulo descreve um omic em particular e as tecnologias associadas à aplicação.

Voltada sobretudo para alunos de pós-graduação e profissionais, a obra tem 26 capítulos, divididos em três partes. A primeira trata do mundo dos omics e aplicações nas ciências biomédicas, a segunda aborda aplicações em plantas e ciências da agricultura, e a última mostra o emprego nas ciências ambientais.

"É uma obra extensa, com 713 páginas, toda escrita em língua inglesa. Mas escrever livros na área de ciências biológicas é uma atividade sem retorno, em todos os sentidos", pondera Azevedo. Ele comenta que na área biológica o CNPq não reconhece livros como atividade acadêmica. Além disso, autores e editores cederam o direito financeiro à editora.

"Trata-se de um trabalho exaustivo, mas fomos estimulados pelo desejo de cobrir um vazio editorial existente na área, já que os livros disponíveis têm focado apenas em ômicas específicas ou aplicações de várias ômicas para uma área em particular", justifica o professor.

(Ana Rita Araújo/Boletim 1832)

https://www.ufmg.br/online/ndc/noticias/professor-do-icb-lanca-livro-que-descreve-universo-das-ciencias-omicas/

B.b: Books with others

1. **Barh D** (Edt): **Omics Technologies and Bio-Engineering: Towards Improving Quality of Life.** ISBN: 9780128046593, First Ed: 2017; *Academic Press Elsevier. (In Press)*
2. **Barh D** (Edt): **Precision Medicine: Prediction, Prevention with Personalization.** ISBN: 978-1498775601, First Ed: 2017; *Taylor & Francis LLC, USA. (In Press)*
3. Verma M, Barh D (Edt): **Progress and Challenges in Precision Medicine**. ISBN: 9780128094112, First Ed: Jan 2017; *Academic Press, Elsevier, USA*
4. Khan MS, Khan IA, **Barh D** (Edt): **Applied Molecular Biotechnology: The Next Generation of Genetic Engineering**. ISBN: 9781498714815, 2016; *Taylor & Francis LLC, USA*
5. **Barh D,** Khan MS, Davies E. (Edt): **PlantOmics: the omics of plant science.** ISBN: 9788132221722; First Ed: 2015; *Springer LLC.*
6. **Barh D**, Gunduz M (Edts): **Noninvasive Molecular Markers in Gynecologic Cancers**. ISBN: 9781466569386, First Ed: 2015; *Taylor & Francis LLC, USA*
7. **Barh D** (Edt): **Omics approaches in breast cancer: towards next-generation diagnosis, prognosis, and therapy**. ISBN: 9788132208426; First Ed: 2014; *Springer LLC*
8. **Barh D**, Carpi A, Verma, M, and Gunduz M (Edts): **Cancer Biomarkers: Non-Invasive Early Diagnosis and Prognosis.** ISBN: 9781466584280; First Ed: 2014; *Taylor & Francis LLC, USA*
9. **Barh D** (Edts): **OMICS Applications in Crop Science**. ISBN: 9781466562813; First Ed: Dec, 2013. *Taylor & Francis LLC, USA*
10. **Barh D**, Dhwan D and Ganguly NK (Edts): **Omics for Personalized Medicine**. ISBN: 978-8132211839, First Ed: 2013; *Springer LLC,*
11. **Barh D,** Blum K, and Madigan MA (Edts): **OMICS: Biomedical Perspectives and Applications.** ISBN: 9781439850084; First Ed: 2011, *CRC Press, USA*.
12. Singh K, Vig AP, and **Barh D**: **VERMICOMPOSTING: a boon for soil, plant, and environment.** ISBN: 9783844334425; 2011; *LAP Lambert Academic Publishing, Germany.*

11. **Barh D,** Blum K, and Madigan MA (Edts): **OMICS: Biomedical Perspectives and Applications.** ISBN: 9781439850084; First Ed: 2011, *CRC Press, USA*.
12. Singh K, Vig AP, and **Barh D**: **VERMICOMPOSTING: a boon for soil, plant, and environment.** ISBN: 9783844334425; 2011; *LAP Lambert Academic Publishing, Germany.*
13. **Barh D**, Das K, and Srivastava HC: **Male Breast Cancer: Identifying and targeting critical disease signaling pathways towards development of personalized phytotherapy**. ISBN: 9783843367592; First Ed: 2010; *LAP Lambert Academic Publishing, Germany.*
14. Srivastava HC, **Barh D**: **Genetics: Fundamentals and Applications.** ISBN: 8181892631; First Ed: 2008; *International Book Distributing Company, INDIA.*

III. APPENDIX

C. Book Chapters

C.a: Book chapters with Prof. Vasco Azevedo

1. de Sousa CS, Corrêa Mendonça MA, Hassan SS, Azevedo VA, **Barh D**. [2016]. **Biotechnology for improved crop productivity and quality**. In "Applied Molecular Biotechnology: The Next Generation of Genetic Engineering". Edt by Khan MS, Khan IA, **Barh D**, ISBN: 9781498714815, First Ed: 2016; pp. 231–248. *Taylor & Francis LLC, USA*
2. Verma M, **Barh D**, Hassan SS, Azevedo V. [2015]. **Next-generation Molecular Markers: Challenges, Applications, and Future Perspectives**. In 'Molecular Biology and Biotechnology", 6th Ed, ISBN: 9781849737951, Edt by Rapley R, Whitehouse D; pp. 420-453, *Royal Society of Chemistry, UK*
3. Verma M, **Barh D**, Tiwari S, Azevedo V. [2015]. **Molecular Biomarkers: Overview, Technologies, and Strategies**. In 'Molecular Biology and Biotechnology", 6th Ed, ISBN: 9781849737951, Edt by Rapley R, Whitehouse D; pp. 369-419, *Royal Society of Chemistry, UK*
4. **Barh D**, Yiannakopoulou UC, SALAWU UO, Chowbina S, Chaitankar V, Ghosh P, Azevedo V**:** [2013]. **In Silico Models: From Simple Networks to Complex Diseases.** In "Animal Biotechnology: Models in Discovery and Translation", ISBN: 9780124160026, Edt by Verma S, Singh A. pp. 385-404; *Elsevier, USA*
5. Chaitankar V, **Barh D**, Zambare V, Azevedo V, Ghosh P: [2013]. **Computational Regulomics: Information Theoretic Approaches Towards Regulatory Network Inference**. In "OMICS: Applications in Biomedical, Agricultural, and Environmental Sciences" ISBN 9781466562813. Edt by **Barh D**, Zambare V, Azevedo V. pp. 225–244; *Taylor & Francis LLC, USA*
6. López-Corrales NL, Miyoshi A, Azevedo V, Stuztman T, **Barh D**: [2011]. **Omics Approaches in Toxicology Research and Biomedical Applications**. In "OMICS: Biomedical Perspectives and Applications". ISBN: 9781439850084. Edt by **Barh D**, et al., pp. 53–76; *CRC Press, USA.*

C.b: Book chapters with other co-authors

7. Jha UC, **Barh D**. [2016]. **Whole genome resequencing: Current status and future prospect in crop improvement**. In "Applied Molecular Biotechnology: The Next Generation of Genetic Engineering". Edt by Khan MS, Khan IA, **Barh D**, ISBN: 9781498714815, First Ed: 2016; pp. 187–211. *Taylor & Francis LLC, USA*
8. Jha UC, Bhat JS, Patil BS, Hossain F, **Barh D**. [2015]. **Functional Genomics: Applications in Plant Science**. in "PlantOmics: The Omics of Plant Science", ISBN: 9788132221722; Edt by **Barh D**, Khan MS, Davies E; pp. 65-111, *Springer, LLC*
9. **Barh D**. Davies E. [2015]. **Plantomics and Futuromics**. in "PlantOmics: The Omics of Plant Science", ISBN: 9788132221722; Edt by **Barh D**, Khan MS, Davies E; pp. 821-825, *Springer, LLC*
10. Bakhtiar SM, Ali A, **Barh D**. [2015]. **Epigenetics in Head and Neck Cancer**. [2015]. In "Cancer Epigenetics: Methods Mol Biol" ISBN: 9781493918034, Edt. by Verma M, pp.751-769; *Humana Press, Springer USA* [PubMed: 25421690]
11. Oznur M, Dede S, **Barh D**, Gunduz M. [2015]. **Cytogenetic Early Markers in Gynecologic Cancers**. In "Noninvasive Molecular Markers in Gynecologic Cancers", ISBN: 9781466569386, Edt. by **Barh D**, Gunduz M, pp. 43-60. *Taylor & Francis, USA*
12. Yilmaz B, Moroski-Erkul CA, Hatipoglu OF, Gunduz E, **Barh D**, Gunduz M. [2015]. **Biomarkers for Early Detection of Familial Breast Cancer**. "Noninvasive Molecular Markers in Gynecologic Cancers", ISBN: 9781466569386, Edt. by **Barh D**, Gunduz M; pp. 168-181. *Taylor & Francis, USA*
13. E Demir, B Atar, D Dhawan, **Barh D**, M Gunduz, E Gunduz. [2014]. **Breast Cancer Stem Cells and Cellomics.** In "Omics Approaches in Breast Cancer", ISBN:9788132208433, Edt. by **Barh D**; pp. 245-263, *Springer, LLC*. DOI: 10.1007/978-81-322-0843-3_12
14. ZN Unal, G Kaya, **Barh D**, E Gunduz, M Gunduz. [2014]. **Omics of Male Breast Cancer**. In "Omics Approaches in Breast Cancer", ISBN:9788132208433, Edt. by **Barh D**; pp. 265-276, *Springer, LLC*. DOI: 10.1007/978-81-322-0843-3_13
15. Verma M, **Barh D**. [2014]. **Breast Cancer Biomarkers for Risk Assessment, Screening, Detection, Diagnosis, and Prognosis**. In "Omics Approaches in Breast Cancer", ISBN:9788132208433, Edt. by **Barh D**; pp. 393-407, *Springer, LLC*. DOI: 10.1007/978-81-322-0843-3_20
16. EC Yiannakopoulou, **Barh D**. [2014]. **Pharmacogenomics–Pharmacoepigenomics of Breast Cancer Therapy: Clinical Implications.** In "Omics Approaches in Breast Cancer", ISBN:9788132208433, Edt. by **Barh D**; pp. 499-518, *Springer, LLC*. DOI: 10.1007/978-81-322-0843-3_25
17. Verma M, **Barh D**, Jain N. [2014]. **Noninvasive Early Markers in Lung Cancer**. In "Cancer Biomarkers: Minimal and Noninvasive Early Diagnosis and Prognosis", Edt by Barh et al. ISBN: 9781466584280, pp. 415-432. *Taylor & Francis LLC, USA*, DOI: 10.1201/b16389-23
18. Bohra A, Jha UC, Singh B, Soren KR, Singh IP, Chaturvedi SK, Nadarajan N, and **Barh D**: [2014]. **Omics Approaches in Pulses**. In "OMICS Applications in Crop Science" ISBN 9781466585256, Edt by **Barh D**. pp. 101–138; *Taylor & Francis LLC, USA* .
19. Blum K, Han D, Giordano J, Lohmann R, Braverman ER, Madigan MA, **Barh D**, Femino J, Hauser M, Downs BW, Simpatico T: [2013]. **Neurogenetics and Nutrigenomics of Reward Deficiency Syndrome (RDS): Stratification of Addiction Risk and Mesolimbic Nutrigenomic Manipulation of Hypodopaminergic Function**. In "Omics for Personalized Medicine" ISBN: 9788132211839 . Edt by **Barh D**, et al. pp. 365-398; *Springer, LLC*
20. **Barh D**, Agte V, Dhawan D, Agte V, Padh H: [2012]. **Cancer Biomarkers for Diagnosis, Prognosis, and Therapy**. In "Molecular and Cellular Therapeutics". ISBN: 9780470748145. Edt by Whitehouse D, Rapley R. pp. 18-68; *John Wiley & Sons Ltd, UK.*
21. **Barh D**, Ahmad S, Bhattacharjee A: [2012]. **In silico and ultra-high throughput screenings (uHTS) in drug discovery: an overview**. In "Pharmaceutical Biotechnology (2nd Ed)". ISBN: 9783527329946. Edt by Oliver Kayser and Herbert Warzecha. pp. 451-490*; Wiley-VCH Verlag GmbH & Co, Germany.*

III. APPENDIX

D. Presentations in Brazil

CERTIFICATE

We hereby certify that the work entitled **BARHL1 is downregulated in Alzheimer's disease and may regulate cognitive functions through ESR1 and multiple pathways** authored by *Debmalya Barh, María E. García-Solano, Neha Jain, Antaripa Bhattacharya, José García-Solano, Daniel Torres-Moreno, Sandeep tiwari, Belén Ferri, Krishna Kant Gupta, Artur Silva, Vasco Azevedo, Preetam Ghosh, Pablo Conesa-Zamora, Kenneth Blum, George Perry*
was presented during the poster session of the "X-Meeting 2015 - 11th International Conference of th AB3C + Brazilian Symposium of Bioinformatics" held in Sao Paulo - Brazil from November 3 to 6 2015.

Glória Regina Franco
Glória R Franco - AB3C President

Mainá Bitar
Mainá Bitar - Poster Chair

Nicole Scherer
Nicole Scherer - Poster Chair

X - meeting + BSB 2015
USP - SÃO PAULO - SP - BRAZIL | NOVEMBER 03 to 06 2015
11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics

CERTIFICATE

We hereby certify that the work entitled **An Integrative in-silico Approach for Therapeutic Target Identification in the Human Pathogen Corynebacterium diphtheria** authored by *Syed Babar Jamal, Syed Shah Hassan, Sandeep Tiwari, Marcus V Viana, Leandro de Jesus Benevides, Asad Ullah , Javed Ali , Adrián G Turjanski , Debmalya Barh, Preetam Gosh, Henrique C P Figueiredo, Artur Silva , Vasco AC Azevedo*
was presented during the poster session of the "X-Meeting 2015 - 11th International Conference of th AB3C + Brazilian Symposium of Bioinformatics" held in Sao Paulo - Brazil from November 3 to 6 2015.

Glória Regina Franco
Glória R Franco - AB3C President

Mainá Bitar
Mainá Bitar - Poster Chair

Nicole Scherer
Nicole Scherer - Poster Chair

X - meeting + BSB 2015
USP - SÃO PAULO - SP - BRAZIL | NOVEMBER 03 to 06 2015
11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics

CERTIFICATE

We hereby certify that the work entitled **The Druggable Pocketome of Corynebacterium diphtheriae as a Tool for Novel Targets Identification** authored by
Syed Shah Hassan, Leandro G Radusky , Syed Babar Jamal , Sandeep Tiwari , Paulo Vinicius Sanches Daltro de Carvalho, Javed Ali , Asad Ullah , Henrique C Figueiredo, Debmalya Barh, Artur Silva, Adrian Gustavo Turjanski , Vasco AC Azevedo
was presented during the poster session of the "X-Meeting 2015 - 11th International Conference of th AB3C + Brazilian Symposium of Bioinformatics" held in Sao Paulo - Brazil from November 3 to 6 2015.

Glória Regina Franco
Glória R Franco - AB3C President

Mainá Bitar
Mainá Bitar - Poster Chair

Nicole Scherer
Nicole Scherer - Poster Chair

- meeting + BSB 2015
USP - SÃO PAULO - SP - BRAZIL | NOVEMBER 03 to 06 2015
11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics

CERTIFICATE

We hereby certify that the work entitled **SIMBA: a web tool for managing bacterial genome assembly** authored by
Diego C. B. Mariano, Felipe L. Pereira, Edgar L. Aguiar, Letícia C. Oliveira, Leandro Benevides, Luís C. Guimarães, Edson L. Folador, Thiago J. Sousa, Preetam Ghosh, Debmalya Barh, Henrique C. P. Figueiredo, Artur Silva, Rommel T. J. Ramos, Vasco A. C. Azevedo
was presented during the poster session of the "X-Meeting 2015 - 11th International Conference of th AB3C + Brazilian Symposium of Bioinformatics" held in Sao Paulo - Brazil from November 3 to 6 2015.

Glória Regina Franco
Glória R Franco - AB3C President

Mainá Bitar
Mainá Bitar - Poster Chair

Nicole Scherer
Nicole Scherer - Poster Chair

- meeting + BSB 2015
USP - SÃO PAULO - SP - BRAZIL | NOVEMBER 03 to 06 2015
11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics

X-meeting 2016

Belo Horizonte | November 16th to 18th

12th International Conference of the AB3C

Certificate of poster presentation

This certifies that the work entitled **In-silico analyses for the discovery of drug and vaccine targets in Corynebacterium camporealensis: A Novel Hierarchical Approach**, authored by **Syed babar Jamal Bacha, sandeep tiwari, Arun Kumar Jaiswal, Daniela Arruda Costa, USUARIO TESTE, Doglas Parise, Henrique CP Figueiredo, Debmalya Barh, Artur Silva and Vasco A de C Azevedo** was presented during the poster session of the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.

Nicole Scherer — Poster Session Chair

Mainá Bitar — Poster Session Co-Chair

Glória Franco — AB3C President

90 ANOS UFMG 1927 - 2017

AB3C

X-meeting 2016

Belo Horizonte | November 16th to 18th

12th International Conference of the AB3C

Certificate of poster presentation

This certifies that the work entitled **Cell cycle and metabolism related candidate human synthetic lethal network**, authored by **sandeep tiwari, Thiago Luiz de Paula Castro, Núbia Seiffert, Debmalya Barh and Vasco A de C Azevedo** was presented during the poster session of the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.

Nicole Scherer — Poster Session Chair

Mainá Bitar — Poster Session Co-Chair

Glória Franco — AB3C President

90 ANOS UFMG 1927 - 2017

AB3C


X-meeting 2016
Belo Horizonte | November 16th to 18th
12th International Conference of the AB3C
This certifies that Dr Debmalya Barh has presented the panel talk entitled "Precision medicine - focus to personalized lifestyle" during the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.
Nicole Scherer
Nicole Scherer
Poster Session Chair
Mainá Bitar
Mainá Bitar
Poster Session Co-Chair
Glória Franco
Glória Franco
AB3C President
UFMG
AB3C