



U F *m* G
UNIVERSIDADE FEDERAL
DE MINAS GERAIS



**INSTITUTE OF BIOLOGICAL SCIENCES
INTERUNIT POST-GRADUATE PROGRAM IN
BIOINFORMATICS**

**“Pan-genomics allied to reverse vaccinology and drug target
Identification for the Syphilis causative agent *Treponema
Pallidum*: a multi-pronged approach towards vaccine and
Drug targets”**

Ph.D. STUDENT: Arun Kumar Jaiswal
SUPERVISOR: Prof. Dr. Siomar de Castro Soares
CO-SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo
CO-SUPERVISOR: Dr. Sandeep Tiwari

**BELO HORIZONTE, MINAS GERAIS, BRAZIL
FEBRUARY-2020
ARUN KUMAR JAISWAL**

**“Pan-genomics allied to reverse vaccinology and drug target
Identification for the Syphilis causative agent *Treponema
Pallidum*: a multi-pronged approach towards vaccine and
Drug targets”**

This thesis is presented as partial requirement
for obtaining a Ph.D. Degree in Bioinformatics
by the Interunit Post-graduate Program in
Bioinformatics at the Institute of Biological
Sciences at Federal University of Minas Gerais.

Ph.D. STUDENT: Arun Kumar Jaiswal

SUPERVISOR: Prof. Dr. Siomar de Castro Soares

CO-SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

CO-SUPERVISOR: Dr. Sandeep Tiwari

**FEDERAL UNIVERSITY OF MINAS GERAIS
INSTITUTE OF BIOLOGICAL SCIENCES
BELO HORIZONTE-MG, BRAZIL**

043

Jaiswal, Arun Kumar.

Pan-genomics allied to reverse vaccinology and drug target Identification for the Syphilis causative agent *Treponema Pallidum*: a multi-pronged approach towards vaccine and Drug targets [manuscrito] / Arun Kumar Jaiswal. – 2020.
57 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Siomar de Castro Soares. Coorientadores: Prof. Dr. Vasco Ariston de Carvalho Azevedo e Dr. Sandeep Tiwari.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Biologia Computacional. 2. *Treponema pallidum*. 3. Sífilis. 4. Genômica. 5. Vacinologia. I. Soares, Siomar de Castro. II. Azevedo, Vasco Ariston de Carvalho. III. Tiwari, Sandeep. IV. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. V. Título.

CDU: 573:004



ATA DA DEFESA DE TESE

Arun Kumar Jaiswal

117/2020

entrada

1º/2016

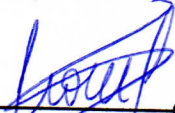
CPF:

702.202.796-08

Às quatorze horas do dia 10 de fevereiro de 2020, 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: "Pan-genomics Allied To Reverse Vaccinology And Drug Target Identification For The Syphilis Causative Agent Treponema Pallidum: A Multi-pronged Approach Towards Better Vaccine And Drug Target", requisito para obtenção do grau de Doutor em Bioinformática. Abrindo a sessão, o Presidente da Comissão, Dr. Siomar de Castro Soares, 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:

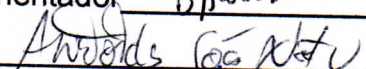
Prof./Pesq.	Instituição	CPF	Indicação
Dr. Siomar de Castro Soares	UFTM	05695182611	APROVADO
Dr. Vasco Ariston de Carvalho Azevedo	UFMG	28314122599	APROVADO
Dr. Sandeep Tiwari	UFMG	02097718604	APROVADO
Dr. Aristóteles Góes Neto	UFMG	54431188520	APROVADO
Dr. José Miguel Ortega	UFMG	059501268-07	APROVADO
Dr. Luis Carlos Guimarães	UFPA	05960854600	APROVADO
Dr. Raghuvir Krishnaswamy Arni	UNESP	138710398-96	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, 10 de fevereiro de 2020.

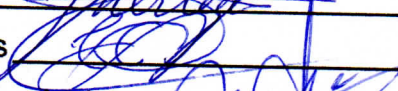
Dr. Siomar de Castro Soares - Orientador 

Dr. Vasco Ariston de Carvalho Azevedo - Coorientador 

Dr. Sandeep Tiwari - Coorientador 

Dr. Aristóteles Góes Neto 

Dr. José Miguel Ortega 

Dr. Luis Carlos Guimarães 


Dr. Raghuvir Krishnaswamy Arni 

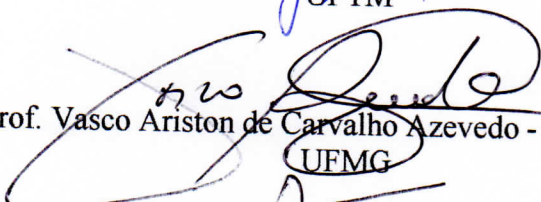


"Pan-genomics Allied To Reverse Vaccinology And Drug Target Identification For The Syphilis Causative Agent Treponema Pallidum: A Multi-pronged Approach Towards Better Vaccine And Drug Target"


Arun Kumar Jaiswal

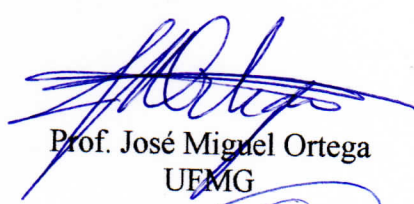
Tese aprovada pela banca examinadora constituída pelos Professores:

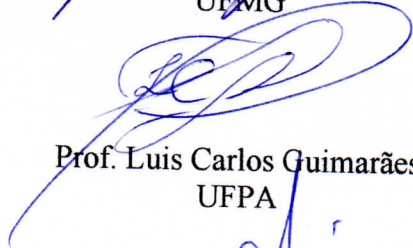

Prof. Siomar de Castro Soares - Orientador
UFTM


Prof. Vasco Ariston de Carvalho Azevedo - Coorientador
UFMG


Prof. Sandeep Tiwari - Coorientador
UFMG

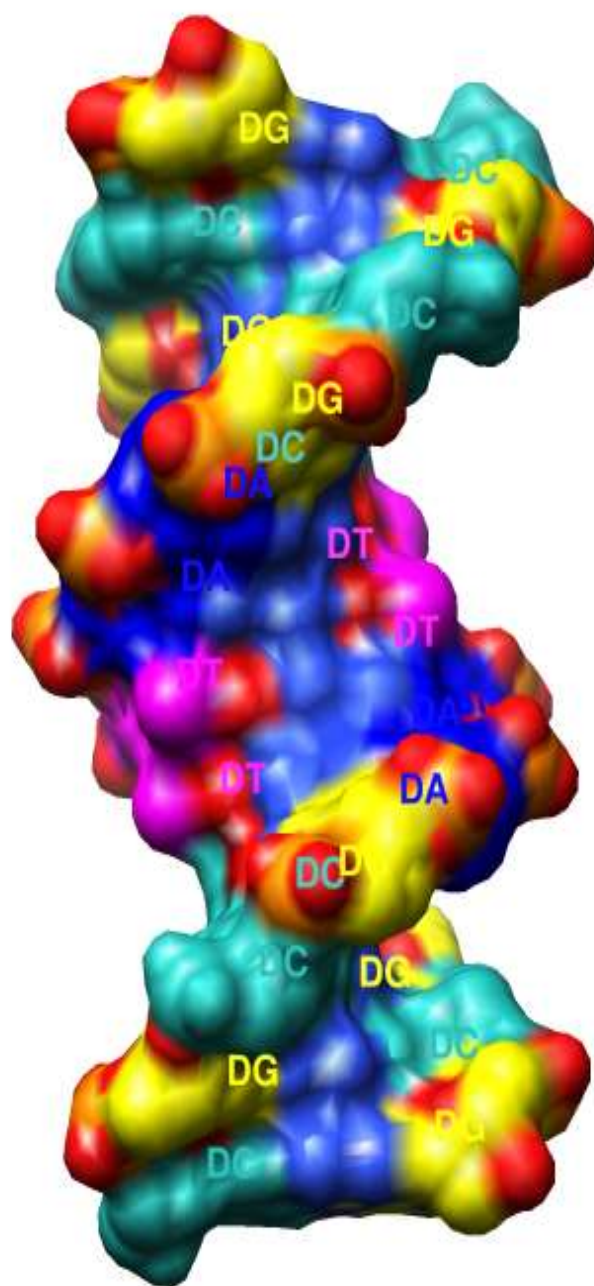
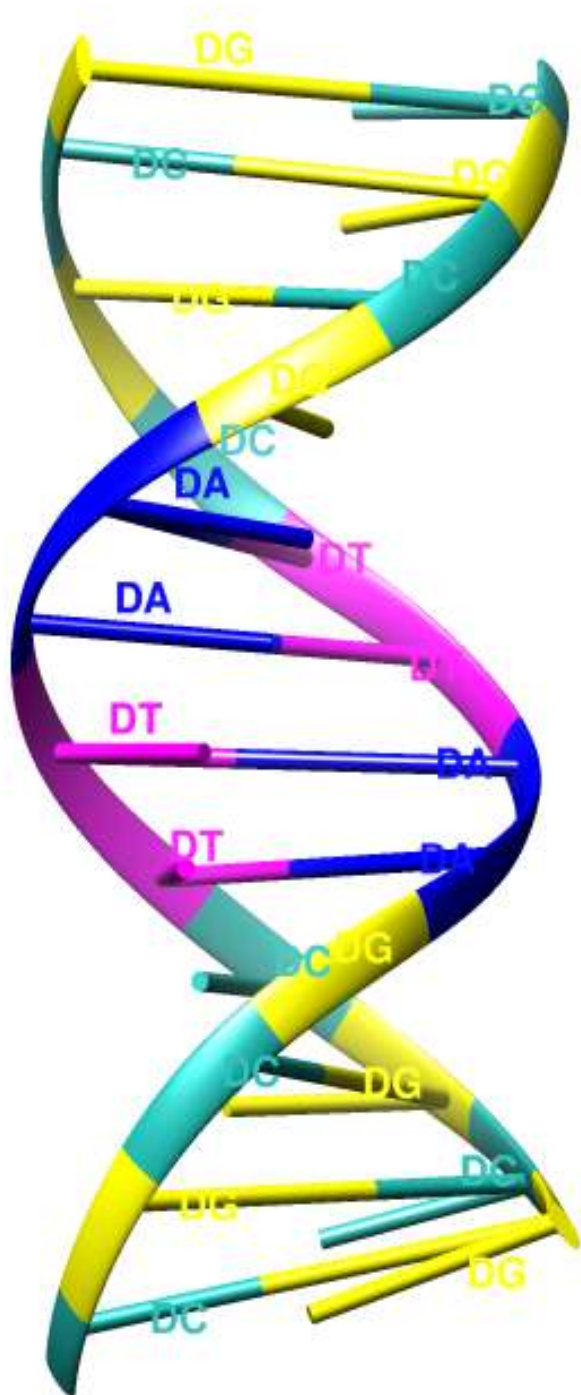

Prof. Aristoteles Góes Neto
UFMG


Prof. José Miguel Ortega
UFMG


Prof. Luis Carlos Guimarães
UFPA


Prof. Raghuvir Krishnaswamy Arni
UNESP

Belo Horizonte, 10 de fevereiro de 2020.



B-DNA

A- Adenine T- Thymine G- Guanine C- Cytosine
 (A T G C)
 AlphAbeTs of Life

!!! जो भी हुआ, अच्छे के लिए हुआ। जो भी हो रहा है, अच्छे के लिए हो रहा है जो भी होगा, वह भी अच्छे के लिए होगा!!!

-श्रीमद् भगवद्गीता

“कर्म करो फल की चिंता मत करो”

भगवद् गीता अध्याय २, श्लोक ४७;
कर्मण्ये वाधिकारस्ते मा फलेषु कदाचन

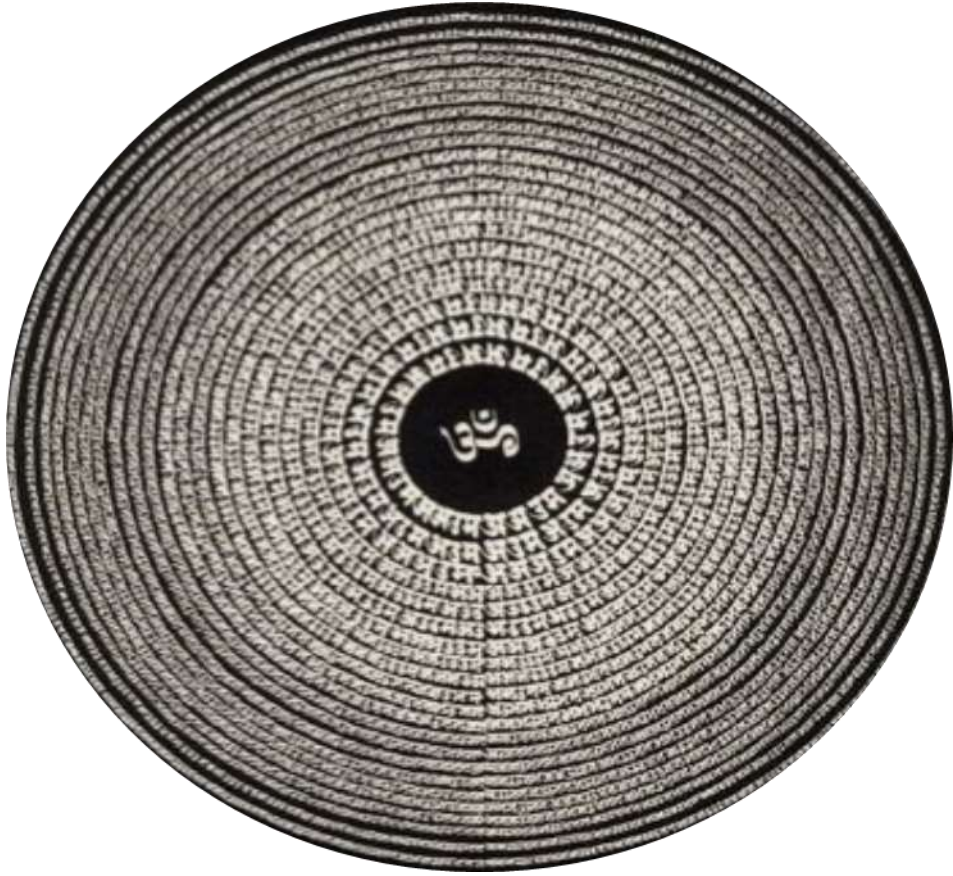
“Whatever happened in the past, it happened for the good; Whatever is happening, is happening for the good; Whatever shall happen in the future, shall happen for the good only.”

-Bhagavad Gita

“You have a right to perform your prescribed duty, but you are not entitled to the fruits of action.”

Bhagavad Gita Chapter 2, Verse 47;
Karmanye Vadhikaraste Ma Phaleshu Kadachana

Thank you, God, for giving me such a beautiful life.



Advaita Vedānta

True self-Non Dualism

Dedication

I would like to dedicate this work; to my beloved family specially my parents and my brothers and my sister, my friends, and every single person who helped me in this work or in any situations which I had been earlier.

Contents

List of Abbreviations	xi
List of Figures	xii
List of tables.....	xiii
ABSTRACT.....	1
RESUMO	3
I. Presentation.....	5
I.A. Collaborations.....	6
I.B. Preface.....	8
I.B.1. History and Origin of Syphilis	8
I.B.2. The Genus <i>Treponema</i> and Subspecies.....	12
I.B.3. Characterization of <i>Treponema pallidum</i> and Syphilis	12
I.B.4. Distribution of Syphilis	15
I.B.5. Symptoms and Stages of Syphilis	15
I.C. Computational Biology and Bioinformatics	15
I.C.1. Pan-genome.....	17
I.C.2. Reverse vaccinology	18
I.C.3. Subtractive Genomics	19
I.C.4. <i>In silico</i> Approaches for Protein Modelling and Drug Discovery	20
I.D. Justification.....	23
I.E. Thesis Delineation.....	25
II. Objectives	26
II.1 General Objective.....	27
II.2 Specific Objective	27
III. Chapters	28
III.1.1. Review Article.....	29
Syphilis: Clinical, Epidemiological and Biological features with future perspectives.	29
III.1.2. Conclusion, Chapter 1.....	30
III.2.1. Book Chapter	31
Pan-omics focused to Crick's Central Dogma	31
III.2.2. Conclusion, Chapter 2.....	32
III.3.1. Research Article	33
The pan-genome of <i>Treponema pallidum</i> reveals differences in genome plasticity between subspecies related to venereal and non-venereal syphilis.	33
III.3.2. Conclusion, Chapter 3.....	34
III.4.1. Research Article	35
An <i>In Silico</i> Identification of Common Putative vaccine Candidates against <i>Treponema pallidum</i> : A Reverse vaccinology and Subtractive Genomics based Approach	35
III.4.2. Conclusion, Chapter 4.....	36

Chapter 5.....	37
III.5.1. General Conclusion.....	37
III.5.2. Future Prospective.....	38
IV. Appendix	39
A. Published, Accepted, Submitted Research articles and Genome assembly, annotation and submission.....	39
B. Book Chapters	48
C. Curriculum Vitae	51
V. Bibliography	52

Acknowledgement

First and foremost, I sincerely thank the almighty God for his graces, strength, sustenance, and above all, His faithfulness and love from the beginning of my academic life up to this doctoral level. His benevolence has made me excel and succeed in all my academic pursuits.

I would like to thank my family: my parents, **Sri Jagdish Prasad Jaiswal** and Smt. **Sela Devi** for giving me birth, unconditional love and support throughout my life. I thank to my brothers **Akash Jaiswal** and **Deepak Jaiswal** and my sisters **Mrs. Neetu Gupta**, **Mrs. Neha Jaiswal**, **Rakhi Jaiswal** and **Chandni Jaiswal** for their eminent support and always believe in me.

I thank to my Uncle **Sri Hari Sharan Prasad Jaiswal** for all the support during graduation.

I thank my brother in Law **Mr. Manoj Gupta** and **Mr. Rajiv Jaiswal** for their all the support, help and encouragement always. It means a lot to me.

I thank my sister in law **Mrs. Rinku Jaiswal** for giving me confidence always. I owe it all to you. Many Thanks!

I would like to express my sincere gratitude to my supervisor **Professor Dr. Siomar de Castro Soares** for the continuous support and guidance during the study and research, for his patience, motivation, enthusiasm and immense knowledge. His guidance helped me all the time throughout the research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my PhD. Study. I have been extremely lucky to have a supervisor who cared so much about me and my work, and who use to respond to me and my questions and queries so promptly. I hold your comments and encouraging words close to my heart, they are more than light to my path. Your encouragement and high degree of freedom to me in the course of this study are highly appreciated.

I would like to express my heartfelt gratitude and admiration to **Professor Dr. Vasco Ariston de Carvalho Azevedo** (UFMG) my co-supervisor for his steady help, guidance, motivation, and concentration.

I thank my Mentor and ex-project supervisor **Mrs. Neha Jain (Indian Institute of Technology-Indore, (IIT-INDORE))** and **Dr. Supriya Ratnaparkhe (Head of Research and Development at Indore Biotech Inputs & Research Pvt Ltd - India)** from India for their guidance and for giving me the confidence to start my scientific journey.

Besides my supervisor and my co-supervisor, I would like to thank my other co-supervisor **Dr. Sandeep Tiwari** and **Dr. Syed Babar Jamal Bacha** (Assistance professor, National University of Medical Sciences, Rawalpindi, Punjab, Pakistan) for their all kind of support, to feel secure, love, and kindness towards me. You people are awesome and I love you.

I thank my fellow lab mates of the **Laboratory of Cellular and Molecular Genetics (UFMG)** and **Laboratório de Imunologia e Bioinformática (UFTM)**, by making exceptional support and guidance in various parts of this work. It was a great sharing laboratory with all of you during the last four years. What a cracking place to work!

I am also grateful to the department and lab staff: **Sheila Santana, Tiago Silva, Natalia Marcia Dutra Ramalho** and **Fernanda Magalhães (UFMG)** and **Monica Miguel Sawan Mendoneca, Betânia Maria Ribeiro (UFTM)** who have always been kind enough in resolving academic and administrative issues.

I thank my friend. **Rodrigo Profeta, Stephane Tosta, Alessandra Lima, Thiago Sousa, Marcela Rezende Lemes, Jonatas da Silva Catarino, Marcus Vinicius Viana** for their various help during this research.

I owe many thanks to my childhood friend **Abhishek Agarwal**, My master's friend **Aseem Kumar Anshu** and my bro **Ramon Macedo** you guys are an example.

Special admiration to **Late Sri Thakur Prasad Jaiswal** for all life lessons.

Last but not least, thanks to all friends in Brazil and India who have always supported and encouraged my decisions and always illuminated me with precious advice. Love you all!!!

I would like to thank the Federal University of Minas Gerais for excellent training; CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior for providing a scholarship for doctoral studies.

Finally, I thank all those who have helped me directly or indirectly in this work. Anyone missed in this acknowledgment is also thanked.

Thanks for all your encouragement!

Arun Kumar Jaiswal

List of Abbreviations

Å	Angstrom = 10^{-10} m
AA	Amino Acids
Bp	Base pairs
BLASTp	Basic Local Alignment Search Tool (protein)
CDS	Coding Sequences
COGs	Database of Clusters of Orthologous Groups of proteins
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
3D	Three-Dimensional
DNA	Deoxyribonucleic Acid
DEG	Database of Essential Genes
GC	Guanine Cytosine
KDa	Kilodaltons (10 ³ Da)
MIGS	Minimum Information About a Genome Sequence
NCBI	National Center for Biotechnology Information
NAS	Non-traceable Author Statement
ORF	Open Reading Frame
PDB	Protein Data Bank
rRNA	Ribosomal Ribonucleic Acid
SIGS	Standards in Genomics Sciences
SDS	Sodium Dodecyl Sulfate
Tp	<i>Treponema pallidum pallidum</i>
TPA	<i>Treponema pallidum</i> subspecies <i>pallidum</i>
TEN	<i>Treponema pallidum</i> subspecies <i>endemicum</i>
TPE	<i>Treponema pallidum</i> subspecies <i>pertenue</i>
TM	Transmembrane (domain)
UniProt	Universal Protein Resource
UFMG	Universidade Federal de Minas Gerais (Federal University of Minas Gerais)
UFTM	Universidade Federal do Triângulo Mineiro (Federal University of Triangulo Mineiro)

List of Figures

Figure 1:- History of Syphilis	11
Figure 2:- Structure of <i>Treponema pallidum</i> (Medcubic, Syphilis: causes, symptoms, diagnosis, treatment, and prevention.)	14
Figure 3:- Application of Bioinformatics and computational biology.	17
Figure 4:- Schematic representation of Reverse Vaccinology.....	19
Figure 5:- The docking Process, with ligand and target.....	22
Figure 6:- <i>In silico</i> drug discovery process.	23

List of tables

Table 1:- Characteristics of pathogenic Treponema species. From Antal <i>et al.</i> , (2002) and Norris <i>et al.</i> , (2006).....	14
---	----

ABSTRACT

Syphilis is caused by the Gram negative spirochete *Treponema pallidum* subspecies *pallidum* and is considered as one of the most imperious and systemic human Sexually Transmitted Infection (STIs). In general, this disease is developed by sexual contact (venereal) especially men (men having sexual relationship with men, MSM) and in some cases a new born may get it from effected mother (congenital syphilis) by transplacental transmission. There are some subspecies of *Treponema pallidum* that are responsible for syphilis without sexual contact (non-venereal), such as, *T. pallidum* subspecies *endemicum* responsible for bejel and *T. pallidum* subspecies. *pertenue* responsible for yaws. These pathogens are quite similar to each other and they cannot be distinguished serologically and morphologically. Every year, the global distribution of syphilis is increasing, reportedly, about six million new cases of syphilis arises around the globe in individuals aged 15 to 49 years. More than 300,000 fetal and newborn deaths are associated to syphilis, with 215,000 further infants placed at high risk of death at early stage of birth. Despite the effort made in developed countries for the elimination of syphilis, the decrease in investment towards public health and STIs control program has over shadowed previous exclusion and control efforts. Furthermore, in developed countries like USA, China and Western Europe, syphilis has been reported in key populations where men are developing sexual relationship with men. In developing or poor countries, syphilis has remained prevalent. The current worldwide prevalence of syphilis combined with the lack of effective vaccine targets and the appearing of antibiotic-resistant strains emphasizes the need for the development of new strategies. In this study, we focused on the comparative genomics studies, Pan-genome, subtractive genomics and reverse vaccinology analysis for the drug and vaccine target identification. In the pan-genomic study, we used 53 strains of *Treponema pallidum* to identify the pan-genome, core genome, and singletons based on subspecies level to reveal the differences in genome plasticity between venereal (Subsp. *pallidum*) and non-venereal (Subsp. *endemicum* and Subsp. *pertenue*) syphilis, which can disclose the close connection among all strains of *Treponema pallidum*. We also used reverse vaccinology, subtractive genomics and molecular docking analysis for the drug and vaccine target identification. In subtractive genomics and reverse vaccinology approaches, we compared 13 strains of *Treponema pallidum* for analysis. We identified 837 core genes and, considering human as host, we identified 567 conserved non-host homologous proteins. Further, using subtractive genomics and reverse vaccinology, 15 putative antigenic proteins, and 6 drug targets were identified which were essential for the bacteria. Identified drug targets were subjected to virtual screening using 28 Natural compounds library. The proposed drug molecules showing favorable interactions, lower energy and high complementarity with predicted targets have also been reported in the present study. Our proposed approach expedites the rapid and efficient selection of *Treponema pallidum* putative proteins for developing a

broad spectrum of novel drugs and vaccines, which can be used as candidate therapeutic targets in the future against syphilis disease.

Key words: *Treponema pallidum*, Syphilis, Pan-genome, Subtractive genomics, Vaccine targets, Drug targets, Reverse vaccinology, Computational and Bioinformatics approaches.

RESUMO

A sífilis é causada pela bactéria *Treponema pallidum* subespécie *pallidum*, uma espiroqueta Gram negativa, e é causada uma das Infecções Sexualmente Transmissíveis (IST) humanas mais prevalentes e sistêmicas. Em geral, esta doença é desenvolvida através do contato sexual (venéreo), especialmente em homens em relações sexuais com outros homens (MSM) e em alguns casos o recém-nascido pode contrair a doença da mãe através da transmissão transplacentária (sífilis congênita). Algumas espécies de *Treponema pallidum* são responsáveis pela sífilis sem contato sexual (não venérea), como a subespécie *endemicum* responsável por causar bejel e a subespécie *pertenue* responsável por causar boubá. Este patógenos são relativamente similares entre eles e eles podem ser distinguidos morfológicamente e sorologicamente. Todos os anos, a distribuição global reportada de sífilis está crescendo em torno de seis milhões de novos casos em indivíduos de 15 a 49 anos em todo o mundo. Mais de 300.000 mortes fetais e de recém-nascidos estão associadas a sífilis, com 215.000 bebês em alto risco de morte nos estágios iniciais do nascimento. Apesar do esforço realizado nos países desenvolvidos para a eliminação da sífilis, a diminuição dos investimentos em saúde pública e programas de controle das ISTs têm sobrepujado os esforços anteriores de controle e exclusão. Além disso, em países desenvolvidos como os Estados Unidos, a China e a Europa Ocidental, a sífilis tem sido reportada em populações específicas nas quais as relações entre homens estão crescendo. Em países em desenvolvimento ou subdesenvolvidos, a sífilis continua prevalente. A prevalência da sífilis combinada com a falta de alvos de vacina efetivos e o aparecimento de linhagens-resistentes a antibióticos enfatizam a necessidade de desenvolvimento de novas estratégias. Neste estudo, nós focamos em estudos de genômica comparativa, pangenoma, genômica subtrativa e vacinologia reversa para a identificação de alvos de drogas e identificação de alvos de vacinas. Nos estudos de pangenoma, nós utilizamos 53 linhagens de *Treponema pallidum* para identificar o genoma pan, genoma central e os genes únicos a nível de subespécie para revelar as diferenças de pastificade genômica entre a sífilis venérea (Subsp. *pallidum*) e não-venérea (Subsp. *endemicum* and Subsp. *pertenue*), o qual pode esclarecer a relação próxima entre as linhagens de *Treponema pallidum*. Nós também utilizamos vacinologia reversa, genômica subtrativa e docking molecular para a identificação de drogas alvos de vacina. Nas abordagens de genômica subtrativa e na vacinologia reversa nós comparamos 13 linhagens de *Treponema pallidum* em cada uma. Nós identificamos 837 genes no genoma central e, considerando o humano como hospedeiro, nós identificamos 567 proteínas conservadas não homólogas ao hospedeiro. Além disso, usando genômica subtrativa e vacinologia reversa, 15 proteínas antigênicas e seis alvos para drogas essenciais para a bactéria foram identificados. Os alvos para drogas identificados foram sujeitos à screening virtual utilizando uma biblioteca de 28 compostos naturais. As moléculas de drogas propostas que mostraram interações favoráveis, baixa energia e alta

complementaridade também foram reportadas no presente estudo. Nossa abordagem provê uma seleção rápida e eficiente de proteínas putativas de *Treponema pallidum* para o desenvolvimento de um amplo espectro de novos alvos e vacinas, as quais podem ser utilizadas como candidatas para alvos terapêuticos contra a sífilis.

Palavras-chave: *Treponema pallidum*, Sífilis, Pangenoma, Genômica subtrativa, Alvos de Vacina, Alvos de droga, Vacinologia reversa, Abordagens computacionais e de bioinformática.

I. Presentation

I.A. Collaborations

This work has been conducted under the supervision of **Prof. Dr. Siomar de Castro Soares**, Laboratório de Imunologia e Bioinformática (UFTM), Universidade Federal do Triângulo Mineiro, Uberaba, Brazil and co-supervision of **Prof. Dr. Vasco Azevedo** and **Dr. Sandeep Tiwari**, Laboratory of Cellular and Molecular Genetics (LGCM), Department of Genetics, Ecology and Evolution, Institute of Biological Sciences (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil and other integral partner researchers/collaborators (national and international) and their respective institutions, among others, are:

- Prof. Dr. Carlo José Freire de Oliveira, Laboratório de Imunologia e Bioinformática (UFTM), Universidade, Universidade Federal do Triângulo Mineiro Uberaba, Brazil.
- Prof. Dr. Virmondes Rodrigues Junior, Laboratório de Imunologia e Bioinformática (UFTM), Universidade, Universidade Federal do Triângulo Mineiro Uberaba, Brazil.
- Prof. Dr. Marcos Vinícius Silva, Laboratório de Imunologia e Bioinformática (UFTM), Universidade, Universidade Federal do Triângulo Mineiro Uberaba, Brazil.
- 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.
- Dr. Debmalya Barh, researcher at the Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, India.
- Dr. Syed Babar Jamal Bacha, Assistant Professor at Department of Biological Sciences, National University of Medical Sciences, Rawalpindi, 46000, Punjab, Pakistan

I have joined the Laboratory of Cellular and Molecular Genetics (LGCM), Department of Bioinformatics in 2016 as full-time Ph.D. student under the supervision of **Prof. Dr. Siomar de Castro Soares** and **Prof. Vasco Ariston de Carvalho Azevedo**.

The research group (LGCM) is actively involved in bioinformatics approaches in genomics of pathogenic bacteria. Being the pioneer in bioinformatics research in Brazil, the group engaged in intensive research projects covering diverse areas of biology like genomics, transcriptomics; and the development of vaccines and diagnostics has made the group a reference point for the study of this microorganism.

So far, the team has successfully sequenced a high number of genomes of *C. pseudotuberculosis* species, isolated from different locations around the world, biovars, and hosts, and deposited to a public database (GenBank). Besides *Corynebacterium* other species like *Campylobacter*, *Leptospira* and *Lactococcus* genome projects were also accomplished.

LGCM has provided the opportunity to participate in various internal and external on-going projects. Starting with previous experience in protein 3D structural modelling and *in silico* structure based drug designing. This lab has strengthened me in computational skills to genome sequencing, assembly and annotation projects and has provided familiarity with the required necessary skills.

The financial support is granted by CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

I.B. Preface

I.B.1. History and Origin of Syphilis

History of Syphilis has been well studied, but the exact origin of the disease is still unknown and it has been the subject of several debates. From the very beginning, syphilis has been taint, disgraceful disease and each country whose population was effected by it, blamed the neighbor's for the outbreak and every country who suffered with syphilis gave different names to this disease. So, the residents of today's Italy, Germany and United Kingdom named syphilis "The French Disease", the French people named it "The Neapolitan Disease", Russian called "Polish" and Polish called it "German Disease". More countries like Danish, Portuguese and African allocate its "Spanish/Castilian Disease" and the Turks assigned the term "Christian Disease". Moreover, in India and some regions of Asia have different belief about the syphilis, it is a religious nuisance. In Northern India, the Muslims blamed the Hindu for the outbreak of the disease and Hindu blames the Muslims, at the end everyone blamed the Europeans (Rothschild, 2005; Tampa et al., 2014). The term "syphilis" was given by "GIROLAMO FRACASTORO", a poet and medical person in Verona. His work "Syphilis sive Morbus Gallicus" in 1530 encompasses three books and present a character named Syphilus. There are two main hypotheses on the origin of syphilis- the pre-Columbian hypothesis and the Columbian hypothesis. The pre-Columbian hypothesis claims, not only syphilis was widely spread in both Old and New World, but also the other treponemal diseases, and in Europe most of these conditions were mistaken for leprosy (de Melo et al., 2010). According to this hypothesis, treponemal disease Pinta occurred in Afro-Asian zone by the year 15,000 BC. Yaws appeared as consequences of the mutation in Pinta around 10,000 BC and spread all over the world except American continent. The endemic syphilis emerged from jaws from the selection of several treponemas, as climate changes around 7000 BC and around 3000 BC the sexually transmitted syphilis emerged from endemic syphilis in South-Western Asia, due to lower temperatures and it spread to Europe and then the rest of the world. Initially, it manifested as a gentle disease but eventually bothered and grew in virulence due to several mutations in 15th century (**Figure 1**) (Forsea et al., 1997; de Melo et al., 2010).

There is one more hypothesis occurs between pre-Columbian and Columbian hypothesis. The Unitarian hypothesis, considered by some authors as variants of pre-Columbian hypothesis. According to this hypothesis, syphilis and non-venereal diseases both are variants of the same infections and the clinical differences happen only because of geographical and climate variations. Briefly, Pinta, Yaws, endemic syphilis and venereal syphilis are considered as an adaptive response of the bacterium *Treponema pallidum* to changes in the environment, cultural differences and contact between various populations (de Melo et al., 2010).

The Columbus hypothesis, according to this hypothesis the navigator of Columbus brought the affliction during their return from the New World in 1493 (Baker and Armelagos, 1988; de Melo et al., 2010). This hypothesis was supported by the documents to “FERNANDEZ de OVIEDO” and “RUY DIAZ de ISLA”, two Spanish origin physicians who were present at the moment when Christopher Columbus returned from America. “RUY DIAZ de ISLA”, the physician acknowledges syphilis as “Unknown disease”, which have not been seen and described and started in Barcelona in 1493, originated Española Island. “RUY DIAZ de ISLA” also states in a manuscript that Pinzon de Palos, the pilot of Columbus, and also other members of the crew already suffered from syphilis on their return from the New World (Morton, 1968; Baker and Armelagos, 1988). In 1494, Columbus returned from his first journey to America after one year, Charles VIII entered in Italy with his army and in 1495 the army of Charles VIII entered in Naples and lead to an alliance made by the Italian princes, including Ludovic Sforaz, who defeated Charles VIII in the battle of Fornovo in July 1495. In this battle, the Italian physicians described for the first time a disease they have seen in French soldier’s bodies consisting of pustules more horrible than leprosy, which could be lethal and was transmitted through the sexual intercourse. The disease proved to be syphilis and the French army was soon blamed for spreading the affliction through Italy and was the first outbreak of syphilis in Italy and after some years of marriage between local females, rapes and prostitution. The disease has spread itself rapidly across the Europe (Baker and Armelagos, 1988; Quételet et al., 1990). People who support the Columbian hypothesis states that the extreme severity of the disease due to its novelty and the population had no time to gain any immunity against the disease and for this reason it became endemic in Europe and strains of *Treponema pallidum* have been selected (Cule, 1992).

From the very beginning, numerous theories on the syphilis origin exist, most of the theories linked syphilis and leprosy together. According to early 16th century, syphilis was the result of a sexual relationship between a Spanish prostitute and leper and the prostitute also infected the soldiers of Charles VIII. Paracelsus (1493-1541) considered that syphilis was the result of sexual intercourse between a prostitute suffering from gonorrhoea and a French leper. Another theory of that time, the disease might be the outcome of the relationship of prostitute having abscess with a leper or the result of poisoning the wine with blood coming from leper (Foa, 1990). However, in 1767 John Hunter, a famous physician of venereal diseases at the time (1728-1793), proved with his experiment that syphilis resulted from gonorrhoea (Qvist, 1977; Forrai, 2011).

In 1831, Ricord designed a larger study on syphilis and gonorrhoea and succeeded to show that it only occurs after the contact with gonorrhoea patient (Forrai, 2011).

In 1905, two scientists Schaudinn (1871-1906) and Hoffman (1868-1959) discovered the etiologic agent of syphilis, whom they named *Spirochaeta pallid*, on various syphilis lesions proving its existence. They also changed the name of the bacteria subsequently to *Treponema pallidum* (Sefton,

2001; Souza, 2005; Forrai, 2011). In 1906, Landsteiner introduced the dark-field microscopy method for the detection of spirochete of syphilis. In 1910 the German bacteriologist August Wasserman (1866-1925) came with the first serologic test for syphilis and in 1949 Nelson and Mayer have conceived *Treponema pallidum* immobilization test (TPI), the first specific test for *T. pallidum*. Their discoveries had a very important role in detecting patient who were suspected of syphilis, as well as in other healthy individuals, and in monitoring syphilis response to treatment (Luger, 1991; Sefton, 2001).

500 years of syphilis



Figure 1:- History of Syphilis (Syphilis: Then and Now, The Scientist (EXPLORING LIFE, INSPIRING INNOVATION))(EXPLORING LIFE, 2014)

I.B.2. The Genus *Treponema* and Subspecies

The genus *Treponema* (phylum Spirochaetes, order Spirochaetes and family Spirochaetaceae) is composed of both Pathogenic and Non-pathogenic bacterial species which are harmful for human and animal and cause serious disease. Microbial species belonging to this genus are Gram-negative, helical, tightly coiled, motile, ranging from 10-20µm in length and 0.1-0.4µm in diameter. The human pathogenic treponemes are very closely related. They are morphologically indistinguishable and they show extensive DNA homology and proteome similarity. The four major human pathogens of this genus are *Treponema pallidum* subsp. *pallidum* responsible for venereal Syphilis (Sexually transmission), *Treponema pallidum* subsp. *pertenue*, *Treponema pallidum* subsp. *endemicum* and *Treponema pallidum* subsp. *carateum* responsible for Non-venereal syphilis (without sexual contact), Yaws, Bejel and Pinta respectively. *T. pertenuae* and *T. endemicum* are classified as subsp. of *T. pallidum* because of their >95% DNA homology and the virtual identity of known sequences from this group of organisms, Also, due to the lack of genetic information about *T. carateum*, it is classified separately ((Norris et al., 2006; Varma et al., 2013).

I.B.3. Characterization of *Treponema pallidum* and Syphilis

The human pathogen treponemes include three antigenically highly related subsp. of *Treponema pallidum* subsp. *pallidum*, (syphilis), subsp. *endemicum* (Non venereal syphilis, Bejel), subsp. *pertenue* (Yaws) and *T. carateum* (Pinta). This classification is based on the distinctive host range, clinical manifestations of each infection and saturation reassociation kinetic study of causative agent of syphilis and yaws. The World Health organization estimates that 12 million new cases arise every year and the total number of cases of yaws, bejel and pinta worldwide is approximately 2.5 million. These treponemes cannot be cultured continuously *in vitro* and may cause lifelong infections in untreated individuals.

Bacterium *Treponema pallidum* consists of three core protein central protoplasmic cylinder surrounded by a cytoplasmic membrane, an overlaying layer of peptidoglycan and an outer membrane (**Figure 2**). *Treponema pallidum* subsp. *pallidum* is obligate parasite and it is very difficult to cultivate this pathogen in culture in laboratory, because unlike most bacteria it cannot survive outside of the mammalian cells (i.e. on petri dish). The survival of bacterium for 1 to 2 hours at 41°C to 42°C is almost impossible, but it can maintain its toxicity, morphology and vigor when preserved at low temperature (-78°C) for many years. It is reproduced by transverse divisions, and it takes 30 to 33 hours' time for generation, thus it is considered as an obligate parasite. It has limited metabolic capacity and cannot synthesize many bio molecules, which is required from a living host. This has posted a hurdle for researchers (Norris et al., 2006; Varma et al., 2013).

Morphologically, *T. pallidum* has double-membrane structure and spirochete is often defined as a Gram-negative bacterium. Though, this analogy is phylogenetically, biochemically, and ultra-structurally inexact. The outer membrane of *T. pallidum* lacks lipopolysaccharide and has a significantly different phospholipid composition compared to outer membranes of classic Gram-negative bacteria. *T. pallidum* are rich in lipoproteins, but these molecules mostly exist in lower surface. Therefore, this rareness of surface-exposed pathogen-associated molecular patterns (PAMPs) helps spirochete to escape triggering host innate surveillance mechanisms, smoothing the process of local replication and early propagation. Its surface antigenicity limitation encourages avoidance of adaptive immune responses, enabling persistence (Peeling et al., 2017). Although availability of sensitive diagnostic tests and effective treatment, it is remaining a serious health problem around the globe. It has two routes of transmission: (1) sexual transmission, which accounts vast majority of cases and (2) vertical transmission from mother to newborn as congenital syphilis.

Table 1:- Characteristics of pathogenic *Treponema* species. From (Norris et al., 2006).

Organism	Disease	Distribution	Predominant age of onset	Transmission	Congenital Infection
<i>T. pallidum</i> subsp. <i>pallidum</i>	Venereal syphilis	Worldwide	Adolescents and adult	Sexual contact	Yes
<i>T. pallidum</i> subsp. <i>pertenue</i>	Yaws	Tropical areas, Africa, South America, Caribbean, Indonesia	Children	Skin contact	No
<i>T. pallidum</i> subsp. <i>endemicum</i>	Bejel (endemic syphilis)	Dry areas, Africa, Middle East,	Children, Adolescents and adult	Mucous membrane,	Rarely
<i>T. pallidum</i> subsp. <i>carateum</i>	Pinta	Semi-dry, warm areas, Central and South America	Children and Adolescents	Skin contact	No

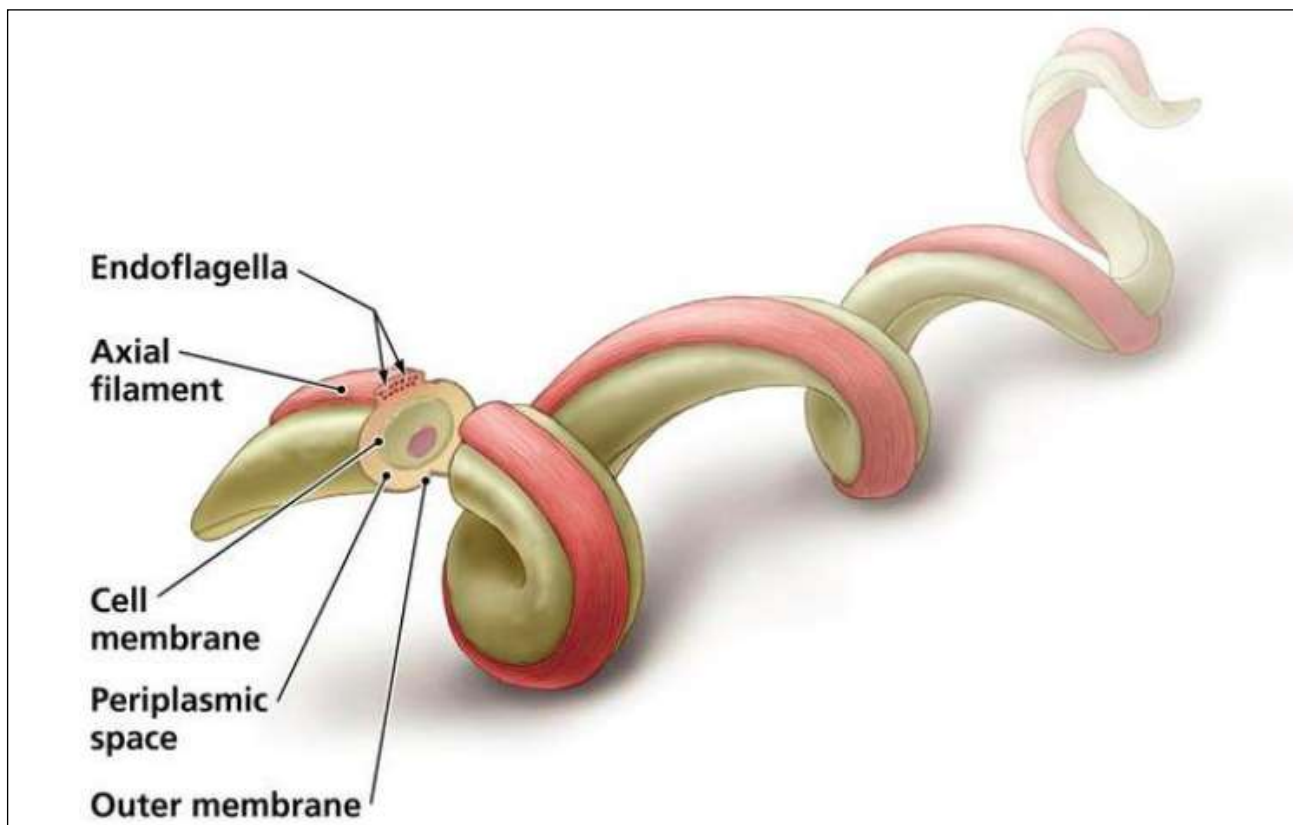


Figure 2:- Structure of *Treponema pallidum* (Syphilis: causes, 2019)

I.B.4. Distribution of Syphilis

The global burden of sexually transmitted infections (STIs) remains high. In 2012, there were an estimated 357.4 million new infections (almost 1 million every day) of the curable STIs- Chlamydia, Gonorrhoea, Trichomoniasis and Syphilis. However, Syphilis data is more robust, every year, the global distribution of syphilis is increasing, reportedly, about six million new cases of syphilis arises around the globe in individuals aged 15 to 49 years. More than 300,000 fetal and newborn deaths are associated to syphilis, with 215,000 further infants placed at high risk of death at early stage of birth (Kojima and Klausner, 2018).

I.B.5. Symptoms and Stages of Syphilis

Initially, microorganism inoculation occurs via visible or microscopic abrasions of the skin or mucous membrane of the mouth and genitals, which can result from the sexual contact. From there, bacteria produce a non-painful ulcer known as chancre. The average incubation period of syphilis (i.e. time from exposure to the development of primary syphilis) is 3 weeks but it can be as long as months and as short as 9-10 days. There are three stages of syphilis - Primary Syphilis, Secondary Syphilis and Tertiary Syphilis (Cherneskie, 2006).

I.C. Computational Biology and Bioinformatics

Bioinformatics is a newly developed interdisciplinary field tying the knot of different sciences including chemistry, biology, mathematics and computer science that develops and improves upon methods for storing, retrieving, organizing and analysing biological data (**Figure 3**). The main goal of bioinformatics is to increase the knowledge of biological processes and entail the analysis and interpretation of various biological data such as nucleotide and amino acids sequences, protein domains and protein structures (Akalin, 2006).

Next generation sequencing (NGS) has made significantly great treads in sequencing technology and integration of different bioinformatics approaches has enable the identification of genes in a high throughput manner in low budget. A number of NGS platforms such as Roche, Illumina, ABI/SOLiD are utilized for wet-lab analysis of NGS data. One of the important implementation of NGS is its usage in diagnosis of disease at early stages. There are many applications of NGS technologies such as DNA-sequencing and assembly, determining unknown genome without search for variations between genome samples (Wadapurkar and Vyas, 2018). The knowledge so obtained from analysis and interpretation of biological data is used to developed new techniques to predict structure and function of protein and to go in depth of molecular processes to gain the knowledge of mechanism of many biological pathways. Comparative genomics is one such approach to compare genomic contents of different genomes. It is the direct comparison of complete genetic material

among different genomes to better understand how species evolved. Genetic contents like gene number, gene location, length of gene, coding regions within genes, the amount of non-coding DNA in the genomes, and conserved regions preserved by both prokaryotic and eukaryotic groups of organisms (Sivashankari and Shanmughavel, 2007).

Well annotated genomes are essential for genome based drug/vaccine target identification against pathogens. There are three main steps involved in computational identification of drug/vaccine targets. In the first step, essential genes for pathogen are identified. Secondly, their host homology is verified. Only non-host homologous genes/proteins are selected for prioritization. (Hofer et al., 2001; Xia, 2017). Third is to reduce the chances of pathogen developing resistance against the drug. It is important to target only pathogenic species but not its phylogenetic relatives that are not pathogenic. To avoid this factor, pathogenic islands (PIs) are identified for pathogenic bacteria (Gal-Mor and Finlay, 2006). Integrative Bioinformatics and computational biology approaches had unveiled drug targets against different human pathogenic bacteria as *Clostridium perfringens* (Bhatia et al., 2014), *Pseudomonas aeruginosa* (Fernandez-Pinar et al., 2015), *Giardia intestinalis* (Cox et al., 2006) similarly, in developing anti-HIV-1 drugs (Xia, 2017).

Bioinformatics and computational biology approaches can be utilized for solving biological problems and understanding of the molecular basis of biological phenomena. The rapid advances in bioinformatics help us in better understanding of drug and target interactions and thus this interaction plays a main role in introducing a computer technology. The idea of cost and time reduction to produce drugs is fulfilled by combinatorial chemistry. A big number of molecules are screened in a systematic manner by using the tools provided by combinatorial chemistry. These computational tools play an important role in the designing and the invention of clinically important chemical entities. Computer software used molecular techniques to design the initiative steps to increase the molecular diversity to virtually synthesize chemical libraries (Altman and Klein, 2007). These concepts are significantly contributing towards the development of research in life sciences, from the level of data generation to its interpretation (Xavier et al., 2008).

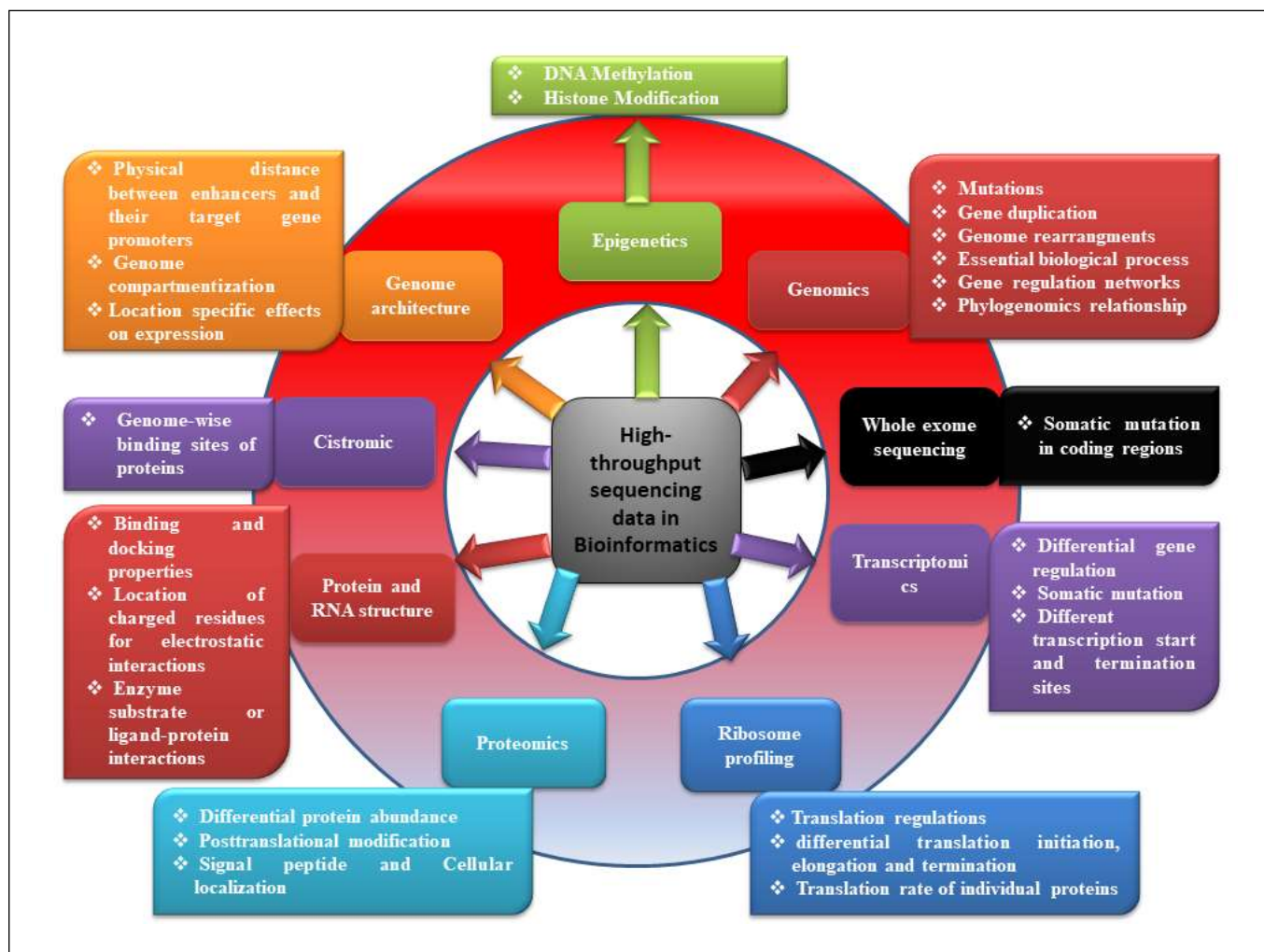


Figure 3:- Application of Bioinformatics and computational biology.

I.C.1. Pan-genome

Soon after the introduction of genomic era, the question of addressing genomic contents of bacterial species was still there. Experimental works were producing new data with novel genes in bacterial species; hence computational approaches were implemented to gather fruitful information out of these newly generated data. One such approach is pan-genomics, which is comprised of complete genomic information regarding bacterial species. Pan-genome is composed of “Core Genome” containing all the commonly present genes in different strains of bacterial species, “shared genome” containing genes from two or more than two strains and “singletons” containing strain specific genes. (Medini et al., 2005).

Tettelin and his colleagues (2005) described the term pan-genome for the first time working with *Streptococcus agalactiae* (Tettelin et al., 2005). Later, other studies were conducted using pan-genomic analysis for different microorganisms, that includes *Corynebacterium pseudotuberculosis* (Soares et al., 2013), *Bacillus cereus* (Rasko et al., 2005), *Streptococcus pneumoniae* (Donati et al., 2010), *Escherichia coli* (Rasko et al., 2008), *Pantoea ananatis* (De Maayer et al., 2014) and

Methanobrevibacter smithii (Hansen et al., 2011). The pan-genomic studies provide vital information about the evolution of bacteria, niche adaptation, host interaction and population structure as well as upshots in more applied issues like vaccine and drug design and the identification of virulent genes (Hansen et al., 2011).

I.C.2. Reverse vaccinology

Reverse vaccinology is the process of antigen discovery that starts with genome information. It combines immunological and genomic information of the pathogen to identify appropriate protein antigens for vaccine purposes (**Figure 4**). In this regard, the identification of the epitopes recognized by CD4+ T cell or CD8+ T cells can be “reversely” used as tool for identification of new antigens (Sette and Rappuoli, 2010). This methodology is evolving with time and accepted as a successful approach for vaccine discovery nowadays, as it can be explored further for the development of vaccines against many types of pathogens, such as, *Streptococcus pneumoniae* (Wizemann et al., 2001), *Porphyromonas gingivalis* (Ross et al., 2001), *Chlamydia pneumoniae* (Montigiani et al., 2002), *Bacillus anthracis* (Ariel et al., 2002), *Staphylococcus aureus* (Etz et al., 2002; Bagnoli et al., 2015).

The first pathogen that was explored for reverse vaccinology was *Neisseria meningitidis* serogroup B (MenB), a Gram-negative bacterium that is responsible for 50% of meningococcal meningitis cases globally (Kelly and Rappuoli, 2005; Stephens et al., 2007). Rino Rappuoli was the first to use this term (Rappuoli, 2001). Reverse vaccinology deals with the rapid and comprehensive assessment of surface protein range of microorganism and has advantages over classical approaches for the identification of candidate antigens. Though RV has advantage over classical approaches, expansion of conventional areas remained significant to support the process, eventually generating vaccines from genome-derived antigens (Kelly and Rappuoli, 2005).

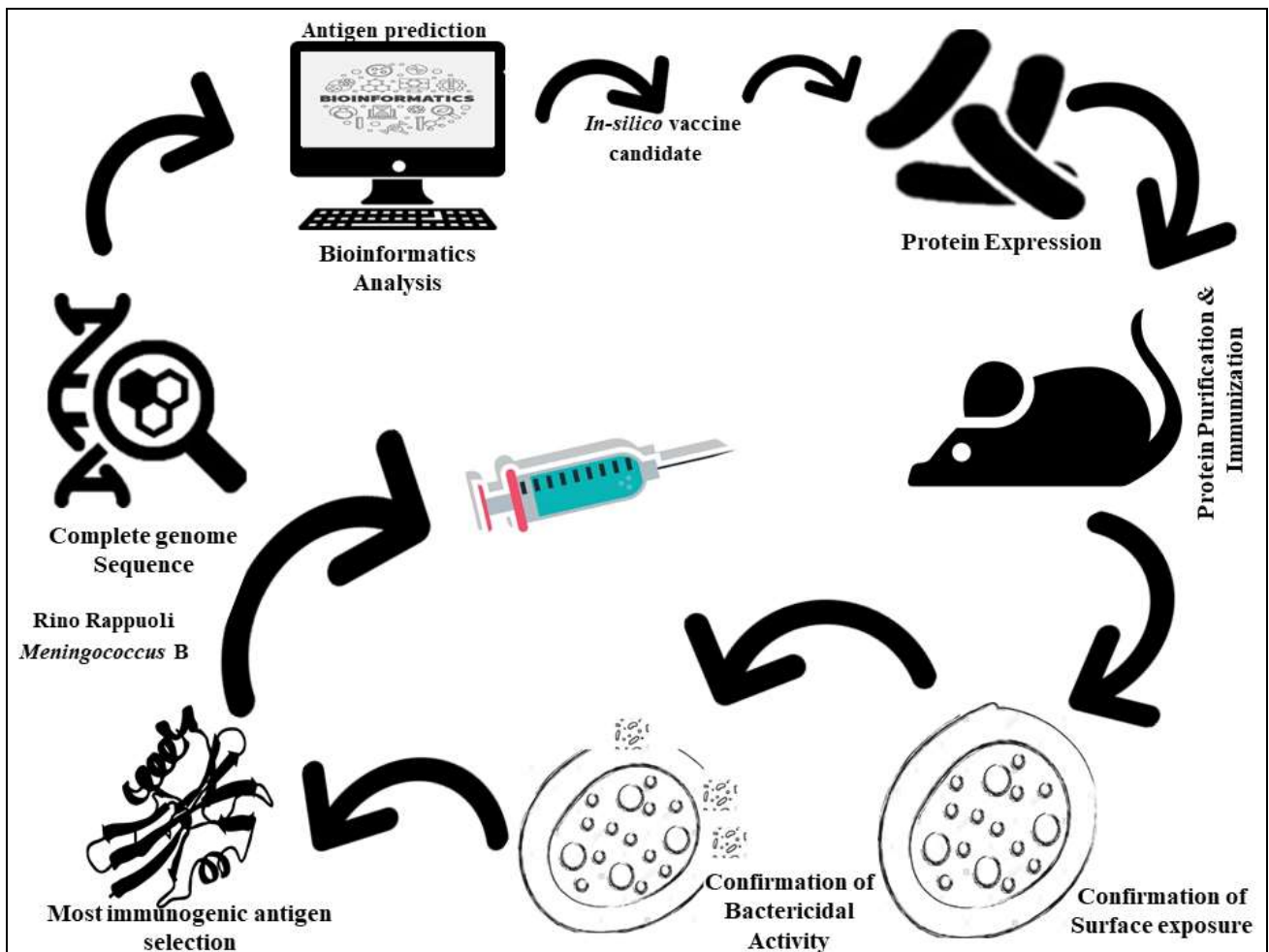


Figure 4:- Schematic representation of Reverse Vaccinology.

I.C.3. Subtractive Genomics

Subtractive genomics is currently a one of the widely used strategy throughout last several years from researchers and many scientific groups for target prediction.

Subtractive genomics is an approach which is applied to detect a novel drug targets in pathogenic organism using the whole genome. This methodology involves the subtraction of sequences between host and the pathogen proteome/genome (proteins/genes) with help implementation of certain rule. This helps in providing information for a set of proteins which are essential to pathogen but are not present in the host.

According to Barh *et al.*, 2011 (Barh et al., 2011), an ideal target should fulfill these properties- **a)** It must be an essential for survival or pathogenesis of the target organism and belongs to the core gene of the pathogen **b)** The target should belong to the pathogen's unique pathway. The pathway related targets are more advantageous and will be best if that it is involved in multiple pathways. The approach of subtractive genomics for target identification has been applied in many pathogens,

Haemophilus ducreyi (de Sarom et al., 2018), *Corynebacterium diphtheria* (Jamal et al., 2017), *Mycobacterium tuberculosis* (Hosen et al., 2014) and others.

I.C.4. *In silico* Approaches for Protein Modelling and Drug Discovery

With the advancements in the post-genomic era, new computational approaches are practiced to handle new amino acid and nucleotide sequences in a faster and efficient manner (Gupta et al., 2014). As of November 13th 2019, **UniProtKB** database reports 561,356 protein sequence entries for SwissProt (statistics, 2019). Algorithms for protein structure and function prediction provides valuable information to addresses many problems of biologists about their proteins of interest (UniProt, 2009; Roy et al., 2010).

It is a fundamental concept of biology that “Sequence implies the Structure and Structure implies the Function”, but with the increase in the number of sequences, it does not replicate any biological implication until unless proteins structures are identified (Gupta et al., 2014). *In silico* protein, modelling is helpful in predicting the 3D protein structures. There are three broadly categorized methods for predicting the 3D structures of protein as; a) **Homology modeling** b) **Threading or Fold recognition** and c) *ab initio* methods (Gupta et al., 2014; Nikolaev et al., 2018). In homology modeling, with the help of computer as experimental tool, the 3D structure of unknown protein is predicted from its amino acids sequence from a protein with known structure as a template. The basis of this is to build suitable models that will presumably closely resemble the unknown protein and to evaluate the quality of the models and choose the best one amongst the built models. Ancestral relationship assumes that proteins from same family share some motifs even if they don't share the same sequence. Protein structure is more conserved than the sequence and hence obvious similarity in sequence show structural similarity. The model's quality is determined by aligning the template and target protein. The decrease in similarity of target and template decrease the quality of model. The hurdle in model building is only the gap present in template structure due to poor NMR experiment. There is a limit in similarity between target and template protein below which the homologous does not produce a suitable model. To formulate the hypothesis about the structure biochemistry of protein, homology modeling is used followed by experimental results to prove the hypothesis. Critical Assessment of Techniques for Protein Structure Prediction [CASP] is used to check the accuracy of protein. Homology model is built on the basis of protein with known 3D structure [template] present in Protein Data Bank [PDB], thus PDB is the basis for homology modeling (John and Sali, 2003; Khan et al., 2016). Threading methods are practiced when sequence identity between target and template is below 30%. In that case, it is difficult to identify the best template for sequence-template alignments and modeling 3D structure (Bowie et al., 1991; Jones et

al., 1992; Peng and Xu, 2010; Yang and Zhang, 2015). Hence in threading method, the query sequence is aligned directly to the 3D structures of other solved proteins. The ultimate aim of this approach is to identify folds that are similar to the query sequence even if there is no evolutionary relationship between target and template protein. The *ab initio* methods solve the cases in which template in PDB library are not available to solve the structure (Simons et al., 1997; Liwo et al., 1999; Wu et al., 2007; Yang and Zhang, 2015). It is a useful technique for short sequences (< 120 amino acids) and considered as one of the toughest methods of solving protein structure (Jauch et al., 2007; Zhang, 2008).

Usually, these model structures are useful for predicting protein and ligand interaction (Docking). Molecular docking explores the interaction pattern of small drug like molecule with protein (**Figure 5**). Computationally, millions of compounds are screened (Virtual Screening, VS) against target protein to find potent ligand molecules among them. Structure-based virtual screening or Structural-based drug designing (SBVS or SBDD) is used to computationally predict how ligands-generally drug-like small molecules- will interact with the binding site of a target receptor, usually a protein structure. Molecular docking and SBVS are extensively used in drug-discovery projects because of their success in identification of hit compounds and lead optimization (**Figure 6**). Thus, reducing time and cost for identification of new drug (Klebe, 2006; Ferreira et al., 2010; Ripphausen et al., 2010; Sousa et al., 2010; Liao et al., 2013).

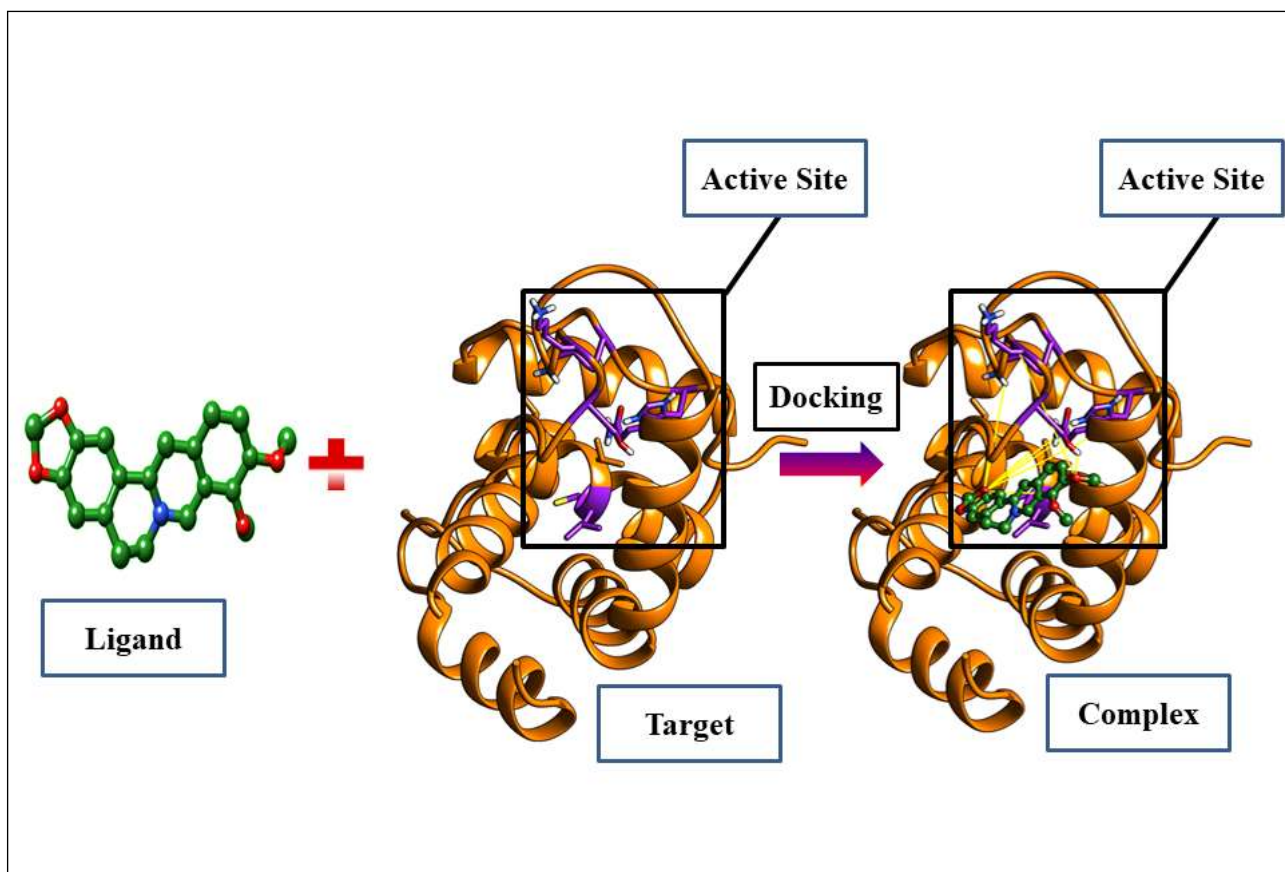


Figure 5:- The docking Process, with ligand and target.

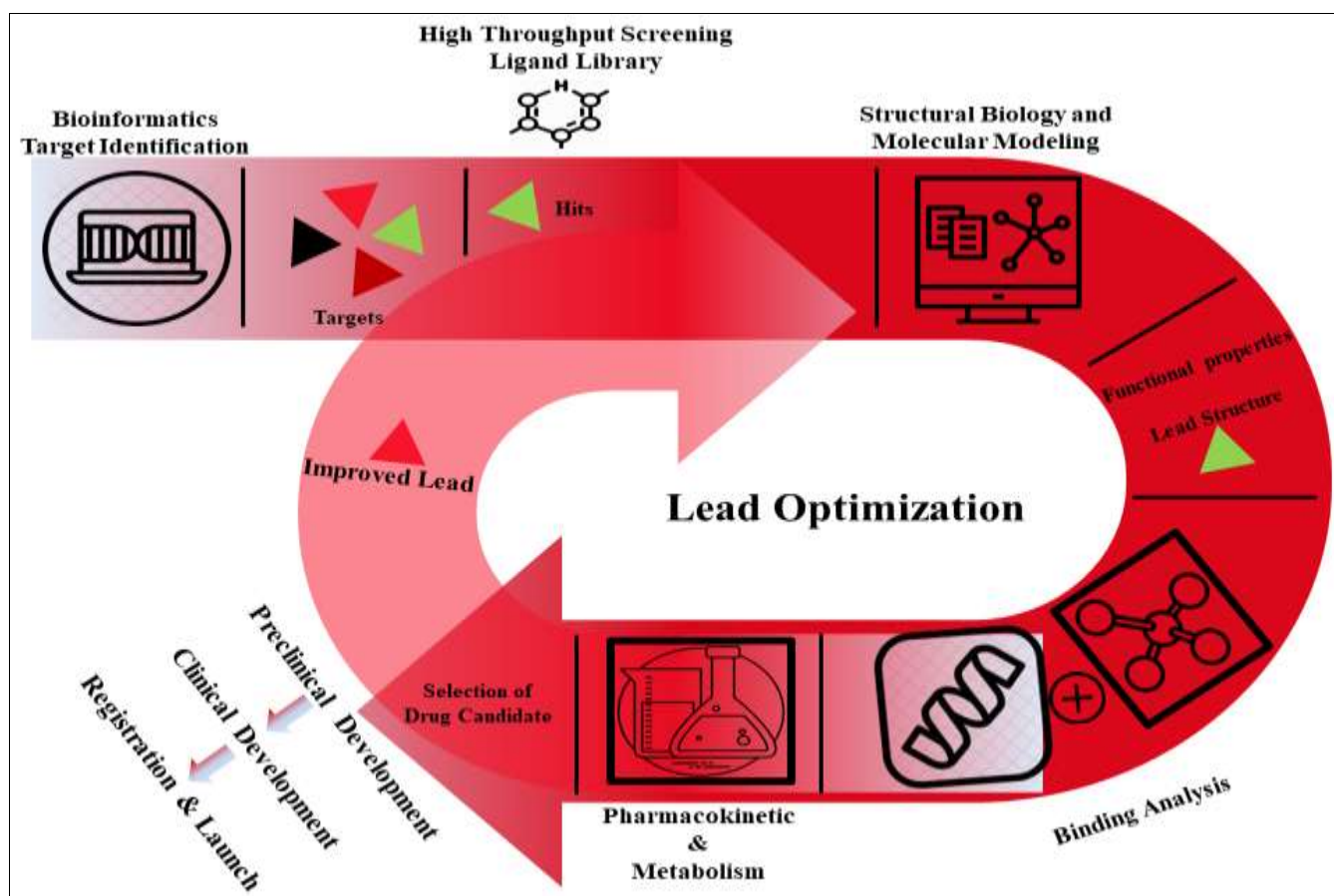


Figure 6:- *In silico* drug discovery process.

I.D. Justification

As many as 50 million people worldwide are being treated for syphilis and about 12 million new cases are diagnosed every year. However, approximately 90% of those infected do not know about the disease, it is the driving force behind the worldwide epidemic. The U.S. Centres for Disease Control (CDC) often refers to syphilis as the “great imitator”, because many of its symptoms are similar to other diseases (Rapid, 2013). According to scientist an accurate and simple approach to the diagnosis of syphilis still mysterious and diagnosis continues to require a comprehensive assessment of the patient, including risk exposure, the presence of compatible clinical symptoms and signs and laboratory tests (Singh and Romanowski, 1999). Because there is no vaccine against the disease for the prevention of syphilis, timely diagnosis and treatment of infected individuals and their sexual partners are keys to control of syphilis and also sex education and promotion uses of condom while having intercourse to prevent infection. A major breakthrough for syphilis treatment occurred in 1943 when Mahoney and colleagues reported the use of Penicillin to successfully cure patients with primary syphilis (Mahoney et al., 1943). Unlike most of the other bacterial microorganisms that rapidly developed resistance to penicillin, *Treponema pallidum* has remained

delicately sensitive to the antibiotic. The U.S. CDC Sexually Transmitted Diseases (STD) Treatment Guidelines, 2015 recommend Benzathine penicillin G (BPG) as the first-line drug for incubating syphilis and for all stages of syphilis (Workowski, 2015).

Oral Azithromycin was shown to be effective for the treatment of early syphilis in the U.S, Africa, China and Madagascar that compared cure rates for this macrolide and penicillin. Azithromycin was used for the treatment of syphilis in Uganda (mid-1990s), in the U.S. (1999 and 2000) and in Canada (2000). However, during 2002-2003 clinical failure was observed for the Azithromycin treatment in San Francisco, CA (Katz and Klausner, 2008).

Syphilis continues to present challenges to the global health, because it increases the risk of transmitting infection with HIV. Syphilis enables the transmission and acquisition of HIV and it has a negative influence on HIV infection, which helps in increasing viral load and decrease CD4 cell counts during infection. In addition, HIV also has impact on syphilis clinical course; patients with HIV are at high risk of neurological problems and treatment failure (Zetola and Klausner, 2007; Salado-Rasmussen, 2015). Nevertheless, serodiagnostic tests, though not perfect are expensive. Early, uncomplicated or first stages of syphilis are treatable with BPG Injection but unfortunately *T. pallidum* also developed resistance against this antibiotic (Stamm, 2010). Finally, 2-6 weeks incubation period of syphilis provides an opportunity to interrupt transmission via prophylactic treatment of sexual contacts. Despite these feature, syphilis has resurged in several developed countries including China, which has recently experienced high burdens of several STIs including syphilis (Stamm, 2016).

I.E. Thesis Delineation

The thesis delineation is articles based and is divided into Objectives, Five chapters, Appendix and Bibliography.

- The Objectives section provides the general and specific aims of this research.
- The chapter first is review article which broadly describes the disease syphilis its pathogenicity, available treatments with future perspectives.
- In chapter second the book chapter that contain literature review about the concept of pan-genomics which is being perfectly applied in the studies of several organisms, diseases, and various research areas.
- In chapter three, the research article describes Pan-genome study of *Treponema pallidum* to identify Genome plasticity between the subspecies level.
- In chapter fourth, the research article describes the *in silico* Reverse vaccinology and Subtractive genomic approach for target identification in *Treponema pallidum*. In this part, research article shows that the identified drug and vaccine candidate could be good targets against syphilis disease.
- The chapter five is representing general conclusions which describe the summary and main outcomes of the entire work represented in this thesis followed by future prospective.
- In Appendix, section A have, published and submitted research articles are listed and Genome assembly, annotation and submission, section B have, submitted book chapters and section C have, Curriculum vitae (CV), National and International Workshops and Courses attended, along with Presented Posters during doctoral study.
- Bibliography has references which are used in preface and Justification.

II. Objectives

II.1 General Objective

The main goal of this study is to find differences between venereal and non-venereal syphilis causing agent *Treponema pallidum* Subsp. *pallidum* (TPA), *Treponema pallidum* Subsp. *pertenue* (TPE) and *Treponema pallidum* Subsp. *endemicum* (TEN) based on the presence or absence of genome plasticity, and also the identification of broad-spectrum potential Drug and vaccine candidates against *Treponema pallidum* through comparative genomics like pan-genomic, subtractive genomics and reverse vaccinology based approach.

II.2 Specific Objective

- Pan-genome analysis for the better understanding of relationship between venereal and non-venereal syphilis of *Treponema pallidum* the subsp. *pallidum*, *pertenue* and *endemicum*.
- Comparative genomics-based approach and identification of core genome, shared genome and singleton genome dataset based on subspecies level.
- Core genome and conserved non-host homologous analysis of all strains of *Treponema pallidum*.
- Reverse vaccinology and Pathogenicity analysis of identified conserved non-host homologous proteomes as vaccine.
- Molecular Docking analysis of identified targets with ligand libraries.

III. Chapters

Chapter 1.

III.1.1. Review Article

Syphilis: Clinical, Epidemiological and Biological features with future perspectives.

Arun Kumar Jaiswal, Sandeep Tiwari, Syed Babar Jamal, Stephane Fraga de Oliveira Tosta, Rodrigo Profeta, Preetam Ghosh, Vasco Azevedo, **Siomar C. Soares**.

Submitted: PeerJ, December 2019

In this review article, we have provided broad information regarding the re-emergence and pathogenesis of *Treponema pallidum* and its interaction with host. We have also covered the global epidemiology of syphilis with its stages, symptoms, persistence and available treatments. The exact mode of infection for this pathogen is still unknown. Hence, more studies are necessary to understand the exact mechanisms of infection to combat against syphilis.

Important declarations

Please remove this info from manuscript text if it is also present there.

Associated Data

Data not supplied by the author for this reason:

This is a Review article. All the required methods and results to reproduce this article are provided as part of this Review.

Required Statements

Competing Interest statement:

The authors declare that they have no competing interests.

Funding statement:

The authors received no funding for this work.

Syphilis: Clinical, Epidemiological and Biological features with future perspectives

Arun Kumar Jaiswal^{1,2}, **Sandeep Tiwari**^{Corresp., 3}, **Syed Babar Jamal**⁴, **Stephane FO Tosta**¹, **Rodrigo Profeta**¹, **Preetam Ghosh**⁵, **Debmalya Barh**⁶, **Vasco Azevedo**¹, **Siomar C. Soares**²

¹ PG Program in Bioinformatics, Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

² Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil

³ Postgraduate Program in Bioinformatics, Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

⁴ Department of Biological Sciences, National University of Medical Sciences, Rawalpindi, Punjab, Pakistan

⁵ Department of Computer Science, Virginia Commonwealth University, Richmond, VA, United States

⁶ Center for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Medinipur, West Bengal, India

Corresponding Author: Sandeep Tiwari

Email address: sandip_sbtbi@yahoo.com

Sexually transmitted infections (STIs) are among the most widely recognized health concerns. The worldwide burden of STIs keeps on being high, as indicated by the World Health Organization (WHO). In 2016, there were an expected 376 million new cases, which resulted in high illness and mortality. Syphilis is an ulcerative, systemic human STI caused by Gram negative spirochete bacteria known as *Treponema pallidum* subspecies *pallidum*. Syphilis disease can likewise be transmitted through sexual contact with infectious sores of the mucous layers or scraped skin, by means of blood transfusion, or vertically transmitted during pregnancy in women to her fetus (Congenital Syphilis). The bacterium *Treponema pallidum* can stay latent for long time periods in which a person lives with no visible sign or symptoms, but the infection remains. In spite of the availability of simple and easy diagnostic tests and an effective single dose of the antibiotic penicillin, syphilis bounced back to the worldwide scene as a reemerging general health concern, especially among men who engage in sexual relations with men (MSM) in developed countries and was also responsible for several hundred thousand still-birth and neonatal death in developing countries. Although many countries are working on eliminating congenital syphilis, there is a rising proliferation of syphilis in HIV infected MSM. Considering the role of hazardous sexual conduct in driving syphilis transmission, there is a requirement for extra lucidity with regards to how best to help and support solid sexual practices among populaces in danger of syphilis. In this review, we portray current discoveries along with bioinformatics approaches that have improved our comprehension of the biological and hereditary structure of the pathogen, novel analytic tests, test techniques and therapeutics that can

improve sickness recognition and treatments, proof-based administration recommendations.

30 Sexually transmitted infections (STIs) are among the most widely recognized health concerns.
31 The worldwide burden of STIs keeps on being high, as indicated by the World Health
32 Organization (WHO). In 2016, there were an expected 376 million new cases, which resulted in
33 high illness and mortality. Syphilis is an ulcerative, systemic human STI caused by Gram
34 negative spirochete bacteria known as *Treponema pallidum* subspecies *pallidum*. Syphilis disease
35 can likewise be transmitted through sexual contact with infectious sores of the mucous layers or
36 scraped skin, by means of blood transfusion, or vertically transmitted during pregnancy in women
37 to her fetus (Congenital Syphilis). The bacterium *Treponema pallidum* can stay latent for long
38 time periods in which a person lives with no visible sign or symptoms, but the infection remains.
39 In spite of the availability of simple and easy diagnostic tests and an effective single dose of the
40 antibiotic penicillin, syphilis bounced back to the worldwide scene as a reemerging general health
41 concern, especially among men who engage in sexual relations with men (MSM) in developed
42 countries and was also responsible for several hundred thousand still-birth and neonatal death in
43 developing countries. Although many countries are working on eliminating congenital syphilis,
44 there is a rising proliferation of syphilis in HIV infected MSM. Considering the role of hazardous
45 sexual conduct in driving syphilis transmission, there is a requirement for extra lucidity with
46 regards to how best to help and support solid sexual practices among populaces in danger of
47 syphilis. In this review, we portray current discoveries along with bioinformatics approaches that
48 have improved our comprehension of the biological and hereditary structure of the pathogen,
49 novel analytic tests, test techniques and therapeutics that can improve sickness recognition and
50 treatments, proof-based administration recommendations.

51

52

53 **Keywords:** Sexually transmitted infections (STIs), Congenital syphilis, Neurosyphilis,
54 Cardiovascular syphilis, Penicillin, *Treponema pallidum*, Syphilis.

55

56

57 **Introduction**

58 Amongst the widely recognized health concerns Sexually transmitted infections (STIs) are one
59 (Newman et al. 2015). The worldwide burden of STIs keeps on being high, with more than 357
60 million new treatable contamination assessed in 2012, as indicated by the World Health
61 Organization (WHO); 376 million new cases were expected in 2016, which resulted in high
62 illness and mortality (Korenromp et al. 2017; WHO 2018). Around the globe, in excess of a
63 million STIs are reported each day (WHO 2017). As per the 2016 report from WHO, the
64 approximately 376 million new cases were related to the four treatable STIs-Chlamydia,
65 Gonorrhea, Syphilis and Trichomoniasis. Among those, 6 million cases were just for syphilis
66 (WHO 2018). Syphilis is an ulcerative, systemic human STI that can likewise be transmitted
67 through sexual contact with infectious sores of the mucous layers or scraped skin, by means of
68 blood transfusion, or from placental of pregnant women to her fetus (Janier et al. 2014; Peeling et
69 al. 2017). The Gram-negative spirochete microorganism called *Treponema pallidum* subspecies
70 *pallidum* (TPA) is responsible for causing Syphilis (Mattei et al. 2012). There are three
71 progressively known species from the same Genus that cause human treponemal infections,
72 namely, *Treponema pallidum* subspecies *pertenue* (TPE) that causes yaws, *Treponema pallidum*
73 subspecies *carateum* causing pinta and *Treponema pallidum* subspecies *endemicum* (TEN)
74 causing bejel or endemic (i.e., nonvenereal-transmission of sickness without sexual contact)
75 syphilis (Peeling et al. 2017; Tampa et al. 2014). These pathogens share exceptional similitudes
76 in morphology, pathogenesis and in clinical manifestations (Giacani & Lukehart 2014). However,
77 they can be separated by their age at obtaining infection, methods of transmission, clinical
78 manifestations, efficiency of central nervous system (CNS) and placenta invasion, genomic
79 sequences, even though the precision of these distinctions remains an issue of discussion (de
80 Melo et al. 2010; Peeling et al. 2017). The genomic mutation rate-based studies suggest that, the
81 causative agents for venereal syphilis and yaws have deviated a thousand years prior from a
82 typical ancestor beginning in Africa (Peeling et al. 2017; Smajs et al. 2012).

83 The U.S. Centers for Disease Control (CDC) often refer to syphilis as the “great imitator” and
84 ‘Great Mimicker’ because many of its symptoms are similar to other diseases (Peeling & Hook
85 2006). Initially, microorganism inoculation occurs via visible or microscopic abrasions of the
86 mucous membrane or skin of the mouth and genitals, which can result from the sexual contact.

87 In this review, we portray current discoveries along with bioinformatics approaches that have
88 improved our comprehension of the biological and hereditary structure of the pathogen, novel

89 analytic tests, test techniques and therapeutics that can improve sickness recognition and
90 treatments, proof-based administration recommendations.

91 **Survey methodology**

92 We used well defined criteria to select a set of articles, their separation methodologies and
93 important information to be analyzed. Two strategies were applied. First, the electronic literature
94 searches were undertaken using Web of Science (Core Collection) looked for these words
95 “(Syphilis, *Treponema pallidum*, *Treponema pallidum genomics OR comparative genome OR*
96 *pan-genomics*)” in the title, abstract or key word. These terms were chosen to collect what the
97 research articles related to Syphilis. In another strategy, only NCBI (National Center for
98 Biotechnology Information) PubMed and Google scholar database were used looking for the
99 most recent research articles related to Syphilis. The first articles were then screened, based on
100 their titles and abstracts. In this phase, all publications in which the terms Syphilis, *Treponema*
101 *pallidum*, *Treponema pallidum genomics OR comparative genome OR pan-genomics* were used
102 to designate recent set of articles, were considered relevant to the proposal for this work, and so
103 the most recent and relevant publications were used.

104 **Stages of syphilis**

105 Initially, syphilis embroils the formation of the painless chancre (some inoculated organisms
106 lodge at the location of entry, and proliferate and synthesize lymphocytes and macrophages). On
107 this level, syphilis is extremely active and contagious. The initial phase of syphilis, also known as
108 primary stage, can last about five weeks and the disease is communicable through contacts with
109 the effected parts of the body. If the infection is on the genital organs, the use of protection *i*;e
110 Condoms, may not be helpful in prevention of transmission. Nevertheless, 25 percent of infected
111 individuals proceed to the secondary phase of the disease. After the hematogenous dissimulation,
112 local skin mucous membrane eruptions occur which are accompanied by lymphadenopathy and
113 legitimate symptoms signaling the secondary syphilis. The lesions of secondary syphilis usually
114 occur between 3-6 weeks once the primary ulcer appears, which can include hair loss, a sore
115 throat fever, headache and a maculopapular rash on the flank, mouth, nose and vagina, shoulders,
116 arm, chest or back and that often involves the hands palm and feet soles (Cherneskie 2006b). In
117 this stage, host immune response suppresses infection enough to eliminate any sign or symptoms
118 of disease, but does not remove the infection completely, resulting in the latent stages infections.
119 When signs and symptoms moderated, patients enter a latent phase or latent stage, which can

120 remain many years (Peeling et al. 2017). There is no clinical manifestation evident during
121 latency, either Early latent (duration of infection ≤ 1 year) or Late latent (> 1 year). The expected
122 history of immune-competent patient of late latent syphilis follows the rule of thirds: 1/3 of the
123 patients will sero-revert to a nonreactive non-treponomal syphilis serology, with no relapse of
124 disease; 1/3 will remain reactive by nontreponomal syphilis serology but will remain free of signs
125 and symptoms of disease; and the remaining 1/3 will proceed to develop tertiary syphilis,
126 sometimes after decades of chronic, persistent, asymptomatic infection. Patients with tertiary
127 syphilis may develop additional diseases like granulomatous lesions (gummas) of the skin or
128 viscera, cardiovascular disease, or neurologic disease. Therefore, if left untreated, syphilis can
129 eventually lead to shocking, irremediable sequelae which include the complications of
130 neurosyphilis and tertiary syphilis. Furthermore, untreated syphilis infection during pregnancy
131 can have tragic consequences for a developing fetus when transmitted in uterus (i.e. congenital
132 disease) (**Figure 1**). It has also become well recognized that STIs such as syphilis interact
133 synergistically with HIV (Cherneskie 2006b).

134 **Figure 1:** Syphilis Transmission, Stages and Types.

135 **Types of Syphilis**

136 **Congenital**

137 Congenital syphilis, a consequence of fetal contamination with *Treponema pallidum*, is an
138 antiquated infection that keeps on plaguing newborn children and remains a noteworthy general
139 medical issue around the world (Cooper & Sanchez 2018). The ailment can cause unsuccessful
140 labor (miscarriage), stillbirth, or demise not long after conveyance. As indicated by the CDC, up
141 to 40 % of children born from women with not medicated for syphilis might be stillborn or die as
142 a neonatal. A few newborn children with disease can seem sound during childbirth yet create life
143 changing entanglements later in life. The treatment for syphilis in pregnancy is indistinguishable
144 to that of grown-ups who are not pregnant; penicillin is the main agent that is proper for use amid
145 pregnancy. Antibiotic medications such as Tetracyclines are contraindicated in pregnancy as a
146 result of their impact on fetal bone and tooth advancement (Arnold & Ford-Jones 2000; Rolfs
147 1995). The WHO prescribes that newborn children with probability of congenital syphilis,
148 including babies who are destined to mothers who are seropositive for syphilis and not treated
149 with penicillin >30 days before conveyance, ought to be treated with aqueous benzyl penicillin or

150 procaine penicillin (Peeling et al. 2017). All neonates presented to syphilis, including babies
151 without signs or manifestations during childbirth, ought to be pursued intently, in a perfect world
152 with NTTs (Non Treponemal Tests) titres. Titres should decline till three months of age and be
153 nonreactive till 6 months. TTs (Treponemal Tests) are not valuable in newborn children due to
154 persistent maternal antibodies (Peeling et al. 2017; Workowski et al. 2015)

155 **Neurosyphilis**

156 Neurosyphilis can occur any time in the span of syphilis on the grounds that no single highly
157 specific diagnostic test exists, and it is categorized into five unique classifications, such as,
158 asymptomatic, meningeal, meningovascular, parenchymatous, and gummatous. Invasion
159 (neuroinvasion) or spread by *T. pallidum* to the CSF and meninges, can happen in early infection,
160 even before the clinical signs of primary syphilis happen (Lukehart 1988; Morshed et al. 2015).
161 Much of the time, the microorganisms are cleared precipitously, yet in others, symptomatic
162 malady can happen. There are no high-quality tests for the determination of neurosyphilis,
163 however authoritative conclusion is normally dependent on clinical discoveries for serologic
164 affirmation of syphilis disease (any stage) together with a receptive cerebrospinal liquid VDRL
165 test, CSF examination, and neuroimaging (Golden et al. 2003; Morshed et al. 2015). Clinical
166 neurosyphilis can show up in a wide scope of ways that generally relate with span of disease,
167 albeit a few discoveries, for example, ocular involvement (uveitis, cranial nerve paralyses, and so
168 on) might happen over the span of untreated syphilis. A few people with secondary syphilis may
169 exhibit an aseptic meningitis disorder of cerebral pain and gentle meningismus (syphilitic
170 meningitis). The great clinical sign proposed by Merritt and colleagues was that extreme
171 syphilitic meningitis was a moderately exceptional type of neurosyphilis (Hook 2017; Simon
172 1985). The other clinical sign of late neurosyphilis, tabes dorsalis, seems to result from inclusion
173 of nerves on the back segments and spinal nerve roots (Gjestland 1955; Hook 2017; Simon
174 1985). Individuals with manifestations and tests showing neurosyphilis ought to get 3 to 4 million
175 units of crystalline penicillin G intravenously every 4 h for 10 to 14 days and infected people
176 with the human immunodeficiency infection (HIV) ought to experience examination for
177 neurosyphilis before treatment (Arnold & Ford-Jones 2000).

178 **Ocular syphilis**

179 The ocular indications of syphilis are changeable and can happen at any phase of the sickness.
180 Various authors have delineated the capacity of syphilis to emulate distinctive visual disarranges,
181 prompting misdiagnoses and a postponement in the suitable antimicrobial treatment. Syphilis can
182 influence the conjunctiva, sclera, cornea, focal point, uveal tract, retina, the retinal vasculature,
183 the optic nerve, pupillomotor pathways, and cranial nerves associated with extraocular
184 developments (Baughn & Musher 2005; Kiss et al. 2009; Mitchell et al. 1992; Singh &
185 Romanowski 1999).

186 Late reports from various world locales recommend ocular syphilis is re-developing, in parallel
187 with an expanding occurrence of the foundational contamination. A current observational
188 medical public investigation in Brazil over a 2.5-year time frame finishing July 2015, considered
189 127 individuals that were treated with ocular syphilis. Out of 127, 104 people were serologically
190 tried for human immunodeficiency infection (HIV), and 34.6% were certain (Furtado et al. 2018).
191 The US (CDC) issued a clinical warning toward the end of 2015 on ocular syphilis; during four
192 months (December 2014 and March 2015), 12 instances of ocular syphilis were accounted from
193 two US metropolis, San Francisco and Seattle (Tuddenham & Ghanem 2016). The analysis of
194 ocular syphilis was made on discoveries of visual irritation by ophthalmological examination and
195 affirmation of systemic contamination with *T. pallidum*. Systemic disease was demonstrated by a
196 receptive treponemal serological test (for example, fluorescent treponemal antibody absorption
197 test [FTA-ABS] or microhemagglutination assay test for *T. pallidum* [MHA-TP]),
198 notwithstanding: (1) a receptive non-treponemal serological test (for example venereal disease
199 research laboratory test [VDRL] or Rapid plasmin reagent test [RPR]); (2) an anomalous CSF
200 (for example receptive VDRL or potentially more prominent than 4 leukocytes/mm³ and protein
201 focus under 40 mg/dl); and additionally (3) predictable clinical signs that settled after intravenous
202 treatment with fluid penicillin G or ceftriaxone (Furtado et al. 2018).

203 **Cardiovascular syphilis**

204 Cardiovascular Syphilis is the dynamic compounding of Syphilis causing aortic aneurysms and
205 spewing forth. Occasionally, even youngsters (as youthful as 15 years) are influenced by
206 congenital syphilis. There are a few hazard factors for Cardiovascular Syphilis that could cause
207 this complexity. Any patient with aortic deficiency or thoracic aortic aneurysm ought to be
208 screened for syphilis. Auscultation must be implemented on individuals with late latent or tertiary

209 syphilis. A chest X-beam is occasionally contributory (Dabis & Radcliffe 2011; Janier et al.
210 2014).

211 **Auricular Syphilis or Oto syphilis**

212 Any patient with unexplained abrupt hearing misfortune ought to be screened for syphilis.
213 Sensorineural hearing misfortune may occur in congenital and acquired shapes. The clinical
214 courses of the early gained and late innate structures are comparative: unexpected or quickly
215 dynamic two-sided sensorineural hearing misfortune with or without mellow vestibular side
216 effects. Luetic internal ear illness (Otologic syphilis) is a moderately uncommon confusion that
217 might be turned around whenever analyzed early and treated forcefully. The determination of
218 inward ear syphilis can be exceptionally troublesome because this substance imitates other
219 internal ear sicknesses, for example, Meniere's ailment, abrupt hearing misfortune, acoustic
220 neuroma, vestibular neuritis, and immune mediated inward ear infection. Otologic syphilis is
221 normally analyzed by positive serologic tests and by avoidance of other conceivable causes
222 (Garcia-Berrocal et al. 2006; Janier et al. 2014).

223 **Available screening and Management methods**

224 Since 2000, the U.S. Preventive Service Task Force (USPSTF) has distributed eight clinical
225 suggestion articulations for STIs screening. The USPSTF suggests that women at expanded
226 danger of infection be screened for diseases *i*;e chlamydia, human immunodeficiency infection
227 (HIV), gonorrhea and syphilis. Men at extended hazard should be screened for HIV and syphilis.
228 Every single pregnant woman ought to be screened for hepatitis B, HIV and syphilis; pregnant
229 women at expanded hazard likewise must to be tested for chlamydia and gonorrhea (Meyers et al.
230 2008). The wide accessibility of successful medicines and coming about decrease in syphilis
231 predominance has prompted a low yield of screening in low-pervasiveness settings; because of
232 that, screening in generally safe adults (for instance, pre-marriage adults or those admitted to
233 clinic) has been relinquished in many spots. Nonetheless, systemic reviews give persuading proof
234 for syphilis screening for pregnant ladies. Adults and young people are at expanded risks of
235 getting the disease, specifically for people donating blood, blood items or solid organs. A few
236 nations additionally prescribe syphilis testing in individuals with unexplained and unexpected
237 visual misfortune, deafness or meningitis as these might be signs of early neurosyphilis (17–18

238 September 2015; Food & Drug Administration 2015; J.B. Tapko 2010; Janier et al. 2014;
239 Owusu-Ofori et al. 2011).

240 **Screening of Population with high risk**

241 The rate of spread of syphilis in a populace is identified with the transmission likelihood per
242 sexual association, i.e., the normal rate of obtaining of sexual accomplices and the span of
243 irresistibility; these rates are ascending among men who engage in sexual relations with men
244 (MSM) (Stoltey & Cohen 2015). Screening of people at high hazard for contracting syphilis is
245 required to recognize infections and end further transmission (Gray et al. 2011; Stoltey & Cohen
246 2015). Observations in HIV-positive MSM point to the fact that increasing number of visits for
247 syphilis screening or more noteworthy screening inclusion of beforehand unscreened people
248 would be economical. Update intercessions have appeared to be viable in expanding testing;
249 these incorporate computer alarms for clinicians to rapidly test high hazard MSM and instant
250 messages to increment STIs (or, sexually transmitted disease) re-testing rates (Bissessor et al.
251 2011; Bourne et al. 2011; Stoltey & Cohen 2015). Studies assessing different mediations, for
252 example, pre-introduction prophylaxis for syphilis, are likewise in progress. One future choice
253 may be to direct pre-introduction prophylaxis at the same time for syphilis and HIV (Dubourg &
254 Raoult 2016; Molina et al. 2015).

255 **Prenatal Screening**

256 As indicated by the WHO, STI rules suggest screening every single pregnant women for syphilis
257 amid the first antenatal care visit in view of the high danger of MTCT and the accessibility of a
258 profoundly powerful preventive intercession against adverse pregnancy outcomes. Powerful
259 counteractive action and recognizable proof of congenital syphilis depend basically on the
260 distinguishing proof of syphilis in pregnant women and so regular screening of every single
261 pregnant woman for syphilis is needed. A precise survey reported that antenatal syphilis
262 intercessions, including screening, could decrease perinatal stillbirth and passing rates by half.
263 Besides that, antenatal screening for syphilis has appeared to be cost-gainful even in developed
264 nations with lower rates of syphilis. Early screening ought (WHO) to preferably be performed in
265 the first trimester and should be repeated at 28 weeks and again at delivery in women at high
266 danger of gaining syphilis (Hawkes et al. 2011; Morshed et al. 2015; Walker & Walker 2002). At
267 the point when access to prenatal care is not ideal or lab limit is restricted, rapid tests are useful in

268 distinguishing and treating syphilis in pregnant women. Systems that improve screening
269 inclusion, for example, expanded utilization of rapid POC (Point-of-Care) testing and
270 incorporating syphilis and HIV screening, will facilitate the worldwide disposal of congenital
271 (inborn) syphilis (Dubourg & Raoult 2016; Hawkes et al. 2011; Mabey et al. 2012; Molina et al.
272 2015; Peeling et al. 2017; Peeling & Ye 2004).

273 **Blood-bank Screening**

274 Uncommon manifestation through blood items and organ donation have also been mentioned
275 (Perkins & Busch 2010; Stoltey & Cohen 2015). Reports of transfusion-transmitted syphilis have
276 turned out to be exceedingly uncommon in the course of recent years as more nations adopt
277 donor selection procedures, widespread serological screening of donor and the utilization of
278 refrigerated items rather than fresh blood elements (Stoltey & Cohen 2015). Blood Screening,
279 components of blood and strong organs for syphilis remains a recommendation in numerous
280 nations (Peeling et al. 2017). Case reports do exist, incorporating one in Ghana, which depicted a
281 seroconversion in a child after receipt of a Rapid Plasma Reagin (RPR)- receptive unit of blood.
282 The authors noticed that the unit had been refrigerated for only 1 day, and that a more extended
283 time of refrigeration was probably going to be important to kill *T. pallidum* (Owusu-Ofori et al.
284 2011; Stoltey & Cohen 2015). Periodic manifestation of transfusion-transmitted syphilis are as
285 yet announced in settings with high syphilis predominance, especially with the transfusion of
286 fresh blood (Owusu-Ofori et al. 2011).

287 It is decisively prescribed that nations should consider refreshed worldwide direction as they set
288 up standard national protocols, adjusted to the local epidemiological circumstance and
289 antimicrobial helplessness information, such as, early discovery, brief treatment with a
290 compelling anti-microbial routine and treating sex accomplices of an individual with irresistible
291 syphilis (primary, secondary or early latent contaminations) (WHO 2016a). As indicated by
292 European guidelines individuals with syphilis are at higher danger of gaining different STIs. All
293 patients with syphilis ought to be examined for HIV and HCV if chance components (as surveyed
294 by local epidemiology) are available. All people with syphilis ought to have a full STI appraisal.
295 Evaluation and inoculation for Hepatitis B ought to likewise be considered (Janier et al. 2014).
296 As indicated by The US (CDC) rules, persons who have had sexual contact with an individual
297 who gets a determination of primary, secondary, or late syphilis inside 90 days going before the
298 finding ought to be dealt with hypothetically for early syphilis, regardless of whether serologic

299 test outcomes are adverse. Long haul sex accomplices of people who have late dormant syphilis
300 ought to be assessed clinically and serologically for syphilis and treated based on the assessment's
301 discoveries (CDC 2015).

302 **Epidemiology and Global Prevalence**

303 Syphilis came back to the worldwide scene as a reemerging general health concern. Every year,
304 there are an expected 6 million new instances of syphilis universally in matured people between
305 15 to 49 years. More than 300,000 fetal and neonatal passing are ascribed to syphilis, with
306 215,000 extra babies put at expanded danger of early demise (Kojima & Klausner 2018).

307 Syphilis infection remains as a global health concern because of its commonness, irresistibility,
308 and toll on both tainted people and health systems (Mutagoma et al. 2016). Overall, public
309 pervasiveness information on syphilis are largely restricted to high-income nations. In LMICs
310 (Low- and Middle-income countries), the heterosexual increase of syphilis has declined in the
311 overall public, although it stays at dangerous levels in some high hazard sub populaces, for
312 example, female sex laborers/workers (FSWs) and their male customers (Peeling et al. 2017).
313 Generally, the appropriation of syphilis varies among LMICs and high-income nations.
314 Numerous LMICs have poor access to syphilis screening, testing and care (Ouedraogo et al. 2018).
315 In the past ten years, a resurgence of syphilis has happened in numerous nations around the world
316 (Halatoko et al. 2017). The predominance of syphilis among FSWs in African nations was
317 reported in (Ouedraogo et al. 2018), and precedent studies in Togo showed a wide scope of
318 syphilis infection among the FSWs, ranging from 1.5 to 42.1% (Halatoko et al. 2017) and 1.5%
319 in the northern zone to 8.9% in the eastern zone of Sudan of Sudan (Elhadi et al. 2013). Studies
320 directed during the 1990s have shown a predominance of 42.1% in South Africa and 15% in
321 Burkina Faso West Africa (Ouedraogo et al. 2018). In 2016, dynamic syphilis predominance was
322 evaluated to be 0.56% in women between 15 to 49 years old and around 21,675 new syphilis
323 diseases have happened in Morocco, a Northern African nation (Bennani et al. 2017). As indicated
324 by a study in 2013, the evaluated commonness of syphilis was higher in the 25– 49-year-maturity
325 than in the 15–24-year-seniority in Rwanda, Central/Eastern Africa (Mutagoma et al. 2016).
326 Counting Tanzania, Eastern Africa in general 2.5 % commonness of syphilis was found among
327 pregnant women. The hazard for syphilis disease was altogether higher among women going to
328 semi-urban and provincial centers and those having 3-4, and 5 past pregnancies (Manyahi et al.
329 2015). In Ghana, there have been instances of syphilis contamination among detainees in

330 restorative offices and different places of the nation. An examination at three prisons
331 demonstrated a predominance of 11%. Likewise, predominance of 7.9% and 4.5% were noted in
332 the overall public at Accra and Kumasi separately. Such pervasiveness rates of syphilis infection
333 demonstrate that the infection might be normal inside the all-inclusive community (Faustina et al.
334 2015). In Ethiopia, as per ANC-based sentinel observations, syphilis predominance expanded
335 from 1% in 2012 to 1.2% in 2014 and as high as 2.9 to 3.7% in ongoing investigations among
336 ANC participants in Gondar. Furthermore, it was reported in 7.3 to 9.8% of HIV patients in
337 Hawassa and Addis Ababa (Amsalu et al. 2018; Assefa 2014; Endris et al. 2015; Melku et al.
338 2015; Shimelis et al. 2015). In Zimbabwe, the estimation depended on the information from
339 seven reviews directed toward women going to ANC from year 2000 to 2012, and yearly routine
340 ANC program information between 2008 and 2015, in which the evaluated syphilis pervasiveness
341 declined from 1.9% to 1.5% from 2000 to 2016 (Korenromp et al. 2017). In 2007, Kenya AIDS
342 Indicator Survey (KAIS) gave appraisals for the first time to study the syphilis disease
343 transmission sero-predominance from a nationwide representative population of grown-ups
344 matured between 15-64 years. A sero-predominance of 1.8% was comparable to roughly 300,000
345 Kenyan grown-ups with active syphilis contamination at the time period of the study (Otieno-
346 Nyunya et al. 2011). Predominance of HIV/STI is high among women associated with high
347 hazard/risk sexual conduct in Kampala, Uganda which demonstrated that 21% were seropositive
348 and 10% were enduring active syphilis (Vandepitte et al. 2011). As per Australia's Division of
349 Health report on STIs, the populace rate of findings of irresistible or infectious syphilis has
350 expanded in the course of the most recent 2 years to achieve 6.7 per 100,000 populace in 2012, in
351 the non-Indigenous populace, where most warnings are because of male to male sex (MSM); the
352 rate expanded 20% from 2008 to 2012, with the most elevated rates in the 30-39 and 40-49 year
353 age gatherings

354 An investigation on information from 2005 to 2014 from China's web-based disease surveillance
355 framework found that instances of syphilis had expanded >3 folds from 135,210 in 2005 to
356 441,818 in 2014 (Kojima & Klausner 2018). A sequential cross-sectional examination in eight
357 urban centers in Shandong, China was directed from 2010 to 2014. The general commonness of
358 syphilis was higher in MSM contrasted with other occupants; migrants took part in anal sex (Hu
359 et al. 2017). From 2010 to 2015 in Guangxi, China, syphilis predominance was similarly found in
360 female sex workers which diminished from 9.2% in 2010 to 7.3% in 2015 among low-level
361 female sex workers and 2.6% in 2010 to 1.4% in 2015 among high level female sex workers

362 (Chen et al. 2016). Between 2010 & 2014, a cross-sectional investigation of 3,859 female drug
363 clients in Beijing found 239 (6.2%) women who tested positive for syphilis serology and an
364 expansion from 6.0% in 2010 to 8.8% in 2014. The commonness of syphilis was greater in
365 Synthetic drug clients (7.9%) when contrasted to conventional drug clients (3.7%) (Kojima &
366 Klausner 2018; Sun et al. 2018). As indicated by STIs in Europe in 2013, 22,237 syphilis cases
367 were accounted in 29 EU/EEA (European) associated States. Most of these cases were reported in
368 individuals more seasoned than 25 years, with youngsters in the range of 15 and 24 years old
369 representing just 14% of cases. Over half (58%) of the syphilis cases with data on transmission
370 class were accounted in MSM. In a reference of resurgent syphilis in France since 2000, reviews
371 contemplated two unique focuses in Montpellier, France: the dermatology division of an open
372 medical clinic and an unknown and free place for arrangement of data, finding and treatment of
373 venereal sicknesses. One hundred and seventy-five instances of syphilis were analyzed: 154 at
374 the CDAG (The dermatology department of a public hospital and anonymous and free centre
375 for provision of information, diagnosis and treatment of venereal disease.) and 21 at the
376 dermatology unit. Majority of these cases (96%) includes men with age range of 36 years. In
377 eighty-two percent of cases, homosexual men were involved. Forty-nine percent of cases were
378 studied in the secondary stage, twenty-two percent in the primary stage and 28% in the latent
379 stage (Amelot et al. 2015). In Germany, the quantity of detailed syphilis cases expanded in the
380 range between 11% and 22% every year from 2010 to 2014. Syphilis notices expanded in 2015
381 by 19% to an aggregate of 6,834. This was for the most part because of expanding warnings in
382 MSM of all age bunches in bigger cities of Germany (Jansen et al. 2016). Among 9,284 female
383 sex workers going to local public health departments in Germany, 1.1% of those were identified
384 with positive for syphilis serology (Jablonka et al. 2016). The prevalence of syphilis also
385 increased in European nations like Netherlands, Norway, Poland, Portugal, Serbia, Spain, Slovak
386 Republic, Ukraine, Malta, Ireland (among pregnant women in an antenatal and peripartum
387 teaching medical clinic in Dublin from 2005 to 2012), and Greenland; the frequency of syphilis
388 in Greenland has expanded from zero cases in 2010 to 95.7 per 100,000 occupants in 2014
389 influencing primarily youngsters with a middle age of 27 years (Albertsen et al. 2015; Azbel et
390 al. 2013; Bjekic et al. 2016; Jakopanec et al. 2013; Kojima & Klausner 2018; Lopes et al. 2016;
391 McGettrick et al. 2016; Padovese et al. 2014; Serwin & Unemo 2016; Svecova et al. 2015). As
392 indicated by retrospective investigation of the records of 1185 patients with an affirmed finding
393 of primary syphilis, the number of infected individuals with primary syphilis has ascended from
394 111 in 2005 to 158 in 2012, an expansion of 42.3%, mostly related to MSM in Greece (Albertsen

395 et al. 2015; Kanelleas et al. 2015; McGettrick et al. 2016; Padovese et al. 2014; Tsachouridou et
396 al. 2016); additionally such cases were reported in Belgium, Bulgaria, Croatia (Bozicevic et al.
397 2012; Kenyon et al. 2014; Tsankova et al. 2016).

398 The United Kingdom (UK) has seen an enduring increment in STIs in the past decade,
399 specifically, in men who engage in sexual relations with men (MSM) (Malek et al. 2015). In 2014
400 in Britain, there were 439,243 findings of STIs, even though this number mirrors a little decay
401 (0.3%) with respect to 2013; quantities of diagnoses of syphilis climbed considerably, by 33%
402 from 3,236 to 4,317. This number is the highest in Britain since 1949 for the syphilis diagnoses
403 (Mohammed et al. 2016). This expansion is believed to be because of large amounts of condom-
404 less sex, especially among men who are co-infected with HIV (Kojima & Klausner 2018). In 2011,
405 2,704 FSWs made 8,411 recorded appearances at genito urinary medicine clinics in England and
406 there were also 17 reported cases of congenital syphilis in new borns in UK from 2010 to 2015
407 (Mc Grath-Lone et al. 2014; Simms et al. 2017).

408 In South America, research shows that syphilis is hyper endemic among men developing sexual
409 relations with men (MSM) and transgender women (male-to-female). In 2008, the predominance
410 of syphilis disease, confirmed by laboratory testing, was evaluated to be 28.9% among MSM in
411 Lima, Peru (Caceres et al. 2008; Kojima et al. 2017). Everybody realizes Brazil is at present
412 enduring a Zika episode. Be that as it may, the genuine general health risk there may be
413 something nobody is discussing: the quantities of pregnant women tainted with syphilis and the
414 infants dying due to it is quietly increasing year after year. In 2016, the administration figures
415 41,762 new syphilis contaminations among pregnant women which is multiple times higher than
416 that revealed 10 years back; unfortunately, the rate of infants tainted in Brazil is devastatingly
417 high (Guimarães August, 2016). As indicated by review associate investigation and VDRL
418 results and the clinical records of 1,150 men with HIV, Buenos Aires, Argentina registered a
419 frequency of 14.9/100 patients every year in MSM (Bissio et al. 2017). As per observation
420 information from 1990 through 2003, there have been studies of disease transmission of primary
421 and secondary syphilis and rates of cases happening among MSM in the US. Amid 1990 till
422 2000, the ratio of primary and secondary syphilis diminished 90% by and large, decreasing 90%
423 and 89% among men and among women, respectively. The general rate expanded 19% from
424 2000 and 2003, mirroring a 62% expansion among men and a 53% decline among women. In
425 2003, an expected 62% of revealed cases happened among MSM (Heffelfinger et al. 2007). In
426 2017, an aggregate of 30,644 instances of Primary and Secondary syphilis were accounted for in

427 the US, yielding a rate of 9.5 cases per 100,000 populaces. This rate shows a 10.5% expansion
428 contrasted with 2016 (8.6 cases per 100,000 populace), and a 72.7% increment compared with
429 2013 (5.5 cases per 100,000 populace). Syphilis reports have expanded quickly in Japan.
430 Although, not at all, like other developed nations where MSM were related with the ascent, the
431 expansion in Japan has been attributed more to men who are involved in sexual relations with
432 women (MSW) and women who have intercourse with men (WSM). In view of observation data,
433 an aggregate of 7,040 (64.0%) of 10,997 cases were primary and secondary; the yearly rate of
434 increment was most noteworthy for primary and secondary and the extent of primary and
435 secondary expanded after some time. Among primary and secondary cases (1,609 MSM, 2,768
436 MSW, and 1,323 WSM), MSW and WSM each outperformed MSM cases in 2016. Congenital
437 syphilis reports expanded from 0.4 (2012) to 1.4 per 100,000 live nativities in 2016. Even though,
438 Tokyo had the most noteworthy detailing rate (3.98 per 100,000 man years) (Takahashi et al.
439 2018).

440

441 **Available Methods for Diagnosis**

442 The diagnosis of syphilis is mostly based on clinical discoveries, suggestive clinical history and
443 serologic tests in strong research facilities since bacterium *T. pallidum* cannot be grown/cultured
444 *in vitro* (Henao-Martinez & Johnson 2014; Peeling et al. 2017). Continued culture of *T. pallidum*
445 is troublesome and typically utilized just in research. Animal models, regularly utilizing rabbit
446 immunization, have been profitable for isolation of *T. pallidum*, and also to examine the response
447 of host to infection. Direct detection of *T. pallidum* from sore exudate gathered from patients with
448 primary and secondary syphilis is the best, however these tests are not easily available in
449 numerous settings (Hook 2017; Lafond & Lukehart 2006). Darkfield microscopy has generally
450 been utilized for distinguishing *T. pallidum*; nonetheless, neither darkfield microscope nor the
451 mastery to utilize them are broadly accessible. Choices for direct detection of *T. pallidum*
452 incorporate fluorescence microscopy and nucleic acid amplification by polymerase chain reaction
453 (PCR); although, these tests are likewise not promptly accessible and are not generally utilized.
454 Though it is not possible to grow *T. pallidum* in culture, there are various tests to detect syphilis
455 directly or indirectly. In any case, there is no single ideal test (Hood et al. 2016; Hook 2017).
456 Serological testing has turned into the most common intends to analyze, screen and follow-up to
457 treatment of syphilis, regardless of whether in individuals with indications of syphilis or in the
458 individuals who have no indications however are recognized through screening. A problem with

459 all syphilis serological tests is their powerlessness to recognize contamination with *T. pallidum*
460 subspecies pallidum and the *T. pallidum* subspecies that cause (non-venereal) yaws, pinta or bejel
461 (Hook 2017; Peeling et al. 2017). Serological tests fall into two classes: non-treponemal tests for
462 screening, and treponemal tests for affirmation and test execution factors, for example,
463 sensitivity, specificity, prescient qualities, and reproducibility sway, can fluctuate depending
464 upon the motivation behind the test. A helpful beginning stage in serological test elucidation is to
465 survey the purpose behind testing and the ideal utilization of test outcomes (Hook 2017)
466 Guaranteeing the precision and reliability of syphilis testing is vital, particularly in
467 nonspecialized research centers, where most patient samples are examined. Syphilis-explicit
468 quality affirmation tactics incorporate the preparation of technologists on explicit strategies and
469 also execution of internal quality control systems, test assessment and intraassay standardization
470 of industrially accessible test kits all the time (Giacani & Lukehart 2014; Peeling et al. 2017). It is
471 particularly critical to give sufficient preparation and standard outer quality appraisal, or
472 capability testing with remedial activity to guarantee the nature of tests and testing for social
473 insurance-suppliers. Performing fast tests in center-based or outreach settings have proven to be
474 inadequate because numerous parts of the world do not have proper laboratory for making an
475 exact analysis; hence, the necessity for lab testing has incredibly obliged the control of syphilis
476 and the disposal of congenital syphilis. Even though, the advancement of inexpensive, rapid tests
477 that can be performed at the POC has enormously expanded access to pre-birth screening and
478 diagnosis, even in under-resourced and remote areas (Parekh et al. 2010; Peeling et al. 2017; Smit
479 et al. 2013).

480 **Crucial Diagnosis by Direct Detection**

481 The syphilis treatment is comparative in HIV-positive or negative patients. Although, HIV-
482 positive patients require increasing numbers of follow-up visits because of high risk of treatment
483 disappointment. The decision of procedure for diagnosing syphilis relies upon the stage of
484 sickness and the clinical introduction. Treatment of syphilis in any stage should consider the
485 dangers of getting different sexually transmitted diseases (2002; Cherneskie 2006a; Peeling et al.
486 2017). When a functioning chancre or condyloma latum and congenital sores is available in
487 patients, direct location strategy which incorporate Darkfield microscopy, fluorescent antibody
488 staining, immunohistochemistry and PCR are the most explicit methods for determination.
489 However, its precision is directly proportional to the skills of the technician, the quantity of live

490 treponemes in the sore, and the nearness of non-pathologic treponemes in oral or anal
491 injuries/lesions (Cherneskie 2006a; Cummings et al. 1996). Microscopy had been utilized for
492 direct discovery and determination since 1920 however is currently utilized rarely. The latest
493 European rules recommended against DFA TP testing in clinical settings, and the reagents are
494 never again accessible (Janier et al. 2014; Peeling et al. 2017). PCR strategies favored techniques
495 for oral and different injuries where tainting with commensal treponemes is likely; they can be
496 performed in tissues, cerebrospinal liquid (CSF), and blood. It is thus exceptionally urgent to
497 choose a carefully approved strategy and dependably use it with fitting quality controls as there is
498 no globally affirmed PCR for *T. pallidum* (Janier et al. 2014).

499

500

501 **Diagnosis through Serological Tests**

502 Serological tests are the main methods for screening asymptomatic people and are the most
503 regularly utilized strategies to determine patients having signs and side effects suggestive of
504 syphilis. There is no serological test for syphilis which separates among non venereal syphilis
505 and the venereal treponematoses because these pathogens are morphologically and antigenically
506 comparative and can be separated just by their method of transmission, the study of disease
507 transmission, or clinical signs. Serological tests fall into two classes: nontreponemal tests (NTTs)
508 for screening, and treponemal tests (TTs) for affirmation (**Figure 2**) (Janier et al. 2014).

509 **Non-Treponemal Tests (NTTs)**

510 All Non-treponemal tests measure both immunoglobulin (Ig) G and IgM antiphospholipid
511 antibodies, shaped by the host response to lipoidal material discharged by harmed host cells at the
512 contamination site and lipid from the cell surfaces of the treponeme itself (Janier et al. 2014;
513 Peeling et al. 2017). The most regularly utilized NTTs are Venereal Diseases Research
514 Laboratory test (VDRL), the Rapid Plasma Reagin test (RPR), and the Tolidine Red Unheated
515 Serum Test (TRUST). All of these tests identify a blend of heterophile IgG and IgM, are manual
516 and cannot be automated, however they are not costly, straightforward and, if properly
517 performed, have a generally high sensitivity. NTTs are valuable in distinguishing dynamic/active
518 syphilis positiveness 10-15 days after the start of the primary chancre (Janier et al. 2014).

519 Additionally, NTTs must be performed physically on serum, and they depend on a subjective
520 interpretation. These tests likewise require prepared faculty from research facility and particular
521 reagents, along these lines, do not satisfy the ASSURED (affordable, sensitive, specific, user-
522 friendly, rapid and robust, equipment-free and deliverable to those who need them) criteria for
523 tests, which can be utilized at the POC (Point-of-Care) (Peeling et al. 2006; Peeling et al.
524 2017). Without treatment, the titer achieves high range of 1 to 2 years following contamination
525 and stays positive with low titers in late infection. Unconstrained seroreversion of NTT alongside
526 tertiary syphilis is exceptionally uncommon (if at all exists). Titers of NTT do not correspond
527 well with malady movement, and hence results ought to be accounted for quantitatively, in order
528 to screen sickness action and adequacy of treatment (Janier et al. 2014).

529

530 **Treponemal Tests (TTs)**

531 Treponemal tests distinguish antibodies to antigenic segments of *T. pallidum* proteins. These tests
532 are utilized principally to affirm the analysis of syphilis in patients with a responsive
533 nontreponemal test (Cherneskie 2006a). *T. pallidum* Haemagglutination test (TPHA), Micro
534 Haemagglutination assay/Examine for *T. pallidum* (MHA TP), Fluorescent Treponemal
535 Antibody absorption test (FTA abs test), *T. pallidum* passive Particle Agglutination test (TPPA),
536 Treponemal Enzyme Immunoassay (EIA), IgG immunoblot test for *T. pallidum*
537 Chemiluminescence Immunoassay (CIA), are accessible for TTs. The vast majorities of these
538 tests utilize recombinant treponemal antigens and distinguish both IgG and IgM (Janier et al.
539 2014). In developed nations, numerous health service establishments rely upon high-throughput
540 screening and use 'reverse' algorithms that screen with a mechanized treponemal EIA or CIA and
541 affirm results with a NTT as opposed to the inverse, traditional methodology (FIG-X). Very less
542 studies until now have been directed to study the precision of these 'reverse testing' algorithms
543 (Bibbins-Domingo et al. 2016; Peeling et al. 2017). As this type of serological reactivity happens
544 in early primary syphilis, in both recently treated sickness and late infection, considerable
545 consideration ought to be given to an exhaustive physical examination of the patient, appraisal of
546 past history and later sexual hazard factors before starting any treatment and accomplice warning
547 exercises.

548 **Rapid Tests**

549 In low- and medium-salary nations, numerous pre-birth facilities that provide screening and
550 treatment to syphilis do not have the ability to perform corroborative symptomatic tests, and
551 hence testing is often done off-site. Besides, patients may neglect to return to the research centers
552 to report outcomes and both the specimens and results can possibly be lost in travel, that may
553 lead to the outcomes of treatment being hampered or missed. Right now, new tests utilizing
554 cloned TP antigens and an immune-chromatographic procedure give an elective stage of rapid TP
555 testing to be performed at point-of-care (POC), which is called immune-chromatographic syphilis
556 (ICS) test (Fears & Pope 2001). Rapid syphilis tests utilize a finger-prick based entire blood test
557 and are normally immuno-chromatographic strip-based TTs examining parts that can be kept at
558 room temperature, require no equipment and negligible preparation, and the test is commonly
559 viewed as genuinely delicate and explicit roughly taking 20min to perform (Peeling et al. 2017).
560 Rapid test for syphilis with prompt outcomes executed in a field setting has the upside of
561 permitting ladies who test positive to be treated nearby at a similar visit, maintaining a strategic
562 distance from misfortunes to catch up for return visits and potential unfriendly results related
563 with delayed treatment. As indicated by the Special Program for Research and Training in
564 Tropical Diseases from the WHO, the perfect attributes of POC tests are that they are moderate,
565 sensitive, explicit, quick and vigorous. On location screening administrations can critically
566 decrease the commonness of maternal and congenital syphilis in some low-pay and middle salary
567 nations (Phang Romero Casas et al. 2018; Schackman et al. 2007).

568 **Figure 2:** Algorithm for Syphilis Screening. The figure adopted from Peeling et., al. 2017 with
569 little modification.

570 **Helpful tests in Neuro and Congenital Syphilis Diagnosis**

571 Neurosyphilis is the most dreaded and extreme appearance of syphilis and diagnosis of
572 neurosyphilis is very tricky and challenging as it alludes to infection of the central nervous
573 system, which may happen at any stage (Rubin et al. 2018; Wong et al. 2015). To analyze
574 neurosyphilis, a patient should initially be affirmed to be infected with *T. pallidum*. To start
575 testing for syphilis, straightforward serum tests can be performed, incorporating venereal disease
576 research laboratory (VDRL) testing and rapid plasma reagin (RPR) testing. It can be identified
577 with serum treponemal tests as well, which incorporate the fluorescent treponemal antibody
578 absorption (FTA ABS) test, Treponema pallidum particle agglutination test (TPPA), and syphilis
579 enzyme immunoassays (EIAs). Unfortunately, nontreponemal tests, for example, RPR and

580 VDRL, can be non-receptive in neurosyphilis (Peeling et al. 2017; Rubin et al. 2018; van der
581 Sluis 1992). Congenital syphilis (CS) is a contamination obtained from a mother with untreated
582 or insufficiently treated syphilis. Most babies in danger for CS can, without much of a stretch, be
583 distinguished by a positive maternal serologic test. The progress of diagnostic tests, for example,
584 enzyme immunoassays, immunoblotting and polymerase chain reaction (PCR) has expanded the
585 affectability and particularity of analyses, yet the identification of explicit IgM is as of now the
586 touchiest serological strategy, and the nearness to explicit IgM ought to be considered as proof of
587 a congenital *T. pallidum* infection. Maternal syphilis disease profoundly corresponds with fetal
588 misfortune; along these lines, the assessment of a stillborn child ought to incorporate an
589 assessment of maternal test results for syphilis (Herremans et al. 2010; Peeling et al. 2017; WHO
590 2016a)

591 **Molecular Mechanism of Pathophysiology**

592 The high level of homologous DNA among subspecies has allowed the utilization of serological
593 tests for syphilis, the conclusion of non-syphilitic treponemal infections, for example, yaws. In
594 territories where the two diseases are found, a patient with a receptive serological positive test
595 has great chances for the contamination of the other (Hook 2017). *Treponema pallidum*
596 subspecies *pallidum* is one of only a handful of main bacterial pathogens that has not been
597 cultured consistently *in vitro*. As of now, *T. pallidum* must be spread by immunization of rabbits
598 for use in research studies or for symptomatic or epidemiological purposes. Rabbits are
599 exceedingly vulnerable to *T. pallidum* disease and it has been reported that upon intradermal (i.d.)
600 contamination with *T. pallidum* subsp. *pallidum*, New Zealand White (NZW) rabbits create
601 clinically and histologically same sores as that of human essential sores. Secondary syphilis sores
602 can be seen after the medication of primary sore in the rabbit and lead to antibody responses like
603 those in people. Contaminated rabbits mount cytokine reactions in response to the pathogen that
604 reflect those of people with similar disease characteristics (Giacani & Lukehart 2014; Peeling et
605 al. 2017). Advancement of auxiliary experimental models of treponemal disease has been studied
606 to conquer the inaccessibility of innate rabbit strains and to encourage immunological
607 investigations, however with moderately low success. Hamsters have shown to be exceptionally
608 great models for yaws and bejel, yet were less helpful for disease with *T. pallidum* subsp.
609 *Pallidum* (Giacani & Lukehart 2014). Guinea pigs create primary ulcers just with inocula far
610 bigger than would normally be appropriate in rabbits or people and with unusual histopathology

611 and cytokine articulation profile contrasted with those of rabbit and human sores (Giacani &
612 Lukehart 2014; Pierce et al. 1983; Wicher & Wicher 1989). Mice can be contaminated with *T.*
613 *pallidum* subsp. *pallidum*, yet they neglect to create ailment signs and are hence of limited utility
614 for the investigation of the malady (Giacani & Lukehart 2014). Although, a local inflammatory
615 response inspired by spirochetes is believed to be the main driver of every clinical sign of
616 syphilis, the tissue damage mechanism and host defense that in the long run adds a proportion of
617 command over the bacterium, are not well characterized. Notwithstanding, the rabbit model
618 inadequately reiterates some clinical and immunological aspects of the human sickness. As
619 anyone might expect, even in the post-genomics period, our comprehension of the pathogenic
620 components in the lingering of syphilis is well behind that of other regular bacterial ailments
621 (Peeling et al. 2017; Radolf et al. 2016). *T. pallidum* has double membrane ultrastructure and
622 frequently has been analogized to enteric Gram-negative microbe. In any case, it is currently
623 acknowledged all around that there are generous contrasts between the external layer of *T.*
624 *pallidum* and those of individuals from the family Enterobacteriaceae. Contrasted with the
625 external layers of enteric Gram-negative microbes, the *T. pallidum* external layer has absence of
626 lipopolysaccharide and has an extraordinarily unique phospholipid structure (**Figure 3**) (Radolf et
627 al. 1995a). Although *T. pallidum* represents various lipoproteins fit for actuating macrophages
628 and dendritic cells (DCs) through Toll-like receptor (TLR) 2-dependent signaling pathways, they
629 are overwhelmingly underneath the surface. The scarcity of surface-exposed pathogen-associated
630 molecular patterns (PAMPs) empowers the bacterium to experience rehashed episodes of
631 scattering that are inadequately identified by innate immunity and furthermore clarifies the
632 absence of systemic inflammatory effects that are normal for the sickness (Peeling et al. 2017;
633 Radolf et al. 2016). *T. pallidum*'s poor surface antigenicity can be ascribed to a pseudo-case of
634 serum proteins and mucopolysaccharides. Examining the properties and arrangement of the outer
635 membrane of *T. pallidum* has been, and stays, laborious. How does *T. pallidum* meet its
636 nourishing prerequisites and complete its complex parasitic life-cycle with a moderate outer
637 membrane? An incomplete answer may lie in the bacterium's moderate (~30 h) rate of
638 replication, which is probably a developmental 'bargain' between the thickness of OMPs required
639 for feasibility and assignment as 'the stealth pathogen' (Peeling et al. 2017; Radolf et al. 2016).
640 Understanding occasions unfurling at the host-pathogen interface requires a point by point
641 learning of *T. pallidum*'s collection of surface-exposed proteins (Peeling et al. 2017).

642

643 **Figure 3:** Molecular Architecture of *Treponema Pallidum* (The figure is adopted from Peeling et
644 al., 2017) (Peeling et al. 2017) with modifications.

645

646

647 **Immune Response and Inflammation**

648 Cell-interceded immune procedures play a noticeable part in the clinical appearances of syphilis.
649 The immune cell type that initiates timely immune response to *T. pallidum* has not been
650 recognized yet. A key part to understand syphilis immunology, its necessary to find recognizable
651 proof of the treponemal factor(s) that induces severe inflammatory response. Of note, *T. pallidum*
652 has absence of lipopolysaccharide (LPS) (Bouis et al. 2001; Hardy & Levin 1983). Be that as it
653 may, treponemes contain copious membrane lipoproteins (Chamberlain et al. 1989). There is
654 presently an extensive collection of proof, obtained from both *in vitro* and *in vivo* examinations,
655 for the hypothesis that *T. pallidum*'s outer layer lipoproteins are the main proinflammatory agents
656 amid syphilitic contamination (Bouis et al. 2001; Norgard et al. 1996; Radolf et al. 1995a; Radolf
657 et al. 2016; Radolf et al. 1995b; Riley et al. 1992; Sellati et al. 1998). In spite of the fact that the
658 lack of PAMPs in the *T. pallidum* outer membrane empowers the bacterium to reproduce locally
659 and experience rehashed episodes of scattering, pathogen detection in the host is inevitably
660 activated. The living bacteria are taken up by dendritic cells, which at that point may traffic to
661 depleting lymph nodes to introduce related treponemal antigens to guileless B cells and T cells
662 (Bouis et al. 2001; Peeling et al. 2017). The part of opsonic antibody acts as an agent for the
663 invulnerability of *T. pallidum* fundamentally improving the redesign and degradation of
664 spirochaetes by phagocytosis and different PAMPs for official to Toll-like receptors triggered by
665 local T-cells (Lukehart 2008; Peeling et al. 2017). Actuated lesional T cells discharge IFN γ ,
666 advancing the leeway by macrophages, but also to strengthen tissue harming cytokines, for
667 example, tumor necrosis factor and IL 6 (Peeling et al. 2017; Stary et al. 2010; Van Voorhis et al.
668 1996). Not long ago, published data of suction rankles of syphilitic skin sores and peripheral
669 blood mononuclear cells demonstrated that *T. pallidum* at the same time evokes likewise innate
670 immune responses, as proved by the direct activation of macrophages by *T. pallidum* determined
671 TLR2 ligands (Salazar et al. 2007; Sellati et al. 1998; Sellati et al. 2001). *T. pallidum* is broadly
672 viewed as an extracellular bacterium. Understanding the reason for the heterogeneity of *T.*

673 *pallidum*'s surface antigenicity is critical to disentangling its system for counteracting agent
674 shirking. Poor target accessibility happens because of the low copy quantities of outer membrane
675 proteins and surface-exposed lipoproteins (Houston et al. 2012). Surface-exposed antigens in *T.*
676 *pallidum* are probably critical destructiveness factors, just as being the atoms that associate with
677 the defensive immune response. A few investigations have demonstrated that *T. pallidum* disease
678 prompts antibodies that restrain cell connection and advance macrophage mediated phagocytosis
679 (Centurion-Lara et al. 1999; Wong et al. 1983). The distinguishing proof of a polymorphic, multi-
680 membered gene family in *T. pallidum* subspecies *pallidum* is Tpr K; it is specially deciphered in
681 the Nichols strain, and its encoded protein works as an objective of opsonic immune response,
682 actuates huge security against irresistible test, and is probably surface-exposed. This gene family
683 seems, by all accounts, to be key to the pathogenesis of syphilis and may add to antigenic decent
684 variety of *T. pallidum* (Centurion-Lara et al. 1999).

685 **Penicillin and Antibiotics**

686 Penicillin has for some time been the medication for treating people in all phases of syphilis (i.e.,
687 benzathine, fluid procaine, or aqueous crystalline) (Hook 2017).

688 In past years, deficiencies have constrained the accessibility of benzathine benzylpenicillin,
689 which is the favored detailing for most syphilis treatment. Long-acting developments of
690 benzathine benzylpenicillin are the most ordinarily prescribed medications for syphilis treatment.
691 Substitute treatment utilizing different portions of procaine penicillin, doxycycline, or ceftriaxone
692 can be utilized when intravenous treatment may be troublesome or on account of feasible
693 penicillin sensitivity (Director-General 2018; Hook 2017; Kingston et al. 2016; Workowski et al.
694 2015). Administration of 2.4 million units of benzathine penicillin G (Bicillin L A)
695 intramuscularly (IM) in a solitary portion/dose are recommended for early syphilis (Primary and
696 Secondary or early) and for susceptible people where penicillin treatment is unfavorable, control
697 regimens of doxycycline 100 mg orally twice every day for 14 days or antibiotic medication
698 (tetracycline) 500 mg multiple (four) times day by day for 14 days (Health 2019) are prescribed.
699 For patients with neurosyphilis, prescribed treatment is higher portions/doses (18– 20 million
700 units for each day in separated dosages) of intravenous aqueous penicillin G regulated every 4 h
701 for 10– 14 days; specialists also prescribe a few dosages of benzathine benzylpenicillin after
702 completion of intravenous treatment to reflect the term of treatment for late contaminations
703 (Hook 2017). Treatment of patients with syphilis who have demonstrated penicillin resistance can

704 be more difficult. Fluoroquinolone, sulphonamides, and aminoglycoside antibiotics are not
705 successful. Doxycycline or tetracycline given for 14– 28 days relying upon the phase of disease
706 can be utilized for treatment of non-pregnant patients with beta-lactam-anti-microbial/antibiotics
707 hypersensitivity, however there are concerns identified with the likelihood of drug non-adherence
708 with the drawn out course of anti-microbials (antibiotics). There are no suggestions to adjust
709 treatment for pregnant women or for patients with HIV contamination. For treatment of syphilis
710 during pregnancy, there are no prescribed exchange regimens for females with penicillin
711 hypersensitivity and desensitization to penicillin is suggested. Azithromycin was assessed as a
712 promising elective beta-lactam antibiotics agent; but, strains conveying a 23S rRNA change for
713 macrolide obstruction (Hook 2017; Hook et al. 2002; Riedner et al. 2005).

714 **Vaccine**

715 Until now, there is no vaccine against syphilis; the best method of counteractive action is
716 immediate treatment to keep away from proceedings with sexual transmission or MTCT, and the
717 treatment of all sex accomplices to maintain a strategic distance from reinfection (Peeling et al.
718 2017).

719 **Implementation of new approaches for the development of new Drug**

720 Normally, anti-microbials, for example, penicillin, erythromycin, and azithromycin are endorsed
721 for treatment of the malady, out of which the intramuscularly controlled penicillin G benzathine
722 is the favored treatment for syphilis, aside from neurosyphilis (Dwivedi et al. 2015). Different
723 antibiotics or anti-toxins, including tetracyclines, macrolides, and cephalosporins, have
724 additionally been utilized as an elective treatment for penicillin hypersensitive patients (Katz &
725 Klausner 2008). Individuals experiencing syphilis are at a two-to five-fold higher hazard to get
726 HIV disease. Amid this, the previous couple of years have seen the rise of an antibiotic-resistant
727 strain of *T. pallidum* (Marra et al. 2006). Reports of resistance from macrolides, the main option
728 in contrast to penicillin for treating *T. pallidum* contaminations, began in the 1960's. South et
729 al.,1964 and Fenton 1976 (Fenton & Light 1976; South et al. 1964) and Light detailed syphilis
730 treatment disappointments utilizing erythromycin as early as 1964. Subsequent reports followed
731 on the resistance against azithromycin, clindamycin, and rifampin. While there is right now
732 worry that *T. pallidum* may create resistance from antibiotic medications (tetracyclines) and its
733 subordinates, so far there is little proof this is going on (Fenton & Light 1976; Ghanem et al.

734 2006; Huigen & Stolz 1974; Katz & Klausner 2008; Krupp & Madhivanan 2015; South et al.
735 1964; Wong et al. 2008; Woznicova et al. 2007). Presently, there is extreme necessity of
736 discovering elective oral treatments. Impetuses for a medication revelation program for syphilis
737 should be set up and, meanwhile, assessment of existing medication blends may be helpful as
738 choices to decrease the danger of creating resistance (Peeling et al. 2017).

739 In the big data period, voluminous datasets are routinely procured and analyzed with the goal of
740 making biomedical revelations and approve theories. Massive data are utilized over the entire
741 drug discovery pipeline from target identification and mechanism to distinguishing proof of novel
742 leads and medication competitors. Such techniques are presented to give a general perspective on
743 computational instruments and accessible databases (Katsila et al. 2016). Computational
744 methodologies for the drug target identifications are used to diminish time and exertion in
745 medication advancement. Genomics and related high-throughput innovations give chances to
746 better comprehend the irresistible ailment component and also identify treatment mechanisms
747 (Kim et al. 2017; Pastuszczak & Wojas-Pelc 2013).

748 Subtractive genomics-based target identification is based on analyses of essential genes and the
749 proteins that are non-host homologous. Essential genes are fundamental genes that are necessary
750 for growth, flexibility and survival of any microorganism. Hence, insufficiency of any such
751 quality ought to be lethal to any microorganism. Also, essential genes probably have common
752 functions in all microorganisms. Subtractive genomics gives new opportunities for finding ideal
753 focuses/targets among unknown cellular functions, based on biological arrangements in bacterial
754 pathogens and their respective hosts (Barh & Kumar 2009; Barh et al. 2011). Currently, working
755 with bacterial pathogens utilizing an *in silico* approach, countless targets have been recognized
756 that are either resistant to drugs or for which no fitting immunization is accessible.
757 Methodologies, for example, comparative and subtractive genomics and differential genome
758 examinations are largely being used for drug target identification in a few human pathogens,
759 including *Mycobacterium tuberculosis* (Asif et al. 2009), *Helicobacter pylori* (Dutta et al. 2006),
760 *Burkholderia pseudomallei* (Chong et al. 2006), *Pseudomonas aeruginosa* (Sakharkar et al.
761 2004), *Salmonella typhi* (Rathi et al. 2009), *Neisseria gonorrhoeae* (Barh & Kumar 2009),
762 *Haemophilus ducreyi* (de Sarom et al. 2018), *Treponema pallidum* (Jaiswal et al. 2020; Kumar
763 Jaiswal et al. 2017) and *Mycoplasma pneumoniae* (Vilela Rodrigues et al. 2019).

764 **Development of vaccine against syphilis**

765 Considering the worldwide ailment weight of syphilis, direct relationship with expanded
766 transmission of HIV, and noteworthy dismalness and mortality related with irresistible syphilis
767 and congenital syphilis, there is a conspicuous requirement for conceptual, key and monetary
768 help for improvement of vaccines against this fatal ailment (Bloom 2011; Cameron & Lukehart
769 2014). Notwithstanding cheap and viable anti-toxin treatment, syphilis remains a common
770 ailment in developing nations and has re-emerged as a general wellbeing risk in developed
771 nations (Cameron & Lukehart 2014). Human-challenge studies have demonstrated that
772 individuals with late dormant syphilis are resistant to symptomatic reinfection with heterologous
773 strains of *T. pallidum*. Aside from a couple of nations, the socioeconomics of syphilis diseases
774 demonstrate an unmistakable separation among developed and developing nations. In most
775 industrialized nations, syphilis is dominated among men who engage in sexual relations with men
776 (MSM), while in developing countries contaminations happen fundamentally among the hetero
777 populace. In the US, both MSM and hetero African American populaces are at high hazard
778 (Cameron & Lukehart 2014). If a viable syphilis vaccine is created, it needs to cater to this
779 statistical profile, working effectively in MSM and other high-hazard populaces. Likewise,
780 defensive antibodies must be created. Although it is necessary to develop protective syphilis
781 vaccine, our comprehension of defensive invulnerability against *T. pallidum* is obstructed by our
782 failure to culture the bacteria *in vitro* (Peeling et al. 2017). Research on syphilis vaccine
783 development (**Table 1**) needs to progress first in the scholastic domain, which may lead to greater
784 industry enthusiasm in the future.

785 **Table 1: List of reported therapeutic targets against *Treponema pallidum***

786

787

788

789

790

791 Reverse vaccinology is a standard and well-known methodology in the post-genomic era for
792 identifying novel vaccine targets (**Figure 4**) (Kumar Jaiswal et al. 2017; Rappuoli 2001). Today,
793 the likelihood of utilizing genomic data enables us to examine vaccine improvements *in silico*,

794 without the need of culturing the pathogen. This methodology decreases the time required to
795 identify viable vaccines and gives new answers for those vaccines which have been difficult to
796 create. This new methodology has led to develop a vaccine against serogroup B meningococcus
797 successfully.

798 **Figure 4:** Schematic representation of Reverse Vaccinology approach

799 ***In silico work on Treponema Pallidum***

800 A recent work was published by Jaiswal et al., in 2017 (Kumar Jaiswal et al. 2017) using
801 subtractive genomics and reverse vaccinology approaches for the identification of new drugs and
802 vaccine candidates for bacterium *T. pallidum*. In the analysis, they compared 13 genomic strains
803 of *T. pallidum* using *T. pallidum* Nichols as the reference genome. As a result, they identified 15
804 putative antigenic proteins as vaccine targets and 6 drug targets.

805 In past, studies have demonstrated the significance of focusing on proteins engaged with the
806 capacity of *T. pallidum* to attack host tissues and to side-step the functional immune response,
807 adding to its steadiness amid the "latency" stage. The vast majority of the depicted gene targets
808 code for proteins accountable for the attachment to extracellular matrix bridges (Tp0136,
809 TP0155, Tp0483, and Tp0751) for example, the low density integral Outer Membrane Proteins
810 (OMPs) (Cameron & Lukehart 2014). Jaiswal *et al.*, 2017 reported vaccine targets
811 Tp_Nichols350 and TpNichols852 with likenesses to two recently depicted OMPs (TP0453 and
812 Tp_0326), alongside two extra OMP domain containing proteins: Tp_Nichols797 and
813 Tp_Nichols141) (Kumar Jaiswal et al. 2017). In their work, with the distinguished six targets,
814 Tp_Nichols130 (UvrB, Uvr ABC System Protein B), Tp_Nichols593 (pfp, Pyrophosphate-
815 fructose 6-phosphate 1-phosphotransferase), Tp_Nichols609 (asnA, Aspartate alkali ligase),
816 Tp_Nichols754 (recA, Protein RecA), Tp_Nichols990 (ndh, NADH dehydrogenase),
817 Tp_Nichols1011 (dxs, 1-deoxy-d-xylulose-5-phosphate synthase), they utilized molecular
818 docking approach with 28 natural compounds and identified the drug molecule Pinosresinol
819 (CID234817) to have best outcomes against four targets uvrB, pfp, asnA, and dxs. Pinosresinol is
820 a lignan, biphenolic compound found in *Araucaria araucana* and *Sambucus williamsii*. It has
821 bactericidal and fungicidal activities and remedial potential as an antifungal agent for the
822 treatment of contagious fungus infections in people (Cespedes et al. 2006; Hwang et al. 2010) .

823 The identification of pinoresinol in their *in silico* study can be possibly utilized as another
824 medication for the treatment of syphilis (Kumar Jaiswal et al. 2017).

825

826

827 **Conclusion**

828 The incidence of syphilis has progressively increased worldwide in the last few decades. It
829 remains as an important public health problem with increasing prevalence. In spite of
830 developments in *in vitro* cultivation methods, the genetic variability of the *T. pallidum* during the
831 infection is yet not understood properly. In this regard, a better understanding of the host
832 pathogen interaction can improve the molecular basis of *T. pallidum* infection. Hence, more
833 studies are necessary to understand the exact mechanisms of infection to combat against syphilis.
834 Furthermore, a robust encouragement and community involvement is needed such that syphilis is
835 given a high priority on the global health agenda. An extra investment is required on the research
836 to find interaction between HIV and syphilis in MSM besides an improved diagnostic, a better
837 test of cure, and a vaccine.

838 **Author Contributions**

839 AKJ, ST, SFOT, SBJ and RP wrote the review. ST, SCS, PG, DB and VA guided and reviewed
840 the work. All authors read and approved the review.

841 **Acknowledgements**

842 We acknowledge the collaboration and assistance of all team members and the Brazilian funding
843 agencies CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil) and
844 CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

845 **Author Conflicts**

846 The authors declare that they have no conflict of interest.

847

848

849 **References**

850

851

852 17–18 September 2015. Antenatal screening for HIV, hepatitis B, syphilis and rubella
853 susceptibility in the EU/EEA – addressing the vulnerable populations. *ECDC*
854 *SCIENTIFIC ADVICE*.

855 Albertsen N, Mulvad G, and Pedersen ML. 2015. Incidence of syphilis in Greenland 2010-2014:
856 the beginning of a new epidemic? *Int J Circumpolar Health* 74:28378.
857 10.3402/ijch.v74.28378

858 Amelot F, Picot E, Meusy A, Rousseau C, Brun M, and Guillot B. 2015. [Syphilis in Montpellier,
859 France, from 2002 to 2011: Survey in a free hospital screening centre for venereal disease
860 and in the dermatology unit of a regional public hospital]. *Ann Dermatol Venereol*
861 142:742-750. 10.1016/j.annder.2015.07.008

862 Amsalu A, Ferede G, and Assegu D. 2018. High seroprevalence of syphilis infection among
863 pregnant women in Yiregalem hospital southern Ethiopia. *BMC Infect Dis* 18:109.
864 10.1186/s12879-018-2998-8

865 Arnold SR, and Ford-Jones EL. 2000. Congenital syphilis: A guide to diagnosis and
866 management. *Paediatr Child Health* 5:463-469.

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

870 Assefa A. 2014. A three year retrospective study on seroprevalence of syphilis among pregnant
871 women at Gondar University Teaching Hospital, Ethiopia. *Afr Health Sci* 14:119-124.
872 10.4314/ahs.v14i1.18

873 Azbel L, Wickersham JA, Grishaev Y, Dvoryak S, and Altice FL. 2013. Burden of infectious
874 diseases, substance use disorders, and mental illness among Ukrainian prisoners
875 transitioning to the community. *Plos One* 8:e59643. 10.1371/journal.pone.0059643

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

878 Barh D, Tiwari S, Jain N, Ali A, Santos AR, Misra AN, Azevedo V, and Kumar A. 2011. In
879 silico subtractive genomics for target identification in human bacterial pathogens. *Drug*
880 *Development Research* 72:162-177. 10.1002/ddr.20413

881 Baughn RE, and Musher DM. 2005. Secondary syphilitic lesions. *Clin Microbiol Rev* 18:205-
882 216. 10.1128/CMR.18.1.205-216.2005

883 Bennani A, El-Kettani A, Hancali A, El-Rhilani H, Alami K, Youbi M, Rowley J, Abu-Raddad
884 L, Smolak A, Taylor M, Mahiane G, Stover J, and Korenromp EL. 2017. The prevalence
885 and incidence of active syphilis in women in Morocco, 1995-2016: Model-based
886 estimation and implications for STI surveillance. *Plos One* 12:e0181498.
887 10.1371/journal.pone.0181498

- 888 Bibbins-Domingo K, Grossman DC, Curry SJ, Davidson KW, Epling JW, García FAR, Gillman
889 MW, Harper DM, Kemper AR, Krist AH, Kurth AE, Landefeld CS, Mangione CM,
890 Phillips WR, Phipps MG, and Pignone MP. 2016. Screening for Syphilis Infection in
891 Nonpregnant Adults and Adolescents. *JAMA* 315. 10.1001/jama.2016.5824
- 892 Bissessor M, Fairley CK, Leslie D, and Chen MY. 2011. Use of a computer alert increases
893 detection of early, asymptomatic syphilis among higher-risk men who have sex with men.
894 *Clin Infect Dis* 53:57-58. 10.1093/cid/cir271
- 895 Bissio E, Cisneros V, Lopardo GD, and Cassetti LI. 2017. Very high incidence of syphilis in
896 HIV-infected men who have sex with men in Buenos Aires city: a retrospective cohort
897 study. *Sex Transm Infect* 93:323-326. 10.1136/sextrans-2016-052893
- 898 Bjekic M, Vlajinac H, and Sipetic-Grujicic S. 2016. Characteristics of gonorrhea and syphilis
899 cases among the Roma ethnic group in Belgrade, Serbia. *Braz J Infect Dis* 20:349-353.
900 10.1016/j.bjid.2016.05.004
- 901 Bloom DE. 2011. The value of vaccination. *Adv Exp Med Biol* 697:1-8. 10.1007/978-1-4419-
902 7185-2_1
- 903 Bouis DA, Popova TG, Takashima A, and Norgard MV. 2001. Dendritic cells phagocytose and
904 are activated by *Treponema pallidum*. *Infect Immun* 69:518-528. 10.1128/IAI.69.1.518-
905 528.2001
- 906 Bourne C, Knight V, Guy R, Wand H, Lu H, and McNulty A. 2011. Short message service
907 reminder intervention doubles sexually transmitted infection/HIV re-testing rates among
908 men who have sex with men. *Sex Transm Infect* 87:229-231. 10.1136/sti.2010.048397
- 909 Bozicevic I, Lepej SZ, Rode OD, Grgic I, Jankovic P, Dominkovic Z, Lukas D, Johnston LG, and
910 Begovac J. 2012. Prevalence of HIV and sexually transmitted infections and patterns of
911 recent HIV testing among men who have sex with men in Zagreb, Croatia. *Sex Transm*
912 *Infect* 88:539-544. 10.1136/sextrans-2011-050374
- 913 Caceres CF, Konda KA, Salazar X, Leon SR, Klausner JD, Lescano AG, Maiorana A, Kegeles S,
914 Jones FR, and Coates TJ. 2008. New populations at high risk of HIV/STIs in low-income,
915 urban coastal Peru. *AIDS Behav* 12:544-551. 10.1007/s10461-007-9348-y
- 916 Cameron CE, and Lukehart SA. 2014. Current status of syphilis vaccine development: Need,
917 challenges, prospects. *Vaccine* 32:1602-1609. 10.1016/j.vaccine.2013.09.053
- 918 CDC. 2015. Syphilis. Available at <https://www.cdc.gov/std/tg2015/syphilis.htm> (accessed
919 October 2019).
- 920 Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Van Voorhis WC, and
921 Lukehart SA. 1999. *Treponema pallidum* major sheath protein homologue Tpr K is a
922 target of opsonic antibody and the protective immune response. *J Exp Med* 189:647-656.
- 923 Cespedes CL, Avila JG, Garcia AM, Becerra J, Flores C, Aqueveque P, Bittner M, Hoeneisen M,
924 Martinez M, and Silva M. 2006. Antifungal and antibacterial activities of *Araucaria*
925 *araucana* (Mol.) K. Koch heartwood lignans. *Z Naturforsch C* 61:35-43.
- 926 Chamberlain NR, Brandt ME, Erwin AL, Radolf JD, and Norgard MV. 1989. Major integral
927 membrane protein immunogens of *Treponema pallidum* are proteolipids. *Infect Immun*
928 57:2872-2877.
- 929 Chen Y, Abraham Bussell S, Shen Z, Tang Z, Lan G, Zhu Q, Liu W, Tang S, Li R, Huang W,
930 Huang Y, Liang F, Wang L, Shao Y, and Ruan Y. 2016. Declining Inconsistent Condom
931 Use but Increasing HIV and Syphilis Prevalence Among Older Male Clients of Female
932 Sex Workers: Analysis From Sentinel Surveillance Sites (2010-2015), Guangxi, China.
933 *Medicine (Baltimore)* 95:e3726. 10.1097/MD.0000000000003726
- 934 Cherneskie T. 2006a. An Update and Review of the Diagnosis and Management of Syphilis.
- 935 Chong CE, Lim BS, Nathan S, and Mohamed R. 2006. In silico analysis of *Burkholderia*
936 *pseudomallei* genome sequence for potential drug targets. *In Silico Biol* 6:341-346.

- 937 Cooper JM, and Sanchez PJ. 2018. Congenital syphilis. *Semin Perinatol* 42:176-184.
938 10.1053/j.semperi.2018.02.005
- 939 Cummings MC, Lukehart SA, Marra C, Smith BL, Shaffer J, Demeo LR, Castro C, and
940 McCormack WM. 1996. Comparison of methods for the detection of treponema pallidum
941 in lesions of early syphilis. *Sex Transm Dis* 23:366-369.
- 942 Dabis R, and Radcliffe K. 2011. Is it useful to perform a chest X-ray in asymptomatic patients
943 with late latent syphilis? *Int J STD AIDS* 22:105-106. 10.1258/ijsa.2010.010248
- 944 de Sarom A, Kumar Jaiswal A, Tiwari S, de Castro Oliveira L, Barh D, Azevedo V, Jose Oliveira
945 C, and de Castro Soares S. 2018. Putative vaccine candidates and drug targets identified
946 by reverse vaccinology and subtractive genomics approaches to control Haemophilus
947 ducreyi, the causative agent of chancroid. *J R Soc Interface* 15. 10.1098/rsif.2018.0032
- 948 Director-General. 2018. Addressing the global shortage of, and access to, medicines and vaccines
949 WHO
- 950 Dubourg G, and Raoult D. 2016. The challenges of preexposure prophylaxis for bacterial
951 sexually transmitted infections. *Clin Microbiol Infect* 22:753-756.
952 10.1016/j.cmi.2016.08.022
- 953 Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, and Bandyopadhyay D. 2006. In silico
954 identification of potential therapeutic targets in the human pathogen Helicobacter pylori.
955 *In Silico Biol* 6:43-47.
- 956 Dwivedi UN, Tiwari S, Singh P, Singh S, Awasthi M, and Pandey VP. 2015. Treponema
957 pallidum putative novel drug target identification and validation: rethinking syphilis
958 therapeutics with plant-derived terpenoids. *OMICS* 19:104-114. 10.1089/omi.2014.0154
- 959 Dwivedi UN, Tiwari S, Singh P, Singh S, Awasthi M, and Pandey VP. 2015. Treponema
960 pallidum Putative Novel Drug Target Identification and Validation: Rethinking Syphilis
961 Therapeutics with Plant-Derived Terpenoids. *OMICS: A Journal of Integrative Biology*
962 19:104-114. 10.1089/omi.2014.0154
- 963 Elhadi M, Elbadawi A, Abdelrahman S, Mohammed I, Bozicevic I, Hassan EA, Elmukhtar M,
964 Ahmed S, Abdelraheem MS, Mubarak N, Elsanousi S, and Setayesh H. 2013. Integrated
965 bio-behavioural HIV surveillance surveys among female sex workers in Sudan, 2011-
966 2012. *Sex Transm Infect* 89 Suppl 3:iii17-22. 10.1136/sextrans-2013-051097
- 967 Endris M, Deressa T, Belyhun Y, and Moges F. 2015. Seroprevalence of syphilis and human
968 immunodeficiency virus infections among pregnant women who attend the University of
969 Gondar teaching hospital, Northwest Ethiopia: a cross sectional study. *BMC Infect Dis*
970 15:111. 10.1186/s12879-015-0848-5
- 971 Faustina N-B, Dankwa K, Ampiah C, Boampong J, and Nuvor S. 2015. Seroprevalence of
972 Syphilis Infection in Individuals at Cape Coast Metropolis, Ghana. *British Journal of*
973 *Medicine and Medical Research* 8:157-164. 10.9734/bjmmr/2015/16267
- 974 Fears MB, and Pope V. 2001. Syphilis fast latex agglutination test, a rapid confirmatory test. *Clin*
975 *Diagn Lab Immunol* 8:841-842. 10.1128/CDLI.8.4.841-842.2001
- 976 Fenton LJ, and Light IJ. 1976. Congenital syphilis after maternal treatment with erythromycin.
977 *Obstet Gynecol* 47:492-494.
- 978 Food, and Drug Administration HHS. 2015. Requirements for blood and blood components
979 intended for transfusion or for further manufacturing use. Final rule. *Fed Regist*
980 80:29841-29906.
- 981 Furtado JM, Arantes TE, Nascimento H, Vasconcelos-Santos DV, Nogueira N, de Pinho Queiroz
982 R, Brandao LP, Bastos T, Martinelli R, Santana RC, Muccioli C, Belfort R, Jr., and Smith
983 JR. 2018. Clinical Manifestations and Ophthalmic Outcomes of Ocular Syphilis at a Time
984 of Re-Emergence of the Systemic Infection. *Sci Rep* 8:12071. 10.1038/s41598-018-
985 30559-7

- 986 Garcia-Berrocal JR, Gorriz C, Ramirez-Camacho R, Trinidad A, Ibanez A, Rodriguez Valiente
987 A, and Gonzalez JA. 2006. Orosyphilis mimics immune disorders of the inner ear. *Acta*
988 *Otolaryngol* 126:679-684. 10.1080/00016480500491994
- 989 Garnett GP, Aral SO, Hoyle DV, Cates W, Jr., and Anderson RM. 1997. The natural history of
990 syphilis. Implications for the transmission dynamics and control of infection. *Sex Transm*
991 *Dis* 24:185-200.
- 992 Giacani L, and Lukehart SA. 2014. The endemic treponematoses. *Clin Microbiol Rev* 27:89-115.
993 10.1128/CMR.00070-13
- 994 Giacani L, Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, and Yang T-C. 2019. Insights into the
995 genetic variation profile of tprK in *Treponema pallidum* during the development of
996 natural human syphilis infection. *PLOS Neglected Tropical Diseases* 13.
997 10.1371/journal.pntd.0007621
- 998 Gjestland T. 1955. The Oslo study of untreated syphilis; an epidemiologic investigation of the
999 natural course of the syphilitic infection based upon a re-study of the Boeck-Bruusgaard
1000 material. *Acta Derm Venereol Suppl (Stockh)* 35:3-368; Annex I-LVI.
- 1001 Golden MR, Marra CM, and Holmes KK. 2003. Update on syphilis: resurgence of an old
1002 problem. *JAMA* 290:1510-1514. 10.1001/jama.290.11.1510
- 1003 Gray RT, Hoare A, McCann PD, Bradley J, Down I, Donovan B, Prestage G, and Wilson DP.
1004 2011. Will changes in gay men's sexual behavior reduce syphilis rates? *Sex Transm Dis*
1005 38:1151-1158. 10.1097/OLQ.0b013e318238b85d
- 1006 Guimarães K. August, 2016. The real infectious disease problem in Brazil isn't actually Zika, it's
1007 syphilis. Available at
1008 <https://qz.com/763105/brazil-zika-syphilis-infant-mortality/#targetText=The%20real%20infectious%20disease%20problem,t%20actually%20Zika%2C%20it's%20syphilis&targetText=In%202016%2C%20the%20government%20forecasts,that%20reported%20a%20decade%20ago.September>, 2019).
- 1009
1010
1011
- 1012 Halatoko WA, Landoh DE, Saka B, Akolly K, Layibo Y, Yaya I, Gbetoglo D, Banla AK, and
1013 Pitche P. 2017. Prevalence of syphilis among female sex workers and their clients in Togo
1014 in 2011. *BMC Public Health* 17:219. 10.1186/s12889-017-4134-x
- 1015 Hardy PH, Jr., and Levin J. 1983. Lack of endotoxin in *Borrelia hispanica* and *Treponema*
1016 *pallidum*. *Proc Soc Exp Biol Med* 174:47-52.
- 1017 Hawkes S, Matin N, Broutet N, and Low N. 2011. Effectiveness of interventions to improve
1018 screening for syphilis in pregnancy: a systematic review and meta-analysis. *The Lancet*
1019 *Infectious Diseases* 11:684-691. 10.1016/s1473-3099(11)70104-9
- 1020 Health MDo. 2019. Syphilis Treatment Protocol.
- 1021 Heffelfinger JD, Swint EB, Berman SM, and Weinstock HS. 2007. Trends in primary and
1022 secondary syphilis among men who have sex with men in the United States. *Am J Public*
1023 *Health* 97:1076-1083. 10.2105/AJPH.2005.070417
- 1024 Henao-Martinez AF, and Johnson SC. 2014. Diagnostic tests for syphilis: New tests and new
1025 algorithms. *Neurol Clin Pract* 4:114-122. 10.1212/01.CPJ.0000435752.17621.48
- 1026 Herremans T, Kortbeek L, and Notermans DW. 2010. A review of diagnostic tests for congenital
1027 syphilis in newborns. *European Journal of Clinical Microbiology & Infectious Diseases*
1028 29:495-501. 10.1007/s10096-010-0900-8
- 1029 Hood JE, Buskin SE, Dombrowski JC, Kern DA, Barash EA, Katz DA, and Golden MR. 2016.
1030 Dramatic increase in preexposure prophylaxis use among MSM in Washington state.
1031 *AIDS* 30:515-519. 10.1097/QAD.0000000000000937
- 1032 Hook EW, 3rd, Martin DH, Stephens J, Smith BS, and Smith K. 2002. A randomized,
1033 comparative pilot study of azithromycin versus benzathine penicillin G for treatment of
1034 early syphilis. *Sex Transm Dis* 29:486-490.

- 1035 Hook EW. 2017. Syphilis. *The Lancet* 389:1550-1557. 10.1016/s0140-6736(16)32411-4
- 1036 Houston S, Hof R, Honeyman L, Hassler J, and Cameron CE. 2012. Activation and proteolytic
1037 activity of the *Treponema pallidum* metalloprotease, pallilysin. *PLoS Pathog* 8:e1002822.
1038 10.1371/journal.ppat.1002822
- 1039 Houston S, Lithgow KV, Osbak KK, Kenyon CR, and Cameron CE. 2018. Functional insights
1040 from proteome-wide structural modeling of *Treponema pallidum* subspecies *pallidum*, the
1041 causative agent of syphilis. *BMC Structural Biology* 18. 10.1186/s12900-018-0086-3
- 1042 Hu J, Gu X, Tao X, Qian Y, Babu GR, Wang G, Liao M, Han L, Kang D, and Tang W. 2017.
1043 Prevalence and Trends of HIV, Syphilis, and HCV in Migrant and Resident Men Who
1044 Have Sex with Men in Shandong, China: Results from a Serial Cross-Sectional Study.
1045 *Plos One* 12:e0170443. 10.1371/journal.pone.0170443
- 1046 Hwang B, Lee J, Liu QH, Woo ER, and Lee DG. 2010. Antifungal effect of (+)-pinoresinol
1047 isolated from *Sambucus williamsii*. *Molecules* 15:3507-3516.
1048 10.3390/molecules15053507
- 1049 International journal of pharmaceutical sciences and research 2. 10.13040/ijpsr.0975-
1050 8232.2(7).1855-59
- 1051 J.B. Tapko BTaLGS. 2010. STATUS OF BLOOD SAFETY IN THE WHO AFRICAN
1052 REGION.
- 1053 Jablonka A, Solbach P, Nothdorft S, Hampel A, Schmidt RE, and Behrens GM. 2016. [Low
1054 seroprevalence of syphilis and HIV in refugees and asylum seekers in Germany in 2015].
1055 *Dtsch Med Wochenschr* 141:e128-132. 10.1055/s-0041-110627
- 1056 Jaiswal AK, Tiwari S, Jamal SB, de Castro Oliveira L, Alves LG, Azevedo V, Ghosh P, Oliveira
1057 CJF, Soares SC. 2020 The pan-genome of *Treponema pallidum* reveals differences in
1058 genome plasticity between subspecies related to venereal and non-venereal syphilis.*BMC*
1059 *Genomics*. 2020 Jan 10;21(1):33.
- 1060 Jakopanec I, Grijibovski AM, Nilsen O, Blystad H, and Aavitsland P. 2013. Trends in HIV
1061 infection surveillance data among men who have sex with men in Norway, 1995-2011.
1062 *BMC Public Health* 13:144. 10.1186/1471-2458-13-144.
- 1063 Janier M, Hegyi V, Dupin N, Unemo M, Tiplica GS, Potocnik M, French P, and Patel R. 2014.
1064 2014 European guideline on the management of syphilis. *J Eur Acad Dermatol Venereol*
1065 28:1581-1593. 10.1111/jdv.12734
- 1066 Jansen K, Schmidt AJ, Drewes J, Bremer V, and Marcus U. 2016. Increased incidence of syphilis
1067 in men who have sex with men and risk management strategies, Germany, 2015. *Euro*
1068 *Surveill* 21. 10.2807/1560-7917.ES.2016.21.43.30382
- 1069 Kanelleas A, Stefanaki C, Stefanaki I, Bezronidii G, Papparizos V, Arapaki A, Kripouri Z,
1070 Antoniou C, and Nicolaidou E. 2015. Primary syphilis in HIV-negative patients is on the
1071 rise in Greece: epidemiological data for the period 2005-2012 from a tertiary referral
1072 centre in Athens, Greece. *J Eur Acad Dermatol Venereol* 29:981-984. 10.1111/jdv.12745
- 1073 Katsila T, Spyroulias GA, Patrinos GP, and Matsoukas MT. 2016. Computational approaches in
1074 target identification and drug discovery. *Comput Struct Biotechnol J* 14:177-184.
1075 10.1016/j.csbj.2016.04.004
- 1076 Katz KA, and Klausner JD. 2008. Azithromycin resistance in *Treponema pallidum*. *Curr Opin*
1077 *Infect Dis* 21:83-91. 10.1097/QCO.0b013e3282f44772
- 1078 Kenyon C, Lynen L, Florence E, Caluwaerts S, Vandenbruaene M, Apers L, Soentjens P, Van
1079 Esbroeck M, and Bottieau E. 2014. Syphilis reinfections pose problems for syphilis
1080 diagnosis in Antwerp, Belgium - 1992 to 2012. *Euro Surveill* 19:20958.
- 1081 Kim B, Jo J, Han J, Park C, and Lee H. 2017. In silico re-identification of properties of drug
1082 target proteins. *BMC Bioinformatics* 18:248. 10.1186/s12859-017-1639-3

- 1083 Kingston M, French P, Higgins S, McQuillan O, Sukthankar A, Stott C, McBrien B, Tipple C,
1084 Turner A, Sullivan AK, Members of the Syphilis guidelines revision g, Radcliffe K,
1085 Cousins D, FitzGerald M, Fisher M, Grover D, Higgins S, Kingston M, Rayment M, and
1086 Sullivan A. 2016. UK national guidelines on the management of syphilis 2015. *Int J STD*
1087 *AIDS* 27:421-446. 10.1177/0956462415624059
- 1088 Kiss S, Damico FM, and Young LH. 2009. Ocular Manifestations and Treatment of Syphilis.
1089 *Seminars in Ophthalmology* 20:161-167. 10.1080/08820530500232092
- 1090 Kojima N, and Klausner JD. 2018. An Update on the Global Epidemiology of Syphilis. *Curr*
1091 *Epidemiol Rep* 5:24-38. 10.1007/s40471-018-0138-z
- 1092 Kojima N, Park H, Konda KA, Joseph Davey DL, Bristow CC, Brown B, Leon SR, Vargas SK,
1093 Calvo GM, Caceres CF, and Klausner JD. 2017. The PICASSO Cohort: baseline
1094 characteristics of a cohort of men who have sex with men and male-to-female transgender
1095 women at high risk for syphilis infection in Lima, Peru. *BMC Infectious Diseases* 17.
1096 10.1186/s12879-017-2332-x
- 1097 Korenromp EL, Mahiane G, Rowley J, Nagelkerke N, Abu-Raddad L, Ndowa F, El-Kettani A,
1098 El-Rhilani H, Mayaud P, Chico RM, Pretorius C, Hecht K, and Wi T. 2017. Estimating
1099 prevalence trends in adult gonorrhoea and syphilis in low- and middle-income countries
1100 with the Spectrum-STI model: results for Zimbabwe and Morocco from 1995 to 2016.
1101 *Sex Transm Infect* 93:599-606. 10.1136/sextrans-2016-052953
- 1102 Krupp K, and Madhivanan P. 2015. Antibiotic resistance in prevalent bacterial and protozoan
1103 sexually transmitted infections. *Indian J Sex Transm Dis AIDS* 36:3-8. 10.4103/0253-
1104 7184.156680
- 1105 Kumar Jaiswal A, Tiwari S, Jamal SB, Barh D, Azevedo V, and Soares SC. 2017. An In Silico
1106 Identification of Common Putative Vaccine Candidates against *Treponema pallidum*: A
1107 Reverse Vaccinology and Subtractive Genomics Based Approach. *Int J Mol Sci* 18.
1108 10.3390/ijms18020402
- 1109 Kumar Jaiswal A, Tiwari S, Jamal SB, Barh D, Azevedo V, and Soares SC. 2017. An In Silico
1110 Identification of Common Putative Vaccine Candidates against *Treponema pallidum*: A
1111 Reverse Vaccinology and Subtractive Genomics Based Approach. *Int J Mol Sci* 18.
1112 10.3390/ijms18020402
- 1113 Lafond RE, and Lukehart SA. 2006. Biological basis for syphilis. *Clin Microbiol Rev* 19:29-49.
1114 10.1128/CMR.19.1.29-49.2006
- 1115 Lopes L, Ferro-Rodrigues R, Llobet S, Lito L, and Borges-Costa J. 2016. [Syphilis: Prevalence in
1116 a Hospital in Lisbon]. *Acta Med Port* 29:52-55. 10.20344/amp.6247
- 1117 Lukehart SA. 1988. Invasion of the Central Nervous System by *Treponema pallidum*:
1118 Implications for Diagnosis and Treatment. *Annals of Internal Medicine* 109.
1119 10.7326/0003-4819-109-11-855
- 1120 Lukehart SA. 2008. Scientific Monogamy: Thirty Years Dancing with the Same Bug. *Sexually*
1121 *Transmitted Diseases* 35:2-7. 10.1097/OLQ.0b013e318162c4f2
- 1122 Mabey DC, Sollis KA, Kelly HA, Benzaken AS, Bitarakwate E, Changalucha J, Chen XS, Yin
1123 YP, Garcia PJ, Strasser S, Chintu N, Pang T, Terris-Prestholt F, Sweeney S, and Peeling
1124 RW. 2012. Point-of-care tests to strengthen health systems and save newborn lives: the
1125 case of syphilis. *PLoS Med* 9:e1001233. 10.1371/journal.pmed.1001233
- 1126 Mahendran R. 2017. "In Silico Metabolic Pathway Analysis and Docking Analysis of *Treponema*
1127 *Pallidum* Subs. *Pallidum* Nichols for Potential Drug Targets". *Asian Journal of*
1128 *Pharmaceutical and Clinical Research* 10. 10.22159/ajpcr.2017.v10i5.17367
- 1129 Malek R, Mitchell H, Furegato M, Simms I, Mohammed H, Nardone A, and Hughes G. 2015.
1130 Contribution of transmission in HIV-positive men who have sex with men to evolving

- 1131 epidemics of sexually transmitted infections in England: an analysis using multiple data
1132 sources, 2009-2013. *Euro Surveill* 20.
- 1133 Manyahi J, Jullu BS, Abuya MI, Juma J, Ndayongeje J, Kilama B, Sambu V, Nondi J, Rabiell B,
1134 Somi G, and Matee MI. 2015. Prevalence of HIV and syphilis infections among pregnant
1135 women attending antenatal clinics in Tanzania, 2011. *BMC Public Health* 15:501.
1136 10.1186/s12889-015-1848-5
- 1137 Marra CM, Colina AP, Godornes C, Tantalo LC, Puray M, Centurion-Lara A, and Lukehart SA.
1138 2006. Antibiotic selection may contribute to increases in macrolide-resistant *Treponema*
1139 *pallidum*. *J Infect Dis* 194:1771-1773. 10.1086/509512
- 1140 Mattei PL, Beachkofsky TM, Gilson RT, and Wisco OJ. 2012. Syphilis: a reemerging infection.
1141 *Am Fam Physician* 86:433-440.
- 1142 Mc Grath-Lone L, Marsh K, Hughes G, and Ward H. 2014. The sexual health of female sex
1143 workers compared with other women in England: analysis of cross-sectional data from
1144 genitourinary medicine clinics. *Sex Transm Infect* 90:344-350. 10.1136/sextrans-2013-
1145 051381
- 1146 McGettrick P, Ferguson W, Jackson V, Eogan M, Lawless M, Ciprike V, Varughese A, Coulter-
1147 Smith S, and Lambert JS. 2016. Syphilis serology in pregnancy: an eight-year study
1148 (2005-2012) in a large teaching maternity hospital in Dublin, Ireland. *Int J STD AIDS*
1149 27:226-230. 10.1177/0956462415580226
- 1150 Melku M, Kebede A, and Addis Z. 2015. Magnitude of HIV and syphilis seroprevalence among
1151 pregnant women in Gondar, Northwest Ethiopia: a cross-sectional study. *HIV AIDS*
1152 (*Auckl*) 7:175-182. 10.2147/HIV.S81481
- 1153 Meyers D, Wolff T, Gregory K, Marion L, Moyer V, Nelson H, Petitti D, Sawaya GF, and
1154 Uspstf. 2008. USPSTF recommendations for STI screening. *Am Fam Physician* 77:819-
1155 824.
- 1156 Mitchell TL, Tornelli JL, Fisher TD, Blackwell TA, and Moorman JR. 1992. Yield of the
1157 screening review of systems: a study on a general medical service. *J Gen Intern Med*
1158 7:393-397.
- 1159 Mohammed H, Mitchell H, Sile B, Duffell S, Nardone A, and Hughes G. 2016. Increase in
1160 Sexually Transmitted Infections among Men Who Have Sex with Men, England, 2014.
1161 *Emerg Infect Dis* 22:88-91. 10.3201/eid2201.151331
- 1162 Molina J-M, Capitant C, Spire B, Pialoux G, Cotte L, Charreau I, Tremblay C, Le Gall J-M, Cua
1163 E, Pasquet A, Raffi F, Pintado C, Chidiac C, Chas J, Charbonneau P, Delaugerre C,
1164 Suzan-Monti M, Loze B, Fonsart J, Peytavin G, Cheret A, Timsit J, Girard G, Lorente N,
1165 Préau M, Rooney JF, Wainberg MA, Thompson D, Rozenbaum W, Doré V, Marchand L,
1166 Simon M-C, Etien N, Aboulker J-P, Meyer L, and Delfraissy J-F. 2015. On-Demand
1167 Preexposure Prophylaxis in Men at High Risk for HIV-1 Infection. *New England Journal*
1168 *of Medicine* 373:2237-2246. 10.1056/NEJMoa1506273
- 1169 Morshed MG, Singh AE, and Papasian CJ. 2015. Recent Trends in the Serologic Diagnosis of
1170 Syphilis. *Clinical and Vaccine Immunology* 22:137-147. 10.1128/cvi.00681-14
- 1171 Mutagoma M, Remera E, Sebuho D, Kanters S, Riedel DJ, and Nsanzimana S. 2016. The
1172 Prevalence of Syphilis Infection and Its Associated Factors in the General Population of
1173 Rwanda: A National Household-Based Survey. *J Sex Transm Dis* 2016:4980417.
1174 10.1155/2016/4980417
- 1175 Newman L, Rowley J, Vander Hoorn S, Wijesooriya NS, Unemo M, Low N, Stevens G, Gottlieb
1176 S, Kiarie J, and Temmerman M. 2015. Global Estimates of the Prevalence and Incidence
1177 of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review
1178 and Global Reporting. *Plos One* 10:e0143304. 10.1371/journal.pone.0143304

- 1179 Norgard MV, Arndt LL, Akins DR, Curetty LL, Harrich DA, and Radolf JD. 1996. Activation of
1180 human monocytic cells by *Treponema pallidum* and *Borrelia burgdorferi* lipoproteins and
1181 synthetic lipopeptides proceeds via a pathway distinct from that of lipopolysaccharide but
1182 involves the transcriptional activator NF-kappa B. *Infect Immun* 64:3845-3852.
- 1183 Otieno-Nyunya B, Bennett E, Bunnell R, Dadabhai S, Gichangi AA, Mugo N, Wanyungu J, Baya
1184 I, Kaiser R, and Kenya AISST. 2011. Epidemiology of syphilis in Kenya: results from a
1185 nationally representative serological survey. *Sex Transm Infect* 87:521-525.
1186 10.1136/sextrans-2011-050026
- 1187 Ouedraogo HG, Meda IB, Zongo I, Ky-Zerbo O, Grosso A, Samadoulougou BC, Tarnagda G,
1188 Cisse K, Sondo A, Sawadogo N, Traore Y, Barro N, Baral S, and Kouanda S. 2018.
1189 Syphilis among Female Sex Workers: Results of Point-of-Care Screening during a Cross-
1190 Sectional Behavioral Survey in Burkina Faso, West Africa. *Int J Microbiol*
1191 2018:4790560. 10.1155/2018/4790560
- 1192 Owusu-Ofori AK, Parry CM, and Bates I. 2011. Transfusion-transmitted syphilis in teaching
1193 hospital, Ghana. *Emerg Infect Dis* 17:2080-2082. 10.3201/eid1711.110985
- 1194 Padovese V, Egidi AM, Melillo TF, Farrugia B, Carabot P, Didero D, Costanzo G, and Mirisola
1195 C. 2014. Prevalence of latent tuberculosis, syphilis, hepatitis B and C among asylum
1196 seekers in Malta. *J Public Health (Oxf)* 36:22-27. 10.1093/pubmed/fdt036
- 1197 Papaleo E, Naqvi AAT, Shahbaaz M, Ahmad F, and Hassan MI. 2015. Identification of
1198 Functional Candidates amongst Hypothetical Proteins of *Treponema pallidum* ssp.
1199 *pallidum*. Plos One 10. 10.1371/journal.pone.0124177
- 1200 Parekh BS, Anyanwu J, Patel H, Downer M, Kalou M, Gichimu C, Keipkerich BS, Clement N,
1201 Omondi M, Mayer O, Ou C-Y, and Nkengasong JN. 2010. Dried tube specimens: A
1202 simple and cost-effective method for preparation of HIV proficiency testing panels and
1203 quality control materials for use in resource-limited settings. *Journal of Virological*
1204 *Methods* 163:295-300. 10.1016/j.jviromet.2009.10.013
- 1205 Pastuszczyk M, and Wojas-Pelc A. 2013. Current standards for diagnosis and treatment of
1206 syphilis: selection of some practical issues, based on the European (IUSTI) and U.S.
1207 (CDC) guidelines. *Postepy Dermatol Alergol* 30:203-210. 10.5114/pdia.2013.37029
- 1208 Peeling RW, and Ye H. 2004. Diagnostic tools for preventing and managing maternal and
1209 congenital syphilis: an overview. *Bull World Health Organ* 82:439-446.
- 1210 Peeling RW, Holmes KK, Mabey D, and Ronald A. 2006. Rapid tests for sexually transmitted
1211 infections (STIs): the way forward. *Sex Transm Infect* 82 Suppl 5:v1-6.
1212 10.1136/sti.2006.024265
- 1213 Peeling RW, Mabey D, Kamb ML, Chen XS, Radolf JD, and Benzaken AS. 2017. Syphilis. *Nat*
1214 *Rev Dis Primers* 3:17073. 10.1038/nrdp.2017.73
- 1215 Perkins HA, and Busch MP. 2010. Transfusion-associated infections: 50 years of relentless
1216 challenges and remarkable progress. *Transfusion* 50:2080-2099. 10.1111/j.1537-
1217 2995.2010.02851.x
- 1218 Phang Romero Casas C, Martyn-St James M, Hamilton J, Marinho DS, Castro R, and Harnan S.
1219 2018. Rapid diagnostic test for antenatal syphilis screening in low-income and middle-
1220 income countries: a systematic review and meta-analysis. *BMJ Open* 8. 10.1136/bmjopen-
1221 2017-018132
- 1222 Pierce CS, Wicher K, and Nakeeb S. 1983. Experimental syphilis: guinea pig model. *Br J Vener*
1223 *Dis* 59:157-168.
- 1224 Radolf JD, Arndt LL, Akins DR, Curetty LL, Levi ME, Shen Y, Davis LS, and Norgard MV.
1225 1995a. *Treponema pallidum* and *Borrelia burgdorferi* lipoproteins and synthetic
1226 lipopeptides activate monocytes/macrophages. *J Immunol* 154:2866-2877.

- 1227 Radolf JD, Deka RK, Anand A, Smajs D, Norgard MV, and Yang XF. 2016. *Treponema*
1228 *pallidum*, the syphilis spirochete: making a living as a stealth pathogen. *Nat Rev*
1229 *Microbiol* 14:744-759. 10.1038/nrmicro.2016.141
- 1230 Radolf JD, Robinson EJ, Bourell KW, Akins DR, Porcella SF, Weigel LM, Jones JD, and
1231 Norgard MV. 1995b. Characterization of outer membranes isolated from *Treponema*
1232 *pallidum*, the syphilis spirochete. *Infect Immun* 63:4244-4252.
- 1233 Rappuoli R. 2001. Reverse vaccinology, a genome-based approach to vaccine development.
1234 *Vaccine* 19:2688-2691.
- 1235 Rathi B, Sarangi AN, and Trivedi N. 2009. Genome subtraction for novel target definition in
1236 *Salmonella typhi*. *Bioinformatics* 4:143-150.
- 1237 Riedner G, Rusizoka M, Todd J, Maboko L, Hoelscher M, Mmbando D, Samky E, Lyamuya E,
1238 Mabey D, Grosskurth H, and Hayes R. 2005. Single-dose azithromycin versus penicillin
1239 G benzathine for the treatment of early syphilis. *N Engl J Med* 353:1236-1244.
1240 10.1056/NEJMoa044284
- 1241 Riley BS, Oppenheimer-Marks N, Hansen EJ, Radolf JD, and Norgard MV. 1992. Virulent
1242 *Treponema pallidum* activates human vascular endothelial cells. *J Infect Dis* 165:484-493.
- 1243 Rolfs RT. 1995. Treatment of syphilis, 1993. *Clin Infect Dis* 20 Suppl 1:S23-38.
- 1244 Rubin AN, Espiridion ED, Truong NH, and Lofgren DH. 2018. Neurosyphilis Presenting with
1245 Anxiety: A Case Report. *Cureus* 10:e3020. 10.7759/cureus.3020
- 1246 Salazar JC, Cruz AR, Pope CD, Valderrama L, Trujillo R, Saravia NG, and Radolf JD. 2007.
1247 *Treponema pallidum* elicits innate and adaptive cellular immune responses in skin and
1248 blood during secondary syphilis: a flow-cytometric analysis. *J Infect Dis* 195:879-887.
1249 10.1086/511822
- 1250 Schackman BR, Neukermans CP, Fontain SN, Nolte C, Joseph P, Pape JW, and Fitzgerald DW.
1251 2007. Cost-effectiveness of rapid syphilis screening in prenatal HIV testing programs in
1252 Haiti. *PLoS Med* 4:e183. 10.1371/journal.pmed.0040183
- 1253 Sellati TJ, Bouis DA, Kitchens RL, Darveau RP, Pugin J, Ulevitch RJ, Gangloff SC, Goyert SM,
1254 Norgard MV, and Radolf JD. 1998. *Treponema pallidum* and *Borrelia burgdorferi*
1255 lipoproteins and synthetic lipopeptides activate monocytic cells via a CD14-dependent
1256 pathway distinct from that used by lipopolysaccharide. *J Immunol* 160:5455-5464.
- 1257 Sellati TJ, Waldrop SL, Salazar JC, Bergstresser PR, Picker LJ, and Radolf JD. 2001. The
1258 cutaneous response in humans to *Treponema pallidum* lipoprotein analogues involves
1259 cellular elements of both innate and adaptive immunity. *J Immunol* 166:4131-4140.
- 1260 Serwin AB, and Unemo M. 2016. Syphilis in females in Bialystok, Poland, 2000-2015. *Przegl*
1261 *Epidemiol* 70:273-280.
- 1262 Shimelis T, Lemma K, Ambachew H, and Tadesse E. 2015. Syphilis among people with HIV
1263 infection in southern Ethiopia: sero-prevalence and risk factors. *BMC Infect Dis* 15:189.
1264 10.1186/s12879-015-0919-7
- 1265 Simms I, Tookey PA, Goh BT, Lyall H, Evans B, Townsend CL, Fifer H, and Ison C. 2017. The
1266 incidence of congenital syphilis in the United Kingdom: February 2010 to January 2015.
1267 *BJOG* 124:72-77. 10.1111/1471-0528.13950
- 1268 Simon RP. 1985. Neurosyphilis. *Archives of Neurology* 42:606-613.
1269 10.1001/archneur.1985.04060060112021
- 1270 Singh AE, and Romanowski B. 1999. Syphilis: review with emphasis on clinical, epidemiologic,
1271 and some biologic features. *Clin Microbiol Rev* 12:187-209.
- 1272 Smit PW, van der Vlis T, Mabey D, Changalucha J, Mngara J, Clark BD, Andreasen A, Todd J,
1273 Urassa M, Zaba B, and Peeling RW. 2013. The development and validation of dried blood
1274 spots for external quality assurance of syphilis serology. *BMC Infectious Diseases* 13.
1275 10.1186/1471-2334-13-102

- 1276 South MA, Short DH, and Knox JM. 1964. Failure of Erythromycin Estolate Therapy in in Utero
1277 Syphilis. *JAMA* 190:70-71.
- 1278 Stary G, Klein I, Bruggen MC, Kohlhofer S, Brunner PM, Spazierer D, Mullauer L, Petzelbauer
1279 P, and Stingl G. 2010. Host defense mechanisms in secondary syphilitic lesions: a role for
1280 IFN-gamma-/IL-17-producing CD8+ T cells? *Am J Pathol* 177:2421-2432.
1281 10.2353/ajpath.2010.100277
- 1282 Stoltey JE, and Cohen SE. 2015. Syphilis transmission: a review of the current evidence. *Sexual*
1283 *Health* 12. 10.1071/sh14174
- 1284 Sun Y, Guo W, Li G, He S, and Lu H. 2018. Increased synthetic drug abuse and trends in HIV
1285 and syphilis prevalence among female drug users from 2010-2014 from Beijing, China.
1286 *Int J STD AIDS* 29:30-37. 10.1177/0956462417715174
- 1287 Svecova D, Part M, and Luha J. 2015. Increasing trend in syphilis. *Bratisl Lek Listy* 116:596-600.
- 1288 Takahashi T, Arima Y, Yamagishi T, Nishiki S, Kanai M, Ishikane M, Matsui T, Sunagawa T,
1289 Ohnishi M, and Oishi K. 2018. Rapid Increase in Reports of Syphilis Associated With
1290 Men Who Have Sex With Women and Women Who Have Sex With Men, Japan, 2012 to
1291 2016. *Sex Transm Dis* 45:139-143. 10.1097/OLQ.0000000000000768
- 1292 Tampa M, Sarbu I, Matei C, Benea V, and Georgescu SR. 2014. Brief history of syphilis. *J Med*
1293 *Life* 7:4-10.
- 1294 Tsachouridou O, Skoura L, Christaki E, Kollaras P, Sidiropoulou E, Zebekakis P, Vakirlis E,
1295 Margariti A, and Metallidis S. 2016. Syphilis on the rise: A prolonged syphilis outbreak
1296 among HIV-infected patients in Northern Greece. *Germs* 6:83-90.
1297 10.11599/germs.2016.1093
- 1298 Tsankova G, Todorova TT, Kostadinova T, Ivanova L, and Ermenlieva N. 2016. Seroprevalence
1299 of Syphilis among Pregnant Women in the Varna Region (Bulgaria). *Acta*
1300 *Dermatovenerol Croat* 24:288-290.
- 1301 Tuddenham S, and Ghanem KG. 2016. Ocular syphilis: opportunities to address important
1302 unanswered questions. *Sexually Transmitted Infections* 92:563-565. 10.1136/sextrans-
1303 2016-052570
- 1304 van der Sluis JJ. 1992. Laboratory techniques in the diagnosis of syphilis: a review. *Genitourin*
1305 *Med* 68:413-419.
- 1306 Van Voorhis WC, Barrett LK, Koelle DM, Nasio JM, Plummer FA, and Lukehart SA. 1996.
1307 Primary and secondary syphilis lesions contain mRNA for Th1 cytokines. *J Infect Dis*
1308 173:491-495.
- 1309 Vandepitte J, Bukonya J, Weiss HA, Nakubulwa S, Francis SC, Hughes P, Hayes R, and
1310 Grosskurth H. 2011. HIV and other sexually transmitted infections in a cohort of women
1311 involved in high-risk sexual behavior in Kampala, Uganda. *Sex Transm Dis* 38:316-323.
- 1312 Vijayakumari Malipatil SMA BB. 2011. Subtractive Genomics Approach For In Silico
1313 Identification Of Novel Drug Targets And Epitopes For Vaccine Design In Treponema
1314 Pallidum Subsp. Pallidum Str. Nichols
- 1315 Vilela Rodrigues TC, Jaiswal AK, de Sarom A, de Castro Oliveira L, Freire Oliveira CJ, Ghosh
1316 P, Tiwari S, Miranda FM, de Jesus Benevides L, Ariston de Carvalho Azevedo V, and de
1317 Castro Soares S. 2019. Reverse vaccinology and subtractive genomics reveal new
1318 therapeutic targets against Mycoplasma pneumoniae: a causative agent of pneumonia. *R*
1319 *Soc Open Sci* 6:190907. 10.1098/rsos.190907
- 1320 Walker DG, and Walker GJA. 2002. Forgotten but not gone: the continuing scourge of congenital
1321 syphilis. *The Lancet Infectious Diseases* 2:432-436. 10.1016/s1473-3099(02)00319-5
- 1322 WHO. 2016a. WHO guidelines for the treatment of Treponema pallidum (syphilis).
- 1323 WHO. 2017. Sexually transmitted infections: implementing the Global STI Strategy. p 8.
- 1324 WHO. 2018. Report on global sexually transmitted infection surveillance, 2018. p 63.

- 1325 Wicher K, and Wicher V. 1989. Experimental syphilis in guinea pig. *Crit Rev Microbiol* 16:181-
1326 234. 10.3109/10408418909104471
- 1327 Wong GH, Steiner B, and Graves S. 1983. Effect of syphilitic rabbit sera taken at different
1328 periods after infection on treponemal motility, treponemal attachment to mammalian cells
1329 in vitro, and treponemal infection in rabbits. *Br J Vener Dis* 59:220-224.
- 1330 Wong T, Fonseca K, Chernesky MA, Garceau R, Levett PN, and Serhir B. 2015. Canadian Public
1331 Health Laboratory Network laboratory guidelines for the diagnosis of neurosyphilis in
1332 Canada. *Can J Infect Dis Med Microbiol* 26 Suppl A:18A-22A.
- 1333 Wong T, Singh AE, and De P. 2008. Primary syphilis: serological treatment response to
1334 doxycycline/tetracycline versus benzathine penicillin. *Am J Med* 121:903-908.
1335 10.1016/j.amjmed.2008.04.042
- 1336 Workowski KA, Bolan GA, Centers for Disease C, and Prevention. 2015. Sexually transmitted
1337 diseases treatment guidelines, 2015. *MMWR Recomm Rep* 64:1-137.
- 1338 Woznicova V, Smajs D, Wechsler D, Matejkova P, and Flasarova M. 2007. Detection of
1339 *Treponema pallidum* subsp. *pallidum* from skin lesions, serum, and cerebrospinal fluid in
1340 an infant with congenital syphilis after clindamycin treatment of the mother during
1341 pregnancy. *J Clin Microbiol* 45:659-661. 10.1128/JCM.02209-06

1342

1343

1344

1345

1346

1347

Figure 1

Syphilis Transmission, Stages and Types.

The different stages of transmission of Syphilis

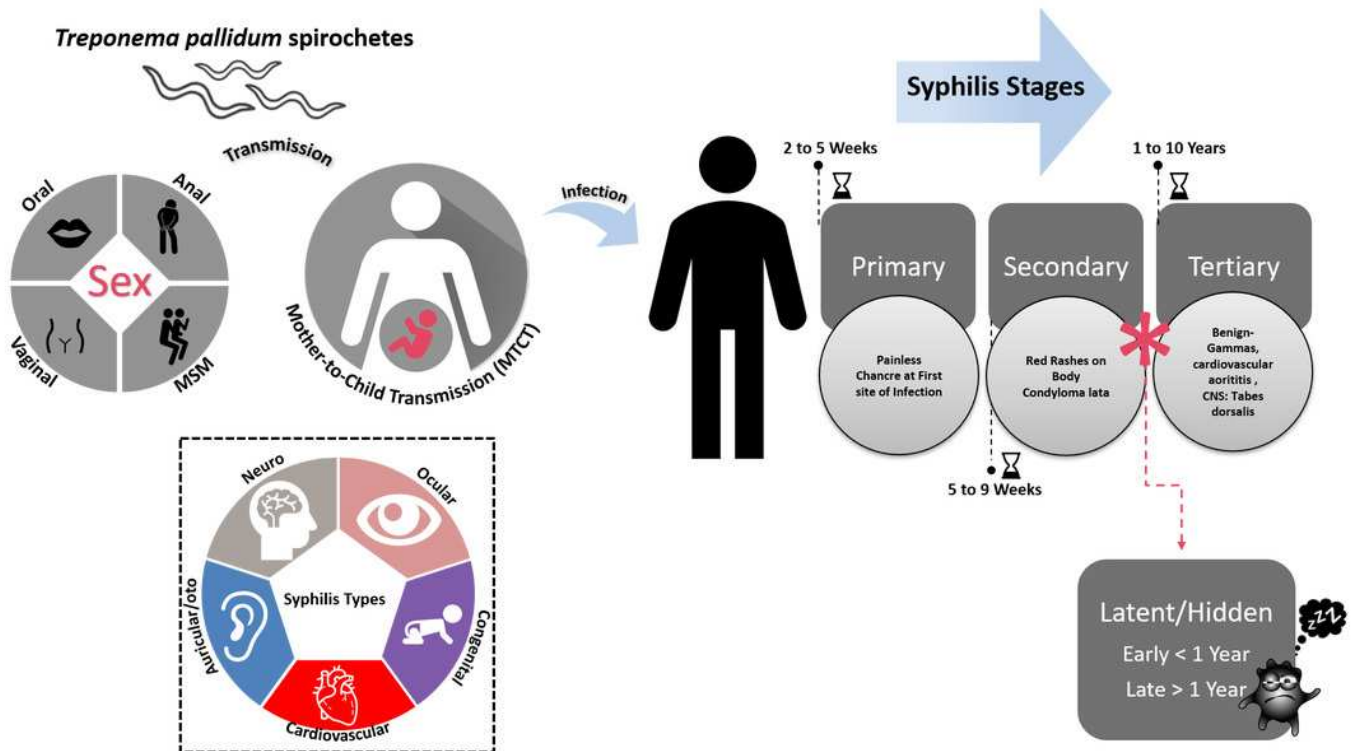


Figure 2

Algorithm for Syphilis Screening

The figure adopted from Peeling et., al. 2017 with little modification.

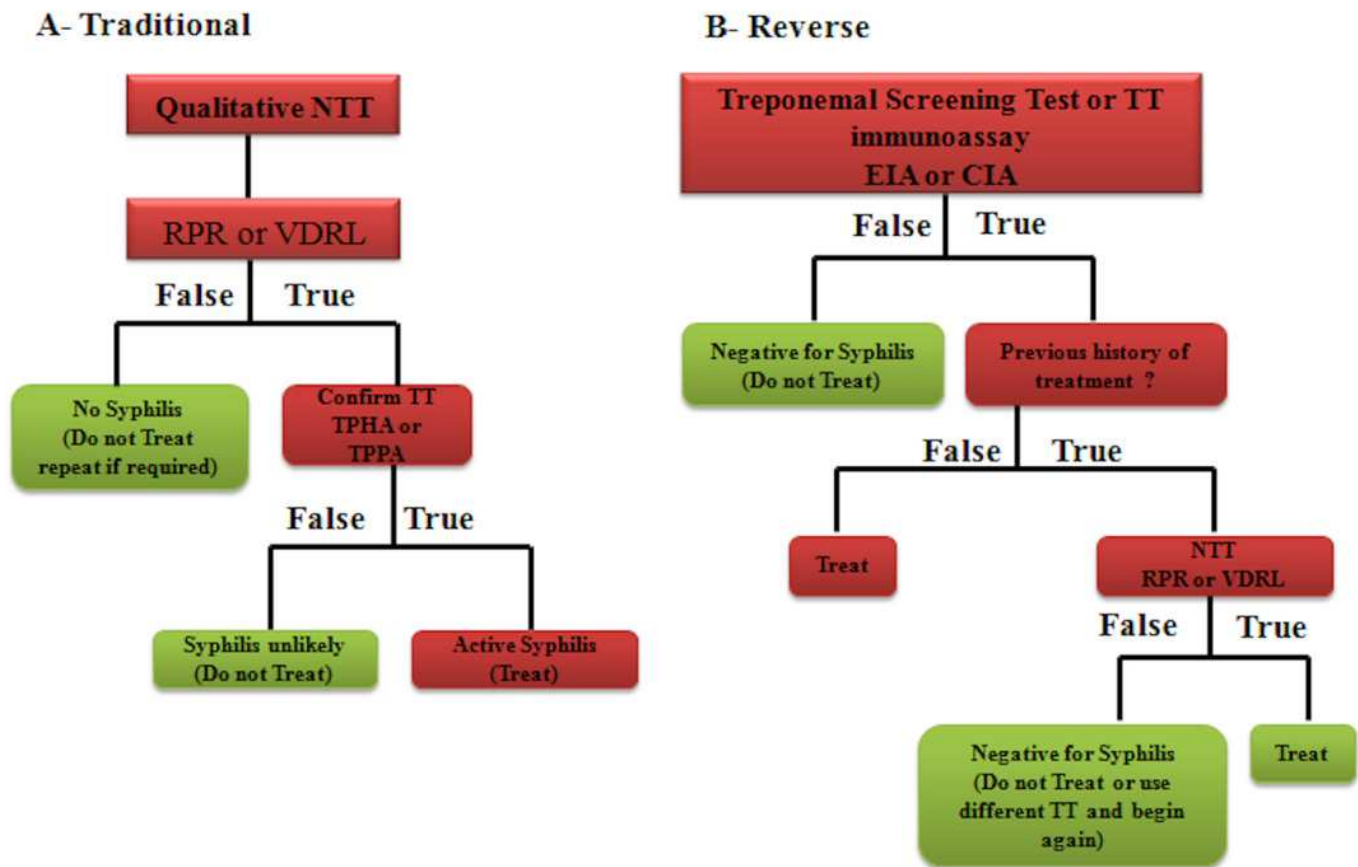


Figure 3

Molecular Architecture of *Treponema Pallidum*

:(The figure is adopted from *Peeling et al., 2017*) (Peeling et al. 2017) with modifications.

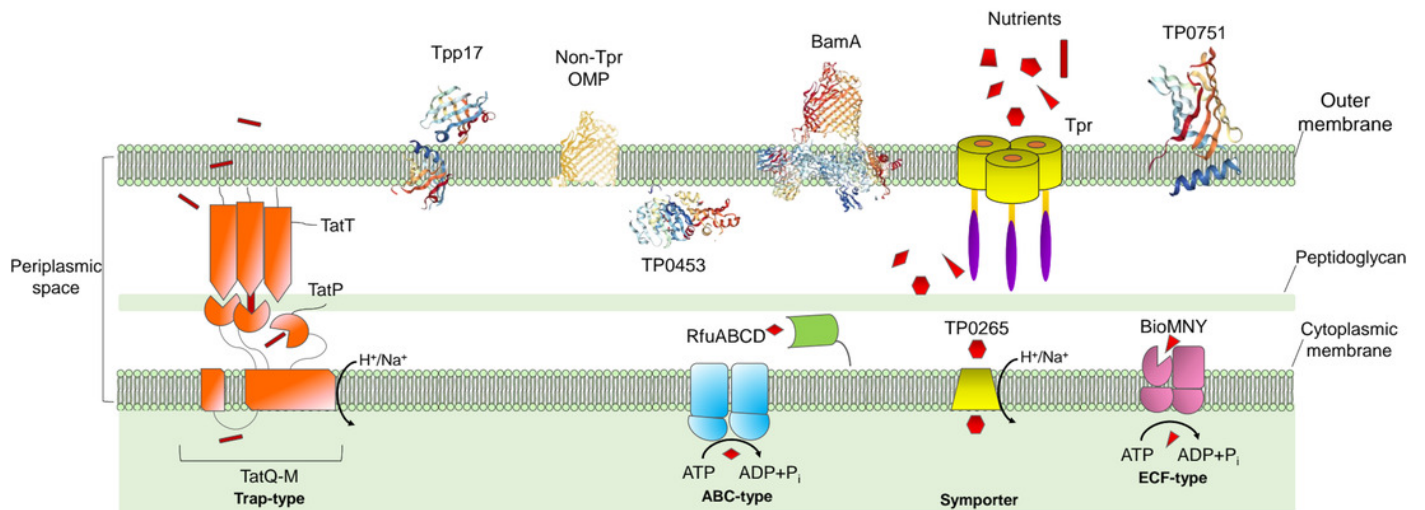


Figure 4

Schematic representation of Reverse Vaccinology approach

Schematic representation of Reverse Vaccinology approach

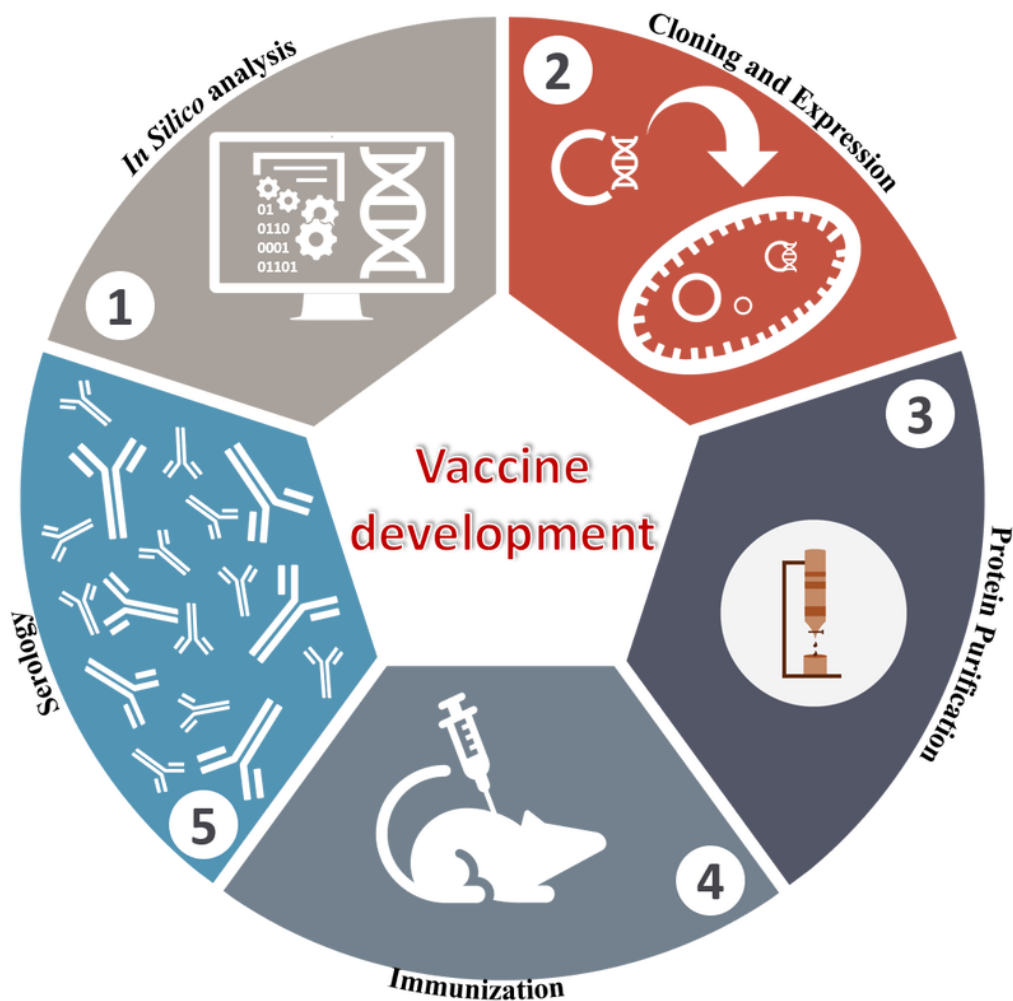


Table 1 (on next page)

List of reported therapeutic targets against *Treponema pallidum*

Table 1: List of reported therapeutic targets against *Treponema pallidum*

Pathogen	No. of Drug	No. of Vaccine	Insilico Techniques	References
<i>Treponema pallidum</i>	6- Drug targets (uvrB, Pfp, asnA, recA, Ndh, Dxs)	15- Vaccine candidates (ntpK, slyD, nlpE, ftr1, TPANIC_0600, TP_0453, tp92, TP_0323, TPANIC_0335)	Subtractive genomics and Reverse vaccinology	(Kumar Jaiswal et al. 2017)
	5- Drug targets (UDPN, DDLA, SECA, CHER, MGLB)	-----	Insilico approach and Molecular docking	(Dwivedi et al. 2015)
	9- Drug targets (TP_0108, TP_0662, TP_0168, TP_0329, TP_0817, TP_0094, TP_0476, TP_0351, TP_0350)	-----	Insilico Metabolic Pathway and Docking	(Mahendran 2017)
	3- Drug targets (Dicarboxylate transporter (dctM), Virulence factor (mviN), Cell division protein (ftsW))	3- Vaccine candidates (Dicarboxylate transporter (dctM), Virulence factor (mviN), Cell division protein (ftsW))	Subtractive genomics	(Vijayakumari Malipatil 2011)
	207- hypothetical proteins with high level of confidence		Functional characterization of Hypothetical Proteins	(Papaleo et al. 2015)
	175 uncharacterized proteins modeled with high confidence. 21- were identified as potential virulence factor		Functional annotation of Whole proteome	(Houston et al. 2018)
		TprK	The highly stable peptides found in V1 of tprK are likely promising potential vaccine components.	(Giacani et al. 2019)

III.1.2. Conclusion, Chapter 1.

A bacterium, *Treponema pallidum*, which is responsible for a sexually transmitted infection called syphilis, has affected humans throughout history, which is endemic in low-income countries and at low rates in middle-income and high-income countries. Every year, an estimated more than 6 million new cases of syphilis occur in the 15-49 age groups globally, especially in men who have sex with men (MSM) and also increase the risk of HIV infection. Considering the role of hazardous sexual conduct in driving syphilis transmission, there is a requirement for extra lucidity with regards to how best to help and support solid sexual practices among populaces in danger of syphilis. We report the rising quantities of syphilis cases in MSM and document the current efforts supporting improved medications against syphilis and also suggest some new techniques like comparative genomics, reverse vaccinology and subtractive genomics approaches which are widely used approaches for the new drug and vaccine candidates for several pathogens. These approaches may eventually lead to the discovery of new medications and immunization for the syphilis curse.

Chapter 2.

III.2.1. Book Chapter

Pan-omics focused to Crick's Central Dogma

[Arun Kumar Jaiswal*](#), Sandeep Tiwari*, Guilherme Campos Tavares, Wanderson Marques, Letícia de Castro Oliveira, Izabela Coimbra Ibraim, Luis Carlos Guimarães, Anne Cybelle Pinto Gomide, Syed Babar Jamal, Yan Pantoja, Basant K. Tiwary, Andreas Burkovski, Faiza Munir, Hai Ha Pham Thi, Nimat Ullah, Amjad Ali, Marta Giovanetti, Luiz Carlos Junior Alcantara, Jaspreet Kaur, Dipali Dhawan, Madangchanok Imchen, Ravali Krishna Vennapu, Ranjith Kumavath, Mauricio Corredor, Henrique César Pereira Figueiredo, Debmalya Barh, Vasco Azevedo, [Siomar C. Soares](#).

Book: Pan-genomics: Applications, Challenges, and Future Prospects. ELSEVIER, ISBN: 9780128170762, Published Date: 17th, January 2020

The Development of the Next-Generation Sequencing (NGS) technologies, the genome sequencing process has become cheaper and faster, making it possible the use of the technology in daily routine and as a result, the number of registered genome projects is increasing rapidly. The Pan-genome analysis is being applied in several prognosis strategies to identify the lead targets which can be used as drugs and vaccines. The pan-genome concept was coined by Tettelin and his colleagues in 2005 for *Streptococcus agalactiae*. The pan-genome consists of core-genome, shared and singletons. The core genomes belong to the set of all genes which is present in all strains of a particular species, shared set of genes belongs to more than one strain but not in all and singletons set of gene belongs to strain-specific. In this book chapter we have broadly explained the use of Pan-genomics approaches in various fields of omics.

CHAPTER 1

Pan-omics focused to Crick's central dogma

Arun Kumar Jaiswal^{*ab}, Sandeep Tiwari^{*a}, Guilherme Campos Tavares^h, Wanderson Marques da Silva^d, Letícia de Castro Oliveira^b, Izabela Coimbra Ibraim^a, Luis Carlos Guimarães^e, Anne Cybelle Pinto Gomide^a, Syed Babar Jamal^c, Yan Pantoja^e, Basant K. Tiwaryⁱ, Andreas Burkovskij^j, Faiza Munir^k, Hai Ha Pham Thi^l, Nimat Ullah^k, Amjad Ali^k, Marta Giovanetti^{a,m}, Luiz Carlos Junior Alcantara^{a,m}, Jaspreet Kaurⁿ, Dipali Dhawan^o, Madangchanok Imchen^p, Ravali Krishna Vennapu^p, Ranjith Kumavath^p, Mauricio Corredor^q, Henrique César Pereira Figueiredo^g, Debmalya Barh^f, Vasco Azevedo^a, **Siomar de Castro Soares^b**

^aPG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

^bDepartment of Immunology, Microbiology and Parasitology, Institute of Biological Science and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Brazil

^cDepartment of Biological Sciences, National University of Medical Sciences, Rawalpindi, Pakistan

^dInstitute of Agrobiotechnology and Molecular Biology, INTA-CONICET, Buenos Aires, Argentina

^eInstitute of Biological Sciences, Federal University of Pará (UFPA), Belém, Brazil

^fCentre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Purba Medinipur, India

^gAQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

^hUniversidade Nilton Lins, Manaus, Brazil

ⁱCentre for Bioinformatics, Pondicherry University, Pondicherry, India

^jFriedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

^kDepartment of Plant Biotechnology, Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

^lFaculty of Biotechnology and Environmental Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam

^mLaboratório de Flavivirus, IOC, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil,

ⁿUniversity Institute of Engineering and Technology (UIET), Department of Biotechnology, Panjab University, Chandigarh, India

^oBaylor Genetics, Houston, TX, United States

^pDepartment of Genomic Science, School of Biological Sciences, Central University of Kerala, Kasaragod, India

^qGEBIOMIC Group, FCEN, University of Antioquia, Medellín, Colombia

1 Introduction

Since the development of the first DNA sequencing technologies, many organisms had their complete DNA repertoire sequenced by Sanger and next-generation sequencing (NGS) technologies, creating the area of genomics, which was originated by the fusion of the words gene and chromosome [1]. In this scenario, a genome is the complete dataset of genes of a given organism. Nowadays, there are more than 200,000 genome projects registered at the Genome Online Database (GOLD), whereas more than 120,000 are

* These authors contributed equally to this work.

genomes isolated from bacteria (<https://gold.jgi.doe.gov/statistics>). Bacteria are widely distributed all over the world and have implications in health, agriculture, industry, and others. Besides, their genomes are small, highly compact, and do not present many repetitions, making them good targets for genome sequencing, once their genomes are easier to sequence than the ones from other organisms. Also, from the genome sequence of bacteria, it is possible to find virulence factors, antibiotic resistance genes, new therapeutic targets for vaccine and drug development, and industrially important genes [2, 3].

Another important point of the development of NGS technologies was the genome sequencing process that has become cheaper and faster, making it possible for small laboratories to use the technology in daily routine. NGS made possible the comparison of several genomes in a multipronged strategy, where phylogenomics, genome plasticity, and whole genome synteny analyses are easier to perform nowadays (Fig. 1). Also, RNA sequencing (RNA-seq) by these platforms and the development of new technologies for sequencing the complete dataset of proteins of an organism created the areas of transcriptomics and proteomics, respectively [4, 5]. Altogether, genomics is responsible for the identification of the complete dataset of genes of a given organism, whereas transcriptomics and proteomics are important for the identification of genes that are differentially expressed

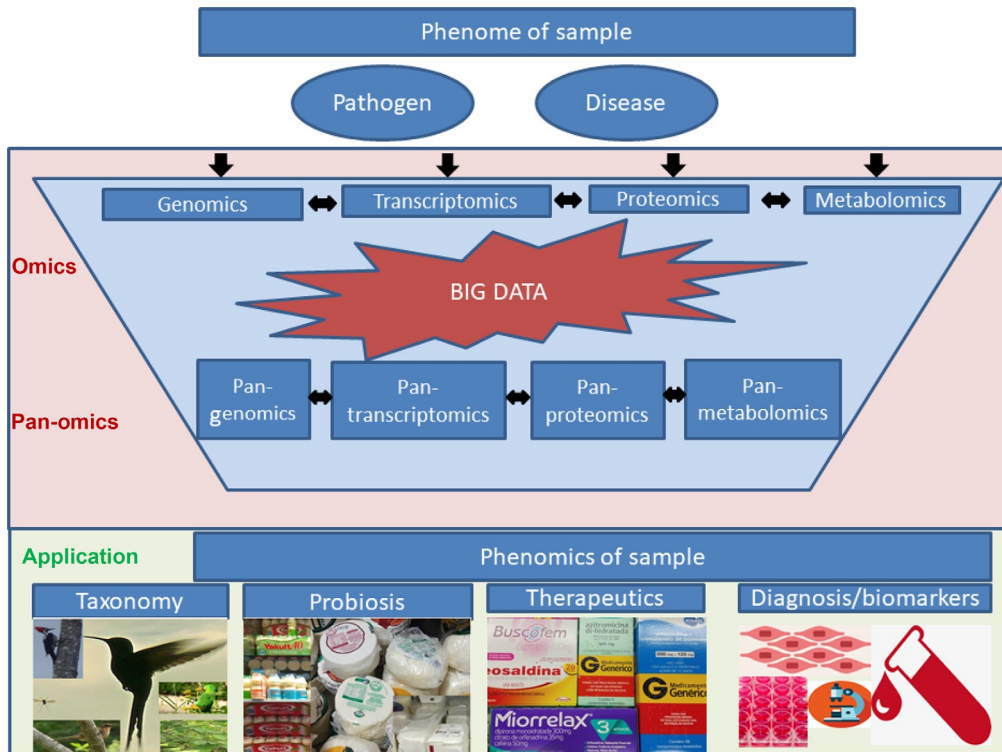


Fig. 1 Pan-omics and its applications.

between strains or species. Finally, the efforts to compare several genomes at once created the area of pan-genomics, which will be further discussed in this book.

1.1 Brief overview of pan-genomics

The term pan-genomics was created by Tettelin and collaborators, in 2005 [6], to describe the complete dataset of genes of a given species through the sequencing of several strains of this species. The pan-genome is composed of the core genome, shared genome, and singletons subsets, whereas the core genome is composed of all the commonly shared genes by all strains of the species; the shared genome contains genes that are present in two or more, but not all strains from a species; and the singletons are strain-specific genes (Fig. 2). From these subsets, one can extrapolate the data to find vaccines and drug targets from the core genome, whereas the shared genes and singletons are responsible for differences between the strains that are normally responsible for the emergence of new pathogens and the adaptation to new traits [6–10].

Normally, the core genome is composed of housekeeping genes and other genes important for metabolism and other important functions of the organism, whereas the shared genes and singletons are the result of genome plasticity. Genome plasticity is the dynamic property of DNA which involves the gain, loss, and rearrangement of genes through plasmids, phages, and genomic islands (GEIs). GEIs are huge blocks of genes acquired through horizontal gene transfer (HGT) that normally share a function in common. They are classified according to the functions of the genes into: pathogenicity islands, harboring virulence factors; metabolic islands, composed of metabolism-related genes; resistance islands, with antibiotic resistant genes; and symbiotic islands, which share in common the presence of symbiotic-related genes [11, 12].

Normally, the subsets of the pan-genome are identified by the use of orthology analyses, which first identify all orthologous genes from the complete dataset using all-vs-all blasts or other alignment search tools. Next, the datasets are classified according to their homology to genes from other strains in the subsets. After the classification, the data is plotted in a chart and mathematical formulas are used to fit the specific curves. Two such

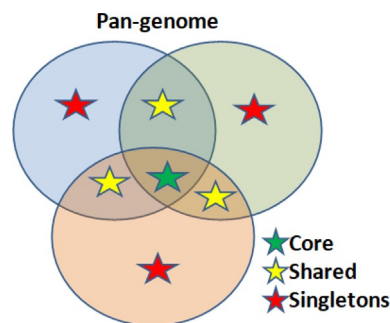


Fig. 2 Schematic representation of the core genome, shared genome, and singleton subsets of pan-genome analysis.

formulas are Heaps' law for the pan-genome development and least-squares fit of the exponential regression decay for the core genome and singleton subsets, which are described respectively as: $n = k \cdot N^{-\alpha}$, where n is the number of genes, N is the number of genomes, and k and α are constants defined by the formula; and $n = k \cdot e^{-x/\tau} + tg\theta$, where n is the number of genes, x is the number of genomes, e is Euler's number, and k , τ , and $tg\theta$ are constants defined by the formula [6, 9].

1.2 Open and closed pan-genomes

According to Heap's law, the α value is representative of the current dynamics of the pan-genome, where an α higher than 1 is representative of a closed pan-genome and an α lower than 1 represents an open pan-genome. A closed pan-genome has all possible genes represented and only few genes will be added to the pan-genome if more genomes are to be sequenced, whereas an open pan-genome is still not fully represented and the sequencing of new genomes will add many genes to the analyses [6, 9]. This definition is controversial, however, once the incorporation of GEIs may change the composition of the pan-genome drastically, even for closed pan-genomes, taking it to be open again. Most important, environmental bacteria and extracellular pathogens normally have open pan-genomes, once they still need to adapt to new traits, whereas obligate intracellular pathogens tend to have closed pan-genomes once they are not in constant contact with other bacteria. Also, intracellular pathogens have lost many genes during evolution, completely adapting to the host organism and, thus, present very compact genomes with a high percentage of essential genes [13].

According to least-squares fit of the exponential regression decay, the $tg\theta$ is representative of the number of genes present in the core genome after stabilization of the core genome curve and, also, of the number of genes that will be added to the analyses after a new genome is sequenced from the singleton development curve. Based on that, researchers may choose the species that need more strains to be sequenced and which do not. Finally, the highest the $tg\theta$ on the singleton development, the lower the α , once a high number of genes will be added to the analyses taking the pan-genome to be more open and the α to be lower (Fig. 3). The opposite is also true, the lower the $tg\theta$, the higher is the α value [6, 7, 10].

1.3 Computational methods used in pan-genomics

Computational methods to find more efficient data structures, algorithms, and statistical methods to perform bioinformatics analyses of pan-genomes have been studied because it is known that in a pan-genome analysis the greater the number of genomes taken to the analysis the greater will be the computational costs, that is, the discovery of a pan-genome content is an NP-hard problem because comparisons between all sets of genes are necessary to solve the task. Furthermore, in an effort to compute standardized pan-genome analysis and minimize computational challenges, several online tools and software suites

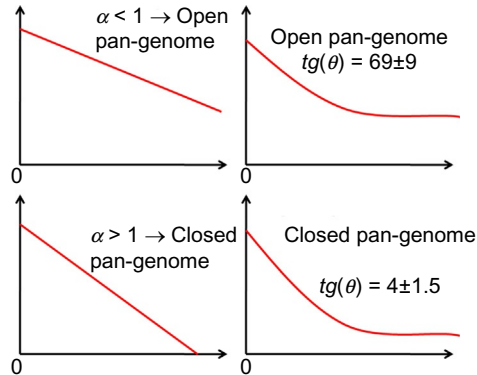


Fig. 3 The concept of open and close pan-genome.

have been developed. Examples of such applications are: PGAP [14], one of the most complete profile available for performing five analysis modules, but the runtime of the analysis grow approximately quadratically with the size of input data and are computationally infeasible with large datasets. The software Roary [15] and BPGA [16] was created to address the computational issues related to performance and execution time. Roary performs a rapid clustering of highly similar sequences, which can reduce the runtime of BLAST. BPGA is an ultrafast computational pipeline with seven functional modules for comprehensive pan-genome studies and downstream analyses. Pan-genome analysis can be applied in many different application domains, such as microbes, metagenomics, viruses, plants, cancer, and others [17]. Nowadays, the processes of similarity search and pan-genome visualization are two of the wide variety of particular computational challenges that need to be considered. For this, novel different computational methods and paradigms are needed over the years, making the computational pan-genomics a subarea of research in rapid extension. Furthermore, new technologies that are emerging in rapid development allow to infer the pan-genome with three-dimensional conformation, which means that possibly in the future three-dimensional pan-genomes will not only represent all sequence variation of the species or genus, but will also encode their spatial organization, as well as their mutual relationships in this regard.

1.4 Applications of pan-genomics in evolutionary studies

The manifestation of rich genetic diversity in the form of a pan-genome in a species is an evolutionary puzzle. These three distinct parts of a pan-genome (core, shared, and singletons) of a particular species may undergo different evolutionary trajectories under the differential influence of evolutionary forces. An ideal pan-genome is expected to be very complete, comprehensive, efficient, and stable [18]. The pan-genome of a species has some evolutionary signatures in the form of gene content and single nucleotide

polymorphism (SNP). These evolutionary signatures are useful in inferring the phylogenetic relationship among different strains of a species based on the pan-genome.

An evolutionary pan-genomic study of microbes provides a holistic picture of all the genomic variations of a species. These genomic variations endow the bacteria with their unique pathogenic properties and subsequent development of resistance to various antibiotics. Thus, a complete mechanistic detail of the processes involved in the pathogenesis and frequent antibiotic resistance in a bacterium will further pave the way for better detection methods and effective control strategies for the pathogen. In addition, evolutionary pan-genomics of a useful bacterium will help us in exploiting maximally the full potential of the microbe in enhancing industrial productivity. In fact, it will be a boom for the industries actively involved in the production of pharmaceuticals and dairy products using microbial cultures. Eukaryotes including crop plants and farm animals have abundant genomic variations in the form of SNP, copy number variants (CNVs), and presence/absence variants (PAVs). The discovery of SNPs associated with productivity or disease resistance in a crop or a farm animal will be much more efficient with the availability of a complete pan-genome of the species [19].

In a recent past, a work published by Benevides et al. [20] utilized 16S rRNA gene phylogeny, whole-genome multilocus sequence typing (wgMLST), phylogenomics, gene synteny, average nucleotide identity (ANI), and pan-genome to explain the phylogenetic relationships in a better way among strains of *Faecalibacterium*. For this, they used 12 newly sequenced, assembled, and curated genomes of *Faecalibacterium prausnitzii*, which were isolated from the feces of healthy volunteers from France and Australia, and combined these with five strains already published, which were downloaded from public databases. The phylogenetic analysis of the 16S rRNA along with the wgMLST profile and the phylogenetic tree based on the comparison of the similarity of genome supports the grouping of *Faecalibacterium* strains in different genospecies [20].

In another work published by Chen et al. [21], the comparison of whole genome and core genome multilocus sequence typing (MLST) and SNP analyses were carried out to show the maximum biased power achieved by using multiple analyses. It was required to differentiate isolates associated with outbreak from a pulsed-field gel electrophoresis (PFGE)-indistinguishable isolate collected in 2012 from a nonimplicated food source. Whole genome sequencing (WGS) has been proven as a powerful subtyping tool for bacteria like *L. monocytogenes*, a foodborne pathogen [21]. A company produced an environmental isolate that was highly similar to all outbreak isolates. The difference observed between unrelated isolates and outbreak isolates was only 7–14 SNPs; consequently, the minimum spanning tree from the analyses of whole genome, phylogenetic algorithm, and usual variant calling approach for core genome-based analyses could not offer the difference between unrelated isolates. This also suggested that the SNP/allele counts should always be pooled with WGS clustering analysis produced by phylogenetically meaningful algorithms on an adequate number of isolates, and the SNP/allele

onset alone does not provide enough evidence to demarcate an outbreak [21]. Hence, it was proposed that the comparison of pan-genome subcategories and their related α value may be utilized as an alternate approach, along with ANI, in the in silico cataloging of new species [20, 22]. We hope that the ever-expanding pan-genome across different species and genera will give impetus to a better data structure of the pan-genome and novel computational methods for a robust evolutionary pan-genomic analysis in near future.

2 Applications of Pan-genomics in Bacteria

2.1 Applications of pan-genomics in model bacteria

Advancement in sequencing technologies and development in sophisticated bioinformatics tools created an overwhelming number of microbial genomic data and allowed the scientific community to estimate the pan-genome of a species. Identification of novel dispensable genes has applications in characterizing novel metabolic pathways, virulence determinants, and molecular fingerprinting targets for epidemiological studies and core genes can be used to predict the evolutionary history of the organism [9]. Therefore, pan-genome analyses are now considered the indispensable and gold standard for bacterial genome comparisons, evolution, and diversity. It is also useful to develop a vaccine against the pathogens of epidemic diseases by filtering different functional genes in the core genome using reverse vaccinology approaches [23].

There are a number of freely accessible tools, pipelines, and web-servers available to estimate the microbial pan-genome including Roary, BPGA, PGAP, PGAPx, Panseq, PanOCT, etc. [16]. A number of model bacterial species pan-genome is determined by researchers and a vast majority of those human pathogens exhibit an open pan-genome, as they colonize multiple environments that facilitate them to exchange genetic materials. These organisms include *Escherichia coli*, *Meningococci*, *Streptococci*, *Salmonellae*, *Helicobacter pylori*, etc. [24]. Therefore, in dealing with such species a reasonable number of genomes is usually required to define the complete gene repertoire of these species. On the other hand, species living in isolated (close) habitats having less possibility to exchange genetic material tend to have closed pan-genome, for example, *Mycobacterium tuberculosis*, *B. anthracis*, and *Chlamydia trachomatis* [25]. Hence, pan-genome analyses serve as a framework to determine and understand the genomic diversity in bacterial species. In Chapter 17, we have discussed the bacterial pan-genome analysis performed till date with specific examples from model organisms along with studying approaches, technical implementations, and their outcome.

2.2 Applications of pan-genomics in *Corynebacterium diphtheriae* and *Corynebacterium ulcerans*

The development of diphtheria toxoid vaccines in the 1920s, the start of mass immunization in the 1940s, and the global introduction of the Expanded Program on

Immunization (EPI) by the World Health Organization (WHO) in 1974 led to a dramatic decrease of diphtheria cases, both in industrialized and developing countries [26]. However, despite this tremendous success story, diphtheria has not been eradicated yet. This has been illustrated dramatically by a diphtheria pandemic connected to the breakdown of the former Union of Socialist Soviet Republics with more than 157,000 cases and more than 5000 deaths reported between 1990 and 1998. Even after the pandemic has finally stopped, local breakouts have been observed constantly during the last years and the reported global cases increased from about 7000 in 2016 to almost 9000 in 2017 with a focus on countries with limited or lacking public health systems, for example India, Indonesia, Nepal, Pakistan, Venezuela, and Yemen. Consequently, *Corynebacterium diphtheriae*, the etiological agent of respiratory and cutaneous diphtheria, is still present on the list of the most important global pathogens [27]. Furthermore, the frequency of human diphtheria-like infections associated with *Corynebacterium ulcerans* appears to be increasing [28]. This species, which was recognized before as a commensal of a large number of animal species, is closely related to *C. diphtheriae* and recognized as an emerging pathogen today [28, 29].

The need of fast and unequivocal identification of especially pathogenic *C. diphtheriae* led to the early development of a number of different methods such as biovar discrimination based on different biochemical reactions, Elek's test to immunologically distinguish between toxigenic and nontoxigenic strains, restriction fragment length polymorphism (RFLP), single-strand conformation polymorphism (SSCP), phage-typing, spoligotyping, ribotyping, MLST and others. This plethora of methods was significantly improved when next-generation sequencing was introduced. The first genome sequence of *C. diphtheriae* was published in 2003 and showed the presence of the *tox* gene on a bacteriophage in addition to a number of other horizontally acquired virulence-associated genes [30]. Subsequent pan-genome studies allowed unraveling the extent of genomic diversity within *C. diphtheriae* and the role of HGT as a source of variation between strains. Furthermore, pan-genomics of *C. ulcerans* helped to estimate the virulence potential of different strains and to verify zoonotic transmission from animals to patients. Today, pan-genomics of *C. diphtheriae* and *C. ulcerans* allow elucidating global transmission traits and local adaptations of pathogenic corynebacteria and, hopefully, a better understanding of population dynamics and strain evolution will help combat diphtheria and other *Corynebacterium*-associated diseases in future.

2.3 Applications of pan-genomics in multidrug-resistant human pathogenic bacteria and pan-resistome

The pan-genome will probably be the largest molecular evolutionary history of the organism ever written. This will integrate all the pan-phenotypes existing on Earth, such as the pan-proteome, the pan-transcriptome, and especially, a portion of pan-genome that has made the organisms successful on Earth: the pan-resistome. The pan-genome

represents the set of all current genes in the genomes of a group of organisms. The basic genome common to all bacteria contains about 250 gene families in the extended core, the specific niche adaptive genome of about 8000 gene families in the character gene pool, and the pan-genomic diversity (accessory genes) of more than 139,000 rare gene families scattered throughout the bacterial genomes [31]. The pan-genome analysis, whereby the size of the gene repertoire accessible to any given species is characterized along with an estimate of the number of whole genome sequences required to proper analysis, and currently it is increasing 10 years after Tettelin et al. [6] publication. Different current models for the pan-genome analysis, accuracy, and applicability depend on the case at hand [32]. The NCBI, EMBL, KEEG, PATRIC, MGD, ENSEMBL, and JGI-IMG/M databases provide complete downloadable genomics information, which can be analyzed for intraspecies diversity, and determine the pan-genome using software tools, currently developed to perform via a personal server [32], or even online resources. The pan-genomics is now a cutting edge of computational genomics field. Pan-genomics is a subarea of computational biology [17]. Therefore, the notion of computational pan-genomics intentionally passes through many other bioinformatics-related disciplines.

The resistome, a term coined by Wright [33], comprises all the genes and their products that contribute to resist whatever environment, substance, or some extreme growth factor. Updated data will close to the metadata available for establishing what part of resistome traits belong both to core-genome as accessory genome inside all bacterial species as well as will offer a broader perspective of bacterial antibiotic resistance. The WHO summarizes antimicrobial resistance (AMR) as the resistance of a microorganism to an antimicrobial drug that was originally effective for the treatment of infections caused by themselves. An adequate approach to solving major questions about the resistome inside of the bacterial genome [34] is to perform a pan-genomics analysis. The updated pan-genome data will be close to the metadata available for establishing the part of resistome traits that belong both to core-genome as accessory genome in bacterial species; as well as a broader perspective of antibiotic resistance in bacteria. The emergent antibiotic-resistant pathogenic bacteria are a current menacing concern. *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, and coliform bacteria are the new emergent antibiotic-resistant bacteria according to the WHO. Pan-genomics has tackled some important concerns, which would be impossible to solve using classical molecular biology or descriptive genomics: it is very important to define the core and accessory genome for establishing the plasticity of resistome. Thousands of unknown bacteria and microorganisms are exposed to manufactured antibiotics, leading us to assume that there are no means to prevent this catastrophe. In opposition, pan-genomics is a powerful approach to prevent such disaster. We must move toward sequencing of known and unknown species, classify them, and establishing its antibiotic-resistant status, their pan-genome, and come out with new alternatives for reducing antibiotic consumption nowadays.

2.4 Applications of pan-genomics in veterinary pathogens

Following the development of NGS, the number of sequenced genomes filed exponentially [35]. Thus, projects aimed at studying groups of organisms became viable, and thus, several studies appeared that are called Omics studies. The studies involving pan-genomes are exposing important information on the differences and similarity between organisms of the same or between species. For concept purposes, we have the Pan-genome as a set of genes in a given group of individuals [10]. This information is being explored and applied by several scientific fronts, for example, in bacteria that infect animals and humans. The main applications of these studies are in the development of prophylactic and diagnostic methods in less time and with less cost, more precise taxonomic studies, studies on genetic variations, and pathogenesis [17]. In this chapter, we describe more recent research involving pan-genomics of the pathogenic bacteria that cause veterinary diseases, including some responsible for zoonoses, they are: *Corynebacterium pseudotuberculosis*; *Corynebacterium ulcerans*; *Streptococcus suis*; *Brachyspira hyodysenteriae*; *Moraxella bovoculi*; *Pasteurella multocida*; *Mannheimia haemolytica*; *Clostridium botulinum*; *Campylobacter*; *Streptococcus agalactiae*; *Francisella tularensis*; *Corynebacterium diphtheriae*; *Brucella* spp. Finally, it is worth highlighting that the influence of the approaches with big data and artificial intelligence are increasing and the influences of these in Pan-genomic studies will bring a new era of studies and discoveries.

2.5 Applications of pan-genomics in aquatic pathogenic bacteria

The sustainability of aquaculture industry is critical both for global food security and economic welfare. However, the massive wealth of pathogenic bacteria poses a key challenge to the development of a sustainable biocontrol method. Recent advances in genome sequencing study combined with pan-genome analysis can be an efficacious management applied to numerous aquatic pathogens [36]. Thus, routine pan genome analyses of genomic-derived aquatic pathogens will deduce the phylogenomic diversity and possible evolutionary trends of aquatic bacterial pathogen strains, elucidate the mechanisms of pathogenesis, as well as estimate patterns of pathogen transmission across epidemiological scales. The whole genome sequencing data is the opportunity to revolutionize the molecular epidemiology of aquaculture pathogens as it has for those pathogens of relevance to public health [37]. Challenges of aquaculture disease management are the biological diversity of pathogens, host-pathogen interactions (e.g., different modes of adaptation and transmission), and shifting environmental pressures, in particular climate change. Hence, analysis of pathogenic phenotype combined with genotype derived from the full potential of genome sequencing data is critical to reconstruct pathogen transmission routes on local and global scales, as well as mitigate disease emergence and spread.

Comparative pan-genome analyses are an effective tool which could possibly be extended to the analysis of aquatic microorganisms and to dynamic characteristics and adaptation to a broad range of their hosts and environmental niches. Conspicuously, our previous pan-genome analysis [38] showed that strain WFLU12 isolated from marine fish exhibited niche-specific characteristics of energy production and conversion, and carbohydrate transport and metabolism by exploring genes in the gene repertoire of strains. Based on the pan-genome categories, the functional annotations of selected genes can be reanalyzed with the Virulence Factors Database (VFDB), Clusters of Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Antibiotic Resistance Genes Database (ARDB). Also, comparative pan-genome has advanced to the point when genes are predicted as belonging to cell surface-exposed proteins (SEPs) from important pathogens, including outer membrane proteins, and extracellular proteins. These predicted genes are serving as vaccine candidates in an animal model called Reversed Vaccinology (RV) [39]. In aquaculture, SEPs from pathogens include several important virulence factors that play key roles in bacterial pathogenesis and host immune responses. For example, the expression of *esa1* from *Edwardsiella tarda*, a D15-like surface antigen, in the Japanese flounder model induced the expression of a broad spectrum of genes possibly involved in both innate and adaptive immunity, as well as a high level of fish survival and produced specific serum antibodies [40]. Vaccination using SEPs results in the development of protective effects against *Aeromonas hydrophila* infection, *Flavobacterium columnare* infection, *Pseudomonas putida* infection, and Edwardsiellosis [as in the review of Abdelgayed [41]]. A recent study [42] has successfully implemented a pan-genome analysis to screen SEPs from 17 representative *Leptospira interrogans* strains covering multiepidemic serovars from around the world, and 118 new candidate antigens were identified in addition to several known outer membrane proteins and lipoproteins. We highly consider that the rapid increase in the number of genome sequencing of aquatic pathogens will allow us to develop a rapid-response infection control protocols, but also be a potential trend for studying aquatic pathogenic bacteria to improve the cross-serotype efficacy of vaccines in farmed fish and stem the disease outbreak when implementing pan-genome analysis (using RV strategy). In the chapter “Pan-genomics of aquatic animal pathogens and its applications,” we reviewed comparative pan-genome analysis with a particular focus on controlling aquatic diseases and give real-world examples by analyzing genome sequencing data derived from aquatic bacterial isolates.

2.6 Pan-genomics applications for therapeutics

The emergence of bacterial resistance is occurring, threatening the ability of antibiotics that have transformed medicine and saved millions of lives around the globe [43, 44]. The occurrence of bacterial resistance has been identified since the beginning of the antibiotic era but the emergence of most dangerous and easily communicated strains has been

reported in past two decades [45, 46]. After several years of the first patient treated with antibiotics, bacterial infections became a threat for society once again. This situation is mainly because of the misuse and/or overuse of antibiotics as well as the inefficiency of pharmaceutical companies for not producing advanced drugs, once economic investments have been reduced [44]. The Centers for Disease Control and Prevention (CDC) has categorized several bacterial strains as an alarming threat that need serious consideration for proper treatment and are already responsible for putting significant burden on the health-care system in the United States (US), ultimately, affecting patients and their families [43, 47, 48]. The infections caused by antibiotic-resistant strains of bacteria are pervasive worldwide [43, 44]. A national survey of infectious-disease specialists led by the IDSA Emerging Infections Network in 2011 found that about two-third (2/3) of the participants had seen a pan-resistant and deadly bacterial infection within the past few years [49]. The rapid emergence of resistant bacteria has been described as a nightmare by several public health organizations that could have disastrous results [50]. The WHO cautioned in 2014 that the disaster of antibiotic resistance is becoming dreadful [51]. Among Gram-positive pathogens, a universal endemic of resistant *S. aureus* and *Enterococcus* species are presently the biggest intimidation [48]. Vancomycin-resistant enterococci (VRE) and additional emergent pathogens are evolving resistance to numerous antibiotics used commonly [43]. The worldwide distribution of common respiratory pathogens includes *Streptococcus pneumoniae* and *Mycobacterium tuberculosis*, which are reported as epidemic [48]. Gram-negative pathogens are in general more troublesome because of the fact that they are becoming more resistant to almost all the available therapeutics, making the conditions evocative to the preantibiotic era [44]. The occurrence of multidrug resistant (MDR) Gram-negative bacilli has outdated all the practice in field of medicine [43]. The most common infections caused by Gram-negative bacteria in health-care settings are usually by *Enterobacteriaceae* (mostly *Klebsiella pneumoniae*), *Acinetobacter*, and *Pseudomonas aeruginosa* [43, 44]. The evolution of bacterial strains and development of antibiotic-resistant genes through HGT make it necessary to look for novel and advanced strategies to cope with the infections [52].

The *in silico* approaches like pan-genome, pan-modelome, subtractive genomics, and reverse vaccinology are playing vital roles in rapid identification of new therapeutic targets in the postgenomic era [53–55]. Comparative microbial genomics approach along with statistical analysis are useful tools for the identification of essential genetic contents commonly present in all pathogenic isolates, based on sequence similarity. In addition to essential genetic contents, it also helps to identify subset of genes encoding virulence and novel functions as the variable genome [56]. A pan-genome is usually divided into three parts, that is, core genes, accessory genes, and strain-specific genes. In the drug and vaccine discovery process, the very first step is always the identification of a suitable target. Subtractive genomics is a widely used process in this regard. In recent past, working with pathogenic bacteria, using computational approaches, a large number of novel

therapeutic targets has been identified, which are either resistant to drugs or no appropriate vaccine is available for these targets [54, 57]. The most popular approach for rapid identification of novel vaccine targets in postgenomic era is reverse vaccinology [54]. Strategies such as comparative genomics, subtractive genomics, and differential genome analyses are being broadly utilized for the identification of targets in several human and animal pathogens (Table 1), that includes *Mycobacterium tuberculosis* [62], *Treponema pallidum* [54], *Corynebacterium diphtheriae* [53, 64], *Hemophilus ducreyi* [52], *Neisseria gonorrhoeae* [59], and *Salmonella typhi* [63]. The basic principle of these approaches is the identification of genes/proteins that are not homologous to gene/protein of the host but are essential for the survival of the pathogen. However, the identified targets might be slightly homologous to host gene/protein but still can be selected for structure-based selective inhibitor development as a supplementary molecular target [54, 64–66].

2.7 Pan-genomics applications for probiotics

The term probiotic has become highlighted in the last few years, but few know that its use is already registered as fermented foods in books such as: the Holy Bible and sacred books of Hinduism [67, 68]. Probiotics are live microorganisms that may provide health to the host [69].

Table 1 Pan-genome studies in bacterial pathogens

Name	Strain/no of strains	No of genes/proteins	Host	Therapeutic drug/vaccine targets	References
<i>Treponema pallidum</i>	13	837	Human	15 vaccine/6 drug	[54]
<i>Haemophilus ducreyi</i>	28	1257	Human	13 vaccine/3 drug	[52]
<i>Chlamydia trachomatis</i>	NC_010287.1	934	Human	63 drug	[58]
<i>Neisseria gonorrhoeae</i>	FA 1090		Human	67 drug	[59]
<i>Ureaplasma urealyticum</i>	ATCC 33699	646	Human	2 drug	[60]
<i>Corynebacterium diphtheriae</i>	13	Not mentioned	Animal	8 drug	[53]
<i>Helicobacter pylori</i>	39	59,958	Human	28 vaccine	[61]
<i>Mycobacterium tuberculosis</i>	H37Rv genome	3989	Human	135 drug	[62]
<i>Salmonella typhi</i>		4718	Human	149	[63]

Its importance gained pace in the medical and biotechnological fields with the results found not only related with inflammatory bowel diseases (IBDs) [70, 71], but also with diabetes [72], multiple sclerosis [73], dermatitis [74], and in the production of heterologous proteins [75]. Many species play a role as probiotic and much more are in the process of testing (Table 2).

Table 2 Probiotics and their effects

Name	Strain	Status	Effect	References
<i>Acinetobacter</i> sp.	BR-12	R	Plant phosphate supply	[76]
<i>Acinetobacter</i> sp.	BR-12	R	Plant phosphate supply	[77]
<i>Acinetobacter</i> sp.	WR922	R	Plant growth	[78]
<i>Bacillus amyloliquefaciens</i>	G1	R	Bacterial infections in animals	[79]
<i>Bacillus amyloliquefaciens</i>	SC06	R	Bacterial infections in animals	[80]
<i>Bacillus clausii</i>	UBBC 07	C	Acute diarrhea	[81]
<i>Bacillus coagulans</i>	–	M	Irritable bowel syndrome (IBS)	[82]
<i>Bacillus coagulans</i>	–	C	Antibiotic-induced diarrhea	[83]
<i>Bacillus licheniformis</i>	2336	M	Acute enteric infections	[84]
<i>Bacillus licheniformis</i>	26L-10/3RA	M	Bacterial infections in animals	[85]
<i>Bacillus licheniformis</i>	8-37-0-1	M	Maintenance of aquatic conditions for animals; Heavy metal accumulation	[86]
<i>Bacillus subtilis</i>	E20	M	Immuno-protection for animals	[87]
<i>Bacteroides fragilis</i>	–	R	Autism spectrum disorders (ASD)	[88]
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i>	BB-12	M	Reduces the risk of infections in early childhood	[89]
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i>	Bb-12	M	<i>H. pylori</i> related	[90]
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i>	Bb-12	C	Atopic dermatitis	[91]
<i>Enterococcus faecalis</i> (<i>Streptococcus faecalis</i>)	SL-5	C	Acne vulgaris	[92]
<i>Enterococcus faecium</i> (<i>Streptococcus faecium</i>)	CTC492	R	Antilisteral effect	[93]

Table 2 Probiotics and their effects—cont'd

Name	Strain	Status	Effect	References
<i>Escherichia coli</i>	M-17	R	Pouchitis	[94]
<i>Escherichia coli</i>	Nissle 1917	C	Ulcerative colitis; Crohn's disease; Inflammatory bowel disease (IBD)	[95–97]
<i>Lactobacillus acidophilus</i>	L-92	C	Atopic dermatitis	[98]
<i>Lactobacillus acidophilus</i>	LA-02 (DSM 21717)	C	Vulvovaginal candidiasis	[99]
<i>Lactobacillus brevis</i>	D7	M	Antioxidation process in animals	[100, 101]
<i>Lactobacillus buchneri</i>	P2	R	Cholesterol removal	[102]
<i>Lactobacillus casei</i>	DN-114001	C	Immune modulation	[103]
<i>Lactobacillus casei</i>	F-19	M	Food digestion	[104]
<i>Lactobacillus crispatus</i>	CTV-05	C	Urinary tract infection	[105]
<i>Lactobacillus delbrueckii</i> subsp. <i>bulgaricus</i>	OLL1073R-1	C	Reduces the risk of infection in the elderly	[106]
<i>Lactobacillus rhamnosus</i>	CGMCC 1.3724	C	Obesity	[107]
<i>Lactobacillus rhamnosus</i>	JCM1136	M	Immuno-protection for animals	[108]
<i>Lactococcus lactis</i> subsp. <i>cremoris</i>	IBB SC1	R	Immunomodulation	[109]
<i>Oxalobacter formigenes</i>	OxCC13	R	Calcium oxalate stone disease	[110]
<i>Propionibacterium freudenreichii</i> subsp. <i>shermanii</i>	–	C	Liver cancer	[111]
<i>Streptococcus salivarius</i>	K12	R	Halitosis	[112]
<i>Weissella koreensis</i>	OK1–6	R	Antiobesity	[113, 114]

R = research; C = Clinical trial; M = Marketed.

The Omics studies allowed an advance in the elucidation and characterization of the properties of these organisms, opening a vast field of application, besides providing new ways to access the information about their genomes. Following the pan-genomic approach, the pan-probiosis analysis consists in comparison of two or more strains, aiming to identify some points in the organism genome that differs or presents similarities related with probiotic characteristics, such as genes coding for adhesion.

In comparative genomics, for example, it is possible to retrieve a high number of genome information in silico—an attractive and cheap way [115]. There are some requirements that are important for an organism to be considered as probiotic which is determined through some mechanisms of action, like surviving to gastric acidity and bile salts [116], competing with other organisms via exclusion mechanisms and antimicrobial activity [117], and modulating the immune system [118], and these features may be used to gather the genome information in silico.

A comparative analysis with *L. lactis* subsp. *lactis* NCDO 2118 was performed aiming to find the potential probiotic characteristics of this strain. The authors found, through comparative genomics, phage regions, GEIs (metabolic and symbiotic), bacteriocins of three different classes, bile salts, and acid stress resistance genes found in other *L. lactis*, adhesion-related, and antibiotic-resistant genes. Besides that, comparing in vitro data of the aforementioned strain with another species, already described as nonprobiotic, they could identify genes encoding proteins (secreted and expressed) that are exclusive of NCDO 2118 [119].

Using a pan-genome microarray with probiotic *E. coli* isolates, Willenbrock and coauthors could characterize the pan-genome of 32 species based in two-control strain: *E. coli* K-12 and O157:H7. Despite they observed different sizes of genomes within the species, they believe they achieved the expected results, one of them being the characterization of the core genome with around 1560 essential genes [120].

Pan-genome approach was also used to discover probiotic characteristics of *L. lactis* WFLU12 [38] that showed resistance against streptococcal infection and improved the growth in olive flounder [121]. They identified some data that supported their previous work, like the identification of bacteriocins and genes involved in stress response. Comparing WFLU12 with other *L. lactis*, there are genes and gene clusters for specific niches based on carbohydrate metabolism, defense mechanisms, and envelope biogenesis [38].

Following the idea about niche-specific, Kant and coauthors worked with 13 *Lactobacillus rhamnosus* from different origins with the pan-genomic analysis. They used *L. rhamnosus* GG as reference, focusing in SEPs that may play a role in niche adaptability. The interesting thing was, they could find uncommon information in lactic acid bacteria, a *spaCBA* operon. This operon may be related with the origin of these strains, maybe of a similar microhabitat, for example [122].

Another species used as probiotic was analyzed via pan-genomics in the study by Smokvina and coauthors, in which 34 different *Lactobacillus paracasei* strains were studied using comparative genomics and pan-genomics. They identified 1800 orthologous groups representing the core genome and these genes were related with cell envelope, pili, hydrolases, or the production of branched short-chain fatty acid (SCFAs). About this, they found genes that encode these SCFAs: *bdkABCD*, only found in *Lactobacillus* until this date [123].

Nowadays, we have a lot of information about potential probiotic organisms, beyond those whose are commonly known in the market, but there is no database concentrating all the information about them, like genes related with bile juice and gastric acid resistance, genes coding adhesion, or secret proteins. A database with those information about known probiotic organism could help in future analysis be them in silico, in vivo, and in vitro. Finally, the comparative and pan-genomic analyses have an important role in the most diverse organism analyses and in the case of probiotic ones, it could be very helpful and elucidating in the precision to characterize new potential probiotics. The diversity inside the genomes may be observed and with this information it is possible to have a better idea of how many genomes will be necessary to characterize fully the organisms in these studies.

3 Pan-genomics of virus and its applications

Advances in DNA sequencing technology have ushered in a new era of pan-genomics and genomic surveillance, in which traditional molecular diagnostics and genotyping methods are being enhanced and even replaced by genomic-based methods to aid epidemiologic investigations of communicable diseases [124]. The ability to compare and analyze entire pathogen's genomes has allowed unprecedented resolution into how and why infectious diseases spread. The rapid development of sequencing technologies has made sequencing routine of viral genomes possible [125]. As these genomic-based methods continue to improve regarding speed, costs, and accuracy, they will increasingly be used to inform and guide infection control and public health practices [125a].

There are currently two major ways in which high-throughput sequencing technologies are used in public health and diagnostic applications: (i) to track outbreaks and epidemics in order to call public health responses and (ii) to characterize individual infections to tailor treatment decisions [126, 127].

Focusing on these aims, genome sequencing has been successfully used to describe unique and detailed insights into the transmission, biology, and epidemiology of many health-care-associated viral pathogens. Considering the improvements on portability and quality of sequencing, and the acceleration and standardization of analytical pipelines, the applicable routine of genome sequencing may soon become the common de facto method for infectious diseases control. Using genomic analysis tools to complement existing genotyping and epidemiologic methods, the future of infection control and prevention will lead to more targeted and successful interventions for outbreaks, which will ultimately result in the reduction of infectious diseases impact.

Next-generation sequencing techniques have transformed genomic studies from the analysis of single or few genomes to an ever-increasing amount of genomic data, bringing with it the need to develop novel techniques to treat efficiently, novel tools to assemble, analyze, and derive useful information from overwhelmingly large datasets. The analysis

of pan-genomes can uncover significant information regarding the genomes of interest. According to Guimaraes et al. [128], pan-genomic studies can help understand pathogen evolution, niche adaptation, population structure, and host interaction. Furthermore, it can help in vaccine and drug design, as well as in the identification of virulence genes.

In the context of virus investigations, pan-genomics and bioinformatics in general face great challenges. Rapid extraction of genomic features with an evolutionary signal will facilitate evolutionary analyses ranging from the reconstruction of species phylogenies to tracing epidemic outbreaks. Improvements on genome assembly using machine learning techniques are proposed by Padovani De Souza et al. [129]. Finally, in order to better use all the information acquired by high-throughput real-time sequencing and its analysis, text mining and knowledge discovery techniques, integrated with medical and scientific literature and gene family and metabolic pathway databases, could help generate new insights and speed up discoveries. High-throughput real-time next-generation sequencing projects have transformed the field of bioinformatics from single-genome studies to pan-genome analyses. The limiting factor now is no longer data rarity, but immense data availability and dimensionality. In this new context, bottom-up analyses stemming from big data provide great challenges and also great rewards.

4 Pan-genomics of plants and its applications

The plants genomes are highly dynamic as compared to many higher eukaryotes due to the presence of transposable elements and frequent genome duplication events [130]. Thus, the identification of such structural variations and dynamics in plant genomes is a prerequisite for subsequent understanding and their applications based on the sequence-trait associations. Several plant genomes were sequenced during the sequencing initiative in 2000 allowing an assembly of their reference genomes [131]. These reference genomes were mainly used to compare genomes of different plant species and to identify the SNPs across populations [132]. These studies increased our understanding regarding the allelic variations associated with phenotypic outcomes in general. However, such studies were not able to capture fully the diversity of sequence variations in plant genomes being themselves dependent on large genetic variations within strains/species. To this end, the advent of high throughput sequencing has played a major role in examining the genetic variations including SNPs, CNV, and presence/absence variations (PAV) comprehensively. The reduced costs of high-throughput sequencing methods have now revolutionized the ways being used for the analyses of plant genomes previously and for asking relevant biological questions. It has made it possible to easily sequence and compare the whole genomes of many individuals of same plants species and thus capturing the interspecies genetic diversity. Accordingly, the full genome content capturing the interspecies genomic diversity is termed as pan-genome [133]. The pan-genome approach allows to predict the number of additional genome sequences

that are necessary to characterize fully the genomic diversity of a species [133]. Analyses of pangenomes of several plants have now revealed the role of structural variations in different plant phenotypes such as flowering times, different stress-resistant mechanisms, etc. [134]. These studies have enhanced our understanding of the diverse applications of these genotypic to phenotypic association such as for increasing the crop production of better varieties in terms of size and flavors, increasing the abiotic stress and pathogens/disease resistances among many others reviewed in this chapter. The pan-genome approach is especially suitable for plant-breeding applications in contrast to the single liner reference genomes because of reduced sampling biases along with the comprehensive representation of genetic diversity [133]. The field of pan-genomics is rapidly evolving based on the underlying sequencing paradigms and the analytical pipelines, tools, and algorithms for sequencing data. The current pangenome assembly approaches can be categorized into a k-mer-based approach, comparative de-novo assembly approach, and iterative assembly approach. One of the challenges associated with the analysis of pan-genome data is related to requiring the increase in precision of the underlying genome assembly approaches. This review chapter aims to describe comprehensively the structural variations in plants genomes, explain the concept of pangenome, and its characterization along with the applications, methods, and approaches to conduct pan-genome analyses for a wide range of plant species.

4.1 Applications of pan-genomics in plant pathogens

The knowledge of plant diseases and host-pathogen interactions is one of the fundamental and active areas of genetic research with a wide array of applications [135]. Previously, linear reference genomes have been widely used for the subsequent analyses of phylogenetic relationships, identification of casual agents, virulence factors, host specificity associations, and pathogenic mechanisms [136]. These studies aided better disease management for economically important crops and plants by counteracting the stress-based resistance factors and better vaccine development. However, there is increasing evidence that the single reference genomes are insufficient in capturing the entire genetic diversity of the strains and subsequent delineation of principles governing the adaptive success of plant pathogens along with the determination of pathogenicity factors [137]. Accordingly, the concept of pan-genome emerged to cater to the interstrain genetic diversity based on different structural variations including CNV, presence/absence variations (PAV), and other allelic transformations. Pan-genome approach is now emerging as an analytical approach for analyzing the genetic diversity of genomes at an unprecedented level of details in contrast to the single reference genome. The strain-specific genome content is especially beneficial for gaining insights into the pathogenic mechanisms of plant pathogens as most of the pathogenic determinants are often strain specific and highly variable. Moreover, the pan-genome analysis allows determining the

genome plasticity through studying the evolutionary impact of HGT. As of yet, pan-genome analyses have already been used to identify and detect new strains along with development of vaccines against many plant pathogens [138]. Several computational pipelines based on tools and software especially designed to conduct a pan-genome analysis are available now. These tools can perform several functions including homologous gene clustering, SNPs identification, pan-genomic profiles visualization, phylogenetic analysis based on orthologous genes or gene families based information, pan-genome visualization, curation, and function-based searching. Most of the established pan-genome analysis methods were initially developed to deal with smaller prokaryotic genomes and thus are beneficial in analyzing most of the plant pathogens including bacteria and fungi. However, there are still certain challenges in assembling and analyzing the pangenomes of the species with complex genome structures [32]. Despite this, the pan-genome analyses is emerging as an important research tool to enhance our understanding about host-pathogen interactions and to develop universal vaccines. Since this approach has a potential for organizing pathogenic diversity, integrating pan-genomics with phylogeny and phylogenomics will be an interesting viewpoint for the future. Overall, we have comprehensively reviewed the studies conducted to assemble the pan-genomes of plant pathogens, its applications, available methods, and tools to conduct a pan-genome analysis in our chapter.

5 Genomics of algae and its applications

Genome sequencing unveils the basis of various fundamental processes and origin as well as the evolution of the organism. Advancement in whole-genome sequencing in the field of algal biomass has answered our queries of ecological and economic importance extending from the adaptation of organisms in diverse environments to synthesizing abundant metabolites of vast economical future. WGS of diverse algal genome has been performed using sequencing approaches ranging from shotgun to high throughput. Shotgun approach includes cloning 1–10 kb g-DNA fragments into pUC18 or pBlue-script II KS (Stratagene). Plasmids have been sequenced using PE BigDye Terminator/ET DYEnamic terminator kit. Sequences have been resolved using PE 377 Automated DNA Sequencers and assembled from end sequences using PHRAP (P. Green) and Consed. Primer walking has been used for gap filling. Glimmer, GeneMarks, and Critica have been used to identify ORFs in the genome. High-throughput sequencing technologies include Illumina HiSeq 2000 technology, Illumina GA II x and Solexa Genome Analyzer (Illumina) and paired reads have been assembled using a DeBruijn method or CLC Genomics Workbench tools.

This development has also initiated metagenomics and metatranscriptomics, maneuvering the expression analysis and functional assays to study intraspecies and interspecies variability among nonmodel and complex biological communities of worth.

Comparative genomics is another approach to identify the essential mechanisms of origin and evolution. Genome analysis showed that a cyanobacterium *Synechococcus* sp. strain WH8102 is nutritionally more adaptable as it has acquired more sodium-dependent transporters for the uptake of organic nitrogen and phosphorus. Reduced gene complement in marine cyanobacterium *P. marinus* SS120 is consistent with the fact that the oligotrophic marine environment where it preferentially thrives is much more stable than freshwaters [139]. There are also examples from other algal genome analysis that unveiled the adaptation strategies to thrive under harsh conditions such as *Ostreococcus tauri* that has adapted costly C4 photosynthetic pathway to acquire critical ecological advantage in the CO₂-limiting conditions of phytoplankton blooms, green alga *Chloroidium* sp. UTEX 3007 is able to survive high temperatures in deserts by accumulation of thermostable palmitic acid [140]. Also, an acidophilic green alga *Chlamydomonas eustigma* NIES-2499 has acquired phytochelatin synthase genes providing it tolerance to toxic metal ions such as cadmium [141]. *Galdieria sulphuraria* and *C. merolae* belong to the Cyanidiophyceae group but at the same time possess many contrasting features. The foremost is the ability of *G. sulphuraria* to adapt to extreme acidic thermophilic environments. It is the only alga in this group with an adaptation of the heterotrophic mode of nutrition with multiple substrates, which indicates how it survives in harsh environments [142]. In the process of evolution of ancestral lineages of red algae, the role of HGT is undeniable. This was indicated in the genome of other red algae, *Porphyridium purpureum*. Along with that, several light-harvesting complexes (LHC) were identified. Genomic analysis revealed evidence for sexual reproduction [143]. To cope with ecological stress, the genome of *P. umbilicalis* reveals the presence of genes coding for high-affinity iron transport complex necessary for the iron uptake processes to obtain nutrients during stressful high tides [144]. The study of gene sequences has also thrown light on the conservation of certain key enzymes such as GDP-mannose 6-dehydrogenase (GMD) required in the process of synthesis of alginates in brown algae *Cladosiphon okamuranus*. Also, *C. okamuranus* holds significant commercial importance as it is cultivated for fucoidan, which is a sulfated polysaccharide, a kind of Japanese seaweed [145]. The information on genomics has opened doors to various other research fields like proteomics, expression analysis, structural biology, metabolomics, etc.

6 Pan-metagenomics and human microbiome

Pan-metagenome is the collective study of all or several metagenomes from all possible units belonging to a particular type of ecosystem or host.

In the past decade, most of the metagenomic studies have aimed at understanding the microbial community from a relatively small set of samples. Such studies could miss out important rare taxa. However, the reduction in cost of gigabyte of NGS data has made the NGS application affordable and widespread [146]. This has given rise to an enormous

amount of publicly accessible data from various types of samples. The application of pan-metagenome ranges from the mosquito gut microbiome [147] to human gut microbiome [148], including various ecosystems [149, 150]. Pan-metagenome primarily aims to explore and redefine the microbial community at a global scale. This will help to capture all the taxonomical variations between samples and understand the shifts in microbial community on a larger scale.

A pan-metagenome comprising thousands of samples pertaining to an ecosystem or host from multiple locations and studies at global level collaborations could be used as a standard reference. Such a reference-based pan-metagenome could serve as a guideline to answer several questions: What types of ecosystems are most vulnerable to global warming? Are rare taxa distributed based on geography?

7 Pan-proteomics and its applications

In the proteomic approach it is possible to identify and quantify a set of proteins synthesized by a determined cell, tissue, or microorganism [151] when exposed to different experimental conditions (such as temperature, osmolarity, antibiotics, nitric oxide, and others), or different steps of the cell growth, or during infection process [151–153]. At a specific condition, the identified proteins from the complex protein mixtures may be characterized in relation to their expression, cellular localization, structure, biological functions, and interactions with other proteins, posttranslation modifications, and metabolic pathways. In this way, proteomic studies contribute to understanding about cellular adaptation in response to external changes, metabolic stresses, or infection, and this response can vary according to time and environment [154]. The proteomic analysis have been considered the most relevant approach to describe a biological system [151].

Proteomic approach in eukaryotic cells is relatively complex due to posttranslational modification, like phosphorylation of proteins, which is involved in protein signaling in different cellular pathways [155]. In humans, datasets from proteome studies have allowed to evaluate the potential methods in diagnosis, prognosis, and treatment for some diseases, including cancer [156]. On the other hand, in prokaryotes the proteomic assays have enabled the investigation of physiological behaviors, mutations, adaptability to different environmental conditions, presence of proteins involved in virulence, and the identification of putative immunogenic proteins [157].

The protein synthesis in eukaryotic and prokaryotic cells can be evaluated by different technologies, such as chromatography-based methods, enzyme-linked immunosorbent assay (ELISA), Western blotting, protein separation using gel-based approaches, especially two-dimensional (2D) polyacrylamide gel electrophoresis, or through the identification and sequencing of polypeptides through mass spectrometry technologies [151]. In chromatography-based techniques, the proteins can be obtained from separation based

on their charge nature and charge strength (ion exchange chromatography), molecular size (size exclusion chromatography), or specificity (affinity chromatography) [158]. On the other hand, ELISA uses antibodies or antigens on the solid surface to detect specific peptides or enzymes from the biological sample, forming enzyme-conjugated antibodies which allow to measure the enzyme activity or protein concentration [159]. Last, Western blotting enables the identification of low abundance proteins after electrophoresis separation, transfer onto nitrocellulose membrane, and detection by enzyme-conjugated antibodies [160]. Nevertheless, these three methodologies allow to evaluate few proteins, and they are unable to determine protein expression level [151]. 2D gel electrophoresis is an efficient and widely used technique in proteomic studies to analyze complex protein mixtures extracted especially from bacterial cells. This methodology involves separation of proteins by isoelectric focusing (proteins with different isoelectric points) and by molecular weight (in polyacrylamide gel electrophoresis). Each spot in a 2D matrix corresponds to a single protein in the sample evaluated. In this way, 2D gel electrophoresis allows to obtain information of several proteins simultaneously as apparent molecular weight, isoelectric point, and quantity of each one [161]. And, mass spectrometry can be defined as the study of matter through the formation of ions in the gas phase and their characterization by mass, charge, structure, or physicochemical properties, using mass spectrometer that measures m/z values and abundance of ions [162].

The association between 2D gel electrophoresis and mass spectrometry was already considered the most appropriate method to recognize and identify proteins from pathogenic microorganisms [163] for being a methodology used for the construction of proteomic databases, due to its greater efficiency and high resolution to investigate the complex mixtures of proteins present in cell or tissues [164]. Nevertheless, with the technical advances achieved in recent years, such as solubilization of complex samples, pH gradient, and detection of proteins present in small quantities, the technique of liquid chromatography associated with mass spectrometry (LC-MS) started to be used and allowed the analysis of complex mixtures of proteins by tryptic digestion without prior gel separation [165]. This technique had the advantage of having a low detection limit for peptides and proteins, capability to identify hundreds to thousands of proteins in a simple experiment as well as allowing the study of membrane proteins, poorly accessible by other methods [166].

LC-MS is divided into two approaches: stable isotopic labeling [167] and label-free quantification [168]. In the first, two solutions containing the proteins to be analyzed are labeled with different molecular mass isotopes, and are mixed, trypsin-digested to obtain peptides and submitted to the LC-MS system [169]. The molecular weight difference allows the identification and quantification of peptides of both samples tested [170], but the labeling occurs after the extraction step, which can lead to a reduction in the precision of the quantification method [171]. Alternatively, label-free quantification allows the evaluation of numerous samples at the same time within the LC-MS system, with

data-independent acquisition, and the concentration of a given peptide is proportional to its chromatographic area [172].

Among the strategies used in proteomic studies in prokaryotic cells surfome and secretome analyses stand out. The bacterial surface has been considered of great importance for understanding the pathogenesis of an infectious disease. On the surface, it can be found that proteins are associated with mechanisms of defense and virulence factors, which can promote adhesion and cellular invasion, culminating consequently in the appearance of clinical signs in an infected host [173]. Therefore, surfome is a proteome-based method, in which allows the identification of bacterial surface proteins [174]. Apart from surface proteins, extracellular and secreted proteins are important in bacterial pathogenesis, since they also mediate the interaction of the bacterium with the host and by stimulating the immune response. Therefore, the secretome has been associated with adhesion, invasion, immune evasion, and spread of bacterium in host tissues. In addition, these proteins can also be used for the development of antibiotics and vaccines [175]. Besides these two methods, comparative proteome analysis has been used for both prokaryotic and eukaryotic cells. This method has also been used to identify virulence factors and to obtain information on physiological and environmental adaptations in different pathogens [176], as well as to compare cells, tissues, and organs from the eukaryotic host in normal and pathological (inflammation, infection, and cancer) conditions [156].

In this context, pan-proteomics is also an approach with characterizes and compares the qualitative and quantitative proteome; however, the comparison occurs across organisms inside a species, with genetic variation and phenotype [177]. Pan-proteomics can be performed using 2D gel electrophoresis or LC-MS; nevertheless, LC-MS by bottom-up/shotgun techniques, from our expertise, is recommended for this type of study, otherwise, we will always have only part of the proteome and not the whole proteome.

Conceptually, pan-proteome refers to the proteins identified from a whole set of samples/strains tested, which are usually more than two samples, under the same experimental conditions. The analysis of two samples is equivalent to comparative proteomic methodology. Pan-proteome can be divided into core proteome, accessory proteome, and orphan (or unique) proteome [177]. The core proteome represents the subset of identified proteins simultaneously in all samples, whereas accessory proteome represents the detected proteins shared by at least two samples, and orphan proteome represents proteins identified exclusively in a single sample.

In the microbiological field, the genetic variation among isolates has been implicated with virulence factors, drug resistance, and environmental adaptation [178]. In this way, understanding about these mechanisms needs the evaluation of several proteomes and not from single proteome analysis [177]. Thus, pan-proteomic analysis may increase knowledge about the adaptation and pathogenicity of a given microorganism, independent of

the genotype. Besides that, this approach can be used to classify bacterial strains in types [179], identify putative vaccine targets from conserved proteins among isolates [178], as well as, to determine drug targets and drug mode of action in analysis with multiple strains [177].

The term pan-proteome and core proteome have been used in different studies of protein identification and quantitation. In this type of study, pan-proteome and core proteome were referenced in the first time from analysis of four epidemic *Salmonella* Paratyphi A strains, with different PFGE types, using 2D gel electrophoresis [180]. From this analysis, the authors verified a high covered (over than 81%) of core proteome among the isolates tested, regardless of the range of pH applied, suggesting a high similarity in protein expression. Proteins involved in metabolic pathways and survival of the bacterium were the most identified within the core proteome. Moreover, the proteome comparison among isolates suggest a geospatial and temporal differentiation of expressed protein profile (spots).

The conserved core proteome was also observed in other works, where this category represented approximately 92% of pan-proteome of five fish-adapted *Streptococcus agalactiae* strains, which belonged to three MLST profiles. This study was performed using a label-free proteomic analysis [178]. The authors suggest that the identified proteins reflect an adaptation to an aquatic environment and fish-pathogen interaction. In addition, in the same study, conserved antigenic proteins were identified and suggested as targets in vaccine design, seeing that the high degree of conservation of these proteins among the isolates would suggest the production of a monovalent vaccine effective against all genetic variants tested.

Another study, despite the conservation of proteins identified simultaneously in avirulent, virulent, and two clinical strains of *Mycobacterium tuberculosis*, the quantitative protein expression profiling revealed a strain-specific variation in proteome patterns of isolates [181]. This study was also performed using label-free analysis, being identified 257 differentially expressed proteins. The differences in virulence among four isolates were suggested to a two-component system, oxidative stress, ribosome biogenesis, energy generation, and transcriptional regulator proteins.

The pan-proteomic analysis of four biotechnological *Lactococcus lactis* strains was performed using label-free analysis and showed a conservation of 52% of core proteome. The identified proteins contribute to physiological adaptation of bacteria, metabolic pathways, microbial metabolism in diverse environment, and proteins involved in post-translational modification, which enable maintenance of cellular integrity and physiological process bacterial during adverse environmental conditions, like temperature and oxidative stress. In this way, the authors suggested that with the results found it would be possible to increase the biotechnological potential of *L. lactis* [182].

On the other hand, in eukaryotic cells, the term pan and core proteome was used in a comparative proteomic analysis of *Gammarus* female reproductive systems (ovaries).

Nevertheless, in this study the authors verified a core proteome relatively low among the three amphipods belonging to *Gammarus* genus [183], identifying proteins involved in cellular process, localization, catalytic activity, and binding. Nevertheless, proteins involved in reproductive process were little found due to the absence of their sequences in the database used.

For the success of pan-proteomic experiments, it is necessary to be attentive as to: sample preparation, being important an optimization of the protocols of protein extraction from the multistrain or multiclinical samples; types of data acquisition from gel-based or gel-free methods; construction of pan-proteome database containing all possible proteins, including the same protein but with sequence variation, to use during searching for peptide identification; and better understand the biological functions of the identified proteins through bioinformatics analysis. All these points were extensively revised in a previous study [177].

8 Pan-transcriptomics and its applications

Transcriptome profiling is a powerful approach to identify and quantify the entire repertoire of transcripts in a cell, including mRNAs, noncoding RNAs, and small RNAs, during specific developmental stages or conditions [184]. Transcriptome analysis has enabled the study of the functional elements of the genome, increasing our understanding of the transcriptional dynamics of biological processes and disease development [185]. Among the various technologies that have been developed for high-throughput transcriptome analyses, microarray and RNA-seq are at the forefront of large-scale genome transcriptome profiling [186]. Microarray is a hybridization-based approach developed in the mid-1990s that measure the abundance of a known set of genes using an array of complementary probes. Microarray is a cost-effective, easy to analyze approach that remains the most extensively used methodology in the scientific community. RNA microarrays are generated using complementary DNA (cDNA) immobilized on a glass slide, where each cDNA fragment represents an individual gene of interest. RNA arrays have been used to identify regulated genes, pathways, networks, biological mechanisms, and processes in a variety of biological conditions [187].

However, since its commercial availability, RNA-seq has been widely applied to identify genes within a genome or to measure the expression of transcripts in an organism in different tissues, conditions, and time points [188]. RNA-seq has many advantages over array-based technology, including a high level of data reproducibility, detection of low abundant transcripts, and identification of isoforms over a wider dynamic range. Moreover, the technology does not depend on existing genome data or annotation, allowing the identification and quantification of novel transcripts [189]. Generating data on RNA transcripts require RNA to be first isolated from the experimental organism, following synthesis of cDNA, PCR amplification of cDNA transcripts, and deep sequencing [188].

Following the increased number of high-throughput RNA data, a wide range of strategies for transcriptome analysis has emerged, ranging from single cell to comparative pan-transcriptomic analysis. The pan-transcriptomics analysis consists of a comparison between complete sets of RNA transcripts, under specific circumstances, aiming to identify genes that are differentially expressed in distinct or related populations, or in response to different treatments to better understand the functional and structural aspect of genes. The integration and collective analysis of transcriptome data has enabled the identification of core and distinct molecular responses that functionally reflect the phenotypical diversity of a specific group or condition including patterns of expression associated with parasitism [190], construction of co-expression networks of differentially expressed genes encoding virulence factors [191], the identification of universal biomarkers of cellular senescence [192], comprehensive analysis of molecular alterations across multiple cancer types [193], and the characterization of tissue-specific expression of long noncoding RNAs (lncRNAs) [194]. Pan-transcriptome analysis is particularly applicable in prokaryotes and has been proven valuable in shedding light on gene expression and transcriptome organization among bacterial groups where the difference in phenotypes cannot be explained by the genome sequences alone [195] (Table 3). Moreover, a comparative approach using high-throughput studies can also show the molecular basis of pathogenicity, orthologous biological features, virulence factors, and signaling pathways responsible for stress tolerance and pathogen resistance of related surrogate bacterial species as well as within larger groups of the bacterial domain (Table 3). In addition, integrated analysis can aid the search for potential targets that can be used in the development of therapeutic strategies against relevant pathogens.

Table 3 Pan-transcriptome studies in prokaryotes

Species	Strains/isolates	Approach	Conditions/remarks	References
<i>Mycobacterium tuberculosis</i> and <i>Mycobacterium bovis</i>	Mtb H37Rv Mtb H37Rv Mtb H37Rv Mbovis AF2122/97 Mbovis AF2122/97 Mtb H37Rv	Microarray	Bacterial response to aerobic chemostat, low oxygen chemostat—0.2% DOT, aerobic rolling, batch culture, aerobic chemostat, aerobic rolling batch culture, harvested from macrophages	[196]
<i>Bacillus subtilis</i>	BR16 BR17 16BCE	Microarray	Bacterial stringent response by mimicking isoleucine and leucine starvation	[197]

Continued

Table 3 Pan-transcriptome studies in prokaryotes—cont'd

Species	Strains/isolates	Approach	Conditions/remarks	References
<i>Acinetobacter baumannii</i>		RNA-seq	Dynamics of gene expression in the transcriptomic response of drug resistance multidrug-resistant strains and sensitive strains	[198]
<i>Campylobacter jejuni</i>	NCTC11168 81-176 81,116 RM1221	RNA-seq	Comparative analysis of regulatory elements between four isolates	[195]
<i>Pseudomonas aeruginosa</i>	PA14	RNA-seq	Identification of phenotypic variability among bacteria dependent on gene expression in response to different environments including growth within biofilms, at various temperatures, growth phases, osmolarities, phosphate, and iron concentrations, under anaerobic conditions, attached to a surface, and conditions encountered within the eukaryotic host	[199]
<i>Mycobacterium tuberculosis</i>	TKK-01-0084 TKK-01-0025 TKK-01-0033 TKK-01-0040	RNA-seq	Identification of novel transcriptional mechanisms of drug resistance in Mtb strains	[200]
<i>Escherichia coli</i>	EPEC1 EPEC5 EPEC7	RNA-seq	Investigate the global transcriptional responses of the enteropathogenic <i>E. coli</i> (EPEC) and enterotoxigenic <i>E. coli</i> (ETEC) using 7 isolates	[201]

9 Pan-cancer analysis and its applications

Pan-cancer analysis has enabled in identifying the molecular aspects underlying cancer thereby benefiting diagnosis, prevention, and therapy for patients. One of the major applications of the pan-cancer data is for drug development by ranking drug targets that can be further exploited to develop targeted therapies for cancer. Further analysis of the data is needed for understanding gene-gene interactions and roles of genetic variants affecting pathways. Extensive research has been done to elucidate the underlying mechanisms of cancer occurrence and progression [202–204]. However, most of the studies are conducted independently on smaller sample sizes, thereby limiting the essence of information that needs to come out of such studies. The numerous projects involved in pan-cancer analysis generated huge volumes of data using various technologies including high-end molecular genetics and cytogenetics techniques. Various web tools have been developed and used to interpret the large amount of data generated by the pan-cancer projects [205]. The International Cancer Genome Consortium hence made a group of researchers conducting such cancer analysis across various tumor types in order to generate a pan-cancer atlas [206]. Data generated through these projects will enable in understanding the molecular aspects of cancer occurrence and further help in cancer prevention and designing cancer therapeutics. There are certain challenges that need to be overcome for the development of clinical trial strategies to connect tumor subsets from diverse tissue types [207].

10 Conclusions

The emergence of NGS technologies and the use of the data generated by these technologies for comparative genomics is a major advancement in understanding the diversity of genomes. There are effective examples of pan-genomic studies in various fields of research. The concept of pan-genomics is so deep that it has been perfectly applied in the studies of several organisms and diseases, for example, in the study of dynamics of biological processes and disease development, identification of therapeutic targets against deadly and emerging pathogens, and in the development of new probiotics. It has great potential, which may bring a closer understanding and help combat prokaryotic and eukaryotic diseases in a better way. Finally, several other fields of research that use pan-genomic idea exist, such as pan-cancer, pan-genomics of plants, virus and fungi, pan-metabolomics, and others. All those fields will be further discussed in the following chapters.

References

- [1] J.M. Heather, B. Chain, The sequence of sequencers: the history of sequencing DNA, *Genomics* 107 (2016) 1–8.
- [2] E.S. Donkor, Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics, *Genes (Basel)* 4 (2013) 556–572.

- [3] M. Land, L. Hauser, S.R. Jun, I. Nookaew, M.R. Leuze, T.H. Ahn, T. Karpinets, O. Lund, G. Kora, T. Wassenaar, S. Poudel, D.W. Ussery, Insights from 20 years of bacterial genome sequencing, *Funct. Integr. Genom.* 15 (2015) 141–161.
- [4] J.W. Prokop, T. May, K. Strong, S.M. Bilinovich, C. Bupp, S. Rajasekaran, E.A. Worthey, J. Lazar, Genome sequencing in the clinic: the past, present, and future of genomic medicine, *Physiol. Genom.* 50 (2018) 563–579.
- [5] J. Zhang, R. Chiodini, A. Badr, G. Zhang, The impact of next-generation sequencing on genomics, *J. Genet. Genom.* 38 (2011) 95–109.
- [6] H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: the bacterial pan-genome, *Curr. Opin. Microbiol.* 11 (2008) 472–477.
- [7] D. Medini, C. Donati, H. Tettelin, V. Masignani, R. Rappuoli, The microbial pan-genome, *Curr. Opin. Genet. Dev.* 15 (2005) 589–594.
- [8] A.J. Van Tonder, S. Mistry, J.E. Bray, D.M.C. Hill, A.J. Cody, C.L. Farmer, K.P. Klugman, A. Von Gottberg, S.D. Bentley, J. Parkhill, K.A. Jolley, M.C.J. Maiden, A.B. Brueggemann, Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Comput. Biol.* 10 (8) (2014) e1003788 <https://doi.org/10.1371/journal.pcbi.1003788>.
- [9] L. Rouli, V. Merhej, P.E. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria, *New Microbes New Infect.* 7 (2015) 72–85.
- [10] H. Tettelin, V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, S.V. Angiuoli, J. Crabtree, A.L. Jones, A.S. Durkin, R.T. Deboy, T.M. Davidsen, M. Mora, M. Scarselli, Y. Margarit, I. Ros, J.D. Peterson, C.R. Hauser, J.P. Sundaram, W.C. Nelson, R. Madupu, L.M. Brinkac, R.J. Dodson, M.J. Rosovitz, S.A. Sullivan, S.C. Daugherty, D.H. Haft, J. Selengut, M.L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K.J. O’connor, S. Smith, T.R. Utterback, O. White, C.E. Rubens, G. Grandi, L.C. Madoff, D.L. Kasper, J.L. Telford, M.R. Wessels, R. Rappuoli, C.M. Fraser, Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome” *Proc. Natl. Acad. Sci. U. S. A.* 102 (2005) 13950–13955.
- [11] S.C. Soares, V.A. Abreu, R.T. Ramos, L. Cerdeira, A. Silva, J. Baumbach, E. Trost, A. Tauch, R. Hirata Jr., A.L. Mattos-Guaraldi, A. Miyoshi, V. Azevedo, PIPS: pathogenicity island prediction software, *PLoS ONE* 7 (2012) e30848.
- [12] S.C. Soares, H. Geyik, R.T. Ramos, P.H. De Sa, E.G. Barbosa, J. Baumbach, H.C. Figueiredo, A. Miyoshi, A. Tauch, A. Silva, V. Azevedo, GIPSy: genomic island prediction software, *J. Biotechnol.* 232 (2016) 2–11.
- [13] M. De Barse, A. Frandi, G. Panis, L. Theraulaz, T. Pillonel, G. Greub, P.H. Viollier, Regulatory (pan-) genome of an obligate intracellular pathogen in the PVC superphylum, *ISME J.* 10 (2016) 2129–2144.
- [14] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, J. Yu, PGAP: pan-genomes analysis pipeline, *Bioinformatics* 28 (2012) 416–418.
- [15] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, M.T. Holden, M. Fookes, D. Falush, J.A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, *Bioinformatics* 31 (2015) 3691–3693.
- [16] N.M. Chaudhari, V.K. Gupta, C. Dutta, BPGA- an ultra-fast pan-genome analysis pipeline, *Sci. Rep.* 6 (2016) 24373.
- [17] C. Computational Pan-Genomics, Computational pan-genomics: status, promises and challenges, *Brief. Bioinform.* 19 (2018) 118–135.
- [18] Computational Pan-Genomics Consortium, Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* 19 (1) (2018) 118–135, <https://doi.org/10.1093/bib/bbw089>.
- [19] B. Hurgobin, D. Edwards, SNP discovery using a pangenome: has the single reference approach become obsolete? *Biology (Basel)* 6 (1) (2017) pii: E21.
- [20] L. Benevides, S. Burman, R. Martin, V. Robert, M. Thomas, S. Miquel, F. Chain, H. Sokol, L.G. Bermudez-Humarán, M. Morrison, P. Langella, V.A. Azevedo, J.M. Chatel, S. Soares, New insights into the diversity of the genus *Faecalibacterium*, *Front. Microbiol.* 8 (2017) 1790.
- [21] Y. Chen, Y. Luo, H. Carleton, R. Timme, D. Melka, T. Muruvanda, C. Wang, G. Kastanis, L.S. Katz, L. Turner, A. Fritzing, T. Moore, R. Stones, J. Blankenship, M. Salter, M. Parish,

- T.S. Hammack, P.S. Evans, C.L. Tarr, M.W. Allard, E.A. Strain, E.W. Brown, Whole genome and core genome multilocus sequence typing and single nucleotide polymorphism analyses of *Listeria monocytogenes* associated with an outbreak linked to cheese, United States, 2013. *Appl. Environ. Microbiol.* 83 (15) (2017) e00633–17 <https://doi.org/10.1128/AEM.00633-17>.
- [22] G. Vernikos, D. Medini, D.R. Riley, H. Tettelin, Ten years of pan-genome analyses, *Curr. Opin. Microbiol.* 23 (2015) 148–154.
- [23] H. Tettelin, The bacterial pan-genome and reverse vaccinology, *Genome Dyn.* 6 (2009) 35–47.
- [24] O. Lukjancenko, T.M. Wassenaar, D.W. Ussery, Comparison of 61 sequenced *Escherichia coli* genomes, *Microb. Ecol.* 60 (2010) 708–720.
- [25] V. Periwal, A. Patowary, S.K. Vellarikkal, A. Gupta, M. Singh, A. Mittal, S. Jeyapaul, R.K. Chauhan, A.V. Singh, P.K. Singh, P. Garg, V.M. Katoch, K. Katoch, D.S. Chauhan, S. Sivasubbu, V. Scaria, Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of *Mycobacterium tuberculosis* pan-genome, *PLoS ONE* 10 (2015) e0122979.
- [26] M.W. Tiwari, Diphtheria toxoid, in: Plotkin's Vaccines, seventh ed., Elsevier, 2017.
- [27] M. Hessling, J. Feiertag, K. Hoenes, Pathogens provoking most deaths worldwide: A review, *Biosci. Biotechnol. Res. Commun.* 10 (2017) 1–7.
- [28] E. Hacker, C.A. Antunes, A.L. Mattos-Guaraldi, A. Burkovski, A. Tauch, *Corynebacterium ulcerans*, an emerging human pathogen, *Future Microbiol.* 11 (2016) 1191–1208.
- [29] A. Burkovski, Pathogenesis of *Corynebacterium diphtheriae* and *Corynebacterium ulcerans*, in: *Human Emerging and Re-emerging Infections*, Wiley, 2015, pp. 699–709 Print ISBN: 9781118644713, Online ISBN: 9781118644843.
- [30] A.M. Cerdeno-Tarraga, A. Efstratiou, L.G. Dover, M.T. Holden, M. Pallen, S.D. Bentley, G.S. Besra, C. Churcher, K.D. James, A. De Zoysa, T. Chillingworth, A. Cronin, L. Dowd, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Moule, M.A. Quail, E. Rabinowitsch, K.M. Rutherford, N.R. Thomson, L. Unwin, S. Whitehead, B.G. Barrell, J. Parkhill, The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129, *Nucleic Acids Res.* 31 (2003) 6516–6523.
- [31] P. Lapierre, J.P. Gogarten, Estimating the size of the bacterial pan-genome, *Trends Genet.* 25 (2009) 107–110.
- [32] J. Xiao, Z. Zhang, J. Wu, J. Yu, A brief review of software tools for pangenomics, *Genom. Proteom. Bioinform.* 13 (2015) 73–76.
- [33] G.D. Wright, The antibiotic resistome: the nexus of chemical and genetic diversity, *Nat. Rev. Microbiol.* 5 (2007) 175–186.
- [34] M.R. Gillings, Evolutionary consequences of antibiotic use for the resistome, mobilome and microbial pangenome, *Front. Microbiol.* 4 (2013) 4.
- [35] M.L. Metzker, Sequencing technologies—the next generation, *Nat. Rev. Genet.* 11 (2010) 31–46.
- [36] S. Ghatak, J. Blom, S. Das, R. Sanjukta, K. Puro, M. Mawlong, I. Shakuntala, A. Sen, A. Goesmann, A. Kumar, S.V. Ngachan, Pan-genome analysis of *Aeromonas hydrophila*, *Aeromonas veronii* and *Aeromonas caviae* indicates phylogenomic diversity and greater pathogenic potential for *Aeromonas hydrophila*, *Antonie Van Leeuwenhoek* 109 (2016) 945–956.
- [37] S.C. Bayliss, D.W. Verner-Jeffreys, K.L. Bartie, D.M. Aanensen, S.K. Sheppard, A. Adams, E.J. Feil, The promise of whole genome pathogen sequencing for the molecular epidemiology of emerging aquaculture pathogens, *Front. Microbiol.* 8 (2017) 121.
- [38] T.L. Nguyen, D.-H. Kim, Genome-wide comparison reveals a probiotic strain *Lactococcus lactis* WFLU12 isolated from the gastrointestinal tract of olive flounder (*Paralichthys olivaceus*) harboring genes supporting probiotic action, *Mar. Drugs* 16 (5) (2018) pii: E140.
- [39] M. Dalsass, A. Brozzi, D. Medini, R. Rappuoli, Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery, *Front. Immunol.* 10 (2019) 113.
- [40] Y. Sun, C.S. Liu, L. Sun, Construction and analysis of the immune effect of an *Edwardsiella tarda* DNA vaccine encoding a D15-like surface antigen, *Fish Shellfish Immunol* 30 (2011) 273–279.
- [41] M.Y. Abdelgayed, Y.G. Alkhateib, A.M. Laila, S.Z. Mona, DNA-based vaccines against bacterial fish diseases: trials and prospective, *Rep. Opinion* 9 (2017) 1–16.
- [42] L. Zeng, D. Wang, N. Hu, Q. Zhu, K. Chen, K. Dong, Y. Zhang, Y. Yao, X. Guo, Y.F. Chang, Y. Zhu, A novel pan-genome reverse vaccinology approach employing a negative-selection strategy for screening surface-exposed antigens against leptospirosis, *Front. Microbiol.* 8 (2017) 396.

- [43] Z. Golkar, O. Bagasra, D.G. Pace, Bacteriophage therapy: a potential solution for the antibiotic resistance crisis, *J. Infect. Dev. Ctries.* 8 (2014) 129–136.
- [44] C.L. Ventola, The antibiotic resistance crisis: part 1: causes and threats, *P T* 40 (2015) 277–283.
- [45] P.C. Appelbaum, 2012 and beyond: potential for the start of a second pre-antibiotic era? *J. Antimicrob. Chemother.* 67 (2012) 2062–2068.
- [46] R.J. Fair, Y. Tor, Antibiotics and bacterial resistance in the 21st century, *Perspect. Medicin. Chem.* 6 (2014) 25–64.
- [47] B.D. Lushniak, Antibiotic resistance: a public health crisis, *Public Health Rep.* 129 (2014) 314–316.
- [48] G.M. Rossolini, F. Arena, P. Pecile, S. Pollini, Update on the antibiotic resistance crisis, *Curr. Opin. Pharmacol.* 18 (2014) 56–60.
- [49] B. Spellberg, D.N. Gilbert, The future of antibiotics and resistance: a tribute to a career of leadership by John Bartlett, *Clin. Infect. Dis.* 59 (Suppl 2) (2014) S71–S75.
- [50] V.K. Viswanathan, Off-label abuse of antibiotics by bacteria, *Gut Microbes* 5 (2014) 3–4.
- [51] C.A. Michael, D. Dominey-Howes, M. Labbate, The antimicrobial resistance crisis: causes, consequences, and management, *Front. Public Health* 2 (2014) 145.
- [52] A. De Sarom, A. Kumar Jaiswal, S. Tiwari, L. De Castro Oliveira, D. Barh, V. Azevedo, C. Jose Oliveira, S. De Castro Soares, Putative vaccine candidates and drug targets identified by reverse vaccinology and subtractive genomics approaches to control *Haemophilus ducreyi*, the causative agent of chancroid, *J. R. Soc. Interface* 15 (142) (2018) 20180032.
- [53] S.B. Jamal, S.S. Hassan, S. Tiwari, M.V. Viana, L.J. Benevides, A. Ullah, A.G. Turjanski, D. Barh, P. Ghosh, D.A. Costa, A. Silva, R. Rottger, J. Baumbach, Azevedo, V.a.C., An integrative in-silico approach for therapeutic target identification in the human pathogen *Corynebacterium diphtheriae*, *PLoS ONE* 12 (2017) e0186401.
- [54] A. Kumar Jaiswal, S. Tiwari, S.B. Jamal, D. Barh, V. Azevedo, S.C. Soares, An in silico identification of common putative vaccine candidates against *Treponema pallidum*: a reverse vaccinology and subtractive genomics based approach, *Int. J. Mol. Sci.* (2017) 18.
- [55] C.D. Rinaudo, J.L. Telford, R. Rappuoli, K.L. Seib, Vaccinology in the genome era, *J. Clin. Invest.* 119 (2009) 2515–2525.
- [56] T. Bhardwaj, P. Somvanshi, Pan-genome analysis of *Clostridium botulinum* reveals unique targets for drug development, *Gene* 623 (2017) 48–62.
- [57] D. Barh, S. Tiwari, N. Jain, A. Ali, A.R. Santos, A.N. Misra, V. Azevedo, A. Kumar, In silico subtractive genomics for target identification in human bacterial pathogens, *Drug Dev. Res.* 72 (2011) 162–177.
- [58] A. Praveena, R. Sindhuja, V. Anuradha, S.K.M. Habeeb, Putative drug target identification for *Chlamydia trachomatis*: an insilico proteome analysis, *Int. J. Biomed. Res.* 2 (2011) 151–160.
- [59] D. Barh, A. Kumar, In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*, *In Silico Biol.* 9 (2009) 225–231.
- [60] S. Madagi, V. Malipatil, Putative drug targets in *Ureaplasma urealyticum* serovar 10 str. ATCC 33699 by insilico genomics approach and virtual screening, *Int. J. Pharma Bio Sci.* 4 (2013) 8.
- [61] A. Ali, A. Naz, S.C. Soares, M. Bakhtiar, S. Tiwari, S.S. Hassan, F. Hanan, R. Ramos, U. Pereira, D. Barh, H.C.P. Figueiredo, D.W. Ussery, A. Miyoshi, A. Silva, V. Azevedo, 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 (2015) 1–17.
- [62] S.M. Asif, A. Asad, A. Faizan, M.S. Anjali, A. Arvind, K. Neelesh, K. Hirdesh, K. Sanjay, Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis, *Bioinformation* 4 (2009) 245–248.
- [63] B. Rathi, A.N. Sarangi, N. Trivedi, Genome subtraction for novel target definition in *Salmonella typhi*, *Bioinformation* 4 (2009) 143–150.
- [64] S.S. Hassan, S.B. Jamal, L.G. Radusky, S. Tiwari, A. Ullah, J. Ali, Behramand, P. De Carvalho, R. Shams, S. Khan, H.C.P. Figueiredo, D. Barh, P. Ghosh, A. Silva, J. Baumbach, R. Rottger, A.G. Turjanski, V.A.C. Azevedo, The druggable pocketome of *Corynebacterium diphtheriae*: a new approach for in silico putative druggable targets, *Front. Genet.* 9 (2018) 44.

- [65] D. Barh, N. Jain, S. Tiwari, B.P. Parida, V. D'afonseca, L. Li, A. Ali, A.R. Santos, L.C. Guimaraes, S. De Castro Soares, A. Miyoshi, A. Bhattacharjee, A.N. Misra, A. Silva, A. Kumar, 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.* 78 (2011) 73–84.
- [66] S.S. Hassan, S. Tiwari, L.C. Guimaraes, S.B. Jamal, E. Folador, N.B. Sharma, S. De Castro Soares, S. Almeida, A. Ali, A. Islam, F.D. Povoia, V.A. De Abreu, N. Jain, A. Bhattacharya, L. Juneja, A. Miyoshi, A. Silva, D. Barh, A. Turjanski, V. Azevedo, R.S. Ferreira, Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Corynebacterium pseudotuberculosis*, *BMC Genom.* 15 (Suppl 7) (2014) S3.
- [67] D.J. Bibel, Elie Metchnikoff's *Bacillus of Long Life*, *ASM News* (1988) 661–665.
- [68] A. Hosono, Fermented milk in the orient, in: Y. Naga Sawa, A. Hosono (Eds.), *Functions of fermented milk. Challenges for the health sciences*, 1992. Elsevier Applied Science.
- [69] A.W. FAO, *Guidelines for the Evaluation of Probiotics in Food*, Food and Agriculture Organization of the United Nations, 2002.
- [70] R. Bibiloni, R.N. Fedorak, G.W. Tannock, K.L. Madsen, P. Gionchetti, M. Campieri, C. De Simone, R.B. Sartor, VSL#3 probiotic-mixture induces remission in patients with active ulcerative colitis, *Am. J. Gastroenterol.* 100 (2005) 1539–1546.
- [71] A. Tursi, G. Brandimarte, A. Papa, A. Giglio, W. Elisei, G.M. Giorgetti, G. Forti, S. Morini, C. Hassan, M.A. Pistoia, M.E. Modeo, S. Rodino, T. D'amico, L. Sebkova, N. Sacca, E. Di Giulio, F. Luzza, M. Imeneo, T. Larussa, S. Di Rosa, V. Annese, S. Danese, A. Gasbarrini, Treatment of relapsing mild-to-moderate Ulcerative Colitis with the probiotic VSL#3 as adjunctive to a standard pharmaceutical treatment: a double-blind, randomized, Placebo-Controlled Study, *Am. J. Gastroenterol.* 105 (2010) 2218–2227.
- [72] F. Calcinaro, S. Dionisi, M. Marinaro, P. Candeloro, V. Bonato, S. Marzotti, R.B. Corneli, E. Ferretti, A. Gulino, F. Grasso, C. De Simone, U. Di Mario, A. Falorni, M. Boirivant, F. Dotta, Oral probiotic administration induces interleukin-10 production and prevents spontaneous autoimmune diabetes in the non-obese diabetic mouse, *Diabetologia* 48 (2005) 1565–1575.
- [73] D. Unutmaz, S. Lavasani, B. Dzhabazov, M. Nouri, F. Fåk, S. Buske, G. Molin, H. Thorlacius, J. Alenfall, B. Jeppsson, B. Weström, A novel probiotic mixture exerts a therapeutic effect on experimental autoimmune encephalomyelitis mediated by IL-10 producing regulatory T cells, *PLoS ONE* 5 (2) (2010) e9009.
- [74] M. Viljanen, E. Pohjavuori, T. Haatela, R. Korpela, M. Kuitunen, A. Sarnesto, O. Vaarala, E. Savilahti, Induction of inflammation as a possible mechanism of probiotic effect in atopic eczema-dermatitis syndrome, *J. Allergy Clin. Immunol.* 115 (2005) 1254–1259.
- [75] A. Miyoshi, E. Jamet, J. Commissaire, P. Renault, P. Langella, V. Azevedo, A xylose-inducible expression system for *Lactococcus lactis*, *FEMS Microbiol. Lett.* 239 (2004) 205–212.
- [76] M.T. Islam, A. Deora, Y. Hashidoko, A. Rahman, T. Ito, S. Tahara, Isolation and identification of potential phosphate solubilizing bacteria from the rhizosphere of *Oryza sativa* L. cv. BR29 of Bangladesh, *Z. Naturforsch. C* 62 (2007) 103–110.
- [77] D. Thakuria, N.C. Talukdar, C. Goswami, S. Hazarika, R.C. Boro, M.R. Khan, Characterization and screening of bacteria from rhizosphere of rice grown in acidic soils of Assam, *Curr. Sci.* 86 (7) (2004) 978–985.
- [78] M. Ogut, F. Er, N. Kandemir, Phosphate solubilization potentials of soil *Acinetobacter* strains, *Biol. Fertil. Soils* 46 (2010) 707–715.
- [79] H. Cao, S. He, R. Wei, M. Diong, L. Lu, *Bacillus amyloliquefaciens* G1: a potential antagonistic bacterium against eel-pathogenic *Aeromonas hydrophila*, *Evid. Based Complement. Alternat. Med.* 2011 (2011) 1–7.
- [80] J. Ji, S. Hu, W. Li, Probiotic *Bacillus amyloliquefaciens* SC06 prevents bacterial translocation in weaned mice, *Indian J. Microbiol.* 53 (2013) 323–328.
- [81] M.R. Sudha, S. Bhonagiri, M.A. Kumar, Efficacy of *Bacillus clausii* strain UBBC-07 in the treatment of patients suffering from acute diarrhoea, *Benefic. Microbes* 4 (2013) 211–216.
- [82] H.A. Hong, L.H. Duc, S.M. Cutting, The use of bacterial spore formers as probiotics: Table 1, *FEMS Microbiol. Rev.* 29 (2005) 813–835.

- [83] M. La Rosa, G. Bottaro, N. Gulino, F. Gambuzza, F. Di Forti, G. Ini, E. Tornambe, Prevention of antibiotic-associated diarrhea with *Lactobacillus sporogens* and fructo-oligosaccharides in children. A multicentric double-blind vs placebo study, *Minerva Pediatr.* 55 (2003) 447–452.
- [84] N.M. Gracheva, A.F. Gavrilov, A.I. Solov'eva, V.V. Smirnov, I.B. Sorokulova, S.R. Reznik, N.V. Chudnovskaia, The efficacy of the new bacterial preparation biosporin in treating acute intestinal infections, *Zh. Mikrobiol. Epidemiol. Immunobiol.* 1 (1996) 75–77.
- [85] P. Pattnaik, S. Grover, V.K. Batish, Effect of environmental factors on production of lichenin, a chromosomally encoded bacteriocin-like compound produced by *Bacillus licheniformis* 26L-10/3RA, *Microbiol. Res.* 160 (2005) 213–218.
- [86] C. Liu, J. Lu, L. Lu, Y. Liu, F. Wang, M. Xiao, Isolation, structural characterization and immunological activity of an exopolysaccharide produced by *Bacillus licheniformis* 8-37-0-1, *Bioresour. Technol.* 101 (2010) 5528–5533.
- [87] D.-Y. Tseng, P.-L. Ho, S.-Y. Huang, S.-C. Cheng, Y.-L. Shiu, C.-S. Chiu, C.-H. Liu, Enhancement of immunity and disease resistance in the white shrimp, *Litopenaeus vannamei*, by the probiotic, *Bacillus subtilis* E20, *Fish Shellfish Immunol.* 26 (2009) 339–344.
- [88] J.A. Gilbert, R. Krajmalnik-Brown, D.L. Porazinska, S.J. Weiss, R. Knight, Toward effective probiotics for autism and other neurodevelopmental disorders, *Cell* 155 (2013) 1446–1448.
- [89] M. Saxelin, S. Tynkkynen, T. Mattila-Sandholm, W.M. De Vos, Probiotic and other functional microbes: from markets to mechanisms, *Curr. Opin. Biotechnol.* 16 (2005) 204–211.
- [90] K.Y. Wang, S.N. Li, C.S. Liu, D.S. Perng, Y.C. Su, D.C. Wu, C.M. Jan, C.H. Lai, T.N. Wang, W.M. Wang, Effects of ingesting *Lactobacillus*- and *Bifidobacterium*-containing yogurt in subjects with colonized *Helicobacter pylori*, *Am. J. Clin. Nutr.* 80 (2004) 737–741.
- [91] C.K. Dotterud, O. Storror, R. Johnsen, T. Øien, Probiotics in pregnant women to prevent allergic disease: a randomized, double-blind trial, *Br. J. Dermatol.* 163 (2010) 616–623.
- [92] B.S. Kang, J.-G. Seo, G.-S. Lee, J.-H. Kim, S.Y. Kim, Y.W. Han, H. Kang, H.O. Kim, J.H. Rhee, M.-J. Chung, Y.M. Park, Antimicrobial activity of enterocins from *Enterococcus faecalis* SL-5 against *Propionibacterium acnes*, the causative agent in *acne vulgaris*, and its therapeutic effect, *J. Microbiol.* 47 (2009) 101–109.
- [93] T. Aymerich, M.G. Artigas, M. Garriga, J.M. Monfort, M. Hugas, Effect of sausage ingredients and additives on the production of enterocin A and B by *Enterococcus faecium* CTC492. Optimization of in vitro production and anti-listerial effect in dry fermented sausages, *J. Appl. Microbiol.* 88 (2000) 686–694.
- [94] B. Olle, Medicines from microbiota, *Nat. Biotechnol.* 31 (2013) 309–315.
- [95] W. Kruis, Maintaining remission of ulcerative colitis with the probiotic *Escherichia coli* Nissle 1917 is as effective as with standard mesalazine, *Gut* 53 (2004) 1617–1623.
- [96] H.A. Malchow, Crohn's disease and *Escherichia coli*. A new approach in therapy to maintain remission of colonic Crohn's disease? *J. Clin. Gastroenterol.* 25 (1997) 653–658.
- [97] A. Sturm, K. Rilling, D.C. Baumgart, K. Gargas, T. Abou-Ghazale, B. Raupach, J. Eckert, R.R. Schumann, C. Enders, U. Sonnenborn, B. Wiedenmann, A.U. Dignass, *Escherichia coli* Nissle 1917 distinctively modulates T-cell cycling and expansion via toll-like receptor 2 signaling, *Infect. Immun.* 73 (2005) 1452–1465.
- [98] Y. Inoue, T. Kambara, N. Murata, J. Komori-Yamaguchi, S. Matsukura, Y. Takahashi, Z. Ikezawa, M. Aihara, Effects of oral administration of *Lactobacillus acidophilus* L-92 on the symptoms and serum cytokines of atopic dermatitis in Japanese adults: a double-blind, randomized, clinical trial, *Int. Arch. Allergy Immunol.* 165 (2014) 247–254.
- [99] F. Murina, A. Graziottin, F. Vicariotto, F. De Seta, Can *Lactobacillus fermentum* LF10 and *Lactobacillus acidophilus* LA02 in a slow-release vaginal product be useful for prevention of recurrent vulvovaginal candidiasis? *J. Clin. Gastroenterol.* 48 (2014) S102–S105.
- [100] Y.-J. Lai, S.-H. Tsai, M.-Y. Lee, Isolation of exopolysaccharide producing *Lactobacillus* strains from sorghum distillery residues pickled cabbage and their antioxidant properties, *Food Sci. Biotechnol.* 23 (2014) 1231–1236.
- [101] N. Waki, N. Yajima, H. Suganuma, B.M. Buddle, D. Luo, A. Heiser, T. Zheng, Oral administration of *Lactobacillus brevis* KB290 to mice alleviates clinical symptoms following influenza virus infection, *Let. Appl. Microbiol.* 58 (2014) 87–93.

- [102] X.Q. Zeng, D.D. Pan, Y.X. Guo, The probiotic properties of *Lactobacillus buchneri* P2. *J. Appl. Microbiol.* 108 (6) (2010) 2059–2066, <https://doi.org/10.1111/j.1365-2672.2009.04608.x>.
- [103] A. Marcos, J. Wärnberg, E. Nova, S. Gómez, A. Alvarez, R. Alvarez, J.A. Mateos, J.M. Cobo, The effect of milk fermented by yogurt cultures plus *Lactobacillus casei* DN-114001 on the immune response of subjects under academic examination stress. *Eur. J. Nutr.* 43 (2004) 381–389.
- [104] R.J. Siezen, G. Wilson, Probiotics genomics, *Microb. Biotechnol.* 3 (2010) 1–9.
- [105] A.E. Stapleton, M. Au-Yeung, T.M. Hooton, D.N. Fredricks, P.L. Roberts, C.A. Czaja, Y. Yarova-Yarovaya, T. Fiedler, M. Cox, W.E. Stamm, Randomized, placebo-controlled phase 2 trial of a *Lactobacillus crispatus* probiotic given intravaginally for prevention of recurrent urinary tract infection, *Clin. Infect. Dis.* 52 (2011) 1212–1217.
- [106] S. Makino, S. Ikegami, A. Kume, H. Horiuchi, H. Sasaki, N. Orii, Reducing the risk of infection in the elderly by dietary intake of yoghurt fermented with *Lactobacillus delbrueckii* ssp. *bulgaricus* OLL1073R-1, *Br. J. Nutr.* 104 (2010) 998–1006.
- [107] M. Sanchez, C. Darimont, V. Drapeau, S. Emady-Azar, M. Lepage, E. Rezzonico, C. Ngom-Bru, B. Berger, L. Philippe, C. Ammon-Zuffrey, P. Leone, G. Chevrier, E. St-Amand, A. Marette, J. Doré, A. Tremblay, Effect of *Lactobacillus rhamnosus* CGMCC1.3724 supplementation on weight loss and maintenance in obese men and women, *Br. J. Nutr.* 111 (2013) 1507–1519.
- [108] S. Chabot, H.-L. Yu, L. De Léséleuc, D. Cloutier, M.-R. Van Calsteren, M. Lessard, D. Roy, M. Lacroix, D. Oth, Exopolysaccharides from *Lactobacillus rhamnosus* RW-9595M stimulate TNF, IL-6 and IL-12 in human and mouse cultured immunocompetent cells, and IFN- γ in mouse splenocytes, *Lait* 81 (2001) 683–697.
- [109] J.P. Madej, T. Stefaniak, M. Bednarczyk, Effect of in ovo-delivered prebiotics and synbiotics on lymphoid-organs' morphology in chickens, *Poult. Sci.* 94 (2015) 1209–1219.
- [110] M.L. Ellis, A.E. Dowell, X. Li, J. Knight, Probiotic properties of *Oxalobacter formigenes*: an in vitro examination, *Arch. Microbiol.* 198 (2016) 1019–1026.
- [111] H.S. El-Nezami, N.N. Polychronaki, J. Ma, H. Zhu, W. Ling, E.K. Salminen, R.O. Juvonen, S.J. Salminen, T. Poussa, H.M. Mykkänen, Probiotic supplementation reduces a biomarker for increased risk of liver cancer in young men from Southern China, *Am. J. Clin. Nutr.* 83 (2006) 1199–1203.
- [112] J.P. Burton, C.N. Chilcott, J.R. Tagg, The rationale and potential for the reduction of oral malodour using *Streptococcus salivarius* probiotics, *Oral Dis.* 11 (2005) 29–31.
- [113] Y.J. Moon, J.R. Soh, J.J. Yu, H.S. Sohn, Y.S. Cha, S.H. Oh, Intracellular lipid accumulation inhibitory effect of *Weissella koreensis* OK1-6 isolated from Kimchi on differentiating adipocyte, *J. Appl. Microbiol.* 113 (2012) 652–658.
- [114] J.A. Park, P.B. Tirupathi Pichiah, J.J. Yu, S.H. Oh, J.W. Daily, Y.S. Cha, Anti-obesity effect of kimchi fermented with *Weissella koreensis* OK1-6 as starter in high-fat diet-induced obese C57BL/6J mice, *J. Appl. Microbiol.* 113 (2012) 1507–1516.
- [115] J. Touchman, *Comparative Genomics* [Online], in: Nature Education Knowledge, 2010. Available: <https://www.nature.com/scitable/knowledge/library/comparative-genomics-13239404>. (Accessed 14 January 2019).
- [116] A. Bezkorovainy, Probiotics: determinants of survival and growth in the gut, *Am. J. Clin. Nutr.* 73 (2001) 399S–405S.
- [117] G. Konuray, Z. Erginkaya, Potential use of *Bacillus coagulans* in the food industry, *Foods* 7 (2018).
- [118] B.R. Johnson, T.R. Klaenhammer, Impact of genomics on the field of probiotic research: historical perspectives to modern paradigms, *Antonie Van Leeuwenhoek* 106 (2014) 141–156.
- [119] L.C. Oliveira, T.D. Saraiva, W.M. Silva, U.P. Pereira, B.C. Campos, L.J. Benevides, F.S. Rocha, H. C.P. Figueiredo, V. Azevedo, S.C. Soares, Analyses of the probiotic property and stress resistance-related genes of *Lactococcus lactis* subsp. *lactis* NCDO 2118 through comparative genomics and in vitro assays. *PLoS ONE* 12 (4) (2017) e0175116. <https://doi.org/10.1371/journal.pone.0175116>.
- [120] H. Willenbrock, P.F. Hallin, T.M. Wassenaar, D.W. Ussery, Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray, *Genome Biol.* 8 (2007).
- [121] T.L. Nguyen, C.-I. Park, D.-H. Kim, Improved growth rate and disease resistance in olive flounder, *Paralichthys olivaceus*, by probiotic *Lactococcus lactis* WFLU12 isolated from wild marine fish, *Aquaculture* 471 (2017) 113–120.

- [122] R. Kant, J. Rintahaka, X. Yu, P. Sigvart-Mattila, L. Paulin, J.-P. Mecklin, M. Saarela, A. Palva, I. Von Ossowski, A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*. PLoS ONE 9 (7) (2014) e102762. <https://doi.org/10.1371/journal.pone.0102762>.
- [123] T. Smokvina, M. Wels, J. Polka, C. Chervaux, S. Brisse, J. Boekhorst, J.E. Van Hylckama Vlieg, R.J. Siezen, *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity, PLoS ONE 8 (2013) e68731.
- [124] J.L. Gardy, N.J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system, Nat. Rev. Genet. 19 (2018) 9–20.
- [125] J. Shendure, H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. 26 (2008) 1135–1145.
- [125a] J. Quick, N.D. Grubaugh, S.T. Pullan, et al., Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat. Protoc. 12 (6) (2017) 1261–1276, <https://doi.org/10.1038/nprot.2017.066>.
- [126] N.R. Faria, J. Quick, I.M. Claro, J. Theze, J.G. De Jesus, M. Giovanetti, M.U.G. Kraemer, S.C. Hill, A. Black, A.C. Da Costa, L.C. Franco, S.P. Silva, C.H. Wu, J. Raghvani, S. Cauchemez, L. Du Plessis, M.P. Verotti, W.K. De Oliveira, E.H. Carmo, G.E. Coelho, A. Santelli, L.C. Vinhal, C. M. Henriques, J.T. Simpson, M. Loose, K.G. Andersen, N.D. Grubaugh, S. Somasekar, C. Y. Chiu, J.E. Munoz-Medina, C.R. Gonzalez-Bonilla, C.F. Arias, L.L. Lewis-Ximenez, S. A. Baylis, A.O. Chieppe, S.F. Aguiar, C.A. Fernandes, P.S. Lemos, B.L.S. Nascimento, H.A. O. Monteiro, I.C. Siqueira, M.G. De Queiroz, T.R. De Souza, J.F. Bezerra, M.R. Lemos, G. F. Pereira, D. Loudal, L.C. Moura, R. Dhalia, R.F. Franca, T. Magalhaes, E.T. Marques Jr., T. Jaenisch, G.L. Wallau, M.C. De Lima, V. Nascimento, E.M. De Cerqueira, M.M. De Lima, D.L. Mascarenhas, J.P.M. Neto, A.S. Levin, T.R. Tozetto-Mendoza, S.N. Fonseca, M. C. Mendes-Correa, F.P. Milagres, A. Segurado, E.C. Holmes, A. Rambaut, T. Bedford, M.R. T. Nunes, E.C. Sabino, L.C.J. Alcantara, N.J. Loman, O.G. Pybus, Establishment and cryptic transmission of Zika virus in Brazil and the Americas, Nature 546 (2017) 406–410.
- [127] J. Theze, T. Li, L. Du Plessis, J. Bouquet, M.U.G. Kraemer, S. Somasekar, G. Yu, M. De Cesare, A. Balmaseda, G. Kuan, E. Harris, C.H. Wu, M.A. Ansari, R. Bowden, N.R. Faria, S. Yagi, S. Messenger, T. Brooks, M. Stone, E.M. Bloch, M. Busch, J.E. Munoz-Medina, C.R. Gonzalez-Bonilla, S. Wolinsky, S. Lopez, C.F. Arias, D. Bonsall, C.Y. Chiu, O.G. Pybus, Genomic epidemiology reconstructs the introduction and spread of zika virus in Central America and Mexico, Cell Host Microbe 23 (855–864) (2018).
- [128] L.C. Guimaraes, J. Florczak-Wyspianska, L.B. De Jesus, M.V. Viana, A. Silva, R.T. Ramos, C. Soares Sde, C. Soares Sde, Inside the pan-genome—methods and software overview, Curr. Genom. 16 (2015) 245–252.
- [129] K. Padovani De Souza, J.C. Setubal, F. Ponce De Leon, A.C. De Carvalho, G. Oliveira, A. Chateau, R. Alves, Machine learning meets genome assembly, Brief Bioinform. (2018) 1–14.
- [130] S.I. Lee, N.S. Kim, Transposable elements and genome size variations in plants, Genom. Inform. 12 (2014) 87–97.
- [131] I. Arabidopsis Genome, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature 408 (2000) 796–815.
- [132] K.L. McNally, K.L. Childs, R. Bohnert, R.M. Davidson, K. Zhao, V.J. Ulat, G. Zeller, R.M. Clark, D.R. Hoen, T.E. Bureau, R. Stokowski, D.G. Ballinger, K.A. Frazer, D.R. Cox, B. Padhukasahasram, C.D. Bustamante, D. Weigel, D.J. Mackill, R.M. Bruskiewich, G. Ratsch, C.R. Buell, H. Leung, J.E. Leach, Genomewide SNP variation reveals relationships among landraces and modern varieties of rice, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 12273–12278.
- [133] A.A. Golicz, J. Batley, D. Edwards, Towards plant pangenomics, Plant Biotechnol. J. 14 (2016) 1099–1105.
- [134] J.D. Montenegro, A.A. Golicz, P.E. Bayer, B. Hurgobin, H. Lee, C.K. Chan, P. Visendi, K. Lai, J. Dolezel, J. Batley, D. Edwards, The pangenome of hexaploid bread wheat, Plant J. 90 (2017) 1007–1013.
- [135] M.G. Milgroom, T.L. Peever, Population biology of plant pathogens: the synthesis of plant disease epidemiology and population genetics, Plant Dis. 87 (2003) 608–617.

- [136] B.M. Tyler, S. Tripathy, X. Zhang, P. Dehal, R.H. Jiang, A. Aerts, F.D. Arredondo, L. Baxter, D. Bensasson, J.L. Beynon, J. Chapman, C.M. Damasceno, A.E. Dorrance, D. Dou, A. W. Dickerman, I.L. Dubchak, M. Garbelotto, M. Gijzen, S.G. Gordon, F. Govers, N. J. Grunwald, W. Huang, K.L. Ivors, R.W. Jones, S. Kamoun, K. Krampis, K.H. Lamour, M.K. Lee, W.H. McDonald, M. Medina, H.J. Meijer, E.K. Nordberg, D.J. Maclean, M. D. Ospina-Giraldo, P.F. Morris, V. Phuntumart, N.H. Putnam, S. Rash, J.K. Rose, Y. Sakihama, A.A. Salamov, A. Savidor, C.F. Scheuring, B.M. Smith, B.W. Sobral, A. Terry, T.A. Torto-Alalibo, J. Win, Z. Xu, H. Zhang, I.V. Grigoriev, D.S. Rokhsar, J.L. Boore, Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis, *Science* 313 (2006) 1261–1266.
- [137] J.O. McInerney, A. McNally, M.J. O’Connell, Why prokaryotes have pangenomes, *Nat. Microbiol.* 2 (2017) 17040.
- [138] C. Plissonneau, F.E. Hartmann, D. Croll, Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome, *BMC Biol.* 16 (2018) 5.
- [139] J.C. Meeks, E.L. Campbell, M.L. Summers, F.C. Wong, Cellular differentiation in the cyanobacterium *Nostoc punctiforme*, *Arch. Microbiol.* 178 (2002) 395–403.
- [140] D.R. Nelson, B. Khraiwesh, W. Fu, S. Alseekh, A. Jaiswal, A. Chaiboonchoe, K.M. Hazzouri, M. J. O’connor, G.L. Butterfoss, N. Drou, J.D. Rowe, J. Harb, A.R. Fernie, K.C. Gunsalus, K. Salehi-Ashtiani, The genome and phenome of the green alga *Chloroidium* sp. UTEX 3007 reveal adaptive traits for desert acclimatization. *Elife* 6 (2017) e25783 <https://doi.org/10.7554/eLife.25783>.
- [141] S. Hirooka, Y. Hirose, Y. Kanesaki, S. Higuchi, T. Fujiwara, R. Onuma, A. Era, R. Ohbayashi, A. Uzuka, H. Nozaki, H. Yoshikawa, S.Y. Miyagishima, Acidophilic green algal genome provides insights into adaptation to an acidic environment, *Proc. Natl. Acad. Sci. U. S. A.* 114 (2017) E8304–E8313.
- [142] G. Barbier, C. Oesterheld, M.D. Larson, R.G. Halgren, C. Wilkerson, R.M. Garavito, C. Benning, A.P. Weber, Comparative genomics of two closely related unicellular thermo-acidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic flexibility of *Galdieria sulphuraria* and significant differences in carbohydrate metabolism of both algae, *Plant Physiol.* 137 (2005) 460–474.
- [143] D. Bhattacharya, D.C. Price, C.X. Chan, H. Qiu, N. Rose, S. Ball, A.P. Weber, M.C. Arias, B. Henrissat, P.M. Coutinho, A. Krishnan, S. Zauner, S. Morath, F. Hilliou, A. Egizi, M.M. Perrineau, H.S. Yoon, Genome of the red alga *Porphyridium purpureum*, *Nat. Commun.* 4 (2013) 1941.
- [144] S. Bose, S.K. Herbert, D.C. Fork, Fluorescence characteristics of photoinhibition and recovery in a sun and a shade species of the red algal genus porphyra, *Plant Physiol.* 86 (1988) 946–950.
- [145] K. Nishitsuji, A. Arimoto, K. Iwai, Y. Sudo, K. Hisata, M. Fujie, N. Arakaki, T. Kushiro, T. Konishi, C. Shinzato, N. Satoh, E. Shoguchi, A draft genome of the brown alga, *Cladosiphon okamuranus*, S-strain: a platform for future studies of ‘mozuku’ biology, *DNA Res.* 23 (2016) 561–570.
- [146] A. Sboner, X.J. Mu, D. Greenbaum, R.K. Auerbach, M.B. Gerstein, The real cost of sequencing: higher than you think!, *Genome Biol.* 12 (2011) 125.
- [147] M. Guegan, K. Zouache, C. Demichel, G. Minard, V. Tran Van, P. Potier, P. Mavingui, C. Valiente Moro, The mosquito holobiont: fresh insight into mosquito-microbiota interactions, *Microbiome* 6 (2018) 49.
- [148] D. Aguirre De Carcer, The human gut pan-microbiome presents a compositional core formed by discrete phylogenetic units, *Sci. Rep.* 8 (2018) 14069.
- [149] M.H. Leung, P.K. Lee, The roles of the outdoors and occupants in contributing to a potential pan-microbiome of the built environment: a review, *Microbiome* 4 (2016) 21.
- [150] P. Vandenkoornhuys, A. Quaiser, M. Duhamel, A. Le Van, A. Dufresne, The importance of the microbiome of the plant holobiont, *New Phytol.* 206 (2015) 1196–1206.
- [151] B. Aslam, M. Basit, M.A. Nisar, M.H. Rasool, M. Khurshid, Proteomics: technologies and their applications, *J. Chromatogr. Sci.* 55 (2017) 182–196.
- [152] W.M. Silva, R.D. Carvalho, S.C. Soares, I.F.S. Bastos, E.L. Folador, G.H.M.F. Souza, Y. Le Loir, A. Miyoshi, A. Silva, V. Azevedo, Label-free proteomic analysis to confirm the predicted proteome of

- Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide, *BMC Genom.* 15 (2014) 1065.
- [153] W.M. Silva, R.D.O. Carvalho, F.A. Dorella, E.L. Folidor, G.H.M.F. Souza, A.M.C. Pimenta, H.C. P. Figueiredo, Y. Le Loir, A. Silva, V. Azevedo, Quantitative proteomic analysis reveals changes in the benchmark *Corynebacterium pseudotuberculosis* biovar equi exoproteome after passage in a murine host. *Front. Cell. Infect. Microbiol.* 7 (2017) 325, <https://doi.org/10.3389/fcimb.2017.00325>.
- [154] T.-C. Chao, N. Hansmeier, The current state of microbial proteomics: where we are and where we want to go, *Proteomics* 12 (2012) 638–650.
- [155] M.A. Moseley, Quantitative proteomics in genomic medicine, in: G.S. Ginsburg, H.F. Willard (Eds.), *Genomic and Personalized Medicine*, second ed., Academic Press, 2013, pp. 155–165 (Chapter 13).
- [156] M.A. Reymond, W. Schlegel, Proteomics in cancer, *Adv. Clin. Chem.* 44 (2007) 103–142.
- [157] M.A. Hussain, F. Huygens, Proteomic and bioinformatics tools to understand virulence mechanisms in *Staphylococcus aureus*, *Curr. Proteom.* 9 (2012) 2–8.
- [158] O. Coskun, Separation techniques: Chromatography, *North. Clin. Istanbul.* 3 (2016) 156–160.
- [159] R.M. Lequin, Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA), *Clin. Chem.* 51 (2005) 2415–2418.
- [160] B.T. Kurien, R.H. Scofield, Western blotting: an introduction, in: B.T. Kurien, R.H. Scofield (Eds.), *Western Blotting: Methods and Protocols*, Springer New York, New York, NY, 2015, pp. 17–30.
- [161] M. D’Innocenzo, Identificação das proteínas por meio da eletroforese 2D, in: R. Verlengia, R. Curi, E. Bevilacqua, P. Newsholme (Eds.), *Análises de RNA, proteínas e metabólitos: metodologia e procedimentos técnicos*, Santos Editora, São Paulo, 2013, pp. 261–280.
- [162] R. Vesecchi, N.P. Lopes, F.C. Gozzo, F.A. Dörr, M. Murgu, D.T. Lebre, R. Abreu, O.V. Bustillos, J.M. Riveros, Nomenclaturas de espectrometria de massas em língua portuguesa, *Quím. Nova* 34 (2011) 1875–1887.
- [163] S.J. Cordwell, A.S. Nouwens, B.J. Walsh, Comparative proteomics of bacterial pathogens, *Proteomics* 1 (2001) 461–472.
- [164] P.M. Bisch, Genômica funcional: proteômica, in: L. Mir (Ed.), *Genômica*, Atheneu, São Paulo, 2004, pp. 139–162.
- [165] E.-H. Jeong, B. Vaidya, S.-Y. Cho, M.-A. Park, K. Kaewintajak, S.R. Kim, M.-J. Oh, J.-S. Choi, J. Kwon, D. Kim, Identification of regulators of the early stage of viral hemorrhagic septicemia virus infection during curcumin treatment, *Fish Shellfish Immunol.* 45 (2015) 184–193.
- [166] N. Solis, S.J. Cordwell, Current methodologies for proteomics of bacterial surface-exposed and cell envelope proteins, *Proteomics* 11 (2011) 3169–3189.
- [167] S. Hölper, A. Ruhs, M. Krüger, Stable isotope labeling for proteomic analysis of tissues in mouse, in: B. Warscheid (Ed.), *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC): Methods and Protocols*, Springer New York, New York, NY, 2014, pp. 95–106.
- [168] K. Cheng, A. Sloan, S. Mccorrister, L. Peterson, H. Chui, M. Drebot, C. Nadon, J.D. Knox, G. Wang, Quality evaluation of LC-MS/MS-based *E. coli* H antigen typing (MS-H) through label-free quantitative data analysis in a clinical sample setup, *Proteom. Clin. Appl.* 8 (2014) 963–970.
- [169] S. Kosono, M. Tamura, S. Suzuki, Y. Kawamura, A. Yoshida, M. Nishiyama, M. Yoshida, Changes in the acetylome and succinylome of *Bacillus subtilis* in response to carbon source, *PLoS ONE* 10 (2015) e0131169.
- [170] S.P. Gygi, B. Rist, T.J. Griffin, J. Eng, R. Aebersold, Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags, *J. Proteome Res.* 1 (2002) 47–54.
- [171] V.J. Patel, K. Thalassinos, S.E. Slade, J.B. Connolly, A. Crombie, J.C. Murrell, J.H. Scrivens, A comparison of labeling and label-free mass spectrometry-based proteomics approaches, *J. Proteome Res.* 8 (2009) 3752–3759.
- [172] D. Chelius, P.V. Bondarenko, Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry, *J. Proteome Res.* 1 (2002) 317–323.
- [173] M.J.G. Hughes, J.C. Moore, J.D. Lane, R. Wilson, P.K. Pribul, Z.N. Younes, R.J. Dobson, P. Everest, A.J. Reason, J.M. Redfern, F.M. Greer, T. Paxton, M. Panico, H.R. Morris, R.

- G. Feldman, J.D. Santangelo, Identification of major outer surface proteins of *Streptococcus agalactiae*, *Infect. Immun.* 70 (2002) 1254–1259.
- [174] F. Doro, S. Liberatori, M.J. Rodríguez-Ortega, C.D. Rinaudo, R. Rosini, M. Mora, M. Scarselli, E. Altindis, R. D'aurizio, M. Stella, I. Margarit, D. Maione, J.L. Telford, N. Norais, G. Grandi, Surface analysis as a fast track to vaccine discovery: identification of a novel protective antigen for group B *Streptococcus hypervirulent* strain COH1, *Mol. Cell. Proteom.* 8 (2009) 1728–1737.
- [175] W.M. Silva, N. Seyffert, A.V. Santos, T.L.P. Castro, L.G.C. Pacheco, A.R. Santos, A. Ciprandi, F.A. Dorella, H.M. Andrade, D. Barh, A.M.C. Pimenta, A. Silva, A. Miyoshi, V. Azevedo, Identification of 11 new exoproteins in *Corynebacterium pseudotuberculosis* by comparative analysis of the exoproteome, *Microb. Pathog.* 61–62 (2013) 37–42.
- [176] W.M. Silva, N. Seyffert, A. Ciprandi, A.V. Santos, T.L.P. Castro, L.G.C. Pacheco, D. Barh, Y. Le Loir, A.M.C. Pimenta, A. Miyoshi, A. Silva, V. Azevedo, Differential exoproteome analysis of two *Corynebacterium pseudotuberculosis* biovar ovis strains isolated from goat (1002) and sheep (C231), *Curr. Microbiol.* 67 (2013) 460–465.
- [177] J.A. Broadbent, D.A. Broszczak, I.U.K. Tennakoon, F. Huygens, Pan-proteomics, a concept for unifying quantitative proteome measurements when comparing closely-related bacterial strains, *Expert Rev. Proteom.* 13 (2016) 355–365.
- [178] G.C. Tavares, F.L. Pereira, G.M. Barony, C.P. Rezende, W.M. Da Silva, G.H.M.F. De Souza, T. Verano-Braga, V.A. De Carvalho Azevedo, Leal, C.a.G., and Figueiredo, H.C.P., Delineation of the pan-proteome of fish-pathogenic *Streptococcus agalactiae* strains using a label-free shotgun approach, *BMC Genom.* 20 (2019) 11.
- [179] J. Rothen, J.F. Pothier, F. Foucault, J. Blom, D. Nanayakkara, C. Li, M. Ip, M. Tanner, G. Vogel, V. Pflüger, C.A. Daubenberger, Subspecies typing of *Streptococcus agalactiae* based on ribosomal subunit protein mass variation by MALDI-TOF MS, *Front. Microbiol.* 10 (2019) 471.
- [180] L. Zhang, D. Xiao, B. Pang, Q. Zhang, H. Zhou, L. Zhang, J. Zhang, B. Kan, The core proteome and pan proteome of *Salmonella* Paratyphi A epidemic strains, *PLoS ONE* 9 (2014) e89197.
- [181] G.D. Jhingan, S. Kumari, S.V. Jamwal, H. Kalam, D. Arora, N. Jain, L.K. Kumar, A. Samal, K.V.S. Rao, D. Kumar, V.K. Nandicoori, Comparative proteomic analyses of avirulent, virulent, and clinical strains of *Mycobacterium tuberculosis* identify strain-specific patterns, *J. Biol. Chem.* 291 (2016) 14257–14273.
- [182] W.M. Silva, C.S. Sousa, L.C. Oliveira, S.C. Soares, G. Souza, G.C. Tavares, C.P. Resende, E.L. Follador, F.L. Pereira, H. Figueiredo, V. Azevedo, Comparative proteomic analysis of four biotechnological strains *Lactococcus lactis* through label-free quantitative proteomics, *Microb. Biotechnol.* 12 (2019) 265–274.
- [183] J. Trapp, C. Almunia, J.-C. Gaillard, O. Pible, A. Chaumot, O. Geffard, J. Armengaud, Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods, *J. Proteome* 135 (2016) 51–61.
- [184] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet.* 10 (2009) 57–63.
- [185] I. Korf, Genomics: the state of the art in RNA-seq analysis, *Nat. Methods* 10 (2013) 1165–1166.
- [186] M.F. Rai, E.D. Tycksen, L.J. Sandell, R.H. Brophy, Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears, *J. Orthop. Res.* 36 (2018) 484–497.
- [187] S.C. Sealfon, T.T. Chu, RNA and DNA microarrays, *Methods Mol. Biol.* 671 (2011) 3–34.
- [188] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, T. Shafee, Transcriptomics technologies, *PLoS Comput. Biol.* 13 (2017) e1005457.
- [189] S. Zhao, W.P. Fung-Leung, A. Bittner, K. Ngo, X. Liu, Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells, *PLoS ONE* 9 (2014) e78644.
- [190] M. Blaxter, S. Kumar, G. Kaur, G. Koutsovoulos, B. Elsworth, Genomics and transcriptomics across the diversity of the Nematoda, *Parasite Immunol.* 34 (2012) 108–120.
- [191] M.S. Kim, H. Zhang, H. Yan, B.J. Yoon, W.B. Shim, Characterizing co-expression networks underpinning maize stalk rot virulence in *Fusarium verticillioides* through computational subnetwork module analyses, *Sci. Rep.* 8 (2018) 8310.

- [192] Z. Wei, H. Guo, J. Qin, S. Lu, Q. Liu, X. Zhang, Y. Zou, Y. Gong, C. Shao, Pan-senescence transcriptome analysis identified RRAAD as a marker and negative regulator of cellular senescence, *Free Radic. Biol. Med.* 130 (2019) 267–277.
- [193] X. Ma, Y. Liu, Y. Liu, L.B. Alexandrov, M.N. Edmonson, C. Gawad, X. Zhou, Y. Li, M.C. Rusch, J. Easton, R. Huether, V. Gonzalez-Pena, M.R. Wilkinson, L.C. Hermida, S. Davis, E. Sioson, S. Pounds, X. Cao, R.E. Ries, Z. Wang, X. Chen, L. Dong, S.J. Diskin, M.A. Smith, J.M. Guidry Auvil, P.S. Meltzer, C.C. Lau, E.J. Perlman, J.M. Maris, S. Meshinchi, S.P. Hunger, D.S. Gerhard, J. Zhang, Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours, *Nature* 555 (2018) 371–376.
- [194] C.R. Cabanski, N.M. White, H.X. Dang, J.M. Silva-Fisher, C.E. Rauck, D. Cicka, C.A. Maher, Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function, *RNA Biol.* 12 (2015) 628–642.
- [195] G. Dugar, A. Herbig, K.U. Forstner, N. Heidrich, R. Reinhardt, K. Nieselt, C.M. Sharma, High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates, *PLoS Genet.* 9 (2013) e1003495.
- [196] B. Sidders, M. Withers, S.L. Kendall, J. Bacon, S.J. Waddell, J. Hinds, P. Golby, F. Movahedzadeh, R. A. Cox, R. Frita, A.M. Ten Bokum, L. Wernisch, N.G. Stoker, Quantification of global transcription patterns in prokaryotes using spotted microarrays, *Genome Biol.* 8 (2007) R265.
- [197] C. Eymann, G. Homuth, C. Scharf, M. Hecker, *Bacillus subtilis* functional genomics: global characterization of the stringent response by proteome and transcriptome analysis, *J. Bacteriol.* 184 (2002) 2500–2520.
- [198] H. Qin, N.W. Lo, J.F. Loo, X. Lin, A.K. Yim, S.K. Tsui, T.C. Lau, M. Ip, T.F. Chan, Comparative transcriptomics of multidrug-resistant *Acinetobacter baumannii* in response to antibiotic treatments, *Sci. Rep.* 8 (2018) 3515.
- [199] A. Dotsch, M. Schniederjans, A. Khaledi, K. Hornischer, S. Schulz, A. Bielecka, D. Eckweiler, S. Pohl, S. Haussler, The *Pseudomonas aeruginosa* transcriptional landscape is shaped by environmental heterogeneity and genetic variation, *MBio* 6 (2015) e00749.
- [200] L. De Welzen, V. Eldholm, K. Maharaj, A.L. Manson, A.M. Earl, A.S. Pym, Whole-transcriptome and -genome analysis of extensively drug-resistant *Mycobacterium tuberculosis* clinical isolates identifies downregulation of etha as a mechanism of ethionamide resistance, *Antimicrob. Agents Chemother.* 61 (2017).
- [201] T.H. Hazen, J. Michalski, Q. Luo, A.C. Shetty, S.C. Daugherty, J.M. Fleckenstein, D.A. Rasko, Comparative genomics and transcriptomics of *Escherichia coli* isolates carrying virulence factors of both enteropathogenic and enterotoxigenic *E. coli*, *Sci. Rep.* 7 (2017) 3513.
- [202] P.A. Northcott, C. Lee, T. Zichner, A.M. Stutz, S. Erkek, D. Kawauchi, D.J. Shih, V. Hovestadt, M. Zapatka, D. Sturm, D.T. Jones, M. Kool, M. Remke, F.M. Cavalli, S. Zuyderduyn, G.D. Bader, S. Vandenberg, L.A. Esparza, M. Ryzhova, W. Wang, A. Wittmann, S. Stark, L. Sieber, H. Seker-Cin, L. Linke, F. Kratochwil, N. Jager, I. Buchhalter, C.D. Imbusch, G. Zipprich, B. Raeder, S. Schmidt, N. Diessl, S. Wolf, S. Wiemann, B. Brors, C. Lawrenz, J. Eils, H.J. Warnatz, T. Risch, M.L. Yaspo, U.D. Weber, C.C. Bartholomae, C. Von Kalle, E. Turanyi, P. Hauser, E. Sanden, A. Darabi, P. Siesjo, J. Sterba, K. Zitterbart, D. Sumerauer, P. Van Sluis, R. Versteeg, R. Volckmann, J. Koster, M.U. Schuhmann, M. Ebinger, H.L. Grimes, G.W. Robinson, A. Gajjar, M. Mynarek, K. Von Hoff, S. Rutkowski, T. Pietsch, W. Scheurlen, J. Felsberg, G. Reifemberger, A.E. Kulozik, A. Von Deimling, O. Witt, R. Eils, R.J. Gilbertson, A. Korshunov, M.D. Taylor, P. Lichter, J.O. Korbel, R.J. Wechsler-Reya, S.M. Pfister, Enhancer hijacking activates GF11 family oncogenes in medulloblastoma, *Nature* 511 (2014) 428–434.
- [203] E. Papaemmanuil, M. Cazzola, J. Boultonwood, L. Malcovati, P. Vyas, D. Bowen, A. Pellagatti, J. S. Wainscoat, E. Hellstrom-Lindberg, C. Gambacorti-Passerini, A.L. Godfrey, I. Rapado, A. Cvejic, R. Rance, C. Mcgee, P. Ellis, L.J. Mudie, P.J. Stephens, S. McLaren, C.E. Massie, P. S. Tarpey, I. Varela, S. Nik-Zainal, H.R. Davies, A. Shlien, D. Jones, K. Raine, J. Hinton, A. P. Butler, J.W. Teague, E.J. Baxter, J. Score, A. Galli, M.G. Della Porta, E. Travaglino, M. Groves, S. Tauro, N.C. Munshi, K.C. Anderson, A. El-Naggar, A. Fischer, V. Mustonen, A.

- J. Warren, N.C. Cross, A.R. Green, P.A. Futreal, M.R. Stratton, P.J. Campbell, Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium, Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts, *N. Engl. J. Med.* 365 (2011) 1384–1395.
- [204] X.S. Puente, M. Pinyol, V. Quesada, L. Conde, G.R. Ordonez, N. Villamor, G. Escaramis, P. Jares, S. Bea, M. Gonzalez-Diaz, L. Bassaganyas, T. Baumann, M. Juan, M. Lopez-Guerra, D. Colomer, J. M. Tubio, C. Lopez, A. Navarro, C. Tornador, M. Aymerich, M. Rozman, J.M. Hernandez, D. A. Puente, J.M. Freije, G. Velasco, A. Gutierrez-Fernandez, D. Costa, A. Carrio, S. Guijarro, A. Enjuanes, L. Hernandez, J. Yague, P. Nicolas, C.M. Romeo-Casabona, H. Himmelbauer, E. Castillo, J.C. Dohm, S. De Sanjose, M.A. Piris, E. De Alava, J. San Miguel, R. Royo, J. L. Gelpi, D. Torrents, M. Orozco, D.G. Pisano, A. Valencia, R. Guigo, M. Bayes, S. Heath, M. Gut, P. Klatt, J. Marshall, K. Raine, L.A. Stebbings, P.A. Futreal, M.R. Stratton, P. J. Campbell, I. Gut, A. Lopez-Guillermo, X. Estivill, E. Montserrat, C. Lopez-Otin, E. Campo, Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia, *Nature* 475 (2011) 101–105.
- [205] Z. Liu, S. Zhang, Toward a systematic understanding of cancers: a survey of the pan-cancer study, *Front. Genet.* 5 (2014) 194.
- [206] T.J. Hudson, W. Anderson, A. Artez, A.D. Barker, C. Bell, R.R. Bernabe, M.K. Bhan, F. Calvo, I. Eerola, D.S. Gerhard, A. Guttmacher, M. Guyer, F.M. Hemsley, J.L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D.P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T.S. Rao, J. Remacle, A.J. Schafer, T. Shibata, M.R. Stratton, J.G. Vockley, K. Watanabe, H. Yang, M.M. Yuen, B.M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S.O. Dyke, Y. Joly, K. Kato, K.L. Kennedy, P. Nicolas, M.J. Parker, E. Rial-Sebbag, C.M. Romeo-Casabona, K.M. Shaw, S. Wallace, G.L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clement, E. De Alava, F. Degos, M.L. Ferguson, P. Geary, D.N. Hayes, T.J. Hudson, A.L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. Van De Vijver, P.A. Futreal, H. Aburatani, M. Bayes, D.D. Botwell, P.J. Campbell, X. Estivill, D.S. Gerhard, S.M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J.D. Mcpherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M.R. Stratton, H.G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R.K. Wilson, H.H. Xue, L. Yang, P.T. Spellman, G.D. Bader, P.C. Boutros, P. J. Campbell, P. Flicek, et al., International network of cancer genome projects, *Nature* 464 (2010) 993–998.
- [207] D.A. Levine, Integrated genomic characterization of endometrial carcinoma, *Nature* 497 (2013) 67–73.

III.2.2. Conclusion, Chapter 2.

The concept of pan-genomics is so deep that it has been perfectly applied in the studies of several organisms and diseases. In the study of dynamics of biological processes and disease development, identification of therapeutics targets against deadly and emerging pathogens and in the development of new probiotics. It has great potential which may bring a closer understanding and combating with prokaryotic and eukaryotic diseases.

Chapter 3.

III.3.1. Research Article

The pan-genome of *Treponema pallidum* reveals differences in genome plasticity between subspecies related to venereal and non-venereal syphilis.

Arun Kumar Jaiswal, Sandeep Tiwari, Syed Babar Jamal, Leticia de Castro Oliveira, Leandro Gomes Alves, Vasco Azevedo, Preetam Ghosh, Carlo Jose Freira Oliveira, **Siomar C. Soares**.

BMC Genomics, 2020, 21:33;

Spirochaetal organisms of the *Treponema* genus are responsible for causing Treponematoses. Pathogenic treponemes cause multi-stage infections like endemic syphilis, venereal syphilis, yaws and pinta. These infections have many similarities, but they can be differentiated based on epidemiological, clinical and geographical criteria. *Treponema pallidum* subsp. *endemicum* (TEN) responsible for bejel (endemic syphilis); *T. pallidum* subsp. *pallidum* (TPA) responsible for venereal syphilis; *T. pallidum* subsp. *pertenue* (TPE) causes yaws; and *T. pallidum* subsp. *carateum* causes pinta. Out of these four high morbidity diseases, venereal syphilis is mediated by sexual contact; the other three diseases are transmitted by close personal contact. Because of re-emergence, the global distribution of syphilis is alarming and there is an increasing need of proper treatment and preventive measures. Unfortunately, effective measures are limited.

In this work, our contribution includes the understanding between venereal and non-venereal syphilis of *Treponema pallidum* based on subspecies level. We have used pan-genomics approach to find the number of pan-genome, core genome and singletons for subsp. *pallidum*, *pertenue* and *endemicum*. Further we used GIPSY (a tool for pathogenicity and genomic island prediction software) for the identification of presence and absence of genomic and pathogenicity island in subsp. *pallidum* (reference strain Nichols), *pertenue* (reference strain SamoaD) and *endemicum* (reference strain BosniaA). The findings of this analysis are very important, as it can help in the understanding of molecular basis of infections from *T. pallidum* subspecies.

RESEARCH ARTICLE

Open Access



The pan-genome of *Treponema pallidum* reveals differences in genome plasticity between subspecies related to venereal and non-venereal syphilis

Arun Kumar Jaiswal^{1,2}, Sandeep Tiwari^{1*} , Syed Babar Jamal³, Leticia de Castro Oliveira^{1,2}, Leandro Gomes Alves², Vasco Azevedo¹, Preetam Ghosh⁴, Carlo Jose Freira Oliveira² and Siomar C. Soares^{2*}

Abstract

Background: Spirochetal organisms of the *Treponema* genus are responsible for causing Treponematoses. Pathogenic treponemes is a Gram-negative, motile, spirochete pathogen that causes syphilis in human. *Treponema pallidum* subsp. *endemicum* (TEN) causes endemic syphilis (bejel); *T. pallidum* subsp. *pallidum* (TPA) causes venereal syphilis; *T. pallidum* subsp. *pertenue* (TPE) causes yaws; and *T. pallidum* subsp. *Ccarateum* causes pinta. Out of these four high morbidity diseases, venereal syphilis is mediated by sexual contact; the other three diseases are transmitted by close personal contact. The global distribution of syphilis is alarming and there is an increasing need of proper treatment and preventive measures. Unfortunately, effective measures are limited.

Results: Here, the genome sequences of 53 *T. pallidum* strains isolated from different parts of the world and a diverse range of hosts were comparatively analysed using pan-genomic strategy. Phylogenomic, pan-genomic, core genomic and singleton analysis disclosed the close connection among all strains of the pathogen *T. pallidum*, its clonal behaviour and showed increases in the sizes of the pan-genome. Based on the genome plasticity analysis of the subsets containing the subspecies *T. pallidum* subsp. *pallidum*, *T. pallidum* subsp. *endemicum* and *T. pallidum* subsp. *pertenue*, we found differences in the presence/absence of pathogenicity islands (PAIs) and genomic islands (GIs) on subsp.-based study.

Conclusions: In summary, we identified four pathogenicity islands (PAIs), eight genomic islands (GIs) in subsp. *pallidum*, whereas subsp. *endemicum* has three PAIs and seven GIs and subsp. *pertenue* harbours three PAIs and eight GIs. Concerning the presence of genes in PAIs and GIs, we found some genes related to lipid and amino acid biosynthesis that were only present in the subsp. of *T. pallidum*, compared to *T. pallidum* subsp. *endemicum* and *T. pallidum* subsp. *pertenue*.

Keywords: Pan-genome, Core genome, Singletons, *Treponema pallidum*, Syphilis

* Correspondence: sandip_sbtbi@yahoo.com; siomars@gmail.com

¹PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

²Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, MG, Brazil

Full list of author information is available at the end of the article



Background

Spirochetal organisms of the *Treponema* genus are responsible for causing Treponematoses. Pathogenic treponemes cause multi-stage infections like endemic syphilis, venereal syphilis, yaws and pinta. These infections have many similarities, but they can be differentiated based on epidemiological, clinical and geographical criteria [1–3]. Primarily, the pathogenic treponemes can be classified based on the clinical symptoms of the respective disease they cause. *Treponema pallidum* subsp. *endemicum* causes endemic syphilis; *T. pallidum* subsp. *pallidum* causes venereal syphilis; *T. pallidum* subsp. *pertenue* causes yaws; and *T. pallidum* subsp. *carateum* causes pinta. Out of these four high morbidity diseases, venereal syphilis is only transmitted by sexual contact; the other three diseases are transmitted by close personal contact [2].

It is estimated by the World Health Organization (WHO) that there are 12 million new cases of syphilis annually and the aggregated cases of yaws, bejel, and pinta (the endemic treponematoses) are approximately 2.5 million globally, although good surveillance data is not available. The infections caused by *T. pallidum* are characterized by periods of active clinical disease interrupted by episodes of asymptomatic latent infection and may cause life-long infections in untreated individuals [4, 5]. *Treponema pallidum* is a Gram-negative, motile, spirochete human pathogen. Syphilis is a multistage infectious disease that can be communicated between sexual partners through active lesions or from an infected woman to her fetus during pregnancy [6, 7]. Syphilis has a worldwide distribution (e.g. Africa has a high incidence), affecting every country and continent except perhaps Antarctica [8–12]. The stages of syphilis have been divided on the basis of clinical findings that lead to treatment and follow-up. Syphilis chancres may go unnoticed primarily due to their well-documented painless nature and if they are present in those parts of the body that are difficult to visualize (e.g. cervix, throat or anus/rectum) [13]. Furthermore, due to pleomorphic appearance and lack of physician familiarity with the expressions of syphilis, their lesions may be misdiagnosed. Secondary, syphilis may manifest itself through severe rashes that may go unobserved by the patient or may mimic an extensive condition [8]. *T. pallidum* is completely sensitive to penicillin treatment, despite the use of this antibiotic for seven decades in treating syphilis infections. Standard treatment of uncomplicated syphilis with parenteral Benzathine penicillin G is highly effective at all stages. Many antibiotics' resistance (e.g. Macrolide and Clindamycin resistance) has been reported in several countries [6]. The ongoing high rate of syphilis worldwide, despite the availability of inexpensive and effective treatment, presents the most convincing

argument for the need of developing new and potent vaccine against syphilis [14]. Despite the WHO's Initiative for the Global Elimination of Congenital Syphilis, an intensive syphilis-targeted public health control has been undertaken to reduce the incidence; however, it has not been achieved yet [14]. Specifically, the reasons for failure are multifactorial; some of the responsibility can be attributed to the difficulty in the diagnosis of syphilis and treatment, and lack of access or use of prenatal screening programs [15]. The advancement in the field of genomics and cost-effective sequencing technologies has transformed the human bacterial pathogens study and helped in the improvement of vaccine designing technologies. A new and emerging methodology to get deep insight of the genome of a species or genus is the pan-genomics approach, which was introduced by Tettelin and collaborators in 2005 working with *Streptococcus agalactiae* [16]. Pan-genome provides us with the complete and non-redundant collection of genes from a species or genus and is composed of three subsets (core genome, shared genome and singletons): the core genome, which is the collection of all the genes commonly shared between all the genomes used as dataset; the shared genome, which contains only the genes shared between two or more strains, which are not present in all strains of the dataset; and, the singletons, which are present only in one strain and are referred to as strain-specific genes.

The first genome of *T. pallidum* subsp. *pallidum* (strain Nichols) was sequenced in 1998. The organism has a comparatively small genome and only 55% of *T. pallidum*'s 1041 open reading frames are recognized to have a biological function, which indicates that it uses host biosynthesis to complete some of its metabolic needs [3]. The DNA-DNA hybridization studies showed homology between DNA of venereal syphilis spirochete and DNA of culturable treponemes (*T. phagedenis* and its biotypes Reiter and Kazan) was less than 5% identical, but was indistinguishable from DNA of the yaws spirochete *T. pallidum* [3, 17, 18]. This study led to the reclassification of the agents of endemic syphilis, venereal syphilis and yaws as *T. pallidum* subsp. *endemicum*, *Treponema pallidum* subsp. *pallidum* and *T. pallidum* subsp. *pertenue*, respectively. Genomic sequencing has recognized these subspecies as clonal, but forming distinct genetic clusters [2, 3].

In this work, we perform a pan-genome approach to better understand the differences of *Treponema pallidum* infections in the broad spectrum and how genome plasticity is related to the symptom patterns. For pan-genomic comparative analyses, we used 53 *T. pallidum* strains. We present phylo-genomic correlations between all *T. pallidum* strains. Furthermore, we describe the “pan-genome”, which is the complete inventory of genes

found in any member of the species; the “core genome”, which is important for basic life processes; and the “singletons”, which are normally related to environmental fitness and adaptation to host. Finally, we provide insights into the specific subsets (singletons and the pan- and core genomes) of 53 genomes of *T. pallidum* strains and correlate these subsets with the plasticity of pathogenicity islands and virulence genes.

Results

Phylogenomics study of *Treponema pallidum* strains

The phylogenomics relationships between *T. pallidum* strains were determined using Gegenees [19]. Furthermore, all genome sequences were cross-compared to generate a phylogenomic tree and to plot a heatmap. According to the generated phylogenomic tree, closely related strains appeared in the same cluster. The subspecies responsible for non-venereal syphilis is *Treponema pallidum* subsp. *endemicum* (TEN) and *T. pallidum* subsp. *pertenue* (TPE) strains appeared in closely related clusters (Fig. 1). The *T. pallidum* subspecies strains responsible for venereal syphilis formed different clusters. Additionally, *T. pallidum* strain BosniaA (subsp. *endemicum*) was positioned between the clusters of *Treponema pallidum* subsp. *Pertenue* and venereal syphilis (*Treponema pallidum* subsp. *pallidum*). According to the heatmap, the non-venereal isolates are 100% similar to each other and many of the venereal isolates are 100% similar to each other, but the two groups show some difference (Additional file 1: Figure S2). Moreover, the heatmap indicated the clonal-like behavior of *T. pallidum* subsp., compared with the isolates other than genital, anal or Neurosyphilitic samples, which showed similarities ranging from 97 to 100%.

The Pan-genome, Core genome and singletons of *Treponema pallidum*

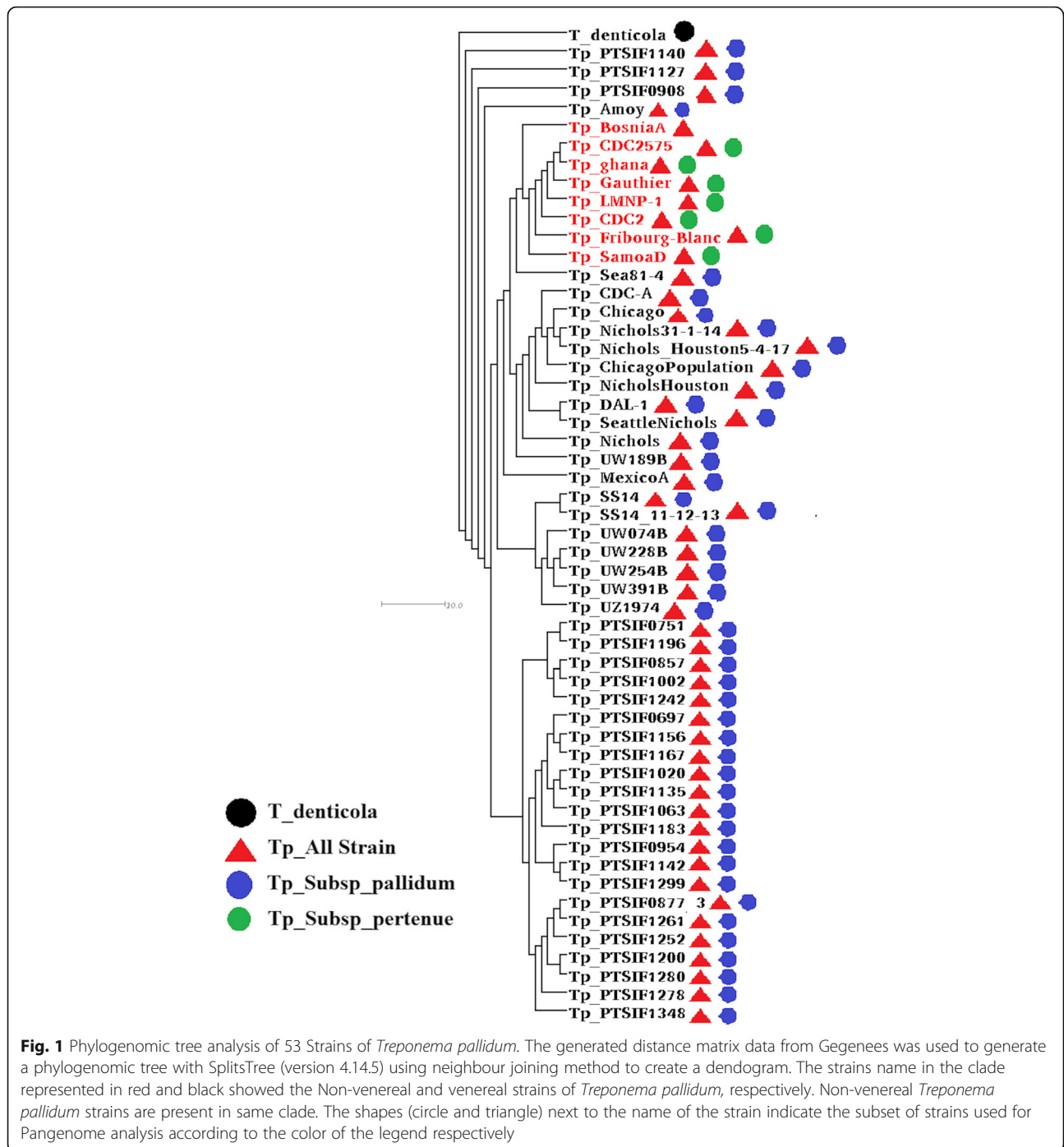
The main goal of the pan-genome is the comparison of different strains of the same species or even genus at the genomic level. The resulting pan-genome of Pan All (Fig. 2A1-A3), Pan Subsp_pallidum (Fig. 3B1-B3), and Pan_subsp_pertenue (Fig. 4C1-C3), of *T. pallidum* contains a total of 2112, 982, and 1049 genes respectively. The formula ($\alpha = 1 - \gamma$) inferred that the pan-genome of *T. pallidum* is increasing with an α of 0.9435. The extrapolation was also separately calculated for all divided subsets for the analysis in this work. The α value for each subset Pan Subsp_pallidum and Pan_subsp_pertenue, were 0.916 and 0.999329 respectively. The α values for all datasets used in this work are less than 1 which indicates that all have an open pan-genome. However, although the pan-genome is still open, it increases at a very low rate [20, 21].

The core genome and singletons of the complete dataset and all the subsets of *T. pallidum* were calculated by the least-squares fit of the exponential regression decay to the mean values, as represented by the formula $n = k * \exp[-x/\tau] + tg(\theta)$, where n is the expected subset of genes for a given number of genomes, x is the number of genomes, \exp is Euler's number, and the other terms are constants defined to fit the specific curve. The resulting core genome of the complete dataset (Pan All), the subsets Pan Subsp_pallidum and Pan Subsp_pertenue, have the following $tg(\theta)$ values, respectively: ~ 318 , ~ 627 , and ~ 1038 . Concerning the Singletons of the complete dataset (Pan All) and the subsets Pan Subsp_pallidum, and Pan Subsp_pertenue, have the following $tg(\theta)$ values, respectively: ~ 1 , ~ 0.1 , and ~ 0.025 . According to the least-squares fit of the exponential regression decay, the $tg(\theta)$ represents the point where the curve stabilizes, which may be translated to the number of genes in the core genome after stabilization and the number of singletons that will be added to the pan-genome for each newly sequenced genome. Considering this rule, the core genome of the subset Subsp_pertenue have higher number of core genes (1038-number of core genes) after stabilization, whereas, the complete dataset has the smallest number of core genes (318-number of core genes). For the Singletons, the $tg(\theta)$ value for all the dataset indicates only one gene will be added, whereas, the subsets from Pan Subsp_pallidum and Pan Subsp_pertenue will have 1 and 0.025 newly added genes respectively.

The core genes of the complete dataset, the subsets Pan Subsp_pallidum and Pan Subsp_pertenue, of *T. pallidum* were classified by COG (Cluster of Orthologous Genes) functional category. According to the chart in Fig. 5a-c, the core genome of all the strains had many genes related to the “Metabolism” and “Information storage and processing” categories. Moreover, the majority of the core genome of all the strains were classified as “poorly characterized” (Additional file 1: Table S2A-C).

Detection of PAIs in the *Treponema pallidum* genome

The presence of pathogenicity islands (PAIs) is generally related to evolution in a different genomic environment [22]. However, it may only be the effect of relaxation of purifying selection genes involved in increasing the range of environmental responses. Interspecies genome plasticity may result from several events, of which horizontal gene transfer is particularly important because it can cause the acquisition of blocks of genes (genomic islands, or GIs), producing evolution by quantum leaps [23]. These genes are often flanked by transposases (insertion elements), have altered G + C content and skew, suggesting their acquisition through Horizontal Gene Transfer (HGT), inter-mediated by phages or recombination [22]. PAIs are



important in this context because they represent a class of GIs that carry virulence genes, i.e., factors that enable or enhance the parasitic growth of an organism inside a host [24]. The genome plasticity of all 53 *T. pallidum* strains was determined by using GIPSY (Genomic Island Prediction Software) on subspecies-based study. The software BRIG (BLAST Ring Image Generator) [25] was used for

the circular genome comparison visualization. Some of the other strains from the representing cluster of the dendrogram were also used for the circular genome visualization. We found differences in the presence/absence of pathogenicity islands (PAIs) and genomic islands (GIs) on subspecies-based study: four Pathogenicity Islands (PAIs) eight genomic islands (GIs) in subsp.

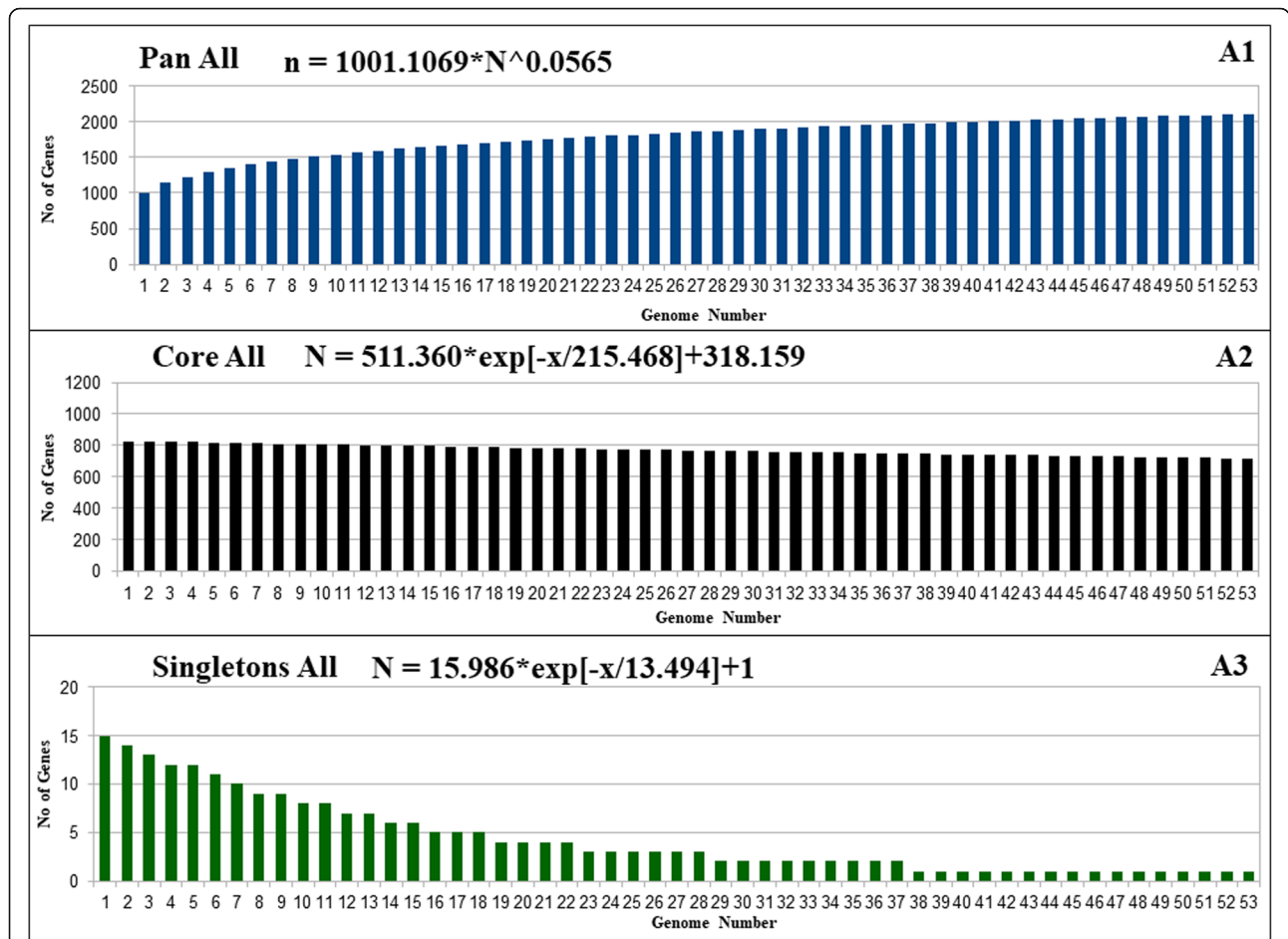


Fig. 2 Pan-genome, core genome and singletons of *T. pallidum*. A1/A2/A3, respectively, showing the pan-genome, core genome and singletons development using 53 strains of *T. pallidum*

pallidum (Fig. 6); three PAIs and seven GIs in subsp. endemicum (Fig. 7); and, three PAIs and eight GIs in subsp. *pertenue* (Fig. 8).

Variations in pathogenicity and Genomic Island in subspecies group

Regarding the presence of genes in PAIs and GIs, we compared the genes in all the subsp. of *T. pallidum* to each other. When compared to each other, we found high similarity of the genes in all the subsp. of *T. pallidum*. The genomic region related to PAIs 2 and PAIs 3 of subsp. *pertenue* and *endemicum* (Non- venereal subsp.) were similar to the PAIs 1 and PAIs 4 of subsp. *Pallidum*. When we compared the genes related to PAIs 2 of subsp. *pertenue* and *endemicum*, there were differences of three genes found that were only present in subsp. *pertenue*. Out of those three genes, two were hypothetical proteins and one was RNA polymerase sigma factor. Furthermore, the genes clusters related to the PAIs 3 of subs. *Pertenue* and *endemicum* were similar to PAIs 4 of subsp. *Pallidum*. Interestingly, we found

the genomic region related to PAIs 1 of subsp. *pertenue* and *endemicum* (Non- venereal subsp.) were not present in any of the GIs or PAIs of subsp. *pallidum*. The list of genes related to PAI 1 of subsp. *pertenue* and *endemicum* is mentioned in Table 1.

On the other hand, we found that the genes present in PAIs 2 of subsp. *pallidum* were not present in any of the GIs or PAIs of subsp. *pertenue* and *endemicum* (Non-venereal subsp.). This may reflect the fact that the genomic signature of those regions has already adapted in subsp. *pallidum* to cause different modes of transmission. The list of genes related to PAI 2 of subsp. *pallidum* is mentioned in Table 2 excluding the hypothetical genes.

Moreover, we also compared GIs of all subspecies; as a result, we found that the genes of some GIs which are present in the GI2 and GI4 in *pallidum* subspecies and are not reported in any of GIs of the subspecies *endemicum* and *pertenue* (Table 3). Most of the genes present in GI2 and GI4 of *pallidum* subspecies are hypothetical genes but some genes are chemotaxis protein (CheA)

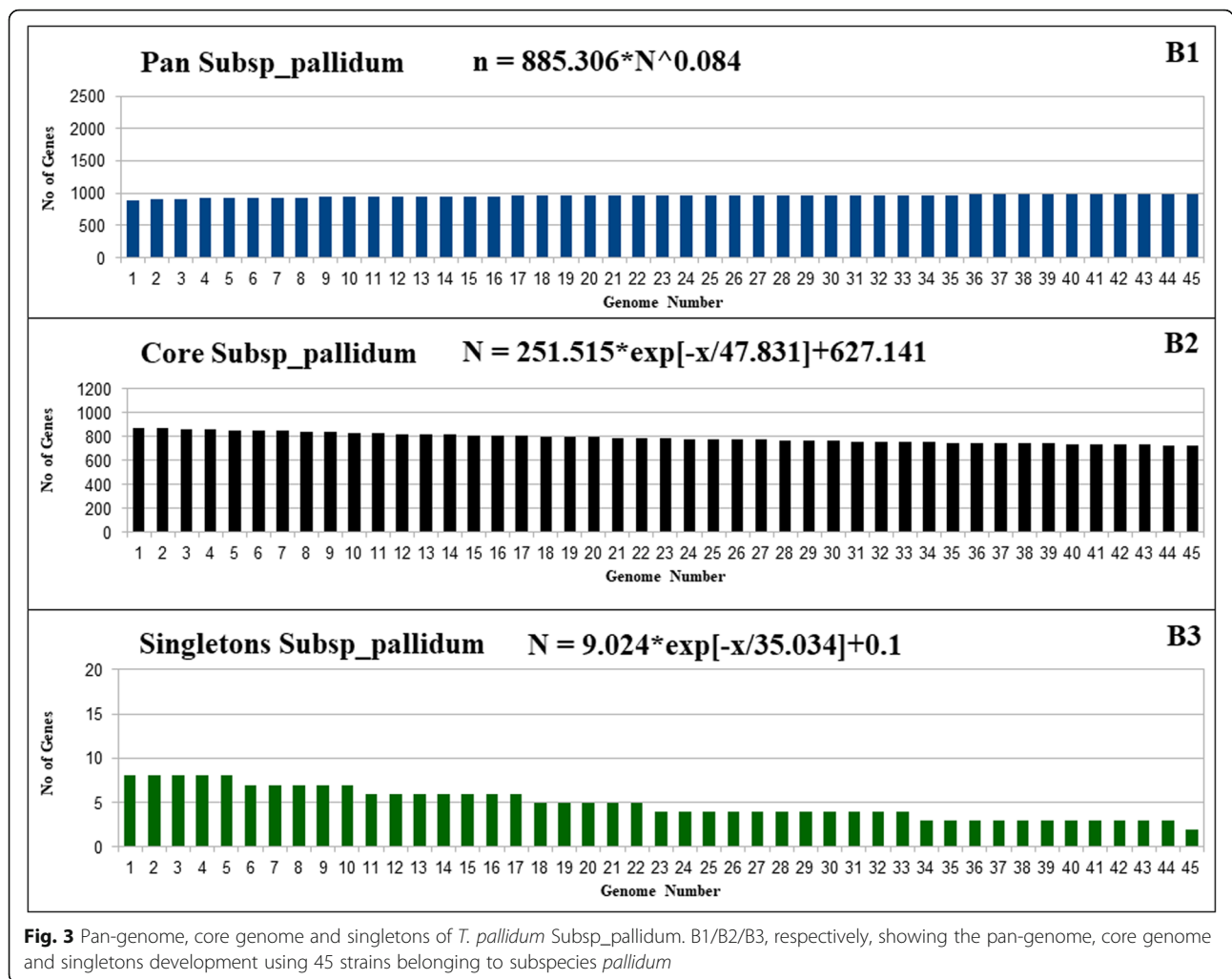


Fig. 3 Pan-genome, core genome and singletons of *T. pallidum* Subsp_pallidum. B1/B2/B3, respectively, showing the pan-genome, core genome and singletons development using 45 strains belonging to subspecies *pallidum*

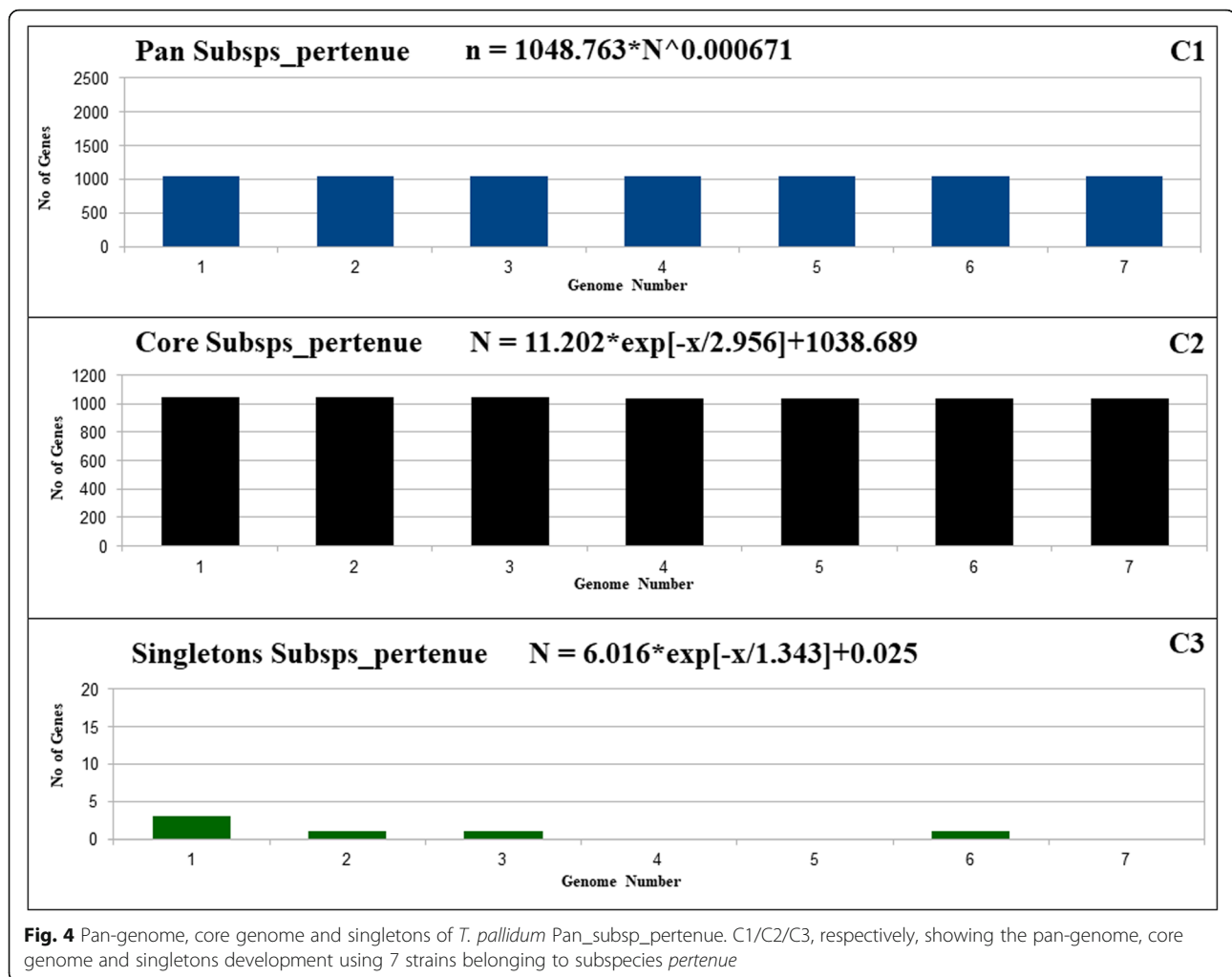
that are associated with the transmission of sensory signals from the chemoreceptors to the flagellar motors [26]. The mechanisms by which *T. pallidum* sense and respond to nutrient gradients help in pathogenic processes such as crossing the endothelial barrier to reach the bloodstream.

Discussion

The subspecies *T. pallidum* subsp. *endemicum* (TEN) and *T. pallidum* subsp. *pertenue* (TPE), are reasons for the diseases bejel and yaws, respectively. In the last few years, *T. pallidum* subsp. *pallidum* (TPA), has been reported as a reemerging pathogen [1, 15]. These three subsp. of *Treponema pallidum* are so close to each other that they cannot be differentiated serologically, their morphology is indistinguishable and are antigenically cross-reactive [27, 28]. Mostly, the disease phenotype caused by these pathogens can only be distinguished clinically and geographically. The distribution of venereal syphilis is global, non-venereal yaws usually effect kids in hot and/or humid regions of Africa and Asia,

endemic syphilis be in dry places like Sahelian Africa and Saudi Arabia [27, 29]. The nature of *T. pallidum* is highly invasive. It circulates through bloodstream and lymphatics and overruns a wide-ranging of tissues and organs. As demonstrated by the widespread clinical manifestations related to syphilis infections, *Treponema pallidum* subsp. *pallidum* crosses placental, endothelial and blood-brain barriers early in infection, the incidence of congenital syphilis and invasion of central nervous system has been observed in almost 40% of early syphilis patients. Though, the understanding of the mechanisms responsible for the widespread distribution capability of *T. pallidum* is still very limited [30, 31].

The transmission of yaws is characterized by direct contact on skin and primary cutaneous lesion. It is facilitated by damaged skin surface. Scratching or rubbing these damaged parts of the body can facilitate the lesions spread across the body [28, 29]. Contrarily, endemic syphilis is an acute infection. Primary lesions of endemic syphilis can be seen in the children of ages between 2 and 15 years in dry and arid climates. While the mode of



transmission is not known, it is believed that it may occur through mucosal and skin contact, even via shared eating utensils or drinking vessels [28, 29].

The defined relationships among the bacteria are still argued. The expansion of next-generation sequencing (NGS) in last few decades influences the fields of treatment and prevention, especially about bacterial diseases [32]. The ability of genomics data of *T. pallidum* gives us better understanding of the biology involving its interaction with its hosts. A comprehensive in silico pan-genome study was carried out for 53 sequenced genomes of *T. pallidum*, which indicates that the pan-genome of *T. pallidum* is still open; however, it is increasing at a very low rate as represented by the α of 0.9435 for the Pan All and the α of 0.916 and 0.999329 for Pan Subsp_pallidum and Pan_subsp_pertenue, respectively. Moreover, the α of 0.999329 indicates that the Pan Subsp_pertenue is almost closed, which is corroborated by the $tg(\theta)$ of ~ 0.025 .

The genome plasticity analysis reveals the differences in the presence and absence of some genome regions

when compared at the subspecies level. Pathogenicity islands carry the genes related to the virulence, which are essential and characterize a class of Genomics Island [33]. The comparative analysis of PAIs and GIs showed the absence of genes at the subspecies level. We found gene clusters, that are related to amino acid and lipid biosynthesis, belonging to PAIs 2 of *T. pallidum* subsp. *pallidum* have not been identified in any PAIs or GIs of *T. pallidum* subsp. *endemicum* and *T. pallidum* subsp. *pertenue*. It might be possible that these genes help bacteria to execute different modes of infection at subsp. level of *T. pallidum*. Acyl carrier protein (ACP) synthase (AcpS) catalyzes the transfer of the 4'-phosphopantetheine moiety from coenzyme A (CoA) onto a serine residue of apo-ACP, to convert apo-ACP to the functional holo-ACP. During the biosynthesis of fatty acids and phospholipids, the holo form of bacterial ACP plays a vital role in mediating the transfer of acyl fatty acid. AcpS is therefore an attractive target for therapeutic interpolation. It has been reported that, AcpS enzymes from *Mycoplasma pneumoniae* and *S. pneumoniae* may

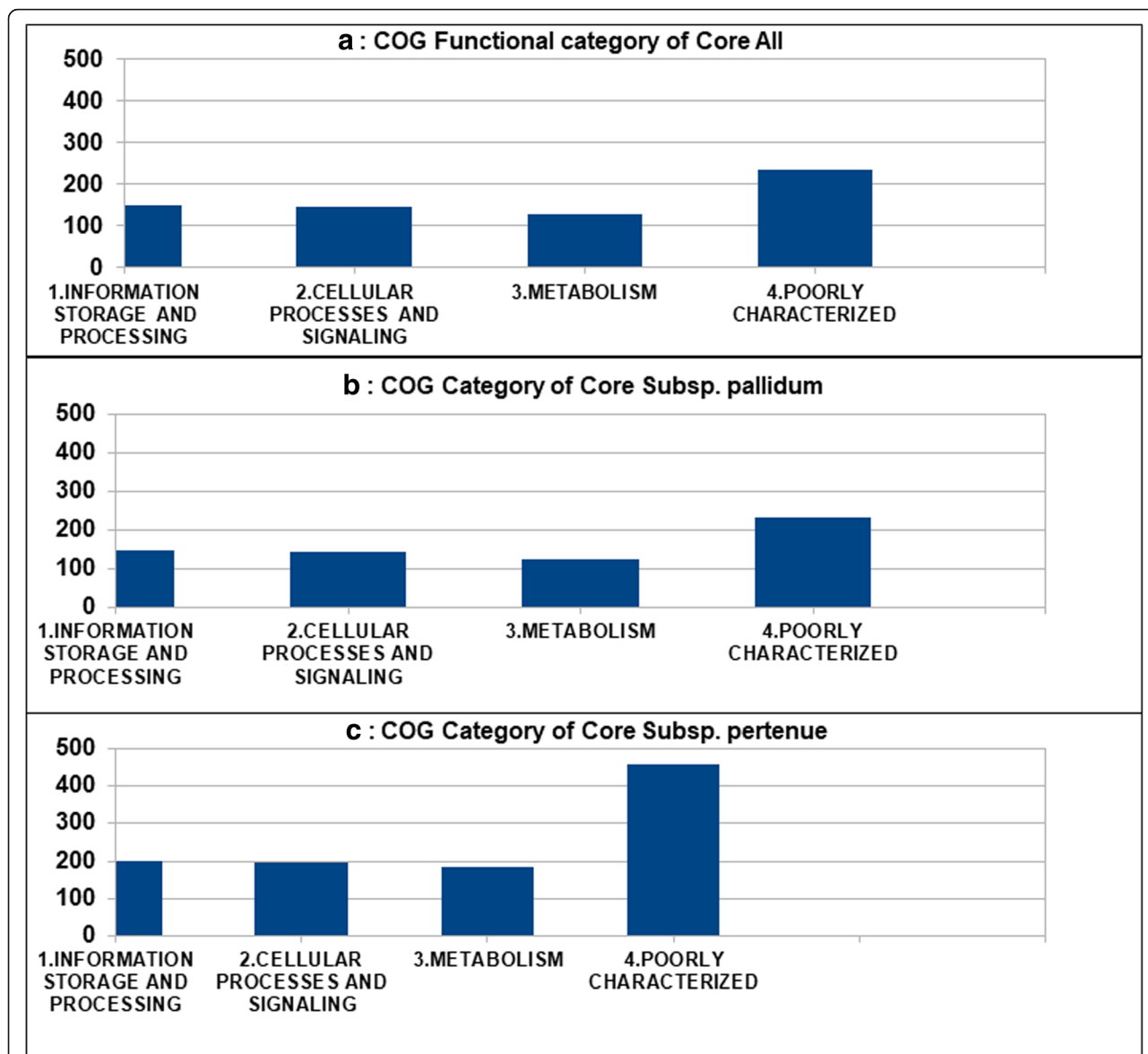


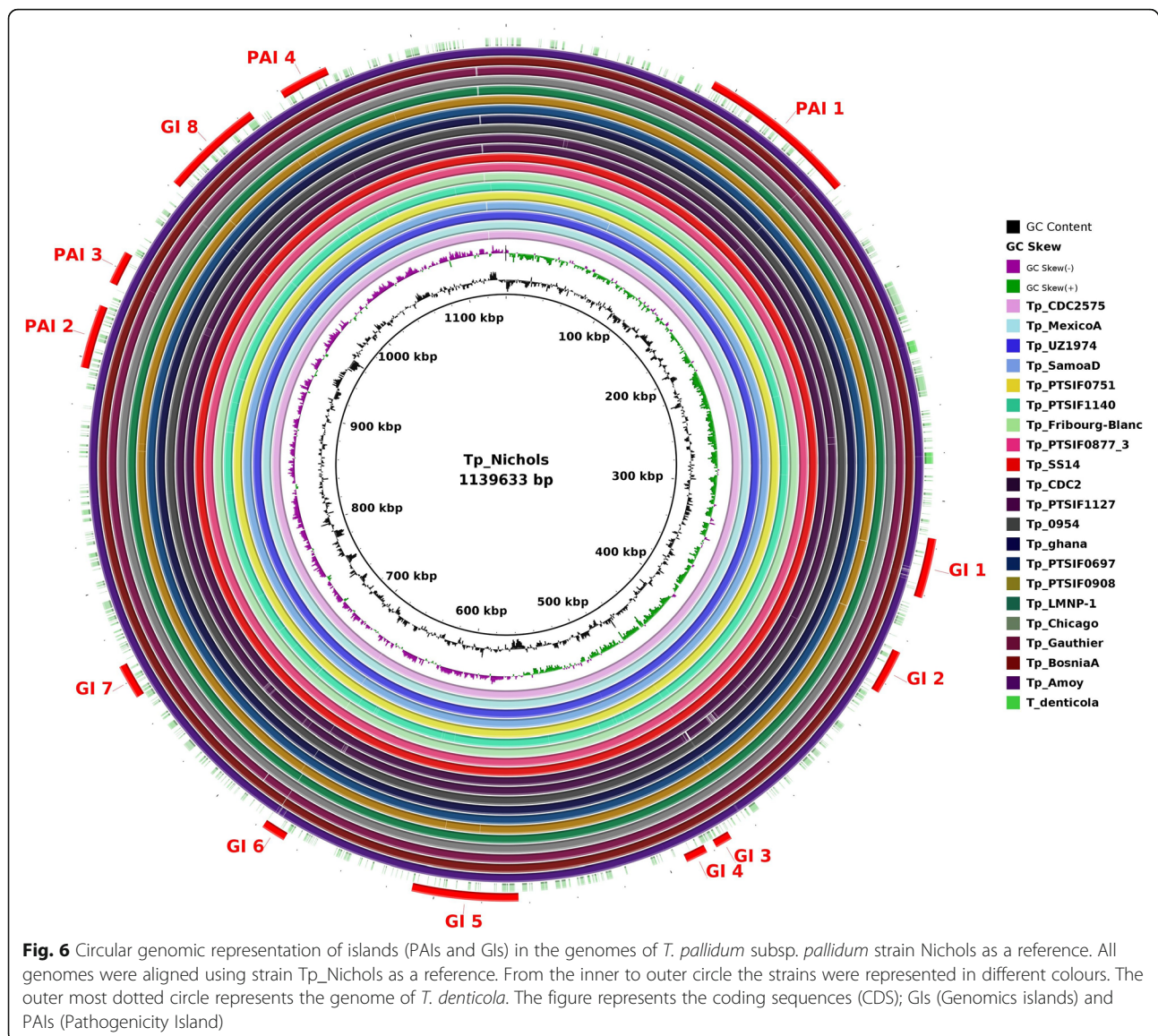
Fig. 5 Graphical representation of COG (Cluster of Orthologous Genes) functional categories of identified core genes. **a, b, c** showing the core genes belonging to the Information storage and processing, cellular processing and signaling, Metabolism and poorly characterized functional categories for complete dataset, the subsets Pan Subsp_pallidum and Pan Subsp_pertenuis of *T. pallidum*, respectively

play a crucial role in the acylation of fatty acids derived from human tissues for their lipid biosynthesis, suggesting that AcpS is a more striking antimicrobial target for discovery of novel antibiotics than bacterial fatty acid biosynthetic enzymes [34, 35].

Moreover, the presence of chemotaxis protein (CheA) in different GIs of *T. pallidum* subsp. *pallidum* might be responsible for different molecular modes of infection as *T. pallidum* genome contains two operons for the Che response regulators [31, 36]. The bacterial transcription-repair coupling factor (TRCF) is a large, multi-domain, SF2 ATPase that is generally conserved. It forms the dual of nucleotide excision repair with transcription by dislodging

inactive RNA polymerase molecules stalled at template DNA lesions, and by increasing the rate at which the Uvr(A) BC exonuclease acts at these sites [37].

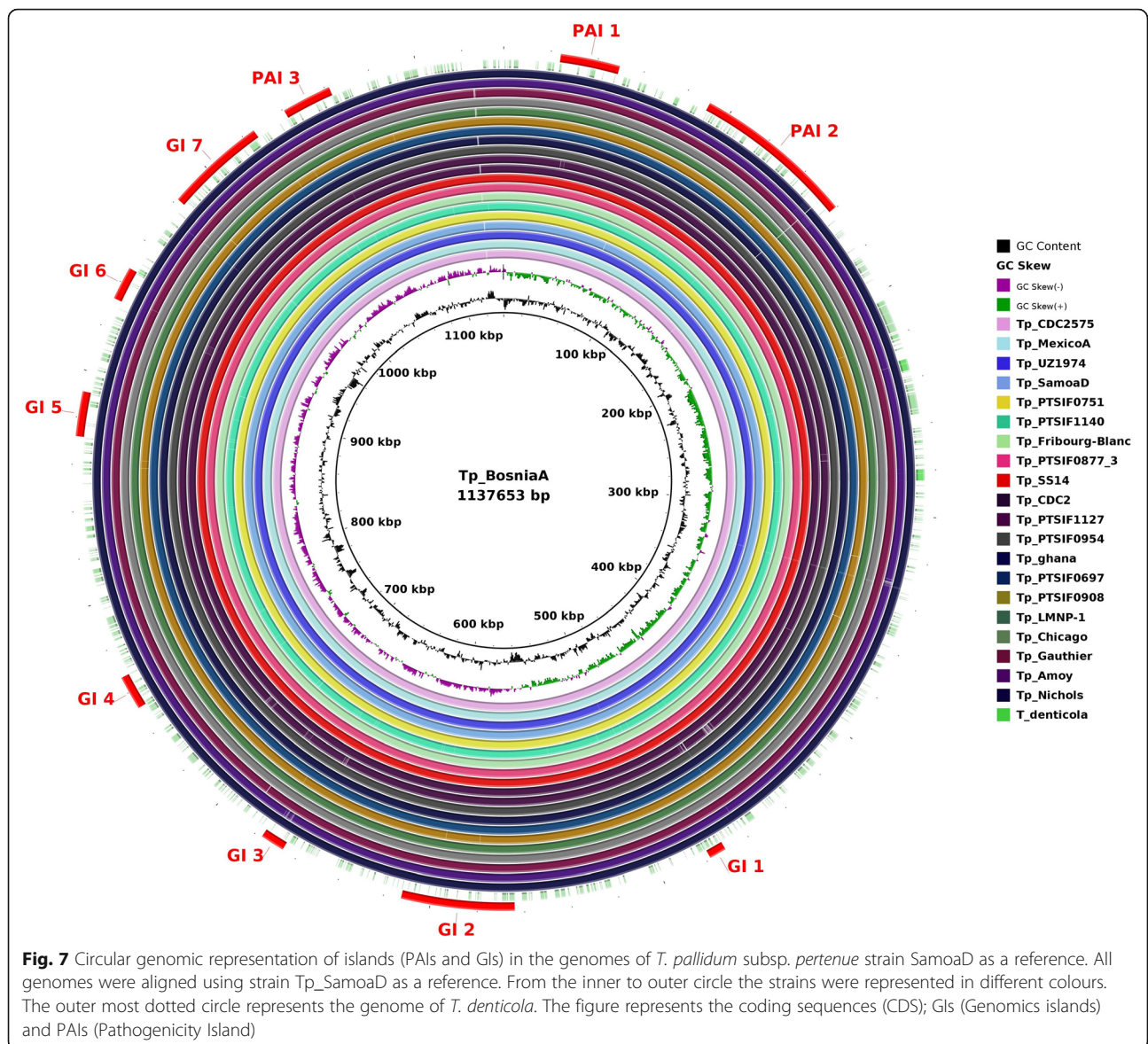
Pathogens are frequently using antigenic variation mechanisms to elude the adaptive immune response that ultimately results in persistent infection [38]. It might be because of the variation in expression of different Tpr proteins in the syphilis spirochete, *Treponema pallidum* subsp. *pallidum*, that have important implications in its ability to elude host immune detection [39]. A 12-membered protein family *Treponema pallidum* repeat (tpr) has been identified in *T. pallidum* subsp. *Pallidum*, which may be concerned in the pathogenesis of *T.*



pallidum [38, 40]. On the basis of amino acid homology, these 12 Tprs are further divided into three subfamilies: subfamily I (TprC, D, F, I), subfamily II (TprE, G, J), and subfamily III (TprA, B, H, K, L) [40] [41].

Despite the host's efforts to eliminate the infection, mechanisms of *T. pallidum*'s persistence include residence within intracellular or immune-privileged positions to hide from the immune effectors. *T. pallidum*'s has the ability to cape its surface with host serum proteins or mucopolysaccharides to dodge immune response and immunosuppression of the host triggered by syphilis infection [38]. Freeze-fracture electron microscopy of *T. pallidum* has revealed lack of integral membrane proteins in the outer membrane (OM) of *T. pallidum*, conceivably accounting for the reasonably poor antigenicity of this spirochete's surface [38, 42, 43].

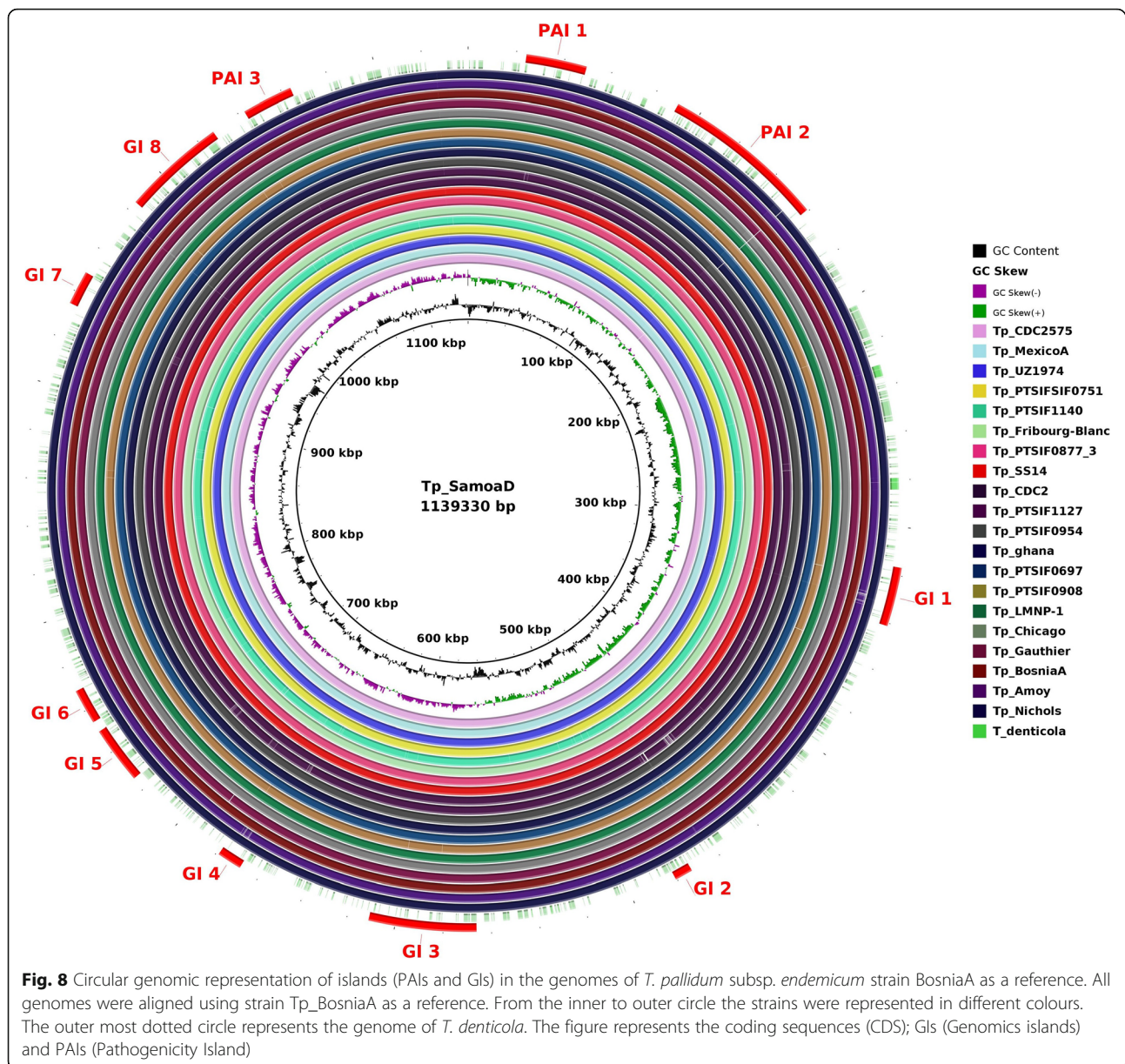
However, as *T. pallidum* could be phagocytized in the presence of opsonic antibody, antibody targets must be present on the surface of the bacterium. Furthermore, the treponemes harvested from the tissues of later stage infections after the elimination of majority of treponemes are resistant to opsonophagocytosis. It raised the likelihood of antigenic variation occurring in *T. pallidum*, but no exact variable antigen was identified [38, 44]. Following the identification and investigation of TprK, provides the first candidate antigen of *T. pallidum* that might function in fudging the immune response. TprK vary among and within *T. pallidum* strains, with diversity of sequence localized in seven distinct regions (V1-V7) bordered by conserved domains [38, 43, 45, 46]. During experimental infection, these V regions are the main targets of the host humoral immune response [38].



Antigenic variation of the TprK antigen has been acknowledged to explain the persistence of *T. pallidum* in the host.

Recent work of Dan Liu et al. [47] has recognized an improved number of variants within these seven V regions of the tprK gene in the samples of secondary syphilis. A 3-bp changing pattern was observed in the sequences within each V region of the protein. However, same pattern of change was observed in variable sequences within the V regions of tprK in the secondary syphilis. Notably, the amino acid sequences IASDGGAIKH and IASEDGSAGNLKH in V1 are not only present in high proportion in inter-strain comparison but also were found at a quite high frequency in the populations. The alignment of all amino acid sequences revealed some really stable pattern

within each V region of the primary and secondary syphilis samples, particularly the amino acid sequences IASDGGAIKH and IASEDGSAGNLKH in V1 region. The highly stable peptides found in V1 region are likely promising vaccine components. The highly heterogenetic regions (e.g., V6) could help to understand the role of tprK in fudging immune response. However, in our analysis, we found that some of tpr genes (*tprC*, *tprD*, *tprF*, *tprI*, *tprJ*) were present in some of PAIs or GIs *T. pallidum* subsp. *endemicum* (TEN) and *T. pallidum* subsp. *pertenue* (TPE). While, the GIs and PAIs related to *T. pallidum* subsp. *pallidum* we only identified some tpr domain proteins. It has been reported by Maděrankova et al. 2019, tpr genes responsible for the adaptive evolution of the pathogen [48].



Conclusions

Apart from establishing phylogenetic relationships among treponemal species and subspecies, the addition of comparative genomics was also required to illuminate the lower degree of virulence associated with *T. pallidum* subsp. *pertenue* than with *T. pallidum* subsp. *pallidum*. Unlike syphilis, it is said that yaws cannot be transmitted vertically or affect the central nervous system. It is rather limited to skin, bones, joints and soft tissues. In the 1980s, a very limited genetic diversity between these pathogens was established when hybridization experiments were carried out with DNA isolated from yaws and syphilis strains [29]. Our work also showed that genomes of syphilis, yaws, and Bejel treponemes share 97–100% overall similarity, as well as the

identical organization. This evidence proposes that small genetic changes in key genes among these organisms could be responsible for the reported differences in disease pathogenesis. Considering the genes in PAIs and GIs, we identified some absence of pathogenicity islands in all subspecies. Genes which are present in *pallidum* subspecies pathogenicity islands (PAIs) or genomic islands (GIs) are absent in the subspecies *endemicum* and *pertenue*. The findings of this analysis are very important, as it can help in the understanding of molecular basis of infections from *T. pallidum* subsps. Furthermore, the core genes represent the most desirable source for the selection of conserved genes; therefore, characterization of such poorly studied proteins helps in understanding the cellular metabolism, mode of infection

Table 1 The list of genes related to PAI 1 of subs. *Pertenuis* and *endemicum*

Protein ID	PAIs Coordinates	Protein Name	Functions (MF: Molecular Function, BP: Biological Process)
WP_012460510.1	22,054–24,177	VWFA domain-ontaining protein	–
WP_010881470.1	25,115–2538	Hypothetical protein	
WP_014342234.1	25,459–27,504	Hypothetical protein	–
WP_010881472.1	27,565–28,896	Sodium-dependent tryptophan transporter	MF: Neurotransmitter:sodium symporter activity BP
WP_010881473.1	29,076–29,786	Potassium transporter Trk	MF: Cation transmembrane transporter activity BP: Potassium ion transport
WP_010881474.1	29,865–32,936	M16C subfamily peptidase	MF: Catalytic activity, Metal ion binding BP: Proteolysis
WP_010881475.1	32,965–33,987	Flagellar motor switch protein FlIG	MF: Motor activity BP: Bacterial-type flagellum-dependent cell motility, Chemotaxis
WP_014342235.1	34,060–35,433	Putative hemolysin HlyC	MF: Flavin adenine dinucleotide binding BP
WP_010881477.1	35,447–36,808	Putative hemolysin HlyC	MF: Flavin adenine dinucleotide binding BP
WP_039487502.1	36,916–38,193	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	–
WP_010881479.1	38,312–39,946	60 kDa chaperonin	MF: ATP binding Source, Unfolded protein binding BP: Protein refolding
WP_014342237.1	40,010–40,288	Hypothetical protein	
WP_010881481.1	40,285–41,085	Ribosomal RNA small subunit methyltransferase E	MF: Methyltransferase activity BP: rRNA processing
WP_014342238.1	41,082–41,750	Hypothetical protein	
WP_039486943.1	41,794–42,924	Zinc (Zn ²⁺) ABC superfamily ATP binding cassette transporter, binding protein	MF: Metal ion binding BP: Cell adhesion, Metal ion transport
WP_010881484.1	42,942–43,658	Zinc (Zn ²⁺) ABC superfamily ATP binding cassette transporter, ABC protein	MF: ATPase activity, ATP binding BP
WP_010881485.1	43,658–44,458	Zinc (Zn ²⁺) ABC superfamily ATP binding cassette transporter, membrane protein	MF: ATPase-coupled transmembrane transporter activity BP:
WP_010881486.1	44,567–45,562	Lactate dehydrogenase	MF: D-lactate dehydrogenase activity, NAD binding BP
WP_014342240.1	45,659–46,711	Putative regulatory protein PfoR	MF: Protein-N (PI)-phosphohistidine-sugar phosphotransferase activity BP: Phosphoenolpyruvate-dependent sugar phosphotransferase system
WP_014342241.1	46,739–46,918	Hypothetical protein	
WP_039486948.1	46,945–49,392	Putative methyl-accepting chemotaxis protein	MF: Transmembrane signaling receptor activity BP: Chemotaxis, Signal transduction
WP_010881491.1	49,513–50,445	Peptidoglycan-binding protein LysM	–

and regulation of gene expression of *Treponema pallidum*. Hence, this study can help to better understand the molecular modes of bacterial infection and are significance for vaccine development for syphilis.

Methods

Genome sequences

The genome sequences of 53 *T. pallidum* strains were retrieved from the NCBI (National Centre for Biotechnology

Table 2 The list of genes related to PAI 2 of subsp. *pallidum*

Protein ID	PAIs Coordinates	Protein Name	Functions (MF: Molecular Function, BP: Biological Process)
WP_010882272.1	895,372–895,749	holo-ACP synthase	MF: holo-[acyl-carrier-protein] synthase activity, magnesium ion binding. BP: Fatty acid biosynthetic process
WP_010882273.1	895,746–896,630	membrane protein	
WP_010882275.1	897,479–899,248	arginine--tRNA ligase	MF: arginine-tRNA ligase activity, ATP binding BP: arginyl-tRNA aminoacylation
WP_010882284.1	910,864–913,113	MFS transporter	
WP_010882286.1	914,944–915,711	methionine aminopeptidase	Aminopeptidase activity, metal ion binding, metalloexopeptidase activity
WP_010882287.1	915,835–916,659	Heat shock protein, putative	
WP_010882288.1	916,752–917,804	glyceraldehyde-3-phosphate dehydrogenase	MF: glyceraldehyde-3-phosphate dehydrogenase (NAD+) phosphorylating activity. BP: glucose metabolic process, glycolytic process.
WP_010882292.1	919,085–919,453	50S ribosomal protein L20	MF: rRNA binding, structural constituent of ribosome. BP: translation.
WP_010882293.1	919,486–919,686	50S ribosomal protein L35	MF: structural constituent of ribosome, BP: translation.
WP_010882295.1	920,441–922,615	DUF4954 domain-containing protein	
WP_010882296.1	922,644–923,642	prolipoprotein diacylglycerol transferase	MF: transferase activity, transferring glycosyl groups. BP: lipoprotein biosynthetic process.

Information) database (<https://www.ncbi.nlm.nih.gov/genome/genomes/741?>) (Accessed June 2018): 46 genomes of *T. pallidum* subsp. *pallidum* were isolated from different parts of human body, rabbits and baboons (USA, China & Portugal). Six genomes from Africa and Australia/ Oceania continents (strain SamoaD, CDC2, Gauthier, CDC2575, Ghana051 and LMNP-1) from subsp. *pertenue* were isolated from humans, baboons and rabbits (Additional file 1: Table S1). One genome of *Treponema pallidum* subsp. *endemicum* (strain BosniaA) was isolated in Europe from human tongue and tonsils. The genome of *Treponema denticola* strain ATCC 35405 was used as non-pathogenic bacteria in this work. The general information about all *T. pallidum* strains and the Complete workflow applied in this work are given in Additional file 1: Table S1 and Figure S1, respectively.

Phylogenomic analysis of all *Treponema pallidum* strains

For phylogenomic analysis of all *Treponema pallidum* strains, Gegenees (version 2.1) [19] was used. The Gegenees software was used to perform an all-versus-all similarity search. It divides the genomes into small sequences and determines the minimum content shared by all the genomes. Subsequently, the obtained minimum shared contents were subtracted from all the genomes resulting in the variable contents, which were eventually compared with all the other strains for the calculation of the percentages of similarity. Finally, these

percentages were plotted in a heatmap chart with a spectrum ranging from low similarity (red) to high similarity (green). The Gegenees data was exported as a distance matrix file in nexus format (.nex) and, further, the generated distance matrix was used as an input file in SplitsTree software (version 4.14.5) [49] using neighbour joining method to create a dendrogram [50, 51].

Prediction of Pan-genome, Core-genome and singleton

We divided 53 strains of *T. pallidum* in 3 subsets for pan-genome calculation. We performed Pan All (with all 53 strains of *T. pallidum*), Pan Subsp_pallidum and Pan_subsp_pertenue (based on subspecies). For the identification of core genome (commonly shared by all strains), shared genome (genes present in two or more than two strains but not shared by all strains) and singletons (strain specific genes), we used OrthoFinder [52]. Briefly, OrthoFinder uses the .faa amino acid sequence file for each genome to perform *all-vs-all* BLASTp for the Orthologous analysis. It uses MCL (Markov Clustering algorithm) program to determine the Orthologous genes [53]. The cut-off value of $1e^{-10}$ was used for Pan-genome, Core-genome and singletons identification for all the subsets. Furthermore, *in-house* scripts were used to estimate the fixed parameters for Heap's Law (pan-genome analyses) [20, 51] and least-squares fit of the exponential regression decay (core-genome and singletons). The extrapolations of the pan-genomes from the

Table 3 The list of genes related to GI 2 and GI 4 of subsp. *pallidum*. The table shows the list of proteins excluding the hypothetical proteins

Protein ID	GI 2 and GI 4 Coordinates	Protein Name	Functions (MF: Molecular Function, BP: Biological Process)
WP_010881790.1	365,067–365,627	cytidylate kinase	MF: ATP binding, cytidylate kinase activity BP: pyrimidine nucleotide metabolic process.
WP_010881791.1	365,621–366,454	adenine glycosylase	MF: catalytic activity. DNA binding. BP: base-excision repair.
WP_010881792.1	366,459–369,881	transcription-repair coupling factor	MF: ATP binding, damaged DNA binding, hydrolase activity. BP: regulation of transcription, DNA-template, transcription-coupled nucleotide-excision repair.
WP_010881793.1	370,034–371,125	phospho-N-acetylmuramoyl-pentapeptide- transferase	MF: phospho-N-acetylmuramoyl-pentapeptide-transferase activity, UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2, transferase activity. BP: cell cycle, cell divisio, cell wall organization.
WP_010881797.1	373,895–374,425	FKBP-type peptidyl-prolyl cis-trans isomerase	MF: Metal ion binding, peptidyl-prolyl sis-trans isomerase activity. BP: protein refolding
WP_010881798.1	374,639–375,925	gamma-glutamyl-phosphate reductase	MF: glutamate-5semialdehyde dehydrogenase activity. BP: L-proline biosynthetic process.
WP_010881799.1	375,922–376,812	glutamate 5-kinase	MF: ATP-binding BP: L-proline biosynthetic process.
WP_010881801.1	377,198–377,707	ribonuclease H	MF: Metal-ion binding, nucleic acid binding
WP_010881802.1	377,970–378,596	thymidylate kinase	MF: ATP binding, thymidylate kinase activity. BP: dTDP biosynthetic process
WP_010881806.1	377,970–378,596	glycosyl hydrolase	MF: catalytic activity. BP: carbohydrate metabolic process.
WP_014342391.1	382,926–383,729	lysophosphatidic acid acyltransferase	MF: transferase activity. BP: metabolic process
WP_014505476.1	384,584–387,016	chemotaxis protein CheA	MF: ATP binding, phosphorelay sensor kinase activity BP: chemotaxis.
WP_010881907.1	488,333–488,911	SMC-Scp complex subunit ScpB	BP: cell division, chromosome separation
WP_014342436.1	488,928–489,725	RsuA family pseudouridine synthase	MF: pseudouridine synthase activity, RNA binding.
WP_010881910.1	490,476–490,835	transcriptional regulator	
WP_010881913.1	492,278–493,030	tRNA (guanine-N(7)-)-methyltransferase	MF: tRNA (guanine-N7)-methyltransferase activity.

complete dataset and all subsets were calculated based on Heap's Law [20, 51], which was used to calculate whether the pan-genome was open or closed. Heap's Law is an empirical law represented by the formula $n = k \cdot N^\gamma$; it describes the number of distinct words in a document (or set of documents) as a function of the document length. In a genetic context, n is the expected number of genes for a given number of genomes, N determines the number of genomes, and the k and γ ($\alpha = 1 - \gamma$) are free parameters that are determined empirically. According to Heap's Law, when $\alpha > 1$ ($\gamma < 0$), the pan-genome is considered to be closed, and there will be no significant increase in the number of genes with the addition of a new genome. On the other hand, when $\alpha < 1$ ($0 < \gamma < 1$), the pan-genome is open and there will be a significant increase in the number of genes for each newly added genome.

Genomic and Pathogenicity Islands prediction

This section describes the analyses that were performed for the prediction of genomic and pathogenicity Islands following three datasets based on the subspecies: A) using *T. pallidum* subsp. *pallidum* strain Nichols as a reference; B) using *T. pallidum* subsp. *pertenue* strain SamoaD as a reference; and C) using *T. pallidum* subsp. *endemicum* strain BosniaA as a reference. The islands predictions for three datasets were determined by using GIPSY (Genomic Island Prediction Software) [33]. GIPSY is a multi-step approach that predicts Genomic islands (GIs) and Pathogenicity islands (PAIs). PAIs and GIs predictions are based on commonly shared features such as genomic signature deviation (anomalous G + C content and codon usage deviation), presence of transposase genes; metabolism, virulence, antibiotic resistance, or symbiosis-related genes; flanking tRNA genes; and

absence in other organisms of the same genus or closely related species [33]. *T. denticola* strain ATCC 35405 was used as a non-pathogenic species from the same *Treponema* genus for GIs and PAIs prediction [54]. The sizes of the islands were compared with all the other strains via ACT (Artemis Comparison Tool) software [55]. PAIs regions were plotted using the software BRIG [25]. Following the curation of the PAIs, the genes of all the islands in each strain were assessed for their presence/absence in all the other strains.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-019-6430-6>.

Additional file 1: Table S1. General information about 53 *Treponema pallidum* Strains used in this work. List of all *Treponema pallidum* strains (with features) retrieved from the NCBI (National Center for Biotechnology Information) database. **Table S2A.** The COG functional categories with detailed description of Core genes: The table showing the number of core genes of the complete dataset were classified by COG (Cluster of Orthologous Genes) functional category. **Table S2B.** The COG functional categories with detailed description of Core genes: The table showing the number of core genes of the Pan Subsp_pallidum dataset were classified by COG (Cluster of Orthologous Genes) functional category. **Table S2C.** The COG functional categories with detailed description of Core genes: The table showing the number of core genes of the Pan Subsp_pertenuis dataset were classified by COG (Cluster of Orthologous Genes) functional category. **Figure S1.** The Complete workflow applied in this work. The figure represent the methodology and software were used in this analysis. **Figure S2.** The heatmap analysis of 53 Strains of *Treponema pallidum*. The figure represents the comparison between the variable content of all strains. The percentages were plotted in the heatmap with a spectrum ranging from red (low similarity) to green (high similarity). The names of the strains on the left side of the figure (vertically) are organized in the same order in the top part of the figure (horizontally). Once Gegenees uses the similarities in the variable contents, the outgroup normally presents a very small percentage of similarity to the other strains.

Abbreviations

CDS: Coding sequences; COG: Cluster of Orthologous Genes; GIs: Genomic islands; PAIs: Pathogenicity islands; TEN: *Treponema pallidum* subsp. endemicum; TPA: *Treponema pallidum* subsp. pallidum; TPE: *Treponema pallidum* subsp. pertenuis; WHO: World Health Organization

Acknowledgments

We acknowledge the collaboration and assistance of all team members and the Brazilian funding agencies CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil), and FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais). Arun Kumar Jaiswal was supported by the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil) fellowship for doctoral studies. Syed Babar Jamal acknowledges the "TWAS-CNPq Postgraduate Fellowship Programme" for granting a fellowship for doctoral studies.

Authors' contributions

AKJ, ST, SBJ, LCO, LGA, SCS conceived, designed the protocol, collected and analysed initial data, wrote the paper: ST, SCS, VA, CJFO coordinated and led the entire project: AKJ, ST, SBJ, SCS, PG, VA, CJFO Cross-checked all data, analysis, wrote the paper: All authors read and approved the manuscript.

Funding

No funding supported this research.

Availability of data and materials

All data generated and analysed during this study are included in this published article and its supplementary information files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. ²Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, MG, Brazil. ³Department of Biological Sciences, National University of Medical Sciences, Abid Majeed Road, Rawalpindi, Punjab 46000, Pakistan. ⁴Department of Computer Science, Virginia Commonwealth University, Richmond VA-23284, USA.

Received: 10 October 2019 Accepted: 24 December 2019

Published online: 10 January 2020

References

1. Peeling RW, Hook EW 3rd. The pathogenesis of syphilis: the great mimicker, revisited. *J Pathol.* 2006;208(2):224–32.
2. Mitja O, Smajs D, Bassat Q. Advances in the diagnosis of endemic treponematoses: yaws, bejel, and pinta. *PLoS Negl Trop Dis.* 2013;7(10):e2283.
3. Radolf JD, Deka RK, Anand A, Smajs D, Norgard MV, Yang XF. *Treponema pallidum*, the syphilis spirochete: making a living as a stealth pathogen. *Nat Rev Microbiol.* 2016;14(12):744–59.
4. Centurion-Lara A, Molini BJ, Godornes C, Sun E, Hevner K, Van Voorhis WC, Lukehart SA. Molecular differentiation of *Treponema pallidum* subspecies. *J Clin Microbiol.* 2006;44(9):3377–80.
5. Nyatsanza F, Tipple C. Syphilis: presentations in general medicine. *Clin Med (Lond).* 2016;16(2):184–8.
6. Stamm LV. Global challenge of antibiotic-resistant *Treponema pallidum*. *Antimicrob Agents Chemother.* 2009;54(2):583–9.
7. Stamm LV. Global challenge of antibiotic-resistant *Treponema pallidum*. *Antimicrob Agents Chemother.* 2010;54(2):583–9.
8. Celum CL. Sexually transmitted infections and HIV: epidemiology and interventions. *Top HIV Med.* 2010;18(4):138–42.
9. Dhawan J, Gupta S, Kumar B. Sexually transmitted diseases in children in India. *Indian J Dermatol Venereol Leprol.* 2010;76(5):489–93.
10. Kumar Jaiswal A, Tiwari S, Jamal SB, Barh D, Azevedo V, Soares SC. An in silico identification of common putative vaccine candidates against *Treponema pallidum*: a reverse vaccinology and subtractive genomics based approach. *Int J Mol Sci.* 2017;18(2):402.
11. Tucker JD, Cohen MS. China's syphilis epidemic: epidemiology, proximate determinants of spread, and control responses. *Curr Opin Infect Dis.* 2011; 24(1):50–5.
12. Uuskula A, Puur A, Toompere K, DeHovitz J. Trends in the epidemiology of bacterial sexually transmitted infections in eastern Europe, 1995–2005. *Sex Transm Infect.* 2010;86(1):6–14.
13. Brown DL, Frank JE. Diagnosis and management of syphilis. *Am Fam Physician.* 2003;68(2):283–90.
14. Cameron CE, Lukehart SA. Current status of syphilis vaccine development: need, challenges, prospects. *Vaccine.* 2014;32(14):1602–9.
15. Peeling RW, Mabey D, Kamb ML, Chen XS, Radolf JD, Benzaken AS. Syphilis. *Nat Rev Dis Primers.* 2017;3:17073.
16. Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A.* 2005;102(39):13950–5.
17. Miao R, Fieldsteel AH. Genetics of *Treponema*: relationship between *Treponema pallidum* and five cultivable treponemes. *J Bacteriol.* 1978; 133(1):101–7.

18. Miao RM, Fieldsteel AH. Genetic relationship between *Treponema pallidum* and *Treponema pertenue*, two noncultivable human pathogens. *J Bacteriol*. 1980;141(1):427–9.
19. Agren J, Sundstrom A, Hafstrom T, Segerman B. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS One*. 2012; 7(6):e39107.
20. Guimaraes LC, Florczak-Wyspianska J, de Jesus LB, Viana MV, Silva A, Ramos RT, Soares Sde C, Soares Sde C. Inside the Pan-genome - methods and software overview. *Curr Genomics*. 2015;16(4):245–52.
21. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008;11(5):472–7.
22. Mira A, Martin-Cuadrado AB, D'Auria G, Rodriguez-Valera F. The bacterial pan-genome: a new paradigm in microbiology. *Int Microbiol*. 2010;13(2):45–57.
23. Schmidt H, Hensel M. Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev*. 2004;17(1):14–56.
24. Karaolis DK, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, Reeves PR. A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proc Natl Acad Sci U S A*. 1998;95(6):3134–9.
25. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*. 2011;12:402.
26. Greene SR, Stamm LV, Hardham JM, Young NR, Frye JG. Identification, sequences, and expression of *Treponema pallidum* chemotaxis genes. *DNA Seq*. 1997;7(5):267–84.
27. Knauf S, Liu H, Harper KN. Treponemal infection in nonhuman primates as possible reservoir for human yaws. *Emerg Infect Dis*. 2013;19(12):2058–60.
28. Gogarten JF, Dux A, Schuenemann VJ, Nowak K, Boesch C, Wittig RM, Krause J, Calvignac-Spencer S, Leendertz FH. Tools for opening new chapters in the book of *Treponema pallidum* evolutionary history. *Clin Microbiol Infect*. 2016;22(11):916–21.
29. Giacani L, Lukehart SA. The endemic Treponematoses. *Clin Microbiol Rev*. 2014;27(1):89–115.
30. Lithgow KV, Hof R, Wetherell C, Phillips D, Houston S, Cameron CE. A defined syphilis vaccine candidate inhibits dissemination of *Treponema pallidum* subspecies pallidum. *Nat Commun*. 2017;8(1):1–10.
31. LaFond RE, Lukehart SA. Biological basis for syphilis. *Clin Microbiol Rev*. 2006;19(1):29–49.
32. Arora N, Schuenemann VJ, Jager G, Peltzer A, Seitz A, Herbig A, Strouhal M, Grillova L, Sanchez-Buso L, Kuhnert D, et al. Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nat Microbiol*. 2016;2:16245.
33. Soares SC, Geyik H, Ramos RT, de Sa PH, Barbosa EG, Baumbach J, Figueiredo HC, Miyoshi A, Tauch A, Silva A, et al. GIPSy: genomic island prediction software. *J Biotechnol*. 2016;232:2–11.
34. Heath RJ, Rock CO. Fatty acid biosynthesis as a target for novel antibacterials. *Curr Opin Investig Drugs*. 2004;5(2):146–53.
35. McAllister KA, Peery RB, Zhao G. Acyl carrier protein synthases from gram-negative, gram-positive, and atypical bacterial species: biochemical and structural properties and physiological implications. *J Bacteriol*. 2006;188(13): 4737–48.
36. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*. 1998; 281(5375):375–88.
37. Deaconescu AM, Savery N, Darst SA. The bacterial transcription repair coupling factor. *Curr Opin Struct Biol*. 2007;17(1):96–102.
38. Giacani L, Molini BJ, Kim EY, Godornes BC, Leader BT, Tantalo LC, Centurion-Lara A, Lukehart SA. Antigenic variation in *Treponema pallidum*: TprK sequence diversity accumulates in response to immune pressure during experimental syphilis. *J Immunol*. 2010;184(7):3822–9.
39. Leader BT, Hevner K, Molini BJ, Barrett LK, Van Voorhis WC, Lukehart SA. Antibody responses elicited against the *Treponema pallidum* repeat proteins differ during infection with different isolates of *Treponema pallidum* subsp. pallidum. *Infect Immun*. 2003;71(10):6054–7.
40. Sun ES, Molini BJ, Barrett LK, Centurion-Lara A, Lukehart SA, Van Voorhis WC. Subfamily I *Treponema pallidum* repeat protein family: sequence variation and immunity. *Microbes Infect*. 2004;6(8):725–37.
41. Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Van Voorhis WC, Lukehart SA. *Treponema pallidum* Major sheath protein homologue TprK is a target of opsonic antibody and the protective immune response. *J Exp Med*. 1999;189(4):647–56.
42. Walker EM, Borenstein LA, Blanco DR, Miller JN, Lovett MA. Analysis of outer membrane ultrastructure of pathogenic *Treponema* and *Borrelia* species by freeze-fracture electron microscopy. *J Bacteriol*. 1991;173(17):5585–8.
43. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, Borrego MJ, Mendonca J, Carpinteiro D, Vieira L, et al. Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nat Microbiol*. 2016;2:16190.
44. Lukehart SAL, Shaffer JM, Zander SAB. A subpopulation of *Treponema pallidum* is resistant to phagocytosis: possible mechanism of persistence. *J Infect Dis*. 1992;166(6):1449–53.
45. Centurion-Lara A, Godornes C, Castro C, Van Voorhis WC, Lukehart SA. The tprK gene is heterogeneous among *Treponema pallidum* strains and has multiple alleles. *Infect Immun*. 2000;68(2):824–31.
46. LaFond RE, Centurion-Lara A, Godornes C, Rompalo AM, Van Voorhis WC, Lukehart SA. Sequence diversity of *Treponema pallidum* subsp. pallidum tprK in human syphilis lesions and rabbit-propagated isolates. *J Bacteriol*. 2003;185(21):6262–8.
47. Giacani L, Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, Yang T-C. Insights into the genetic variation profile of tprK in *Treponema pallidum* during the development of natural human syphilis infection. *PLoS Negl Trop Dis*. 2019; 13(7):e0007621.
48. Maderankova D, Mikalova L, Strouhal M, Vadjak S, Kuklova I, Pospisilova P, Krbkova L, Koscova P, Provaznik I, Smajcs D. Identification of positively selected genes in human pathogenic treponemes: syphilis-, yaws-, and bejel-causing strains differ in sets of genes showing adaptive evolution. *PLoS Negl Trop Dis*. 2019;13(6):e0007463.
49. Klopper TH, Huson DH. Drawing explicit phylogenetic networks and their integration into SplitsTree. *BMC Evol Biol*. 2008;8:22.
50. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006;23(2):254–67.
51. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, et al. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. *PLoS One*. 2013;8(1): e53818.
52. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16:157.
53. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30(7):1575–84.
54. Seshadri R, Myers GS, Tettelin H, Eisen JA, Heidelberg JF, Dodson RJ, Davidsen TM, DeBoy RT, Fouts DE, Haft DH, et al. Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc Natl Acad Sci U S A*. 2004;101(15):5646–51.
55. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis comparison tool. *Bioinformatics*. 2005;21(16):3422–3.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Additional file 1

Table S1: General information about 53 *Treponema pallidum* Strains used in this work. List of all *Treponema pallidum* strains (with features) retrieved from the NCBI (National Center for Biotechnology Information) database.

Strain	Size(Mb)	Subspecies	Geographical Location	Harvested	GenBank Accession No	GC%	Gene	Protein
Tp_Nichols	1.13	pallidum	Washington DC	Human- Neurosyphilitic patient	AE000520.1	52.80	1044	970
Tp_Sea81-4	1.13	pallidum	USA: Seattle. WA	Human	CP003679.1	52.80	1032	931
Tp_SS14	1.13	pallidum	San DiegoCA: USA	Human- Skin	CP004011.1	52.80	1066	1002
Tp_Chicago	1.13	pallidum	San DiegoCA: USA	Rabbit- testis	CP001752.1	52.80	1030	969
Tp_SamoaD	1.13	pertenue	Australia/Oceania	Rabbit-testis	CP002374.1	52.80	1064	1005
Tp_CDC2	1.13	pertenue	Africa	Human	CP002375.1	52.80	1030	973
Tp_Gauthier	1.13	pertenue	Africa	Human- Skin	CP002376.1	52.80	1029	971
Tp_DAL1	1.13	pallidum	Africa	Human-	CP003115.1	52.80	1030	969
Tp_MexicoA	1.14	pallidum	Mexico	Human- Skin	CP003064.1	52.80	1029	968
Tp_Fribourg-Blanc	1.14	pertenue	West Africa	Baboons- Skin	CP003902.1	52.80	1030	970
Tp_Nichols(2013/06/11)	1.13	pallidum	Washington DC	Human- Skin	CP004010.2	52.80	1065	1004
Tp_SS14(11.12.2013)	1.13	pallidum	USA: Seattle. WA	Human- Skin	CP000805.1	52.80	1088	1028
Tp_BosniaA	1.13	endemicum	Europe	Human-Tongue & tonsils.	CP007548.1	52.80	1065	1003
Tp_pallidum Amoy	1.13	pallidum	China: Xiamen	Not Available	CP015162.1	52.70	1033	964
Tp_Chicago population	1.13	pallidum	USA: Seattle	Human-Bacteria harveste from rabbit testes	CP010558.1	52.80	1034	971
Tp_CDC-A	1.13	pallidum	USA: Seattle	Human-Bacteria harveste from rabbit testes	CP010559.1	52.80	1033	969
Tp_Nichols Houston, Clone E	1.13	pallidum	USA: Seattle	Human-Bacteria harveste from rabbit testes	CP010560.1	52.80	1031	966
Tp_Nichol Houston, Clone J	1.13	pallidum	USA: Seattle	Human-Bacteria harveste from rabbit testes	CP010561.1	52.80	1031	969
Tp_UW074B	1.13	pallidum	USA: Seattle	Human-Blood	CP010562.1	52.80	1033	969
Tp_UW189B	1.13	pallidum	USA: Seattle	Human-Blood	CP010563.1	52.80	1033	970
Tp_UW228B	1.13	pallidum	USA: Seattle	Human-Blood	CP010564.1	52.80	1034	973
Tp_UW254B	1.13	pallidum	USA: Seattle	Human-Blood	CP010565.1	52.80	1033	970

Tp_UW391B	1.13	pallidum	USA: Seattle	Human-Blood	CP010566.1	52.80	1025	962
Tp_PT_SIF0697	1.13	pallidum	Portugal	Human- Penile	CP016045.1	52.80	1035	970
Tp_PT_SIF0857	1.13	pallidum	Portugal	Human-Oropharyngeal	CP016047.1	52.80	1035	973
Tp_PT_SIF0908	1.13	pallidum	Portugal	Human- Vaginal	CP016049.1	52.80	1034	933
Tp_PT_SIF0954	1.13	pallidum	Portugal	Human- Scrotum	CP016050.1	52.80	1034	970
Tp_PT_SIF1002	1.13	pallidum	Portugal	Human- Penile	CP016051.1	52.80	1033	970
Tp_PT_SIF1020	1.13	pallidum	Portugal	Human- Penile	CP016052.1	52.80	1033	969
Tp_PT_SIF1063	1.13	pallidum	Portugal	Human- Anal	CP016053.1	52.80	1035	972
Tp_PT_SIF1127	1.13	pallidum	Portugal	Human- Penile	CP016054.1	52.80	1035	917
Tp_PT_SIF1135	1.13	pallidum	Portugal	Human- Anal	CP016055.1	52.80	1036	972
Tp_PT_SIF1140	1.13	pallidum	Portugal	Human- Anal	CP016056.1	52.80	1032	863
Tp_PT_SIF1142	1.13	pallidum	Portugal	Human- Penile	CP016057.1	52.80	1035	971
Tp_PT_SIF1156	1.13	pallidum	Portugal	Human- Penile	CP016058.1	52.80	1035	971
Tp_PT_SIF1167	1.13	pallidum	Portugal	Human- Penile	CP016059.1	52.80	1035	971
Tp_PT_SIF1183	1.13	pallidum	Portugal	Human- Penile	CP016060.1	52.80	1034	971
Tp_PT_SIF1196	1.13	pallidum	Portugal	Human- Penile	CP016061.1	52.80	1037	974
Tp_PT_SIF1200	1.13	pallidum	Portugal	Human- Penile	CP016062.1	52.80	1035	973
Tp_PT_SIF1242	1.13	pallidum	Portugal	Human- Penile	CP016063.1	52.80	1034	971
Tp_PT_SIF1252	1.13	pallidum	Portugal	Human- Penile	CP016064.1	52.80	1035	972
Tp_PT_SIF 1261	1.13	pallidum	Portugal	Human- Anal	CP016065.1	52.80	1034	971
Tp_PT_SIF1278	1.13	pallidum	Portugal	Human- Anal	CP016066.1	52.80	1034	972
Tp_PT_SIF1280	1.13	pallidum	Portugal	Human- Anal	CP016067.1	52.80	1034	972
Tp_PT_SIF1299	1.13	pallidum	Portugal	Human- Penile	CP016068.1	52.80	1035	972
Tp_PT_SIF1348	1.13	pallidum	Portugal	Human- Penile	CP016069.1	52.80	1032	967
Tp_PT_SIF0751	1.13	pallidum	Portugal	Human- Tongue	CP016046.1	52.80	1033	969

Tp_PT_SIF0877_3	1.13	pallidum	Portugal	Human- Tongue	CP016048.1	52.80	1034	971
Tp_seattle Nichols	1.13	Pallidum	USA Seattle WA	Human-Bacteria harveste from rabbit testes	CP010422.1	52.80	1065	1000
Tp_Ghana-051	1.13	pertenue	Ghana	Human	CP020365.1	52.80	1067	1006
Tp_CDC 2575	1.13	pertenue	Ghana	Human	CP020366.1	52.80	1067	1006
Tp_UZ1974	1.13	pallidum	Czeck Republic (Europe)	Human	CP028438.1	52.80	1066	1000
Tp_LMNP-1	1.13	pertenue	Tanzania	Papio anubis	CP021113.1	52.80	1064	1000

Table S2A: The COG functional categories with detailed description of Core genes: The table showing the number of core genes of the complete dataset were classified by COG (Cluster of Orthologous Genes) functional category.

Code	Description	No of Genes	Percentage
Information Storage and processing			
[A]	RNA processing and modification	0	0.0
[B]	Chromatin structure and dynamics	0	0.0
[J]	Translation, ribosomal structure and biogenesis	93	14.29
[K]	Transcription	19	2.91
[L]	Replication, recombination and repair	30	4.70
Cellular processes and signalling			
[D]	Cell cycle control, cell division, chromosome partitioning	11	1.69
[M]	Cell wall/membrane biogenesis	39	6.00
[N]	Cell motility	9	1.40
[O]	Posttranslational modification, protein turnover, chaperones	29	4.50
[T]	Signal transduction mechanisms	15	3.30
[U]	Intracellular trafficking and secretion	6	0.90
[V]	Defense mechanisms	5	0.80
[W]	Extracellular structures	0	0.0
[Y]	Nuclear structure	0	0.0
[Z]	Cytoskeleton	0	0.0
Metabolism			
[C]	Energy production and conversion	17	2.61
[E]	Amino acid transport and metabolism	11	1.70
[F]	Nucleotide transport and metabolism	10	1.53
[G]	Carbohydrate transport and metabolism	18	2.80
[H]	Coenzyme transport and metabolism	15	2.30
[I]	Lipid transport and metabolism	13	1.90
[P]	Inorganic ion transport and metabolism	21	3.22
[Q]	Secondary metabolites biosynthesis, transport and catabolism	0	0.0
Poorly Characterized			
[R]	General function prediction only	52	8.00
[S]	Function unknown	35	5.38
-	Not in COGs	203	31.18

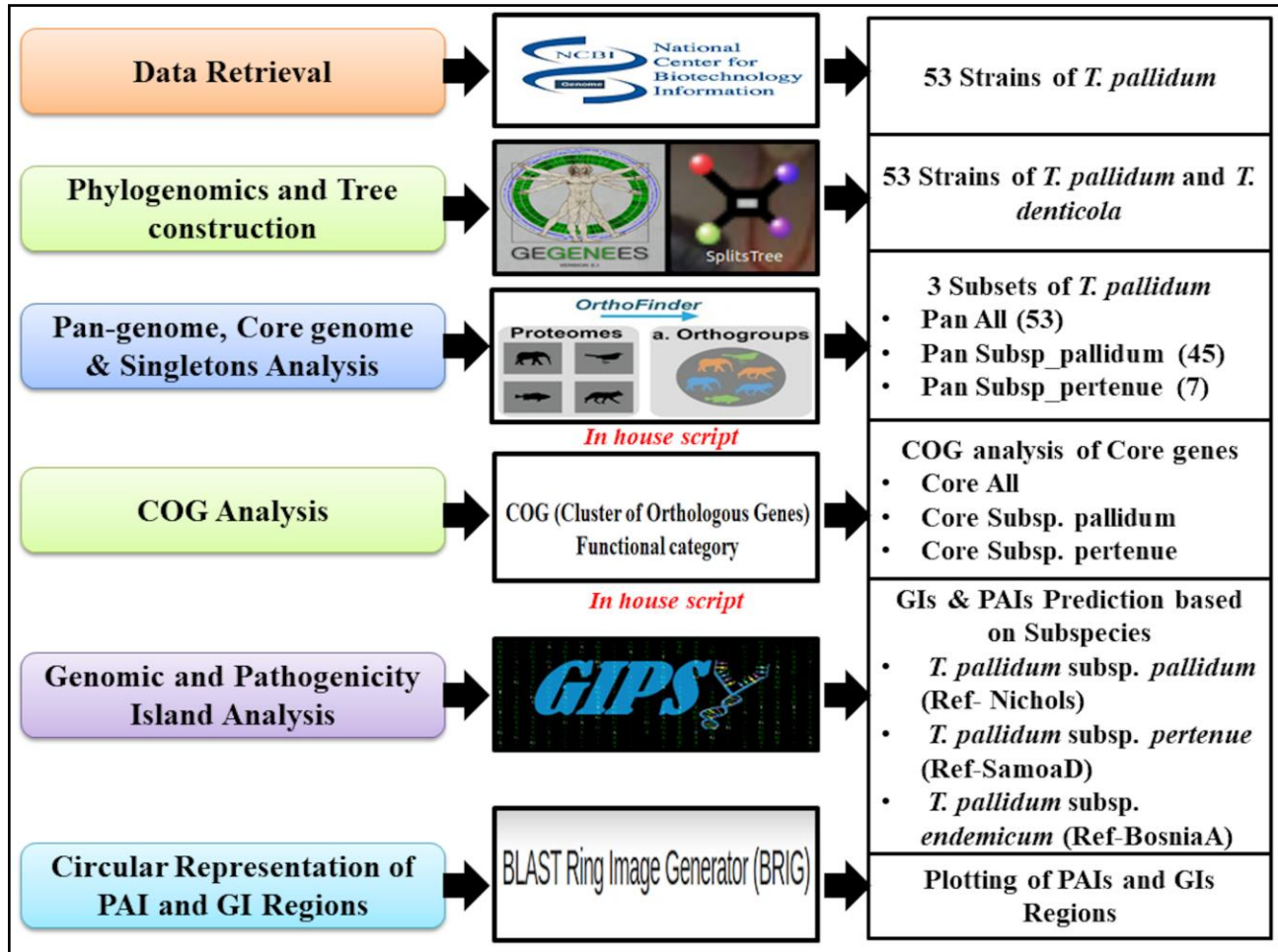
Table S2B: The COG functional categories with detailed description of Core genes: The table showing the number of core genes of the Pan Subsp_pallidum dataset were classified by COG (Cluster of Orthologous Genes) functional category.

Code	Description	No of Genes	Percentage
Information Storage and processing			
[A]	RNA processing and modification	0	0.0
[B]	Chromatin structure and dynamics	0	0.0
[J]	Translation, ribosomal structure and biogenesis	95	14.54
[K]	Transcription	24	3.67
[L]	Replication, recombination and repair	33	5.05
Cellular processes and signalling			
[D]	Cell cycle control, cell division, chromosome partitioning	11	1.68
[M]	Cell wall/membrane biogenesis	43	6.58
[N]	Cell motility	27	4.13
[O]	Posttranslational modification, protein turnover, chaperones	36	5.51
[T]	Signal transduction mechanisms	28	4.28
[U]	Intracellular trafficking and secretion	18	2.75
[V]	Defense mechanisms	6	0.91
[W]	Extracellular structures	0	0.0
[Y]	Nuclear structure	0	0.0
[Z]	Cytoskeleton	0	0.0
Metabolism			
[C]	Energy production and conversion	25	3.82
[E]	Amino acid transport and metabolism	15	2.29
[F]	Nucleotide transport and metabolism	13	1.99
[G]	Carbohydrate transport and metabolism	25	3.82
[H]	Coenzyme transport and metabolism	18	2.75
[I]	Lipid transport and metabolism	15	2.29
[P]	Inorganic ion transport and metabolism	24	3.67
[Q]	Secondary metabolites biosynthesis, transport and catabolism	1	0.15
Poorly Characterized			
[R]	General function prediction only	61	9.34
[S]	Function unknown	36	5.51
-	Not in COGs	99	15.59

Table S2C: The COG functional categories with detailed description of Core genes: The table showing the number of core genes of the Pan Subsp_pertenuis dataset were classified by COG (Cluster of Orthologous Genes) functional category.

Code	Description	No of Genes	Percentage
Information Storage and processing			
[A]	RNA processing and modification	0	0.0
[B]	Chromatin structure and dynamics	0	0.0
[J]	Translation, ribosomal structure and biogenesis	117	11.25
[K]	Transcription	22	2.11
[L]	Replication, recombination and repair	51	4.90
Cellular processes and signalling			
[D]	Cell cycle control, cell division, chromosome partitioning	13	1.25
[M]	Cell wall/membrane biogenesis	61	5.86
[N]	Cell motility	42	4.03
[O]	Posttranslational modification, protein turnover, chaperones	47	4.51
[T]	Signal transduction mechanisms	32	3.07
[U]	Intracellular trafficking and secretion	29	2.78
[V]	Defense mechanisms	7	0.67
[W]	Extracellular structures	0	0.0
[Y]	Nuclear structure	0	0.0
[Z]	Cytoskeleton	0	0.0
Metabolism			
[C]	Energy production and conversion	38	3.65
[E]	Amino acid transport and metabolism	27	2.59
[F]	Nucleotide transport and metabolism	21	2.01
[G]	Carbohydrate transport and metabolism	43	4.13
[H]	Coenzyme transport and metabolism	21	2.01
[I]	Lipid transport and metabolism	19	1.82
[P]	Inorganic ion transport and metabolism	26	2.5
[Q]	Secondary metabolites biosynthesis, transport and catabolism	3	0.28
Poorly Characterized			
[R]	General function prediction only	87	8.36
[S]	Function unknown	136	13.07
-	Not in COGs	198	19.03

Figure S1. The Complete workflow applied in this work. The figure represent the methodology and software were used in this analysis.



III.3.2. Conclusion, Chapter 3.

The genomic information was used in this work with the aim to determine differences between the *Treponema pallidum* different strains based on their subspecies. We calculated pan-genome, core genome and singletons sets of each subspecies. Core genome showed the conserved genes and characterization of such poorly studied proteins helps in understanding the cellular metabolism, mode of infection and regulation of gene expression of *Treponema pallidum*. This study may help to better understand the molecular modes of bacterial infection and its persistence in the host, which are significant for vaccine development for syphilis.

Chapter 4.

III.4.1. Research Article

An *In Silico* Identification of Common Putative vaccine Candidates against *Treponema pallidum*: A Reverse vaccinology and Subtractive Genomics based Approach

Arun Kumar Jaiswal, Sandeep Tiwari, Syed Babar Jamal, Debmalya Barh, Vasco Azevedo and **Siomar C. Soares**.

International Journal of Molecular Science, 2017, 18(2), 402;

In this study, we focus on the in silico Comparative genomics, Reverse vaccinology, Subtractive genomics and Molecular docking analysis for the drug and vaccine target identification. We compared 13 strains of *Treponema pallidum* for comparative analysis. We identified 837 Core-genes and considering human as a host, 567 conserved non-host homologous proteins were identified. Further, using subtractive genomics and reverse vaccinology approach we identified 15 putative antigenic proteins and 6 drug targets which were essential for the bacteria. Identified drug target were subjected to virtual screening using 28 Natural compounds library. This can be used as candidate therapeutic targets in future against syphilis.



Article

An In Silico Identification of Common Putative Vaccine Candidates against *Treponema pallidum*: A Reverse Vaccinology and Subtractive Genomics Based Approach

Arun Kumar Jaiswal^{1,2}, Sandeep Tiwari¹, Syed Babar Jamal¹, Debmalya Barh^{1,3}, Vasco Azevedo¹ and Siomar C. Soares^{2,*}

¹ Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte 31270-901, MG, Brazil; arunjaiswal1411@gmail.com (A.K.J.); sandip_sbtbi@yahoo.com (S.T.); syedbabar.jamal@gmail.com (S.B.J.); dr.barh@gmail.com (D.B.); vascoariston@gmail.com (V.A.)

² Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba 38025-180, MG, Brazil

³ Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal 721137, India

* Correspondence: siomars@gmail.com

Academic Editor: Christopher Woelk

Received: 14 November 2016; Accepted: 27 January 2017; Published: 14 February 2017

Abstract: Sexually transmitted infections (STIs) are caused by a wide variety of bacteria, viruses, and parasites that are transmitted from one person to another primarily by vaginal, anal, or oral sexual contact. Syphilis is a serious disease caused by a sexually transmitted infection. Syphilis is caused by the bacterium *Treponema pallidum* subspecies *pallidum*. *Treponema pallidum* (*T. pallidum*) is a motile, gram-negative spirochete, which can be transmitted both sexually and from mother to child, and can invade virtually any organ or structure in the human body. The current worldwide prevalence of syphilis emphasizes the need for continued preventive measures and strategies. Unfortunately, effective measures are limited. In this study, we focus on the identification of vaccine targets and putative drugs against syphilis disease using reverse vaccinology and subtractive genomics. We compared 13 strains of *T. pallidum* using *T. pallidum* Nichols as the reference genome. Using an in silico approach, four pathogenic islands were detected in the genome of *T. pallidum* Nichols. We identified 15 putative antigenic proteins and six drug targets through reverse vaccinology and subtractive genomics, respectively, which can be used as candidate therapeutic targets in the future.

Keywords: sexually transmitted infections (STIs); drug target; vaccine target

1. Introduction

Sexually transmitted infections (STIs) are triggered by a number of bacteria, viruses, and parasites that are transferred mainly by vaginal, anal, or oral sexual contact between people. Different STIs can be existent or transmitted instantaneously, and such infections can trigger other STIs [1]. The World Health Organization (WHO) has reported more than 30 different bacteria, viruses, and parasites that are responsible for disease transmission through sexual contact.

Syphilis is among the most severe sexually transmitted infections (STIs) caused by the *Treponema pallidum* subspecies *pallidum*, a motile, gram-negative spirochete bacterium [2]. The annual estimated frequency of infectious syphilis is 36 million cases and over 11 million new infections; thus, it is an important public health burden globally [3]. Furthermore, the number of cases increased 10-fold

in the last 15 years, with 4317 newly reported infections in 2014. This number is the highest it has been in the last 40 years and was mainly observed among men who have sex with men (MSM) [2].

If not properly treated, syphilis can cause long-term problems. It is important to screen women for syphilis during pregnancy to provide rapid treatment and to avoid congenital infections. Syphilis is a globally reemerging infection, as recently observed in the United States and Italy. Asian, African, and Latin American countries have high syphilis occurrences and are motivated to control prenatal care [4,5]. According to the Ministry of Health, in Brazil, 50,000 pregnant women are diagnosed with syphilis annually. The prevalence ranges from 1.1% to 11.5%, depending on maternal schooling and prenatal care. As a result, almost 12,000 infants are born with congenital syphilis each year [4]. In Brazil, the regulation of syphilis is one of the goals of the Pact for Health project initiated by the World Health Organization (WHO) for the elimination of congenital syphilis [4].

Despite sevent decades of penicillin use for the treatment of syphilis infections, *T. pallidum* exhibits complete sensitivity to this antibiotic. An increase in treatment complexity has led to the use of azithromycin as an oral antibiotic. However, over the last few decades, resistance against macrolides has been reported in many countries and at present, macrolides are not recommended for the cure or prophylaxis of syphilis [6]. The recent global prevalence of syphilis elicits a need for sustained preventive measures and strategies. Unfortunately, effective measures are inadequate. Relevant application of chemicals, antibiotics, lotions, creams, and thorough washing with soap and water after sexual contact are ineffective. The development of an effective vaccination appears to be the only alternative for the control of syphilis in the future. In spite of intense research for developing proper syphilis treatments, restricted progress has been noticed [7]. There are recent cases of emergence reported in several countries including Norway [8], China [9], the United States, Western Europe [10], and Martinique [11]. Although in today's drug discovery process, high-throughput techniques and synthetic chemistry accelerate the process dramatically, it still takes 10–15 years to introduce a new drug to the market and therefore, a large investment is required [12].

The first step in the drug and vaccine discovery process is target identification. With the advent of new sequencing technologies and the deluge of genomic data, scientists are able to use computational methods to rapidly identify new targets, which are more time and cost effective than old approaches. Computational methods (i.e., subtractive genomics) are broadly used in this process. Recently, working with bacterial pathogens using an *in silico* approach, a large number of targets have been identified that are either resistant to drugs or for which no appropriate vaccine is available [13]. Reverse vaccinology is a conventional and popular approach in the post-genomic era for the prompt identification of novel vaccine targets [14,15]. Approaches, such as comparative and subtractive genomics and differential genome analyses [16], are being widely utilized for target identification in several human pathogens, including *Mycobacterium tuberculosis* [17], *Helicobacter pylori* [18], *Burkholderia pseudomallei* [19], *Pseudomonas aeruginosa* [20], *Salmonella typhi* [21], and *Neisseria gonorrhoeae* [22]. Generally, the principle behind these approaches is the identification of gene/protein targets that are essential for the survival of the pathogen but are not homologous to genes/proteins of the host [23]. Nevertheless, the identified targets may have a certain degree of homology with the host protein and are essential for the survival of the pathogen; hence, they can also be selected for structure-based selective inhibitor development as an additional molecular target. The differences in the active sites or other pockets with suitable druggability of the pathogenic protein could play an important role when compared to the host protein [24,25]. In this study, we mainly focus on the *in silico* identification of putative vaccine and drug targets against syphilis disease using reverse vaccinology and subtractive genomics. The goal was to identify plant-derived new lead antimicrobial compounds, and the proposed drug molecules show favorable interactions, lowered energy values, and high complementarity with the predicted targets.

2. Result and Discussion

The total number of proteins described in each of the following sections and all the methodologies used in our work are described on the workflow in Figure 1.

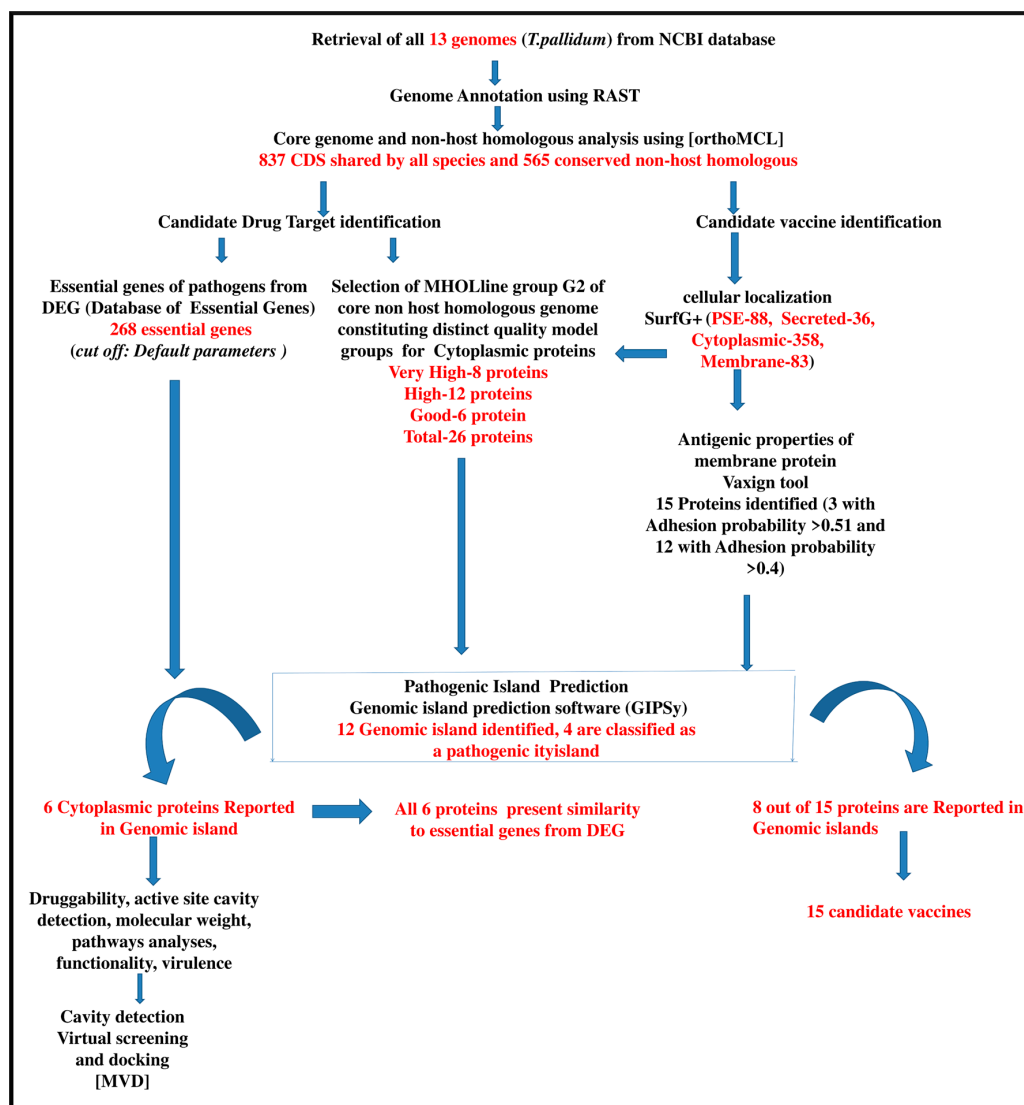


Figure 1. Complete workflow with the number of genes selected in each step and methodologies used. The sentences in black describe the analyses made and the software used in each step. The sentences in red represent the number of proteins selected in each step. CDS = coding DNA sequence; MVD = Molegro Virtual Docker.

2.1. Identification of Intra-Species Conserved Non-Host Homologous Proteins and Pathogenicity Islands

We compared 13 *Treponema pallidum* strains (Table 1) using *Treponema pallidum* Nichols as the reference using the orthoMCL software [26]. Coding DNA sequences (CDSs) shared by all species were considered a part of the core genome. Considering the human genome as the host genome, a set of 565 conserved non-host homologous proteins were identified. The prediction of genomic islands (GIs) was subsequently performed. GIs are gene clusters, usually >8 kb in size, likely acquired via horizontal gene transfers (HGT), and often playing a role in the environmental or host adaptation of bacteria. GIs significantly influence bacterial evolution and provide further insight in differentiating bacterial species and strains. For *T. pallidum* Nichols strains, 10 putative GIs were identified through

the Genomic Island Prediction Software (GIPSy) [27], using *Treponema denticola* as a closely related, non-pathogenic organism. Of the 10 GIs, four are classified as pathogenicity islands (PAIs), i.e., they present high concentrations of virulence factors and are absent in the aforementioned closely related non-pathogenic organism (Figure 2).

Table 1. Genomic features of all *T. pallidum* (Tp) strains.

Strain	Size (Mb)	GC%	Gene	Protein
Tp_Nichols	1.13	52.80	1044	970
Tp_Sea81-4	1.13	52.80	1032	931
Tp_SS14	1.13	52.80	1042	971
Tp_Chicago	1.13	52.80	1030	969
Tp_SamoaD	1.13	52.80	1027	971
Tp_CDC2	1.13	52.80	1030	973
Tp_Gauthier	1.13	52.80	1029	971
Tp_DAL1	1.13	52.80	1030	969
Tp_MexicoA	1.14	52.80	1029	968
Tp_Fribourg-Blanc	1.14	52.80	1030	970
Tp_SS14 (14.8.2015)	1.13	52.80	1029	970
Tp_BosniaA	1.13	52.80	1027	970
Tp_pallidum	1.13	52.70	1033	964

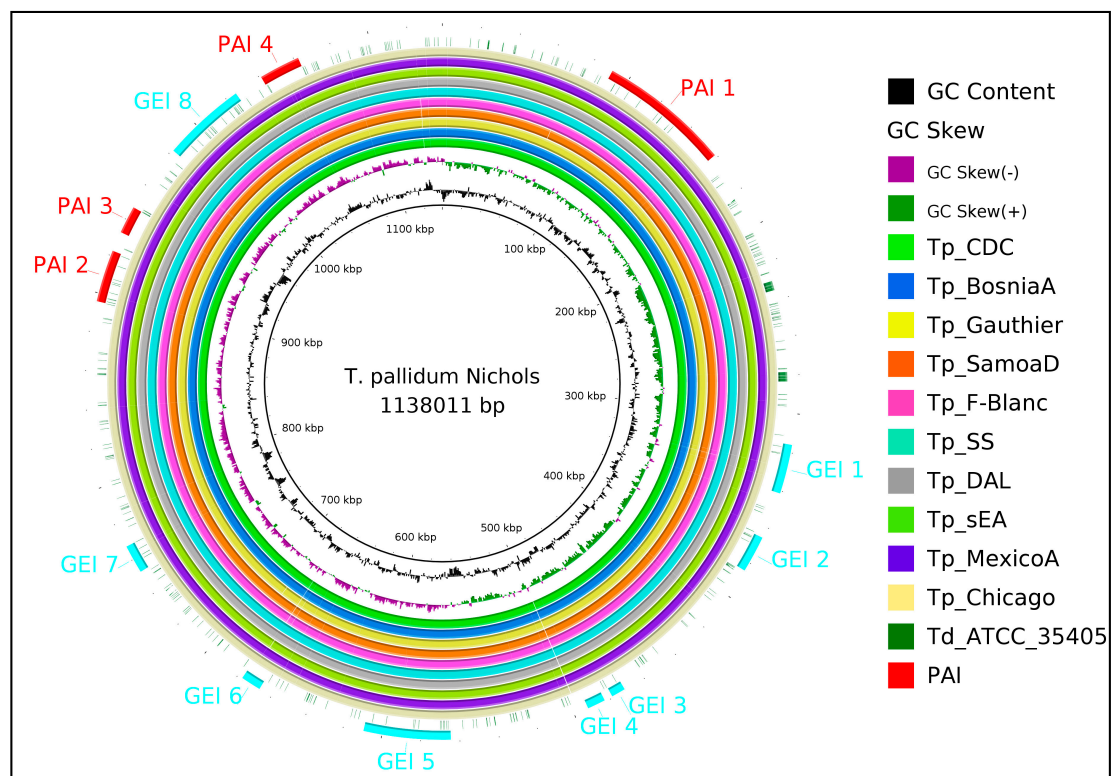


Figure 2. Genomic islands (GIs) of *T. pallidum* Nichols strains as predicted by the genomic island prediction software (GIPSy) using *Treponema denticola* as a closely related non-pathogenic organism. The outermost circle highlighted in red shows the four pathogenicity islands from 10 GIs. Guanine-Cytosine (GC) content is shown in black.

2.2. Assessment of Essential Genes

Essentiality analysis identifies significant genes required for pathogen survival such as adhesion, entry into the host, infection, and persistence in the host [13]. The conserved 565 non-hosts homologous

proteins were subjected to the Database of Essential Genes (DEG) for the identification of essential proteins, through which a final set of 268 proteins was obtained (Table S1). Essential proteins are necessary for the survival of pathogen within the host. When these essential proteins are declared to be virulent, they can be of vital significance to unveil novel therapeutic targets. There is a probability of essential proteins to be conserved among various populations and species because of their vital roles in various pathways for pathogen survival [13,28]. Virulence is the characteristic of a pathogen responsible for causing severe human diseases. In the present study, these properties have been given high priority to identify potential vaccine candidates computationally. Although only 268 proteins were identified as essential by DEG, we considered all 565 proteins for our analyses.

2.3. Prediction of Candidate Vaccine Target for *T. pallidum*

The subcellular localization of conserved non-hosts homologous proteins of *T. pallidum* strains were predicted with the SurfG+ software [29]. We classified 207 gene products as putative surface-exposed (PSE) proteins, secreted proteins, or membrane proteins (Table 2). The proteins predicted by SurfG+ were further analyzed with the software Vaxign [30] for antigenic properties with adhesion probabilities greater than 0.51, resulting in the detection of three proteins in the *T. pallidum* strains Nichols (Table 3). We found that out of these three proteins, Tp_Nichols141 and Tp_Nichols797 were hypothetical proteins. Tp_Nichols141 belongs to the pathogenicity island 1 (Figure 2). When the adhesion probability threshold was >0.4, we also identified 12 more proteins that can also be considered potential vaccine candidates against *T. pallidum*.

Table 2. Subcellular location of *Treponema pallidum* (Tp) strain proteins.

Localization	Number of Proteins
Cytoplasmic Protein	358
Membrane Protein	83
PSE ^a	88
Secreted Protein	36

^a Putative Surface Exposed.

Previous studies have shown the importance of targeting proteins involved in the capability of *T. pallidum* to invade host tissues and to evade the functional immune response, contributing to its persistence during the “latency” stage. Most of the described gene targets code for proteins responsible for the attachment to extracellular matrix bridges (Tp0136, TP0155, Tp0483, and Tp0751), such as the low density integral Outer Membrane Proteins (OMPs) [6]. Briefly, in our predictions of good vaccine targets, we have identified Tp_Nichols350 and TpNichols852 with similarities to two previously described OMPs (TP0453 and Tp_0326), along with two additional OMP domain containing proteins: Tp_Nichols797 and Tp_Nichols141. Interestingly, both Tp_Nichols797 and Tp_Nichols141 presented adhesion probabilities higher than 0.5 and should be given priority in in vitro assays.

Table 3. Putative antigenic proteins of *Treponema pallidum* (Tp) identified using Vaxign.

Tp_Nichols	Protein ID	Gene Name	Subcellular Localization	SignalP Result (Cleavage Site)	TMHMM Result	InterProScan (Domain)	Gene Product	Adhesion Probability
Tp_Nichols797	WP_010882178.1	-	SEC	Yes (between 25 and 26)	TMH = 0	Outer membrane protein/outer membrane enzyme PagP, beta-barrel—IPR011250 (65–219)	Hypothetical protein	0.552
Tp_Nichols141	WP_014342713.1	-	PSE	No	TMH = 1	Outer membrane protein/outer membrane enzyme PagP, beta-barrel—IPR011250 (100–225)	Hypothetical protein	0.525
Tp_Nichols466	WP_010881878.1	<i>ntpK</i>	MEM	No	TMH = 4	V-ATPase proteolipid subunit C-like domain—IPR002379 (76–138)	Two-sector ATPase, V(0) subunit K	0.590
Tp_Nichols930	WP_010882306.1	<i>slyD</i>	PSE	No	TMH = 1	Peptidyl-prolyl cis-trans isomerase, FKBP-type, N-terminal—IPR000774 (66–143)	FKBP-type peptidyl-prolyl cis-trans isomerase SlyD	0.488
Tp_Nichols471	WP_010881883.1	<i>nlpE</i>	SEC	Yes (between 23 and 24)	TMH = 0	No	Copper resistance lipoprotein NlpE	0.475
Tp_Nichols650	WP_010882040.1	-	PSE	No	TMH = 2	Domain of unknown function DUF2147—IPR019223 (71–193)	Hypothetical Protein	0.474
Tp_Nichols1046	WP_010882416.1	<i>ptr1</i>	MEM	No	TMH = 6	No	Conserved hypothetical integral membrane protein	0.44
Tp_Nichols52	WP_010881498.1	<i>TPANIC_0600</i>	PSE	No	TMH = 1	Duplicated hybrid motif—Ipr011055 (196–355)	Zinc metalloprotease	0.428
Tp_Nichols610	WP_010882004.1	-	SEC	No	TMH = 1	Zinc finger, CHCC-type—IPR019401 (8–34)	Hypothetical Protein	0.425
Tp_Nichols323	WP_010881746.1	-	SEC	No	TMH = 1	Sporulation-related domain—IPR007730 (172–252)	Hypothetical Protein	0.41
Tp_Nichols852	WP_010882234.1	<i>TP_0453</i>	SEC	Yes (between 23 and 24)	TMH = 0	No	Outer membrane protein TP0453	0.408
Tp_Nichols350	WP_014342788.1	<i>tp92</i>	SEC	Yes (between 37 and 38)	TMH = 1	Bacterial surface antigen (D15)—IPR000184 (478–849)	Putative outer membrane protein assembly factor TP_0326	0.405
Tp_Nichols98	WP_010881537.1	-	PSE	No	TMH = 0	No	Hypothetical Protein	0.401
Tp_Nichols347	WP_010881771.1	<i>TP_0323</i>	MEM	No		No	Ribose/galactose ABC transporter, permease protein (RbsC-2)	0.401
Tp_Nichols362	WP_010881783.1	<i>TPANIC_0335</i>	MEM	No	TMH = 2	No	Putative membrane protein	0.401

SEC = secreted; PSE = Putative surface exposed; MEM = Membrane; TMH = Transmembrane Helix, TMHMM = Transmembrane Helix prediction server, based on a hidden Markov model.

2.4. High Throughput Structural Modeling

The main focus of this study was to find candidate vaccine targets. However, according to Caroline et al., 2014 [6], the difficulty in curing syphilis is due to the vilification of many antibiotics for treatment or prophylaxis. Our contributions include the prediction of some novel drug targets against *Treponema pallidum*. For this, the identified 565 conserved non-host homologous *Treponema pallidum* proteins were submitted to MHOLline [31] an online web tool, to predict the modelome. MHOLline utilizes multi-fasta files of amino acids as an input data and then uses HMMTOP, BLAST, BATS, MODELLER, and PROCHECK programs for the detailed analyses. The program HMMTOP detects transmembrane regions. The BLAST algorithm is used to identify the template structure by performing a random search against the Protein Data Bank. BATS (Blast Automatic Targeting for Structures) carries out the refinement in the template search and it is a key step for the model construction. BATS refinement identifies sequences that make the modeling possible by selecting a template from a BLAST output file using their BATS scores, expectation values, identity, and sequence similarity as criteria, as well as considering the number of gaps and the alignment coverage. BATS selects the best template for 3D model generation and performs automated alignment using the MODELLER program. Furthermore, it gathers all the BLAST output files into four distinctive groups (i.e., G0, G1, G2, and G3) according to the following criteria: G0 = unaligned sequence; G1 = E-value $> 10 \times 10^{-5}$ or identity $< 15\%$; G2 = E-value $\leq 10 \times 10^{-5}$ and identity $\geq 25\%$ AND LVI ≤ 0.7 ; G3 = E-value $\leq 10 \times 10^{-5}$ and identity $\leq 15\%$ and $< 25\%$ OR LVI (Length Variation Index) > 0.7 . Only the first three distinct quality G2 model groups were taken into consideration in this study; these were: 1—very high quality model sequences (identity $\geq 75\%$) (LVI ≤ 0.1), 2—high quality model sequences (identity $\geq 50\%$) and $< 75\%$) (LVI ≤ 0.1), and 3—good quality model sequences (identity $\geq 50\%$) (LVI > 0.1 and ≤ 0.3) [31]. Therefore, all the considered protein 3D models were constructed from sequences for which their template is available with identity $\geq 50\%$. We found 26 proteins (8 very high, 12 high, and 6 good) in the first 3 distinct quality G2 model groups.

The membrane and cell wall associated proteins are, theoretically, more exposed as targets than the cytoplasmic drug targets. However, membrane proteins are difficult to purify and assay [32]. Cytoplasmic membrane proteins are also very important for the physiology of bacteria, as they are involved in many important metabolic functions. Therefore, the membrane, putative surface exposed, and secreted proteins are better applicable as targets for reverse vaccinology, whereas the pivotal role of cytoplasmic proteins in maintenance of cell viability makes them more favorable as drug targets [33]. Out of the 26 proteins, only cytoplasmic proteins that were present in any GIs were selected as candidate drug targets. Six proteins that were also present in the 268 proteins were identified as essential in the DEG analyses and were considered for the target prioritization and docking studies (Table 4).

The outer membrane may pose a barrier for drugs to gain access to cytoplasmic targets. However, small molecules are able to gain access to the periplasm through porins and reach the cytoplasm. In previous studies, it was shown that one of the pore forming OMPs, OmpF, has an exclusion limit of 600 Daltons, for example, which is used by ions, amino acids, and small sugars as a means to reach the periplasm [34]. The molecular weight of the compounds used here varies from ~ 275.1 g/mol (liriodenine) to ~ 488.7 g/mol (jacarandic acid) and they may also be able to use porins to gain access to the periplasm. Alternatively, the use of nanoparticles as delivery systems or a combined treatment, such as with polymyxins and derivatives that increase the permeability of the outer membrane, may also help in overcoming the outer membrane barrier [35].

Table 4. Drug target prioritization parameters and functional annotation of the six non-homologous putative targets.

Locus Tag, Gene, and Protein ID	Official Full Name	Mol. Wt (KDa) ^a	Functions ^b	Cellular Component ^c	Pathways ^d	Virulence ^e	DEG Analyses
Tp_Nichols130, uvrB, WP_010881565.1	UvrABC system protein B	76.19	MF: ATP (Adenosine triphosphate) binding, DNA binding, excinuclease ABC activity, helicase activity. BP: nucleotide-excision repair, SOS response.	Cytoplasm	Unknown	Yes	Essential gene
Tp_Nichols593, Pfp, WP_010881989.1	Pyrophosphate-fructose 6-phosphate 1-phosphotransferase	62.43	–	Cytoplasm	Glycolysis	Yes	Essential gene
Tp_Nichols609, asnA, WP_010882003.1	Aspartate-ammonia ligase	36.86	MF: Aminoacyl-tRNA ligase activity, aspartate-ammonia ligase activity, ATP binding. BP: L-asparagine biosynthetic process, tRNA aminoacylation for protein translation.	Cytoplasm	L-asparaginebiosynthesis	Yes	Essential gene
Tp_Nichols754, recA, WP_010882137.1	Protein RecA	45.33	MF: ATP binding, damaged DNA binding, DNA-dependent ATPase activity, single stranded DNA binding. BP: DNA recombination, DNA repair, SOS response.	Cytoplasm	Unknown	Yes	Essential gene
Tp_Nichols990, Ndh, WP_010882364.1	NADH (Nicotinamide adenine dinucleotide) dehydrogenase	48.64	MF: flavin adenine dinucleotide binding, NADH dehydrogenase activity. BP: cell redox homeostasis.	Cytoplasmic	Unknown	Yes	Essential gene
Tp_Nichols1011, Dxs, WP_010882382.1	1-deoxy-D-xylulose-5-phosphate synthase	129.82	MF: 1-deoxy-D-xylulose-5-phosphate synthase activity, magnesium ion binding, thiamine pyrophosphate binding. BP: 1-deoxy-D-xylulose-5-phosphate biosynthetic process, terpenoid biosynthesis process, Thiamine biosynthesis process.	Cytoplasmic	1-deoxy-D-xylulose 5-phosphate biosynthesis	Yes	Essential gene

^a Molecular weight was determined using the ProtParam tool [36]; ^b Molecular function (MF) and biological process (BP) for each target protein was determined using UniProt; ^c Cellular localization of pathogen targets was performed using SurfG+; ^d KEGG (Kyoto Encyclopedia of Genes and Genomes) was used to find the role of these targets in different cellular pathways; ^e PAIDB (PAthogenesis Island DataBase) and GIPSy were used to check if the putative targets are involved in pathogen virulence. DEG = Database of Essential Genes; MF = Molecular function; BP = Biological process.

2.5. Analyses of Non-Host Homologous Targets and Molecular Docking

In molecular docking, lower energy scores represent better protein-ligand bindings compared to higher energy values [37]. We considered the lower MolDock score and the interaction with the residues that were involved in the active site of the target for the prediction of therapeutic candidates. For each target protein (uvrB, pfp, asnA, recA, ndh, and dxs), a library of 28 natural compounds were docked to examine each molecule one-by-one for the selection of the final set of promising molecules that showed favorable interactions with the active site residues of targets. The biological importance for each target is described here (Table 4) along with an analysis of the predicted protein-ligand interaction(s). The name of the molecules, MolDock scores for the selected ligands, and the number of predicted hydrogen bonds with the active residues involved in these interactions are shown below for each target protein (Table 5). The predicted configurations of one of the best-docked molecules are also shown for each pathogen target in Figure 3A–F.

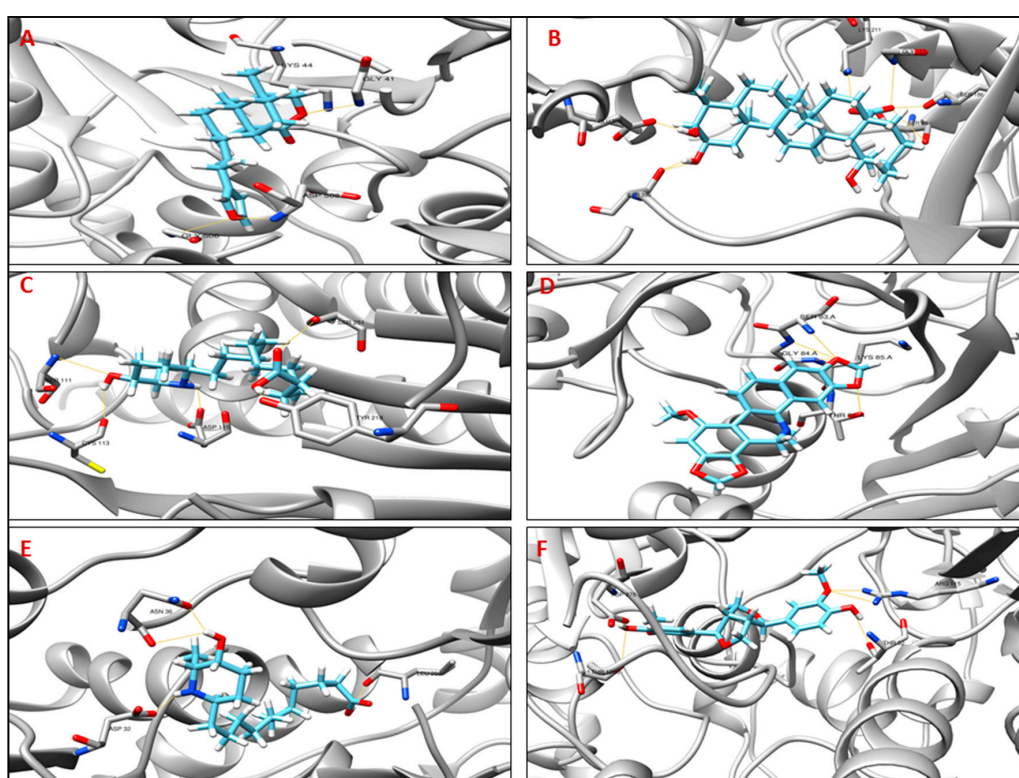


Figure 3. 3D graphic representation of the docking analyses for the most druggable protein cavity of drug target. (A) Tp_Nichols130 (uvrB, Uvr ABC system protein B) with potamogetonin (CID 5742898); (B) Tp_Nichols593 (pfp, pyrophosphate—fructose 6-phosphate 1-phosphotransferase) with jacarandic acid (CID 73645); (C) Tp_Nichols609 (asnA, aspartate-ammonia ligase) with leptophyllin B (CID 10447482); (D) Tp_Nichols754 (recA, RecA protein) with dihydrochelirubine (CID 440589); (E) Tp_Nichols9904 (ndh, NADH dehydrogenase) with leptophyllin B (CID 10447482); (F) Tp_Nichols1011 (dxs 1-deoxy-D-xylulose-5-phosphate synthase) with pinoresinol (CID 234817).

Based on a structural comparison with a crystallographic structure of the uvrB template (2d7d, uvrB from *Bacillus subtilis*), the active site residues involved in H-bond interactions with the crystallographic ligand adenosine-5'-diphosphate are Phe10, Gln11, Gln16, Gly41, Gly43, and Arg541. One of these residues, Gly41, was predicted to make hydrogen bonds to the ligand potamogetonin (CID 5742898) with a MolDock score of -97.81 . Similarly, for the target pfp template (2F48, *Borrelia burgdorferi*), the active site residues involving in H-bond interactions are Lys211, Pro210, Asp214, Gly90, Tyr434, Arg154, Met259, Arg261, and Glu320. The residue Lys211 interacts with

jacarandic acid (CID 73645) and pinoresinol (CID 234817) with MolDock scores of -62.15 and -112.67 , respectively. The compound leptophyllin B (CID 10447482) interacts with the identified active site residues Ser111, Cys113, Asp115, Tyr218, and Ser251 of *asnA* (PDB ID: 12AS from *Escherichia coli*) and Leu298, Asp32, and Asn36 of *ndh* (PDB Template ID: 2BC0 from *Streptococcus pyogenes*).

Interestingly, the drug molecule pinoresinol (CID234817) was predicted to show good results against four of our targets *uvrB*, *pfp*, *asnA*, and *dxs*. Pinoresinol is a lignan, biphenolic compound found in *Araucaria araucana* and *Sambucus williamsii*. It possesses bactericidal and fungicidal activities and therapeutic potential as an antifungal agent for the treatment of fungal infectious diseases in humans [38,39]. Thus, the identification of pinoresinol in our in silico study strengthens our protocol and can be potentially used as a new drug for the treatment of syphilis.

Table 5. The MolDock scores of natural compounds and predicted hydrogen bonds for the selected best-ranked molecules against each drug target.

Compounds Name	MolDock Score	Number of H-Bond	Residues Interacting
Tp_Nichols130 (<i>UvrB</i> , <i>UvrABC</i> System Protein B)			
Diospyrin (CID 308140) MW: ~374.3 g/mol	-119.83	4	Gly506, Asp508
Pinoresinol (CID 234817) MW: ~358.4 g/mol	-114.82	2	His64, Asp508
Potamogetonin (CID 5742898) MW: ~314.4 g/mol	-97.81	4	Gly41, Lys44, Gly506, Asp508
Tp_Nichols593 (<i>pfp</i> , Pyrophosphate-fructose 6-phosphate 1-phosphotransferase)			
Pinoresinol (CID 234817) MW: ~358.4 g/mol	-112.67	5	Ser88, Lys211, Gly260, Glu320
Jacarandic acid (CID 73645) MW: ~488.7 g/mol	-62.15	7	Ser88, Ser186, Gly183, Lys211, Glu320, Ser396
Texalin (CID 473253) MW: ~266.3 g/mol	-91.57	4	Gly90, Thr212, Ser186, Ile213
Tp_Nichols609 (<i>asnA</i> , Aspartate-ammonia ligase)			
Leptophyllin B (CID 10447482) MW: ~299.4 g/mol	-141.21	5	Ser111, Cys113, Asp115, Tyr218, Ser251
Pinoresinol (CID 234817) MW: ~358.4 g/mol	-132.814	5	Ser49, Lys77, Ser251, Arg255
Liriodenine (CID 10144) MW: ~275.1 g/mol	-95.65	2	Lys77, Arg255
Tp_Nichols754 (<i>recA</i> , Protein RecA)			
Dihydrochelirubine (CID 440589) MW: ~363.4 g/mol	-138.94	4	Gly84, Lys85, Ser83, Thr86
Piperine (CID 638024) MW: ~285.3 g/mol	-17.14	5	Ser83, Gly84, Lys84, Gln207, Gly279
Rhein (CID 10168) MW: ~284.2 g/mol	-96.11	7	Ser83, Gly84, Thr86, Tyr116, Asn254, Gly279
Tp_Nichols990 (<i>ndh</i> , NADH dehydrogenase)			
Leptophyllin B (CID 10447482) MW: ~299.4 g/mol	-122.62	4	Leu298, Asp32, Asn36
Dicentrinone (CID 177744) MW: ~335.3 g/mol	-111.09	4	Arg33, Ala11
Isosakuranetin (CID 160481) MW: ~286.3 g/mol	-109.35	3	Arg33, Ala11, Cyc42
Tp_Nichols1011 (<i>dxs</i> , 1-deoxy-D-xylulose-5-phosphate synthase)			
Pinoresinol (CID 234817) MW: ~358.4 g/mol	-146.18	5	Asp978, Thr1006, Thr32, Arg115
Piperine (CID 638024) MW: ~285.3 g/mol	-131.40	3	Thr32, Arg115, Trp980
Berberine (CID 2353) MW: ~336.4 g/mol	-115.94	3	Thr32, Gly979, Asn1011

MW = molecular weight; CID = PubChem Compound Identifier.

3. Materials and Methods

3.1. Selection of Data

The genome sequences of all 13 strains of *T. pallidum* were retrieved from the NCBI (National Center for Biotechnology Information) server [40]. For homogeneity in the functional annotation, all genomes were annotated using the RAST server (Rapid Annotations using Subsystems Technology) [41]. Furthermore, these annotated genome sequences were used for analysis.

3.2. Identification of Intra-Species Conserved Non-Host Homologous Proteins

In comparative genomics, the orthologous genes are clustered to obtain a framework to integrate information from multiple genomes, highlighting the conservation and divergence of gene families and biological processes. For pathogens, clustering orthologs can facilitate drug and/or vaccine targets identification. We compared 13 strains of *Treponema pallidum* using *Treponema pallidum* Nichols as the reference genome, using orthoMCL software [26] with an E-value of 1×10^{-50} . CDSs shared by all strains were considered a part of the core genome. The possible candidates for drugs and/or vaccines should be non-homologues to human proteins; thus, autoimmunity is avoided, and an accurate immune response is elicited against the targeted pathogen. Accordingly, these core genes were subjected to orthoMCL software (E-value = 1×10^{-50}) against the human genome for the identification of non-host homolog targets.

3.3. Identification of Pathogenicity Islands

Knowledge about pathogenicity islands, the virulence factors they encode, their mobility, and their structure is not only helpful in understanding the bacterial evolution and their interactions with eukaryotic host cells, but may also facilitate in providing delivery systems for vaccination and tools for the development of new approaches for treating bacterial infections [28]. The identification of pathogenicity islands in the genome of *T. pallidum* Nichols was performed with GIPSY (Genomic Island Prediction Software) [27] through the detection of regions presenting: deviations in genomic signature (i.e., anomalous G+C and/or codon usage deviation); presence of transposase, virulence or flanking tRNA genes; and absence in the non-pathogenic organism *Treponema denticola*.

3.4. Assessment of Essential Genes

A subtractive genomics approach was followed to identify conserved targets that were essential to the bacteria [13]. The set of core conserved proteins of *T. pallidum* Nichols was subjected to the Database of Essential Genes (DEG) [42] for homology analyses. The DEG contains experimentally validated data from bacteria, archaea, and eukaryotes that are comprised of currently reported essential genomic elements including protein-coding genes that are indispensable to support cellular life. The cut-off values used for BLASTp were: E-value = 0.0001, bit score = 100, and identity = 25% [15,18,30].

3.5. Reverse Vaccinology Approach for Prediction of Putative *T. pallidum* Vaccine Targets

For potential vaccine targets, subcellular localization and the secretion of pathogenic proteins are important factors for consideration, where secreted and membrane proteins are the first to be in contact with the host, eliciting an immune response. Therefore, the prediction of the exoproteome or secretome, composed of the proteins localized in the extracellular matrix or outer membrane of the organism, is highly valuable for reverse vaccinology strategies. In combination with subtractive proteomics, reverse vaccinology can provide a more reliable output compared to screening of the whole data set without considering prioritizing parameters [13]. The non-host homologous conserved proteome of *T. pallidum* Nichols was screened using SurfG+ software [29] to identify secreted proteins, membrane proteins, and putative surface exposed proteins. We searched for cleavage sites and transmembrane helices in all 15 proteins using SignalP [43] and TMHMM (Transmembrane Helix prediction server, based on

a hidden Markov model) [44], respectively, and we also predicted the presence of functional domains for all the 15 proteins with InterProScan, which uses several databases for domain prediction [45]. The dataset was screened by Vaxign [30] by searching for proteins with the following features: major histocompatibility complex (MHC I) and (MHC II) binding properties, an adhesion probability greater than 0.51, and no similarity to host proteins.

3.6. High Throughput Structural Modeling

MHOLine [31] was used to predict the modelome (complete set of protein 3D models for the whole conserved core non-host homologous proteome). MHOLine utilizes multi-fasta files of amino acids as input data and then uses HMMTOP, BLAST, BATS, MODELLER, and PROCHECK programs for the detailed analyses. The program HMMTOP detects transmembrane regions [46]. The BLAST algorithm is used to identify template structure by performing random searches against the Protein Data Bank [47]. BATS (Blast Automatic Targeting for Structures) performs the refinement in the template search; its use represents a key step for the model construction. BATS refinement identifies sequences that make the modeling possible by selecting templates from the BLAST output file using their BATS scores, expectation values, identity, and sequence similarity as criteria as well as considering the number of gaps and the alignment coverage. BATS selects the best template for 3D model generation and performs automated alignment used by the MODELLER program. The adopted methodology was revised accordingly from the original work by Hassan et al. [46].

3.7. Ligand Libraries and Docking Analyses

The ligand libraries of 28 natural compounds presented by Tiwari et al., 2014 [48] were used for the docking analysis. The 3D structures of all target proteins were carefully examined for structural errors (wrong bonds, missing atoms, and protonation states) in the MVD (Molegro Virtual Docker) [37]. The active side residues of the target proteins were identified by comparing its 3D structure to the respective templates. Furthermore, taking identified cavities from a template used in a grid for molecular docking. The program includes three search algorithms for molecular docking analyses, namely MolDock Optimizer [37], MolDock Simplex Evolution (SE), and Iterated Simplex (IS). We employed the MolDock Optimizer search algorithm, which is based on a differential evolutionary algorithm, using the default parameters, that are (a) population size = 50; (b) scaling factor = 0.5; and (c) crossover rate = 0.9. The 3D poses of docked molecules were analyzed in Chimera [49]. Molecular function (MF) and biological process (BP) for each target protein were determined using UniProt [41]. The biochemical pathway of these proteins were checked using KEGG (Kyoto Encyclopedia of Genes and Genomes) [50], SurfG+ software [29], and virulence using GIPSy [31]. The final list of targets was based on 12 criteria, as described earlier in [13,46].

4. Conclusions

Here, the genomic information was used with the aim of determining the conserved proteome of 13 strains of *Treponema pallidum* in a search for regions of genome plasticity. Moreover, we used reverse vaccinology and subtractive genomics to predict new antigenic/drug targets, which can be used in the development of new vaccines and drugs for *Treponema pallidum*. After a detailed in silico analysis between host and pathogen proteins, we suggest that the identified non-host homologous proteins could be considered for prophylaxis of syphilis due to further experimental validations.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/18/2/402/s1.

Acknowledgments: We acknowledge the support of all team members and financing agencies. Arun Kumar Jaiswal was supported by the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil) fellowship for doctoral studies. Sandeep Tiwari, Syed Babar Jamal acknowledges the “TWAS-CNPq Postgraduate Fellowship Programme” for granting a fellowship for doctoral studies. Debmalya Barh acknowledges the “TWAS-CNPq Postdoctoral Fellowship Programme” for granting a fellowship for postdoctoral studies.

The authors also thank the funding agency FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais) for financial support.

Author Contributions: Arun Kumar Jaiswal, Sandeep Tiwari, and Siomar C. Soares planned the entire work; Arun Kumar Jaiswal, Sandeep Tiwari, Syed Babar Jamal, and Siomar C. Soares analyzed the data; Arun Kumar Jaiswal, Sandeep Tiwari, Syed Babar Jamal, and Siomar C. Soares drafted the manuscript; Siomar C. Soares, Vasco Azevedo, Sandeep Tiwari, Debmalya Barh, and Arun Kumar Jaiswal reviewed and analyzed the manuscript.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Wagenlehner, F.M.; Brockmeyer, N.H.; Discher, T.; Friese, K.; Wichelhaus, T.A. The presentation, diagnosis, and treatment of sexually transmitted infections. *Dtsch. Arztebl. Int.* **2016**, *113*, 11–22. [[PubMed](#)]
2. Nyatsanza, F.; Tipple, C. Syphilis: Presentations in general medicine. *Clin. Med.* **2016**, *16*, 184–188. [[CrossRef](#)] [[PubMed](#)]
3. Newman, L.; Rowley, J.; Vander Hoorn, S.; Wijesooriya, N.S.; Unemo, M.; Low, N.; Stevens, G.; Gottlieb, S.; Kiarie, J.; Temmerman, M. Global estimates of the prevalence and incidence of four curable sexually transmitted infections in 2012 based on systematic review and global reporting. *PLoS ONE* **2015**, *10*, e0143304. [[CrossRef](#)] [[PubMed](#)]
4. Lafeta, K.R.; Martelli Junior, H.; Silveira, M.F.; Paranaíba, L.M. Maternal and congenital syphilis, underreported and difficult to control. *Rev. Bras. Epidemiol.* **2016**, *19*, 63–74. [[PubMed](#)]
5. Deperthes, B.D.; Meheus, A.; O'Reilly, K.; Broutet, N. Maternal and congenital syphilis programmes: Case studies in Bolivia, Kenya and South Africa. *Bull World Health Organ* **2004**, *82*, 410–416. [[PubMed](#)]
6. Cameron, C.E.; Lukehart, S.A. Current status of syphilis vaccine development: Need, challenges, prospects. *Vaccine* **2014**, *32*, 1602–1609. [[CrossRef](#)] [[PubMed](#)]
7. Radolf, J.D. Treponema. In *Medical Microbiology*, 4th ed.; Baron, S., Ed.; University of Texas Medical Branch at Galveston: Galveston, TX, USA, 1996.
8. Jakopanec, I.; Grjibovski, A.M.; Nilsen, O.; Aavitsland, P. Syphilis epidemiology in Norway, 1992–2008: Resurgence among men who have sex with men. *BMC Infect. Dis.* **2010**, *10*. [[CrossRef](#)] [[PubMed](#)]
9. Tucker, J.D.; Cohen, M.S. China's syphilis epidemic: Epidemiology, proximate determinants of spread, and control responses. *Curr. Opin. Infect. Dis.* **2011**, *24*, 50–55. [[CrossRef](#)] [[PubMed](#)]
10. Abara, W.E.; Hess, K.L.; Neblett Fanfair, R.; Bernstein, K.T.; Paz-Bailey, G. Syphilis trends among men who have sex with men in the United States and Western Europe: A systematic review of trend studies published between 2004 and 2015. *PLoS ONE* **2016**, *11*, e0159309. [[CrossRef](#)] [[PubMed](#)]
11. Cabie, A.; Rollin, B.; Pierre-Francois, S.; Abel, S.; Desbois, N.; Richard, P.; Hochedez, P.; Theodose, R.; Quist, D.; Helenon, R.; et al. Reemergence of syphilis in Martinique, 2001–2008. *Emerg. Infect. Dis.* **2010**, *16*, 106–109. [[CrossRef](#)] [[PubMed](#)]
12. Plotkin, S.A. Why certain vaccines have been delayed or not developed at all. *Health Aff.* **2005**, *24*, 631–634. [[CrossRef](#)] [[PubMed](#)]
13. Barh, D.; Tiwari, S.; Jain, N.; Ali, A.; Santos, A.R.; Misra, A.N.; Azevedo, V.; Kumar, A. In silico subtractive genomics for target identification in human bacterial pathogens. *Drug Dev. Res.* **2011**, *72*, 162–177. [[CrossRef](#)]
14. Barh, D.; Gupta, K.; Jain, N.; Khatri, G.; Leon-Sicairos, N.; Canizalez-Roman, A.; Tiwari, S.; Verma, A.; Rahangdale, S.; Shah Hassan, S.; et al. 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. *Integr. Biol.* **2013**, *5*, 495–509.
15. Perumal, D.; Lim, C.S.; Sakharkar, K.R.; Sakharkar, M.K. Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. *In Silico Biol.* **2007**, *7*, 453–465. [[PubMed](#)]
16. Pizza, M.; Scarlato, V.; Masignani, V.; Giuliani, M.M.; Arico, B.; Comanducci, M.; Jennings, G.T.; Baldi, L.; Bartolini, E.; Capecchi, B.; et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **2000**, *287*, 1816–1820. [[CrossRef](#)] [[PubMed](#)]
17. Asif, S.M.; Asad, A.; Faizan, A.; Anjali, M.S.; Arvind, A.; Neelesh, K.; Hirdesh, K.; Sanjay, K. Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis. *Bioinformatics* **2009**, *4*, 245–248. [[CrossRef](#)] [[PubMed](#)]

18. Dutta, A.; Singh, S.K.; Ghosh, P.; Mukherjee, R.; Mitter, S.; Bandyopadhyay, D. In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. *In Silico Biol* **2006**, *6*, 43–47. [[PubMed](#)]
19. Chong, C.E.; Lim, B.S.; Nathan, S.; Mohamed, R. In silico analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In Silico Biol* **2006**, *6*, 341–346. [[PubMed](#)]
20. Sakharkar, K.R.; Sakharkar, M.K.; Chow, V.T. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. *In Silico Biol* **2004**, *4*, 355–360. [[PubMed](#)]
21. Rathi, B.; Sarangi, A.N.; Trivedi, N. Genome subtraction for novel target definition in *Salmonella typhi*. *Bioinformatics* **2009**, *4*, 143–150. [[CrossRef](#)] [[PubMed](#)]
22. Barh, D.; Kumar, A. In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*. *In Silico Biol* **2009**, *9*, 225–231. [[PubMed](#)]
23. Barh, D.; Jain, N.; Tiwari, S.; Parida, B.P.; D'Afonseca, V.; Li, L.; Ali, A.; Santos, A.R.; Guimaraes, L.C.; de Castro Soares, S.; et al. 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. [[CrossRef](#)] [[PubMed](#)]
24. Aronov, A.M.; Verlinde, C.L.; Hol, W.G.; Gelb, M.H. Selective tight binding inhibitors of trypanosomal glyceraldehyde-3-phosphate dehydrogenase via structure-based drug design. *J. Med. Chem.* **1998**, *41*, 4790–4799. [[CrossRef](#)] [[PubMed](#)]
25. Singh, S.; Malik, B.K.; Sharma, D.K. Molecular modeling and docking analysis of entamoeba histolytica glyceraldehyde-3 phosphate dehydrogenase, a potential target enzyme for anti-protozoal drug development. *Chem. Biol. Drug Des.* **2008**, *71*, 554–562. [[CrossRef](#)] [[PubMed](#)]
26. Li, L. Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **2003**, *13*, 2178–2189. [[CrossRef](#)] [[PubMed](#)]
27. Soares, S.C.; Geyik, H.; Ramos, R.T.; de Sa, P.H.; Barbosa, E.G.; Baumbach, J.; Figueiredo, H.C.; Miyoshi, A.; Tauch, A.; Silva, A.; et al. Gipsy: Genomic island prediction software. *J. Biotechnol.* **2016**, *232*, 2–11. [[CrossRef](#)] [[PubMed](#)]
28. Naz, A.; Awan, F.M.; Obaid, A.; Muhammad, S.A.; Paracha, R.Z.; Ahmad, J.; Ali, A. Identification of putative vaccine candidates against *Helicobacter pylori* exploiting exoproteome and secretome: A reverse vaccinology based approach. *Infect. Genet. Evol.* **2015**, *32*, 280–291. [[CrossRef](#)] [[PubMed](#)]
29. Barinov, A.; Loux, V.; Hammani, A.; Nicolas, P.; Langella, P.; Ehrlich, D.; Maguin, E.; van de Guchte, M. Prediction of surface exposed proteins in streptococcus pyogenes, with a potential application to other gram-positive bacteria. *Proteomics* **2009**, *9*, 61–73. [[CrossRef](#)] [[PubMed](#)]
30. He, Y.; Xiang, Z.; Mobley, H.L.T. Vaxign: The first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J. Biomed. Biotechnol.* **2010**, *2010*, 1–15. [[CrossRef](#)] [[PubMed](#)]
31. Capriles, P.V.S.Z.; Guimarães, A.C.R.; Otto, T.D.; Miranda, A.B.; Dardenne, L.E.; Degraeve, W.M. Structural modelling and comparative analysis of homologous, analogous and specific proteins from *Trypanosoma cruzi* versus *Homo sapiens*: Putative drug targets for chagas' disease treatment. *BMC Genomics* **2010**, *11*, 610. [[CrossRef](#)] [[PubMed](#)]
32. Mondal, S.I.; Ferdous, S.; Jewel, N.A.; Akter, A.; Mahmud, Z.; Islam, M.M.; Afrin, T.; Karim, N. Identification of potential drug targets by subtractive genome analysis of *Escherichia coli* O157:H7: An in silico approach. *Adv. Appl. Bioinform. Chem.* **2015**, *8*, 49–63. [[CrossRef](#)] [[PubMed](#)]
33. Duffield, M.; Cooper, I.; McAlister, E.; Bayliss, M.; Ford, D.; Oyston, P. Predicting conserved essential genes in bacteria: In silico identification of putative drug targets. *Mol. Biosyst.* **2010**, *6*, 2482–2489. [[CrossRef](#)] [[PubMed](#)]
34. Delcour, A.H. Outer membrane permeability and antibiotic resistance. *Biochim. Biophys. Acta* **2009**, *1794*, 808–816. [[CrossRef](#)] [[PubMed](#)]
35. Vaara, M. Agents that increase the permeability of the outer membrane. *Microbiol. Rev.* **1992**, *56*, 395–411. [[PubMed](#)]
36. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.E.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein identification and analysis tools on the expasy server. In *The Proteomics Protocols Handbook*; Walker, J.M., Ed.; Humana Press: New York, NY, USA, 2005; pp. 571–607.
37. Thomsen, R.; Christensen, M.H. Moldock: A new technique for high-accuracy molecular docking. *J. Med. Chem.* **2006**, *49*, 3315–3321. [[CrossRef](#)] [[PubMed](#)]

38. Hwang, B.; Lee, J.; Liu, Q.H.; Woo, E.R.; Lee, D.G. Antifungal effect of (+)-pinoselinol isolated from *Sambucus williamsii*. *Molecules* **2010**, *15*, 3507–3516. [[CrossRef](#)] [[PubMed](#)]
39. Cespedes, C.L.; Avila, J.G.; Garcia, A.M.; Becerra, J.; Flores, C.; Aqueveque, P.; Bittner, M.; Hoeneisen, M.; Martinez, M.; Silva, M. Antifungal and antibacterial activities of *Araucaria araucana* (Mol.) K. Koch heartwood lignans. *Z. Naturforsch. C* **2006**, *61*, 35–43. [[CrossRef](#)] [[PubMed](#)]
40. Coordinators, N.R. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2016**, *44*, 7–19.
41. Brettin, T.; Davis, J.J.; Disz, T.; Edwards, R.A.; Gerdes, S.; Olsen, G.J.; Olson, R.; Overbeek, R.; Parrello, B.; Pusch, G.D.; et al. *Rasttk*: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **2015**, *5*. [[CrossRef](#)] [[PubMed](#)]
42. Zhang, R.; Ou, H.Y.; Zhang, C.T. Deg: A database of essential genes. *Nucleic Acids Res.* **2004**, *32*, D271–D272. [[CrossRef](#)] [[PubMed](#)]
43. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. Signalp 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785–786. [[CrossRef](#)] [[PubMed](#)]
44. Sonnhammer, E.L.; von Heijne, G.; Krogh, A. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1998**, *6*, 175–182. [[PubMed](#)]
45. Mitchell, A.; Chang, H.Y.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S.; et al. The interpro protein families database: The classification resource after 15 years. *Nucleic Acids Res.* **2015**, *43*, D213–D221. [[CrossRef](#)] [[PubMed](#)]
46. Hassan, S.S.; Tiwari, S.; Guimaraes, L.C.; Jamal, S.B.; Folador, E.; Sharma, N.B.; de Castro Soares, S.; Almeida, S.; Ali, A.; Islam, A.; et al. Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Corynebacterium pseudotuberculosis*. *BMC Genom.* **2014**, *15*. [[CrossRef](#)]
47. Gutmanas, A.; Alhroub, Y.; Battle, G.M.; Berrisford, J.M.; Bochet, E.; Conroy, M.J.; Dana, J.M.; Fernandez Montecelo, M.A.; van Ginkel, G.; Gore, S.P.; et al. PDBE: Protein data bank in Europe. *Nucleic Acids Res.* **2014**, *42*, D285–D291. [[CrossRef](#)] [[PubMed](#)]
48. Tiwari, S.; da Costa, M.P.; Almeida, S.; Hassan, S.S.; Jamal, S.B.; Oliveira, A.; Folador, E.L.; Rocha, F.; de Abreu, V.A.; Dorella, F.; et al. *C. pseudotuberculosis* PhoP confers virulence and may be targeted by natural compounds. *Integr. Biol.* **2014**, *6*, 1088–1099. [[CrossRef](#)] [[PubMed](#)]
49. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. Ucsf chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [[CrossRef](#)] [[PubMed](#)]
50. Kanehisa, M.; Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [[CrossRef](#)] [[PubMed](#)]



Query Protein	No. of homologs in DEG
Tp_Nichols84	4
Tp_Nichols85	3
Tp_Nichols87	9
Tp_Nichols89	1
Tp_Nichols91	7
Tp_Nichols95	6
Tp_Nichols100	12
Tp_Nichols102	12
Tp_Nichols105	9
Tp_Nichols107	3
Tp_Nichols110	2
Tp_Nichols111	6
Tp_Nichols112	4
Tp_Nichols118	1
Tp_Nichols125	5
Tp_Nichols127	2
Tp_Nichols128	2
Tp_Nichols129	10
Tp_Nichols130	5
Tp_Nichols238	2
Tp_Nichols239	1
Tp_Nichols243	1
Tp_Nichols251	3
Tp_Nichols252	15
Tp_Nichols259	23
Tp_Nichols261	22
Tp_Nichols262	25
Tp_Nichols271	7
Tp_Nichols278	25
Tp_Nichols280	28
Tp_Nichols297	7
Tp_Nichols298	24
Tp_Nichols299	16
Tp_Nichols300	19
Tp_Nichols303	1
Tp_Nichols305	4
Tp_Nichols310	25
Tp_Nichols316	4
Tp_Nichols318	5
Tp_Nichols322	10
Tp_Nichols327	1
Tp_Nichols332	31
Tp_Nichols347	1
Tp_Nichols350	21

Tp_Nichols355	28
Tp_Nichols360	2
Tp_Nichols361	1
Tp_Nichols367	16
Tp_Nichols369	42
Tp_Nichols373	36
Tp_Nichols377	4
Tp_Nichols379	9
Tp_Nichols388	3
Tp_Nichols395	2
Tp_Nichols400	7
Tp_Nichols404	23
Tp_Nichols406	5
Tp_Nichols407	29
Tp_Nichols412	33
Tp_Nichols157	3
Tp_Nichols416	7
Tp_Nichols419	50
Tp_Nichols422	37
Tp_Nichols423	33
Tp_Nichols426	2
Tp_Nichols428	3
Tp_Nichols434	6
Tp_Nichols440	24
Tp_Nichols442	6
Tp_Nichols445	39
Tp_Nichols446	21
Tp_Nichols448	16
Tp_Nichols449	2
Tp_Nichols454	4
Tp_Nichols135	1
Tp_Nichols160	2
Tp_Nichols462	46
Tp_Nichols463	46
Tp_Nichols467	11
Tp_Nichols472	13
Tp_Nichols479	26
Tp_Nichols480	1
Tp_Nichols484	24
Tp_Nichols486	4
Tp_Nichols488	63
Tp_Nichols496	2
Tp_Nichols497	9
Tp_Nichols502	5
Tp_Nichols510	6

Tp_Nichols514	6
Tp_Nichols533	5
Tp_Nichols534	27
Tp_Nichols536	2
Tp_Nichols538	21
Tp_Nichols539	15
Tp_Nichols550	2
Tp_Nichols555	3
Tp_Nichols557	5
Tp_Nichols558	13
Tp_Nichols560	17
Tp_Nichols561	5
Tp_Nichols562	1
Tp_Nichols564	19
Tp_Nichols565	34
Tp_Nichols567	32
Tp_Nichols171	2
Tp_Nichols570	17
Tp_Nichols172	1
Tp_Nichols590	1
Tp_Nichols592	64
Tp_Nichols593	11
Tp_Nichols594	8
Tp_Nichols597	12
Tp_Nichols598	21
Tp_Nichols173	2
Tp_Nichols606	1
Tp_Nichols609	1
Tp_Nichols613	2
Tp_Nichols615	2
Tp_Nichols619	1
Tp_Nichols621	1
Tp_Nichols175	2
Tp_Nichols624	2
Tp_Nichols630	56
Tp_Nichols632	62
Tp_Nichols638	1
Tp_Nichols640	1
Tp_Nichols642	9
Tp_Nichols652	24
Tp_Nichols656	11
Tp_Nichols657	25
Tp_Nichols658	31
Tp_Nichols669	100
Tp_Nichols672	45

Tp_Nichols673	5
Tp_Nichols681	1
Tp_Nichols682	3
Tp_Nichols685	1
Tp_Nichols181	13
Tp_Nichols694	35
Tp_Nichols705	23
Tp_Nichols706	3
Tp_Nichols709	1
Tp_Nichols714	5
Tp_Nichols715	17
Tp_Nichols716	2
Tp_Nichols722	24
Tp_Nichols726	2
Tp_Nichols729	41
Tp_Nichols184	100
Tp_Nichols742	15
Tp_Nichols751	75
Tp_Nichols753	4
Tp_Nichols186	2
Tp_Nichols754	6
Tp_Nichols187	2
Tp_Nichols767	1
Tp_Nichols768	14
Tp_Nichols777	57
Tp_Nichols778	1
Tp_Nichols780	1
Tp_Nichols784	3
Tp_Nichols190	6
Tp_Nichols788	2
Tp_Nichols789	1
Tp_Nichols794	3
Tp_Nichols808	21
Tp_Nichols809	4
Tp_Nichols194	1
Tp_Nichols817	4
Tp_Nichols818	42
Tp_Nichols837	1
Tp_Nichols839	1
Tp_Nichols840	1
Tp_Nichols841	1
Tp_Nichols842	15
Tp_Nichols849	100
Tp_Nichols853	3
Tp_Nichols859	3

Tp_Nichols863	2
Tp_Nichols867	2
Tp_Nichols868	3
Tp_Nichols872	3
Tp_Nichols874	1
Tp_Nichols201	4
Tp_Nichols886	2
Tp_Nichols887	1
Tp_Nichols890	33
Tp_Nichols892	10
Tp_Nichols894	12
Tp_Nichols203	10
Tp_Nichols204	2
Tp_Nichols911	19
Tp_Nichols918	30
Tp_Nichols919	2
Tp_Nichols920	21
Tp_Nichols922	1
Tp_Nichols930	3
Tp_Nichols934	4
Tp_Nichols944	22
Tp_Nichols945	16
Tp_Nichols952	3
Tp_Nichols957	25
Tp_Nichols959	6
Tp_Nichols961	29
Tp_Nichols969	3
Tp_Nichols971	28
Tp_Nichols976	26
Tp_Nichols981	1
Tp_Nichols986	6
Tp_Nichols990	32
Tp_Nichols991	3
Tp_Nichols997	11
Tp_Nichols1003	34
Tp_Nichols1005	2
Tp_Nichols1007	6
Tp_Nichols1011	2
Tp_Nichols214	33
Tp_Nichols1018	1
Tp_Nichols1020	1
Tp_Nichols1021	27
Tp_Nichols215	31
Tp_Nichols1027	1
Tp_Nichols1031	1

Tp_Nichols1050	4
Tp_Nichols1053	3
Tp_Nichols1057	6
Tp_Nichols1058	2
Tp_Nichols1062	1
Tp_Nichols1070	3
Tp_Nichols1075	5
Tp_Nichols220	30
Tp_Nichols1076	2
Tp_Nichols1077	26
Tp_Nichols1083	2
Tp_Nichols221	28
Tp_Nichols1090	31
Tp_Nichols1094	16
Tp_Nichols222	34
Tp_Nichols1097	4
Tp_Nichols1101	4
Tp_Nichols1120	2
Tp_Nichols1123	16
Tp_Nichols1	30
Tp_Nichols2	27
Tp_Nichols3	4
Tp_Nichols5	55
Tp_Nichols226	25
Tp_Nichols14	2
Tp_Nichols16	9
Tp_Nichols227	33
Tp_Nichols31	26
Tp_Nichols34	2
Tp_Nichols37	100
Tp_Nichols38	2
Tp_Nichols42	7
Tp_Nichols46	1
Tp_Nichols49	2
Tp_Nichols229	29
Tp_Nichols52	16
Tp_Nichols63	33
Tp_Nichols65	5
Tp_Nichols66	20
Tp_Nichols67	22
Tp_Nichols231	33
Tp_Nichols71	4
Tp_Nichols73	1
Tp_Nichols74	1

DEG AC Number

DEG10270504; DEG10070115; DEG10060157; DEG10140014;
DEG10250450; DEG10350141; DEG10350142
DEG10240265; DEG10310059; DEG10350106; DEG10070101; DEG10410479; DEG10360122; DEG1040001
DEG10150183;
DEG10030266; DEG10410274; DEG10220041; DEG10340155; DEG10230304; DEG10260026; DEG1028003
DEG10180477; DEG10020243; DEG10240033; DEG10020295; DEG10050416; DEG10020060;
DEG10030043; DEG10370123; DEG10210106; DEG10410438; DEG10200343; DEG10400434; DEG1028045
DEG10160051; DEG10030497; DEG10230118; DEG10340177; DEG10250222; DEG10190152; DEG1028046
DEG10140080; DEG10170051; DEG10340084; DEG10200353; DEG10050446; DEG10060240; DEG1023023
DEG10340273; DEG10220094; DEG10390051;
DEG10410333; DEG10140073;
DEG10280307; DEG10050531; DEG10100069; DEG10410056; DEG10250105; DEG10010166;
DEG10080040; DEG10410465; DEG10400083; DEG10100070;
DEG10350224;
DEG10180476; DEG10220248; DEG10340510; DEG10230134; DEG10390151;
DEG10110223; DEG10300086;
DEG10260103; DEG10300085;
DEG10130141; DEG10280018; DEG10100054; DEG10070085; DEG10270078; DEG10250080; DEG1020038
DEG10050448; DEG10060055; DEG10020228; DEG10020066; DEG10270219;
DEG10130374; DEG10400645;
DEG10400645;
DEG10280385;
DEG10410180; DEG10240229; DEG10200162;
DEG10330271; DEG10240354; DEG10350415; DEG10150311; DEG10290352; DEG10400176; DEG1018056
DEG10390197; DEG10240346; DEG10160263; DEG10360211; DEG10080222; DEG10330266; DEG1035041
DEG10170041; DEG10120338; DEG10060064; DEG10240344; DEG10410418; DEG10280261; DEG1034048
DEG10170042; DEG10120337; DEG10240343; DEG10160261; DEG10410417; DEG10060295; DEG1033026
DEG10200239; DEG10350049; DEG10290076; DEG10050029; DEG10170229; DEG10240052; DEG1031016
DEG10200477; DEG10380087; DEG10120357; DEG10140097; DEG10360316; DEG10240272; DEG1022034
DEG10170170; DEG10120048; DEG10290176; DEG10160080; DEG10150154; DEG10130054; DEG1027049
DEG10340398; DEG10230115; DEG10180135; DEG10310048; DEG10220461; DEG10080042; DEG1031012
DEG10100608; DEG10170191; DEG10380106; DEG10010155; DEG10070190; DEG10210108; DEG1022017
DEG10240406; DEG10200475; DEG10230274; DEG10270685; DEG10390016; DEG10020017; DEG1010061
DEG10200476; DEG10030779; DEG10180197; DEG10270326; DEG10260032; DEG10340116; DEG1035049
DEG10400418;
DEG10400141; DEG10400602; DEG10110109; DEG10410042;
DEG10370161; DEG10130351; DEG10200198; DEG10240027; DEG10160282; DEG10150013; DEG1040050
DEG10140113; DEG10060221; DEG10180129; DEG10140085;
DEG10350322; DEG10410550; DEG10030351; DEG10030188; DEG10270161;
DEG10290066; DEG10130055; DEG10150266; DEG10390179; DEG10010210; DEG10170251; DEG1021014
DEG10060096;
DEG10390232; DEG10290055; DEG10370222; DEG10140283; DEG10230143; DEG10060252; DEG1040037
DEG10060096;
DEG10400239; DEG10390027; DEG10350253; DEG10120316; DEG10290158; DEG10190036; DEG1016004

DEG10190184; DEG10400074; DEG10080192; DEG10390002; DEG10120163; DEG10200390; DEG1010056
DEG10360085; DEG10150148;

[DEG10200112;](#)

DEG10060176; DEG10050080; DEG10050199; DEG10340089; DEG10180012; DEG10260015; DEG1005002
DEG10290363; DEG10330019; DEG10030478; DEG10050409; DEG10290366; DEG10230206; DEG1019001
DEG10130456; DEG10290367; DEG10330015; DEG10240107; DEG10010244; DEG10400038; DEG1019001
DEG10180514; DEG10030148; DEG10310010; DEG10150269;

DEG10130307; DEG10410038; DEG10400544; DEG10240123; DEG10100390; DEG10280435; DEG1025047
DEG10250599; DEG10270549; DEG10100493;

DEG10110115; DEG10300014;

DEG10350262; DEG10020122; DEG10070066; DEG10140282; DEG10240232; DEG10010121; DEG1041052
DEG10150282; DEG10160086; DEG10400078; DEG10010011; DEG10230027; DEG10350026; DEG1012002
DEG10390023; DEG10170028; DEG10250189; DEG10340167; DEG10220047;

DEG10130378; DEG10340189; DEG10030433; DEG10010016; DEG10380003; DEG10200391; DEG1016004
DEG10400071; DEG10260035; DEG10390172; DEG10270340; DEG10030473; DEG10270575; DEG1015024
DEG10140046; DEG10380036; DEG10060261;

DEG10060188; DEG10340068; DEG10330009; DEG10160008; DEG10250416; DEG10140189; DEG1010033
DEG10130457; DEG10380166; DEG10340078; DEG10330014; DEG10290368; DEG10290369; DEG1024010
DEG10290361; DEG10400077; DEG10030476; DEG10340058; DEG10350377; DEG10310087; DEG1028045
DEG10340057; DEG10290360; DEG10400616; DEG10350376; DEG10030475; DEG10070204; DEG1031008
DEG10080058; DEG10320221;

DEG10350013; DEG10180554; DEG10100295;

DEG10270526; DEG10240293; DEG10100470; DEG10030353; DEG10350216; DEG10250570;

DEG10400612; DEG10010201; DEG10270244; DEG10160266; DEG10340092; DEG10170123; DEG1032032
DEG10330129; DEG10320102; DEG10020122; DEG10160127; DEG10060185; DEG10010121;

DEG10130097; DEG10170232; DEG10270478; DEG10050106; DEG10150077; DEG10270477; DEG1039021
DEG10130097; DEG10170232; DEG10190048; DEG10270477; DEG10180074; DEG10050106; DEG1025050
DEG10170303; DEG10370121; DEG10380128; DEG10270603; DEG10160225; DEG10360270; DEG103202:

DEG10110061; DEG10050085;

DEG10290292; DEG10390008; DEG10340117; DEG10220025;

[DEG10050119;](#)

DEG10050061; DEG10030729;

DEG10030559; DEG10380085; DEG10200418; DEG10380087; DEG10060328; DEG10150334; DEG1020041
DEG10030559; DEG10380085; DEG10200418; DEG10380087; DEG10060328; DEG10150334; DEG1020041
DEG10220316; DEG10240021; DEG10390185; DEG10340150; DEG10150029; DEG10400453; DEG1036021
DEG10060159; DEG10070076; DEG10270505; DEG10390149; DEG10140198; DEG10170007; DEG1034040
DEG10120306; DEG10030174; DEG10280205; DEG10320219; DEG10160185; DEG10230262; DEG1025035

[DEG10070066;](#)

DEG10130094; DEG10400193; DEG10270515; DEG10010176; DEG10310085; DEG10220209; DEG1024027
DEG10060021; DEG10020233; DEG10400425; DEG10140121;

DEG10010033; DEG10130313; DEG10340492; DEG10230160; DEG10060366; DEG10170048; DEG1017004
DEG10010167; DEG10140034;

DEG10050444; DEG10220254; DEG10270327; DEG10100297; DEG10250360; DEG10230257; DEG1018021
DEG10270034; DEG10250032; DEG10100016; DEG10030084; DEG10060285;

DEG10060174; DEG10100231; DEG10250271; DEG10020104; DEG10270258; DEG10180295;

DEG10140081; DEG10060294; DEG10180359; DEG10030589; DEG10220242; DEG10020202;
DEG10080094; DEG10220077; DEG10410367; DEG10250500; DEG10290246;
DEG10120215; DEG10130353; DEG10240367; DEG10100370; DEG10010179; DEG10210086; DEG100602
DEG10220437; DEG10230042;
DEG10200186; DEG10340449; DEG10160241; DEG10400077; DEG10060246; DEG10330244; DEG1019019
DEG10400076; DEG10200187; DEG10190197; DEG10240006; DEG10160240; DEG10290350; DEG1034044
DEG10400655; DEG10300030;
DEG10250692; DEG10200090; DEG10280383;
DEG10260067; DEG10050201; DEG10120193; DEG10300004; DEG10380035;
DEG10020250; DEG10380237; DEG10120209; DEG10360106; DEG10370225; DEG10060151; DEG1035019
DEG10160121; DEG10200017; DEG10190085; DEG10350053; DEG10150272; DEG10290308; DEG1027068
DEG10250511; DEG10310228; DEG10290232; DEG10400206; DEG10180312;

[DEG10170140;](#)

DEG10200087; DEG10300095; DEG10200130; DEG10300076; DEG10200128; DEG10200422; DEG101700
DEG10170021; DEG10160163; DEG10380161; DEG10030223; DEG10220190; DEG10320058; DEG1014001
DEG10330018; DEG10290364; DEG10030479; DEG10120272; DEG10050408; DEG10270376; DEG1023020
DEG10220359; DEG10390216;
DEG10140194; DEG10350297; DEG10290217; DEG10180585; DEG10410268; DEG10270470; DEG1020010

[DEG10400199;](#)

[DEG10400534;](#)

DEG10200195; DEG10410374; DEG10030494; DEG10170210; DEG10130096; DEG10170194; DEG1039000
DEG10070151; DEG10380154; DEG10010211; DEG10340339; DEG10170253; DEG10370140; DEG1021008
DEG10350352; DEG10170234; DEG10220179; DEG10030355; DEG10410191; DEG10390102; DEG1030002
DEG10080174; DEG10190195; DEG10270215; DEG10380241; DEG10180482; DEG10160238; DEG1029034
DEG10290304; DEG10280274; DEG10230107; DEG10190003; DEG10150268; DEG10130450; DEG1020040
DEG10050140; DEG10260015;

[DEG10310181;](#)

[DEG10050178;](#)

DEG10060305; DEG10140285;
DEG10220343; DEG10390204;

[DEG10170091;](#)

[DEG10240141;](#)

DEG10290256; DEG10180128;
DEG10220344; DEG10390205;
DEG10340132; DEG10160089; DEG10230286; DEG10340070; DEG10400547; DEG10370126; DEG1001024
DEG10230049; DEG10210115; DEG10390019; DEG10240026; DEG10330294; DEG10280268; DEG1006023

[DEG10110050;](#)

[DEG10170216;](#)

DEG10280042; DEG10240032; DEG10060067; DEG10380067; DEG10020068; DEG10140279; DEG1037000
DEG10100608; DEG10170191; DEG10380106; DEG10010155; DEG10070190; DEG10210108; DEG1022017
DEG10350254; DEG10290157; DEG10300060; DEG10190035; DEG10160040; DEG10100455; DEG1003044
DEG10330038; DEG10120317; DEG10270516; DEG10010133; DEG10320037; DEG10050291; DEG1029015
DEG10290155; DEG10340119; DEG10270517; DEG10390001; DEG10130199; DEG10120046; DEG1001013
DEG10030734; DEG10100215; DEG10340077; DEG10330114; DEG10210103; DEG10050394; DEG1002008
DEG10340054; DEG10360165; DEG10030158; DEG10060277; DEG10350470; DEG10250595; DEG1005040

DEG10170087; DEG10100245; DEG10010233; DEG10400446; DEG10250286;

[DEG10220021;](#)

DEG10410550; DEG10030188; DEG10270161;

[DEG10060185;](#)

DEG10170233; DEG10410192; DEG10160083; DEG10350353; DEG10250510; DEG10070009; DEG1033008
DEG10380081; DEG10230046; DEG10030205; DEG10100488; DEG10290269; DEG10250593; DEG1040006
DEG10010135; DEG10060022; DEG10200003; DEG10130165; DEG10220382; DEG10360078; DEG1037020
DEG10280390; DEG10200022; DEG10410536;

[DEG10350155;](#)

DEG10350141; DEG10140158; DEG10060031; DEG10060030; DEG10180247;

DEG10270436; DEG10410115; DEG10050061; DEG10400633; DEG10140157; DEG10100383; DEG1006003
DEG10180187; DEG10050163;

DEG10130236; DEG10210186; DEG10170299; DEG10150018; DEG10250065; DEG10230088; DEG1016020
DEG10400250; DEG10060314;

DEG10020135; DEG10120128; DEG10230121; DEG10210088; DEG10400545; DEG10130253; DEG1006021
DEG10030734; DEG10100216; DEG10270457; DEG10340077; DEG10330114; DEG10180114; DEG1005039
DEG10070209; DEG10030067; DEG10380208; DEG10010072; DEG10080161; DEG10270600; DEG103303
DEG10130096; DEG10400041; DEG10380059; DEG10030494; DEG10130311; DEG10150286; DEG1020019
DEG10140033; DEG10010168; DEG10410311; DEG10350064;

DEG10340109; DEG10230120;

DEG10160189; DEG10400097; DEG10330191; DEG10280079; DEG10020142; DEG10080023;

DEG10180134; DEG10250527;

DEG10200177;

DEG10170188; DEG10340270; DEG10220170; DEG10080098; DEG10410256; DEG10370167; DEG1041023
DEG10230049; DEG10350369; DEG10170147; DEG10380142; DEG10330294; DEG10210115; DEG1028026

[DEG10350484;](#)

[DEG10320139;](#)

DEG10290275; DEG10030408; DEG10180329;

DEG10310012; DEG10120127; DEG10320030; DEG10180035; DEG10370055; DEG10290125;

DEG10030188; DEG10270161;

[DEG10180321;](#)

DEG10380239; DEG10250369; DEG10270341;

DEG10190187; DEG10220350; DEG10160230; DEG10280432; DEG10320252; DEG10130364; DEG1017023
DEG10140244; DEG10020092; DEG10060348; DEG10200181;

[DEG10310176;](#)

DEG10020060; DEG10180477; DEG10020295; DEG10050416;

DEG10240003; DEG10330331; DEG10060300; DEG10350179; DEG10010113; DEG10340404; DEG1040053

[DEG10050207;](#)

[DEG10070241;](#)

[DEG10050153;](#)

[DEG10050590;](#)

DEG10240075; DEG10270444; DEG10340344; DEG10130350; DEG10390169; DEG10220285; DEG1040002
DEG10030734; DEG10320118; DEG10330114; DEG10330295; DEG10100381; DEG10050394; DEG1035002
DEG10270081; DEG10250086; DEG10100060;
DEG10220354; DEG10350126; DEG10050077;

DEG10410383; DEG10280071;
DEG10110224; DEG10240131;
DEG10370073; DEG10170181; DEG10380072;
DEG10050100; DEG10400449; DEG10030542;
[DEG10270431;](#)
DEG10240051; DEG10290010; DEG10200410; DEG10350047;
DEG10020025; DEG10020076;
[DEG10350118;](#)
DEG10160166; DEG10170174; DEG10130437; DEG10240365; DEG10220460; DEG10100424; DEG1002009
DEG10370120; DEG10140152; DEG10390053; DEG10170304; DEG10220098; DEG10380127; DEG1034008
DEG10200194; DEG10030493; DEG10320204; DEG10160054; DEG10070210; DEG10180394; DEG1001007
DEG10060046; DEG10220298; DEG10340345; DEG10410283; DEG10300083; DEG10170083; DEG1014013
DEG10120310; DEG10170097;
DEG10060168; DEG10390221; DEG10310159; DEG10100040; DEG10170213; DEG10100375; DEG1002017
DEG10030595; DEG10100277; DEG10060164; DEG10010207; DEG10220197; DEG10400013; DEG1038010
DEG10390230; DEG10220393;
DEG10230179; DEG10340546; DEG10120279; DEG10100270; DEG10160196; DEG10180429; DEG1032023
[DEG10400645;](#)
DEG10030719; DEG10260104; DEG10280426;
DEG10100461; DEG10250563; DEG10050403; DEG10270521;
DEG10130279; DEG10030055; DEG10120082; DEG10200432; DEG10190289; DEG10010073; DEG1031008
DEG10240228; DEG10380213; DEG10320168; DEG10160084; DEG10150091; DEG10030387; DEG1033008
DEG10220128; DEG10400138; DEG10390076;
DEG10290307; DEG10240063; DEG10220205; DEG10360250; DEG10160002; DEG10120109; DEG1037013
DEG10020138; DEG10130060; DEG10160221; DEG10330224; DEG10120364; DEG10030136;
DEG10190183; DEG10130058; DEG10010136; DEG10350217; DEG10080305; DEG10370178; DEG1033022
DEG10050580; DEG10140109; DEG10140292;
DEG10240108; DEG10290366; DEG10330016; DEG10400420; DEG10200347; DEG10050407; DEG1019001
DEG10100464; DEG10060361; DEG10130447; DEG10380105; DEG10320214; DEG10330182; DEG1016018
[DEG10340104;](#)
DEG10390012; DEG10140153; DEG10400314; DEG10100183; DEG10200313; DEG10220030;
DEG10140010; DEG10100061; DEG10120319; DEG10050440; DEG10100453; DEG10270511; DEG1025065
DEG10230242; DEG10340451; DEG10220045;
DEG10350299; DEG10160052; DEG10030495; DEG10320207; DEG10230190; DEG10190151; DEG1024020
DEG10240108; DEG10290369; DEG10240105; DEG10130458; DEG10220249; DEG10400378; DEG1019000
DEG10270192; DEG10250198;
DEG10340195; DEG10390037; DEG10110045; DEG10180408; DEG10220068; DEG10110108;
DEG10270450; DEG10250480;
DEG10360198; DEG10030531; DEG10150038; DEG10060128; DEG10400569; DEG10050271; DEG1008026
[DEG10020299;](#)
[DEG10070107;](#)
DEG10240349; DEG10310124; DEG10360335; DEG10120008; DEG10230024; DEG10050355; DEG1017029
DEG10110186; DEG10290039; DEG10360197; DEG10120060; DEG10150039; DEG10030530; DEG1005027
[DEG10350155;](#)
[DEG10050371;](#)

DEG10020030; DEG10300079; DEG10050604; DEG10190180;
DEG10100254; DEG10120107; DEG10250307;
DEG10410169; DEG10350500; DEG10130392; DEG10170068; DEG10270248; DEG10350034;
DEG10390109; DEG10220188;
[DEG10080230;](#)
DEG10080322; DEG10150265; DEG10050013;
DEG10340360; DEG10050610; DEG10220143; DEG10390083; DEG10200118;
DEG10240326; DEG10410397; DEG10150044; DEG10120065; DEG10400563; DEG10360196; DEG1005027
DEG10110061; DEG10050085;
DEG10320089; DEG10250414; DEG10270494; DEG10250047; DEG10130148; DEG10190073; DEG1037005
DEG10220282; DEG10400340;
DEG10240325; DEG10130418; DEG10410395; DEG10180497; DEG10120067; DEG10360195; DEG104005
DEG10240028; DEG10130358; DEG10120234; DEG10150283; DEG10060065; DEG10400054; DEG1005013
DEG10310037; DEG10290301; DEG10280285; DEG10150059; DEG10320055; DEG10220178; DEG1034016
DEG10240324; DEG10170319; DEG10130417; DEG10010050; DEG10360194; DEG10400561; DEG102301
DEG10010142; DEG10080128; DEG10220323; DEG10060105;
DEG10160164; DEG10050571; DEG10410354; DEG10330167;
DEG10050485; DEG10070194;
DEG10380047; DEG10250484; DEG10270453; DEG10350099; DEG10410356; DEG10250483; DEG1005023
DEG10340394; DEG10170001; DEG10270001; DEG10050351; DEG10060380; DEG10020001; DEG1012000
DEG10170002; DEG10270002; DEG10380001; DEG10050350; DEG10120002; DEG10410002; DEG102000
DEG10340135; DEG10270003; DEG10020002; DEG10110201;
DEG10060003; DEG10390044; DEG10200196; DEG10290228; DEG10320187; DEG10170177; DEG1012012
DEG10390237; DEG10170315; DEG10240320; DEG10010054; DEG10180493; DEG10410390; DEG1006014
DEG10050478; DEG10410354;
DEG10170220; DEG10250200; DEG10050182; DEG10070073; DEG10140210; DEG10310029; DEG1002017
DEG10290052; DEG10170314; DEG10390236; DEG10130412; DEG10180492; DEG10010055; DEG1006014
DEG10400194; DEG10230099; DEG10400055; DEG10290344; DEG10190189; DEG10410533; DEG1032025
DEG10150316; DEG10300090;
DEG10340077; DEG10330114; DEG10330295; DEG10050394; DEG10020083; DEG10350028; DEG1005054
DEG10340109; DEG10230120;
DEG10350128; DEG10030096; DEG10030094; DEG10350187; DEG10180242; DEG10300013; DEG1035049
[DEG10050301;](#)
DEG10390189; DEG10220324;
DEG10170310; DEG10150051; DEG10290053; DEG10380018; DEG10060146; DEG10130411; DEG1036019
DEG10350112; DEG10050255; DEG10230271; DEG10260033; DEG10300077; DEG10340373; DEG1040016
DEG10170008; DEG10330337; DEG10010264; DEG10050567; DEG10230028; DEG10190283; DEG1040033
DEG10010265; DEG10030065; DEG10140197; DEG10060075; DEG10020006;
DEG10280019; DEG10030064; DEG10170019; DEG10050172; DEG10240259; DEG10340399; DEG1006007
DEG10330339; DEG10170018; DEG10020223; DEG10190284; DEG10120020; DEG10010268; DEG1007007
DEG10390231; DEG10290056; DEG10060148; DEG10100546; DEG10230142; DEG10250676; DEG1003051
DEG10280275; DEG10050260; DEG10180524; DEG10290385;
[DEG10030373;](#)
[DEG10290284;](#)

III.4.2. Conclusion, Chapter 4.

The genomic information was used with the aim of determining the conserved core proteome data of 13 strains of *Treponema pallidum*, together with the 3D structural information. Reverse vaccinology and Subtractive genomic based approach was used for the identification of vaccine and drug targets against the syphilis disease. The analysis of protein models used in this study was carried out very carefully. The data presented here can effectively contribute to future research for the development of drugs and vaccines. The measurement for the target selection in *T. pallidum* was kept strict, resulting in a small set of prioritized putative drug/vaccine targets. After the detailed structural comparison between host and pathogen proteins, we suggested 15 non host homologous proteins and 6 non-host homologous proteins for vaccine and drug targets, respectively. We expect that the *in silico* computational approaches adopted in this study might aid in the development of novel therapeutic targets against *T. pallidum* which can be used as candidate in future for the treatment of syphilis disease.

Chapter 5.

III.5.1. General Conclusion

A bacterium, *Treponema pallidum*, which is responsible for a sexually transmitted infection called syphilis, has affected humans throughout history, which is endemic in low-income countries and at low rates in middle-income and high-income countries. Every year, an estimated more than 6 million new cases of syphilis occur in the 15-49 age group globally, especially in men who have sex with men (MSM) and also increase the risk of HIV infection. Syphilis is responsible for more than 3 hundred thousand fatal and new-born mortality and also for more than 2 hundred thousand infants' risk of early death due to congenital syphilis. The disease occurs basically in three stages, primary secondary and tertiary with latent stage (early and late). The average incubation period of syphilis (from the beginning of exposure to the primary stage) is 3 weeks, but it can be long as a month and short to 9-10 days. It can also have possibilities to do not show any clinical manifestations and symptoms during the latent stages and may develop many asymptomatic infections and other complications. Available treatment with the antibiotic penicillin in the mid-twentieth century, infection rates got decreased dramatically. But unfortunately, the disease has been re-emerging globally in last few decades. The re-emergence of new cases of syphilis has concerned the scientific community to rethink the currently available treatments and to develop novel drugs and vaccine targets to cure the disease. Many research works are continuing to understand the biology of *T. pallidum* and host response to infection as part of efforts to develop new drugs and vaccines to stop syphilis. The availability of genomic data provides means to better understanding the molecular and genetic basis of virulence of this bacterium, enabling a detailed investigation of *T. pallidum*. In the long run, providing a new gateway for development and/or improvement of a potent vaccine.

In this study, the genomic information of *T. pallidum* was used with aim to perform pan-genome analysis based on subspecies level to distinguish the differences and presence or absence of genomic islands between venereal and non-venereal syphilis and also to calculate pan-genome, core genome and singletons data of subsp. *pallidum*, *pertenue* and *endemicum*.

Furthermore we extrapolated our work determining the conserved core proteome data of 13 strains of *T. pallidum*, together with the 3D structural information. we used reverse vaccinology and subtractive genomics and molecular modelling and docking based approach for the identification of vaccine and drug targets against the syphilis disease. The analysis of protein models used in this study was carried out very carefully. The data presented here can effectively contribute to future research for the development of drugs and vaccines. After the detailed structural comparison between host and pathogen proteins, we suggested that 15 non-host homologous proteins and 6 non-host homologous proteins as vaccine candidates and drug targets. We expect that the *in silico*

computational approaches adopted in this study might aid in the development of novel therapeutic targets against *T. pallidum* which can be used as candidate in future for the treatment of syphilis disease.

III.5.2. Future Prospective

The future works of this research works are:

- An immunoinformatics based approach for multi-epitope vaccine designing with our identified vaccine targets.
- Experimental validation of *in silico* identified drug molecules in *Treponema pallidum* infected model organism.
- In South America, syphilis is hyper endemic especially in Brazil among men who engage in sexual relations with men (MSM), male-to-female transgender women and in pregnant women. We are also working on identification and sequencing of *Treponema pallidum* in biological samples from syphilis patients in Brazil and comparative genomic analysis.
- Software for automatic analyses of vaccine candidates and drug targets.

IV. Appendix

A. Published, Accepted, Submitted Research articles and Genome assembly, annotation and submission

**Comparative genomics, *In-silico* Drug target identifications,
Homology Modelling, Molecular docking and Reverse vaccinology
(My contributions to these papers)**

Genomics is a broad discipline, which may be divided into three main areas; Structural genomics, Functional genomics & Comparative genomics. Structural genomics deals with the physical nature of genome. Its primary objective is to determine and analyze the genomic DNA sequence.

Functional genomics is concerned with the way genome functions. That is, it examines the transcript produced by genome and the collection of proteins they encode. The third and relatively new area of genomics is **comparative genomics** in which genome from different organisms are compared to look for significant similarity and difference. This helps in the identification of the important, conserved portion of genome and discriminate patterns in functions, regulations and therapeutic target identifications. After joining the LGCM as a Ph.D. student, I experienced working with different projects at the same time. It was good for my learning process to relate my previous experiences in the scientific field with the on-going projects in LGCM and Laboratório de Imunologia e Bioinformática (UFTM). I contributed as a structural biologist to give flow to these projects by applying protein homology modeling and molecular docking approaches. Simultaneously, I worked with the analyzing the data and drafting the manuscript.

Research



Cite this article: Vilela Rodrigues TC *et al.* 2019 Reverse vaccinology and subtractive genomics reveal new therapeutic targets against *Mycoplasma pneumoniae*: a causative agent of pneumonia. *R. Soc. open sci.* **6**: 190907. <http://dx.doi.org/10.1098/rsos.190907>

Received: 17 May 2019

Accepted: 4 July 2019

Subject Category:

Genetics and genomics

Subject Areas:

bioinformatics/genomics/immunology

Keywords:

vaccinology, *Mycoplasma pneumoniae*, bioinformatics, genomic, molecular docking, pneumonia

Author for correspondence:

Siomar de Castro Soares
e-mail: siomars@gmail.com

[†]These authors contributed equally to this study.

Electronic supplementary material is available online at <https://dx.doi.org/10.6084/m9.figshare.c.4593944>.

Reverse vaccinology and subtractive genomics reveal new therapeutic targets against *Mycoplasma pneumoniae*: a causative agent of pneumonia

Thaís Cristina Vilela Rodrigues^{1,†}, **Arun**

Kumar Jaiswal^{1,2,†}, Alissa de Sarom¹, Letícia de Castro Oliveira^{1,2}, Carlo José Freire Oliveira¹, Preetam Ghosh³, Sandeep Tiwari², Fábio Malcher Miranda², Leandro de Jesus Benevides⁴, Vasco Ariston de Carvalho Azevedo² and **Siomar de Castro Soares**¹

¹Department of Microbiology, Immunology and Parasitology, Federal University of Triângulo Mineiro, Minas Gerais, Brazil

²Department of Genetics, Ecology and Evolution, Federal University of Minas Gerais, Minas Gerais, Brazil

³Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

⁴Bioinformatics Laboratory - LABINFO, National Laboratory of Scientific Computation - LNCC/MCTI, Rio de Janeiro, Brazil

TCVR, 0000-0002-2048-5522; AdS, 0000-0002-3146-7784; LdCO, 0000-0002-2036-4456; PG, 0000-0003-3880-5886; ST, 0000-0002-8554-1660; FMM, 0000-0002-6823-5995

Pneumonia is an infectious disease caused by bacteria, viruses or fungi that results in millions of deaths globally. Despite the existence of prophylactic methods against some of the major pathogens of the disease, there is no efficient prophylaxis against atypical agents such as *Mycoplasma pneumoniae*, a bacterium associated with cases of community-acquired pneumonia. Because of the morphological peculiarity of *M. pneumoniae*, which leads to an increased resistance to antibiotics, studies that prospectively investigate the development of vaccines and drug targets appear to be one of the best ways forward. Hence, in this paper, bioinformatics tools were used



Acetate Kinase (AcK) is Essential for Microbial Growth and Betel-derived Compounds Potentially Target AcK, PhoP and MDR Proteins in *M. tuberculosis*, *V. cholerae* and Pathogenic *E. coli*: An *in silico* and *in vitro* Study



Sandeep Tiwari^{1,2,#}, Debmalya Barh^{1,2,#,*}, M. Imchen³, Eswar Rao³, Ranjith K. Kumavath³, S. Prabu Seenivasan⁴, Arun K. Jaiswal^{2,5}, Syed B. Jamal², Vanaja Kumar⁴, Preetam Ghosh⁶ and Vasco Azevedo²

¹Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India; ²PG Program in Bioinformatics (LGCM), Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil; ³Department of Genomic Sciences, School of Biological Sciences, Central University of Kerala, Kasaragod, India; ⁴Department of Bacteriology, National Institute for Research in Tuberculosis (ICMR) (Formerly Tuberculosis Research Centre), Chetpet, Chennai- 600031, India; ⁵Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, MG, Brazil; ⁶Department of Computer Science, Virginia Commonwealth University, Richmond, VA-23284, USA

Abstract: Background: *Mycobacterium tuberculosis*, *Vibrio cholerae*, and pathogenic *Escherichia coli* are global concerns for public health. The emergence of multi-drug resistant (MDR) strains of these pathogens is creating additional challenges in controlling infections caused by these deadly bacteria. Recently, we reported that Acetate kinase (AcK) could be a broad-spectrum novel target in several bacteria including these pathogens.

Methods: Here, using *in silico* and *in vitro* approaches we show that (i) AcK is an essential protein in pathogenic bacteria; (ii) natural compounds Chlorogenic acid and Pinoreosinol from *Piper betel* and Piperidine derivative compound 6-oxopiperidine-3-carboxylic acid inhibit the growth of pathogenic *E. coli* and *M. tuberculosis* by targeting AcK with equal or higher efficacy than the currently used antibiotics; (iii) molecular modeling and docking studies show interactions between inhibitors and AcK that correlate with the experimental results; (iv) these compounds are highly effective even on MDR strains of these pathogens; (v) further, the compounds may also target bacterial two-component system proteins that help bacteria in expressing the genes related to drug resistance and virulence; and (vi) finally, all the tested compounds are predicted to have drug-like properties.

Results and Conclusion: Suggesting that, these *Piper betel* derived compounds may be further tested for developing a novel class of broad-spectrum drugs against various common and MDR pathogens.

Keywords: Infectious disease, Multi-drug resistant, Natural compounds, *Piper betel*, Tuberculosis, ACK.

1. INTRODUCTION

M. tuberculosis, pathogenic *E. coli*, and *V. cholerae* are deadly human pathogenic bacteria that cause tuberculosis (TB), food poisoning, and cholera, respectively. Although, several drugs have been introduced to control these pathogens, infections from these bacteria frequently remain uncontrolled and epidemics are reported globally due to emerging multi-drug resistance (MDR) of these pathogens [1-3]. Hence, there is a need to develop promising next-generation drugs that can counter these ever-evolving pathogens.

In our previous reports, we showed that natural compounds from *Piper betel* are effective against these pathogens. Piperdardine inhibited pathogenic *E. coli* O157:H7 growth like ampicillin [4]. Piperdardine at 60 mM concentration exhibited a similar anti-*Vibrio* effect as 100 mg/ml of Chloramphenicol [5]. We also reported that Pinoreosinol, Guineensine, Dehydropiperonaline, Piperrolcin-B, Eugenyl acetate and Chlorogenic acid from *Piper betel* may have target specificity in *V. cholerae* [5]. Similarly, Acetate kinase (AcK) could be a common target in *C. pseudotuberculosis*, *Y. pestis*, *M. tuberculosis*, *C. diphtheriae*, *C. ulcerans*, and *E. coli* that may be targeted by *Piper betel* compounds [5].

For the generation of Adenosine triphosphate (ATP) from the excess of acetyl-CoA, the enzymes acetate kinase (AcK) and phosphotransacetylase (PTA) form an important

*Address correspondence to this author at the Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India; Tel./ Fax +91 944 955 0032; E-mail- dr.barh@gmail.com

[#]These authors contributed equally to this work.

ARTICLE HISTORY

Received: November 14, 2018

Revised: November 22, 2018

Accepted: December 14, 2018

DOI:

10.2174/1568026619666190121105851



CrossMark

Research



Cite this article: de Sarom A, Kumar Jaiswal A, Tiwari S, de Castro Oliveira L, Barh D, Azevedo V, Jose Oliveira C, de Castro Soares S. 2018 Putative vaccine candidates and drug targets identified by reverse vaccinology and subtractive genomics approaches to control *Haemophilus ducreyi*, the causative agent of chancroid. *J. R. Soc. Interface* **15**: 20180032. <http://dx.doi.org/10.1098/rsif.2018.0032>

Received: 12 January 2018

Accepted: 30 April 2018

Subject Category:

Life Sciences – Chemistry interface

Subject Areas:

bioinformatics

Keywords:

Haemophilus ducreyi, reverse vaccinology, vaccine candidates, drug targets, molecular docking, chancroid

Author for correspondence:

Siomar de Castro Soares

e-mail: siomars@gmail.com

Putative vaccine candidates and drug targets identified by reverse vaccinology and subtractive genomics approaches to control *Haemophilus ducreyi*, the causative agent of chancroid

Alissa de Sarom¹, Arun Kumar Jaiswal², Sandeep Tiwari²,
Leticia de Castro Oliveira², Debmalya Barh³, Vasco Azevedo²,
Carlo Jose Oliveira¹ and Siomar de Castro Soares^{1,2}

¹Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil

²Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

³Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India

AdS, 0000-0002-3146-7784; LdCO, 0000-0002-2036-4456

Chancroid is a sexually transmitted infection (STI) caused by the Gram-negative bacterium *Haemophilus ducreyi*. The control of chancroid is difficult and the only current available treatment is antibiotic therapy; however, antibiotic resistance has been reported in endemic areas. Owing to recent outbreaks of STIs worldwide, it is important to keep searching for new treatment strategies and preventive measures. Here, we applied reverse vaccinology and subtractive genomic approaches for the *in silico* prediction of potential vaccine and drug targets against 28 strains of *H. ducreyi*. We identified 847 non-host homologous proteins, being 332 exposed/secreted/membrane and 515 cytoplasmic proteins. We also checked their essentiality, functionality and virulence. Altogether, we predicted 13 candidate vaccine targets and three drug targets, where two vaccines (A01_1275, ABC transporter substrate-binding protein; and A01_0690, Probable transmembrane protein) and three drug targets (A01_0698, Purine nucleoside phosphorylase; A01_0702, Transcription termination factor; and A01_0677, Fructose-bisphosphate aldolase class II) are harboured by pathogenicity islands. Finally, we applied a molecular docking approach to analyse each drug target and selected ZINC77257029, ZINC43552589 and ZINC67912117 as promising molecules with favourable interactions with the target active site residues. Altogether, the targets identified here may be used in future strategies to control chancroid worldwide.

1. Introduction

Chancroid is a sexually transmitted infection (STI) caused by the bacterium *Haemophilus ducreyi*. It is endemic in poor countries of Asia, Africa and Latin America [1], indicating that there is a close relationship between the social economic situation and the incidence of chancroid in a given population.

The World Health Organization estimated the global prevalence of the disease in around 7 million, in the 1990s. However, it is difficult to assess the current epidemiology of chancroid because of syndromic management of genital ulcer diseases and the lack of reporting and diagnostic tools [2–4].

Haemophilus ducreyi is a fastidious non-motile Gram-negative cocobacillus, negative catalase, D-glucose fermenter, has fine pili, and does not form endospores [5]. Moreover, it does not have a known animal or environmental reservoir and infects preferably the human mucosal epithelium, but it may

Genomic Islands: an overview of current software tools and future improvements

Siomar de Castro Soares^{1,2*}, **Letícia de Castro Oliveira²**, **Arun Kumar Jaiswal²**,
Vasco Azevedo²

¹Department of Microbiology, Immunology and Parasitology, Institute of Biological and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba - MG, Brazil

²Laboratory of Cellular and Molecular Genetics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte - MG, Brazil

Summary

Microbes are highly diverse and widely distributed organisms. They account for ~60% of Earth's biomass and new predictions point for the existence of 10^{11} to 10^{12} species, which are constantly sharing genes through several different mechanisms. Genomic Islands (GI) are critical in this context, as they are large regions acquired through horizontal gene transfer. Also, they present common features like genomic signature deviation, transposase genes, flanking tRNAs and insertion sequences. GIs carry large numbers of genes related to specific lifestyle and are commonly classified in Pathogenicity, Resistance, Metabolic or Symbiotic Islands. With the advent of the next-generation sequencing technologies and the deluge of genomic data, many software tools have been developed that aim to tackle the problem of GI prediction and they are all based on the prediction of GI common features. However, there is still room for the development of new software tools that implements new approaches, such as, machine learning and pan-genomics based analyses. Finally, GIs will always hold a potential application in every newly invented genomic approach as they are directly responsible for much of the genomic plasticity of bacteria.

1 Living in the "Age of Bacteria"

Stephen Jay Gould, a renowned paleontologist, once said, "We live now in the 'Age of Bacteria.' Our planet has always been in the 'Age of Bacteria' ever since the first fossils, bacteria, of course, were entombed in rocks more than three and a half billion years ago" [1]. Microbes are highly diverse organisms responsible for approximately 60% of the Earth's biomass. They were the first organisms on Earth, they are distributed worldwide, from volcanos to salt water, and they play a pivotal role in several medical, biotechnological and industrial applications. Although their importance is widely known, less than 1% of the previously estimated 2-3 billion microbial species are identified so far [2]. Much of this lack of knowledge on microbes is due to the use of culture-dependent identification and characterization of microbes. Microbiological culture media are usually intended for selective growing and, thus, the microorganisms recovered using these methods are not representative of the microbial community inside the sample [3]. However, with the advent of the next-generation sequencing (NGS) technologies and the widespread of metagenomics methodologies, scientists are now able to determine the complete gene set off an entire community, transcending the idea of a single species genomics to a complete view of the

*To whom correspondance should be addressed. Email: siomars@gmail.com




Multi-epitope based vaccine against Yellow fever virus applying immunoinformatics approaches

Stephane Fraga de Oliveira Tosta, Mariana Santana Passos, Rodrigo Kato, Álvaro Salgado, Joilson Xavier, Arun Kumar Jaiswal, Siomar C. Soares, Vasco Azevedo, Marta Giovanetti, Sandeep Tiwari & Luiz Carlos Junior Alcantara



To cite this article: Stephane Fraga de Oliveira Tosta, Mariana Santana Passos, Rodrigo Kato, Álvaro Salgado, Joilson Xavier, Arun Kumar Jaiswal, Siomar C. Soares, Vasco Azevedo, Marta Giovanetti, Sandeep Tiwari & Luiz Carlos Junior Alcantara (2019): Multi-epitope based vaccine against Yellow fever virus applying immunoinformatics approaches, Journal of Biomolecular Structure and Dynamics, DOI: [10.1080/07391102.2019.1707120](https://doi.org/10.1080/07391102.2019.1707120)

To link to this article: <https://doi.org/10.1080/07391102.2019.1707120>

 View supplementary material 

 Accepted author version posted online: 19 Dec 2019.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 

Multi-epitope based vaccine against Yellow fever virus

applying immunoinformatics approaches

Stephane Fraga de Oliveira Tosta¹, Mariana Santana Passos², Rodrigo Kato¹, Álvaro Salgado¹, Joilson Xavier², Arun Kumar Jaiswal^{1,3}, Siomar C. Soares³, Vasco Azevedo¹, Marta Giovanetti^{4*}, Sandeep Tiwari^{1*}, Luiz Carlos Junior Alcantara^{4*}

¹Postgraduate Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte 31270-901, MG, Brazil

²Department of Genetics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte 31270-901, MG, Brazil

³Department of Immunology, Microbiology and Parasitology, Institute of Biological and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba 38025-180, MG, Brazil

⁴Laboratório de Flavivírus, Instituto Oswaldo Cruz, FIOCRUZ, Manguinhos, Rio de Janeiro 21040-900 Rio de Janeiro, Brazil.

***Corresponding Authors:** Sandeep Tiwari (sandip_sbtbi@yahoo.com), Marta Giovanetti (giovanetti.marta@gmail.com) and Luiz Carlos Junior Alcantara (luiz.alcantara@ioc.fiocruz.br)

Abstract

Yellow fever disease is considered a re-emerging major health issue which has caused recent outbreaks with a high number of deaths. Tropical countries, mainly African and South American are the most affected by Yellow fever outbreaks. Despite the availability of an attenuated vaccine, its use is limited for some groups such as pregnant and nursing women, immunocompromised and immunosuppressed patients, elderly people >65 years, infants < 6 months and patients with biological disorders like thymus disorders. In order to achieve new preventive measures, we applied immunoinformatics approaches to develop a multi-epitope based subunit vaccine for Yellow fever virus. Different epitopes, related to humoral and cell-mediated immunity, were predicted for complete polyproteins of two Yellow fever strains (Asibi and 17D vaccine). Those epitopes common for both strains were mapped into a set of 137 sequences of Yellow fever virus, including 77 sequences from a recent outbreak at the state of Minas Gerais, southeast Brazil. Therefore, the present work uses robust bioinformatics approaches for the identification of a multi-epitope vaccine against the Yellow fever virus. Our results indicate that the identified multi-epitope vaccine might stimulate humoral and cellular immune responses and could be a potential vaccine candidate against Yellow fever virus infection. Hence, it should be subjected to further experimental validations.

Keywords: Immunoinformatics, flavivirus, multi-epitope vaccine, vaccine, and Yellow fever virus.

Abbreviations: AEs – adverse events, C – Capsid, E – Envelope, NHPs – Non-human primates, NS – Non-structural, PAHO – Pan American Health Organization, prM- Premembrane, SAEs – serious adverse events, YFV – Yellow fever virus.

An *In Silico* identification of common putative vaccine candidates and drug targets against *Mycoplasma genitalium*, the sexually transmitted causative agent of pelvic inflammatory disease (PID) and several inflammatory reproductive tract syndromes

Wylerson G. Nogueira ^{a,*}, **Arun Kumar Jaiswal ^{a,*}**, Sandeep Tiwari, Syed Jamal Bacha, Rommel T. J. Ramos, Vasco Azevedo, **Siomar C. Soares**

Submitted: Genomics Journal ScienceDirect

Abstract

Mycoplasma genitalium is a sexually transmitted pathogen characterized as a pleomorphic, flask shaped, slow growing and obligate intracellular bacterium. It is one of the STI (sexually transmitted infections) pathogens associated with non-gonococcal urethritis in men and several inflammatory reproductive tract syndromes in women such as cervicitis, pelvic inflammatory disease (PID) and infertility. Here, we applied reverse vaccinology and subtractive genomic approaches for the *in silico* prediction of potential vaccine and drug targets against five strains of *M. genitalium*. We identified 403 genes shared by all five strains, from which 104 non-host homologous proteins, being 44 exposed/secreted/membrane and 60 cytoplasmic proteins. We also checked their essentiality, functionality and structure-based binding affinity. Altogether, we predicted 19 candidate vaccine targets – an ABC transporter permease; a PTS glucose EIICBA component; an adhesion P1 protein; a diacylglycerol transferase; an IgG blocking protein M; and 14 hypothetical proteins – and 7 drug targets – Type I restriction-modification protein (WP_009885596.1); Akyl hydro peroxide reductase, Ahp (WP_009885605.1); ribosome-binding factor A, RbfA (WP_009885829.1); 50S ribosomal protein L32 (WP_009885820.1); Class Ib ribonucleoside diphosphate reductase, NrdI; a DUF3217 domain containing protein (WP_009885939.1); and a hypothetical protein (WP_009885876.1). Furthermore, we performed a molecular docking analysis for each drug target against a 5008 antimicrobial natural-compounds database retrieved from Zinc database and selected ZINC08636510, ZINC04235924, ZINC15709489, ZINC04237087, ZINC04236001 and ZINC35415766 as promising molecules with favorable interactions with the target's active site residues. Finally, we predicted 14 potential vaccine targets that have not yet been described as on its antigenic properties or biological functions, and may be used in further studies as novel vaccine strategies against the pathogen. Concerning our predicted drug targets, no DUF3217 domain containing proteins regarding their potential antibiotic activity, strategical binding sites for compounds or signed as drug targets could be found in the literature. Hence, our study was the first to reveal both predicted hypothetical protein (WP_009885876.1) and DUF3217 domain containing

protein (WP_009885939.1) as novel putative drug targets against *Mycoplasma genitalium*. Altogether, both vaccine candidates and drug targets identified here may contribute in the future development of therapeutic strategies to control the spread of *M. genitalium* worldwide.

In silico* identification of new targets for diagnosis, vaccine and drug candidates against *Trypanosoma cruzi

Rafael Obata Trevisan, Malú Mateus Santos, Chamberttan Souza Desidério, Leandro Gomes Alves, Thiago de Jesus Sousa, Letícia de Castro Oliveira, [Arun Kumar Jaiswal](#), Sandeep Tiwari, Juliana Cristina Costa-Madeira, Lúcio Roberto Cançado Castellano, Marcos Vinicius Da Silva, Virmondes Rodrigues Junior, Carlo José Freire de Oliveira, [Siomar de Castro Soares](#).

Submitted: Disease Markers Hindawi, November 2019.

Abstract

Chagas disease is a neglected tropical disease caused by the parasite *Trypanosoma cruzi*. Despite the efforts and distinct methodologies, search of antigens for diagnosis, vaccine and drug targets for the disease are still needed. The aim of the present study was to identify possible antigens that could be used for diagnosis, vaccine and drugs targets against the *T. cruzi* using Reverse Vaccinology and Molecular Docking. The genomes of 28 *T. cruzi* strains available in the GeneBank (NCBI) were used to obtain the genomic core. Then, the subtractive genomics was carried out to identify non-homologous genes to the host in the core. A total of 2630 conserved proteins in 28 strains of *T. cruzi* were predicted using Orthofinder and Diamond software, in which 515 showed no homology to the human host. These proteins were evaluated for its subcellular localization, from which 214 are cytoplasmic, and 117 are secreted or present in the plasma membrane. To identify the antigens for diagnosis and vaccine targets, we used the VAXIJEN software, and it was selected 14 non-homologous proteins showing high binding efficiency with MHC I and MHC II with potential for *in vitro* and *in vivo* tests. When these 14 non-homologous molecules were compared against other trypanosomatids it was found that the Retrotransposon Hot Spot (RHS) Protein is specific only for *T. cruzi* parasite suggesting it could be used for Chagas diagnosis. Such 14 proteins were analyzed using IEDB software to predict their epitopes in both B and T lymphocytes. Furthermore, molecular docking analysis was performed using the software MHOLLline. As a result, we identified 6 possible *T. cruzi* drug targets that could interact with 4 compounds already known as antiparasitic activities. These 14 proteins targets, along with 6 potential drugs' candidates can be further validated in future studies, *in vivo*, regarding Chagas disease.

Design of a broad-spectrum candidate multi-epitope vaccine against Diphtheria: An integrative immunoinformatics approach

Mariana Santana, Stephane Tosta, Rodrigo Kato, [Arun Kumar Jaiswal](#), [Siomar C. Soares](#), Luiz Carlos Junior Alcantara, Preetam Ghosh, Debmalya Barh, Vasco Azevedo, Anderson Miyoshi, Sandeep Tiwari

Submitted: MDPI, August 2019

Abstract

Corynebacterium diphtheriae is the etiological agent of diphtheria, a life-threatening disease that was considered as main cause of child death; however, the development of the diphtheria vaccine ameliorates the problematic. The vaccine comprises a detoxified form of the diphtheria toxin absorbed into an alum adjuvant. Despite the high successful rate of the vaccine, outbreaks and isolated cases are still reported worldwide, especially in developing countries. This may be linked with poor vaccination programs, the emergence of multidrug resistant and non-toxigenic strains, and socioeconomic issues. The World Health Organization in association with main stakeholders manage programs to improve vaccine coverage and diphtheria awareness. Nonetheless, a few countries still hold a coverage below 50% (children immunized with three doses of the vaccine), which could be associated with the underprivileged sanitation and health infrastructure conditions, and lack of trained medical personal. Bearing all, the development of a cutting-edge treatment could overcome the difficulties and decrease the number of *C. diphtheriae* infections. Peptide-based vaccine are customized hence can target different strains or microorganism life cycle (multi-epitope), reproducible, fast, cost effective, stable under simple storage conditions (i.e. generally do not require ‘cold chain’), and removes problems with contamination, autoimmune and non-allergenic responses, but the use of adjuvant/delivery system is recommended to induce a proficient immune reaction. Thus, we selected seven *C. diphtheriae* proteins relevant for the bacterium colonization and virulence (i.e. DIP0222, DIP0411, DIP0733, DIP1084, DIP1281, DIP2193, and DIP2379) and the common predicted MHC-I and MHC-II epitopes were used to construct the diphtheria multi-epitope vaccine. The vaccine was predicted as stable, with non-allergenic and antigenic properties, and predicted B cell and IFN-gamma inducing epitopes. Finally, the tertiary multi-epitope vaccine structure and the interaction with Toll-receptor 2 were predicted and then the best docking structure was considered stable with favorable interactions. The results indicated that the designed diphtheria multi-epitope vaccine can stimulate a humoral and cellular response and would be a considerable advantage against this deadly pathogen.

A reverse vaccinology and immunoinformatics based approach for multi-epitope vaccine against *Clostridioides difficile* infection.

Mariana Santana, [Arun Kumar Jaiswal](#), Stephane Fraga de Oliveira Tosta, Rodrigo Kato, [Siomar C. Soares](#), Luiz Carlos Junior Alcantara, Anderson Miyoshi, Vasco Azevedo, Sandeep Tiwari*

Submitted: PeerJ, Organic Chemistry, November 2019.

Abstract

Clostridioides difficile is a potentially life-threatening bacillus that causes hospital and community-acquired gastrointestinal illness. *C. difficile* infection (CDI) clinical symptoms range from diarrhea and vomiting to toxic megacolon and pseudo membrane colitis. The clostridial toxin A (TcdA or ToxA) and the clostridial toxin B (TcdB or ToxB) are *C. difficile* primary mediators of inflammation that triggers host cellular response. Over the past few years, *C. difficile* has become one of the major causes of nosocomial infections and the continual/misuse use of antibiotics, especially during hospitalization, increases the risk of CDI. The antibiotic resistance problematic and the growing number of CDI outbreaks and asymptomatic patients expose the need for a valid and acceptable alternative treatment for *C. difficile*. Considering these, we used 53 complete genomes of *C. difficile* for comparison taking *C. difficile* ATCC 9689/DSM 1296 as the reference genome to predict putative vaccine and drug targets against *C. difficile* using reverse vaccinology and subtractive genomics. After the detailed in silico analysis, four non-host homologous protein drug targets were identified. The identified vaccine targets were analyzed for multi-epitope based vaccine design against *C. difficile*. This study will facilitate future in vitro and in vivo tests for the production of a peptide vaccine, drugs and prophylactic targets against a bacterium with high clinical relevance.

(My contributions to the Genome assembly, Annotation and submission)

Our research groups have an extensive involvement on ongoing genomics projects nationally and internationally. I have actively participated in these genomics projects and got an opportunity to enhance my knowledge of learning assembly and annotation of genomes. Working with genomics projects, I got familiar with different Next Generation Sequencing platforms and its application in various aspects of biological analysis.

We sequenced multi drug-resistant *Klebsiella pneumoniae* strain B31 (**GenBank under accession number CP035929**) from Brazil. In work entitled “**Whole-genome analyses with a novel multi drug resistant *Klebsiella pneumoniae* strain from Brazil**”.

Our group is also working on Genome assembly, annotation and submission of 69 Brazilian isolates of *Aeromonas* from aquatic and environment. The objective of this project is to carry out the characterization of bacterial pathogens obtained from aquatic organisms and their environment, in order to determine the virulence profile and sensitivity to antimicrobial drug patterns.

I am responsible for these two isolates *Aeromonas hydrophila* strain 10M (**Bio project number- PRJNA590950**) and *Aeromonas dhakensis* strain 26M (**Bio project number- PRJNA590952**).

Comparative genomics with a multidrug-resistant *Klebsiella pneumoniae* isolate reveals the panorama of unexplored diversity in Northeast Brazil

Rodrigo Profeta, Nbia Seyffert, Sandeep Tiwari, Marcus V. C. Viana, **Arun Kumar Jaiswal**, Ana Carolina Caetano, Daniel Henrique Bcker, Luciana Tavares Oliveira, Roselane Santos, Alfonso Gala-Garcia, Rodrigo B. Kato, Francine F. Padilha, Isabel B. Lima-Verde, Preetam Ghosh, Debmalya Barh, Aristteles Ges-Neto, Henrique C. P. Figueiredo, **Siomar C. Soares**, Bertram Brenig, Pablo I. P. Ramos, Vasco Azevedo⁺, and Thiago L. P. Castro^{*,+}

Submitted: Genomics, Journal-Elsevier.

Abstract

Klebsiella pneumoniae is one of the most important human pathogens, emerging as an agent of severe community infections. A comprehensive analysis of the genome of this microorganism could provide a better understanding of the molecular basis of its virulence and pathogenesis, contributing to new forms of disease control. In this study, we sequenced the genome of a multidrug resistant *K. pneumoniae* strain, herein named B31, isolated from a patient in Brazil. The genome sequencing of B31 resulted in a 5.27-Mb sized chromosome, one multireplicon plasmid (IncFIIk/ IncFIBk), and an IncII replicon likely to belong to another plasmid. The characterization of its genome confirmed that B31 is a multi-drug resistant (MDR) strain and belongs to the sequence type ST15, reported in Brazil for the first time in the present work. Comparative genomic analyses were conducted including B31 and 172 *K. pneumoniae* genomes from different parts of the world. We provide relevant information on the antibiotic resistance profiles and distribution of virulence-associated genetic elements in B31 and other strains.

IV. Appendix

B. Book Chapters

(My contributions to the Book Chapters)

Our research group has been invited a number of times to write book chapters for graduates and masters students. I actively participated in this work and collected data from authentic sources; finally, I have made a manuscript of the assigned theme. This activity helped me in improving data mining and manuscript writing skills. Furthermore, it was really useful in understanding the research work going on in different research areas around the world. Here I have mentioned the booked chapters published with different groups of publishers.

CHAPTER 12

Pan-genomics of fungi and its applications

Rodrigo Bentes Kato^{a,*}, Arun Kumar Jaiswal^{b,*}, Sandeep Tiwari^b,
Debmalya Barh^c, Vasco Azevedo^b, Aristóteles Góes-Neto^a

^aMolecular and Computational Biology of Fungi Laboratory, Department of Microbiology, Institute of Biological Sciences (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

^bPG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

^cCentre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Purba Medinipur, India

1 Introduction

Fungi are an evolutionary lineage within Opisthokonta, comprising one of the largest and most diverse groups of Eukarya on planet Earth [1]. Their multicellular non-motile bodies (mycelia) are constructed of apically growing, walled, and tubular cells (hyphae) or can be unicellular in which each adult individual is a single cell. These two distinct morphological groups are the so-called mycelial (or filamentous) fungi and yeasts, respectively [2]. The fungi are eukaryotic and chemoheterotrophic organisms, exhibiting osmotrophic nutrition with partial external digestion [3].

Fungi play a key role in the global carbon cycle especially in terrestrial biomes [4], and survive by using three basic trophic modes: (i) as saprotrophs, breaking down dead organic matter, (ii) as parasitic (and pathogens), or (iii) mutualistic symbionts with other living organisms [5]. Fungi and their by-products have a great economic importance. Fungi can be used to produce fermented food (e.g., beers, wines, breads, and cheeses) [6], primary and secondary metabolites (e.g., ethanol, hydrolases, and oxidoreductases enzymes, organic acids, and many vitamins, hypocholesterolemic, antineoplastics) [7], and inoculants and biocides (for mycorrhization and as biological control agents) [8]. Furthermore, fungi can be used for bioremediation of solid residues, effluents, and gaseous emissions [9], as well as to produce new biomaterials, such as mycocomposites [10] and nanomycomaterials [11].

Fungi constitute one of the major clades of organisms with approximately 145,000 species already described [12], however, there must be many more species since estimates suggest the existence of 5.1 million species, despite only about 10% were described [13]. Although there is still no consensus in the hierarchical taxonomical classification of the

* These authors contributed equally to this work.

main groups inside of Kingdom Fungi since many recent classification schemes proposed different number of phyla [14–16], the great majority of the species are in the well-established and consensual phyla Ascomycota and Basidiomycota, which form the sub-kingdom Dikarya [17].

With the advancement and the low cost of high-throughput sequencing technology, these days leads in excess of 250,000 genome projects registered at the Genome Online Database (GOLD) (<https://gold.jgi.doe.gov/measurements>) [18]. These endeavors have made big change in the study of fungal genes and genomic association. The fungal genomic datasets can be exploited for adaptive and environmental behavior study. Substantial genomics and transcriptomics dataset of fungi have empowered the utilization of novel strategies and molecular evolution studies of fungi [19,20]. The combinations of experimental and computational methods have a great potential for point-by-point investigation of the fungal evolution and its biology [20].

The pan-genomics is a comparative genomics-based methodology that identifies the core and the dispensable genomes. The dispensable genome is composed of genes that present in some but not in all the strains studied, as well as the strain-specific genes. The dispensable genome helps to its fundamental way of life but rather present particular points of interest including antifungal resistance, niche adaptation, and the capacity to colonize new hosts [21]. As gene content and genome copy number can vary in distinct populations of a single species, the inventory of the variation at genomic level in different isolates is crucial to characterize the complete set of genes (core and accessory) that exists in a fungal species [20]. In this chapter, our work aimed to perform an extensive literature review and meta-analysis of this customized database in order to depict the state of the art of fungal pan-genomics.

2 Application of pan-genomics of fungi based on meta-analysis

The metadata related to genomics, comparative genomics and pan-genomics analyses on fungi, were mined through the literature. The NCBI (National Center for Biotechnology Information) genome and JGI (Joint Genome Institute) genome databases were also used for the search of data related to these areas. Thereafter, abstract and full-text level manual curations were performed. For the description of the data, histograms, and pie charts were constructed using the R version 3.5.0 software [22] and the ggplot2 version 2.2.1 package [23]. The localization on global map was done using the GPS Visualizer tool [24]. As a result, among the obtained 159 articles, only the most cited articles were considered for manual text curation. From the metadata from the databases, we found 97 species and 16.5% of these species has more than 16 distinct isolates fully sequenced. Then, we used this threshold (~16%) to our analyses. The obtained metadata related to fungal pan-genomics from published articles were divided in four groups.

- (1) *Technological*: This group contains fungal isolates that are used to produce several products related to pharmaceuticals, food, dairy products, and alcohol industries. In our search, we found many papers using fungi to produce foods as fermented products and alcohol beverages.
- (2) *Environmental*: It is a collection of published research results on the study of fungal diversity in the natural environment. We found some works that use fungi in biological process of biodegradation and biological treatment of lagoons or rivers. Industry and biodegradation areas have interesting in fungi that act in enzyme activities. These fungi are very important in agricultural and ecological contexts because they maintain the balance of the environment decomposing plant debris, degrade toxic substances, help plants to grow, and protect themselves against enemies.
- (3) *Host pathogen*: In this group, we discussed some studies on the fungal diversity in the host-pathogen interactions. Some of these fungi contribute to pharmacy industry to cure some disease.
- (4) *Laboratory*: some fungi were comparable with model, laboratory strains, and then we created this group.

Fig. 1 shows the frequency of number of works in each group. We found that most of the works were related to the technological importance fungal group (~61%). The reason behind the several studies in this area may be the presence of many industries, which invest large amount of money to its research and development. On the other hand, we found that host-pathogen related fungal research were the second bigger group with around 30%, mainly supported by agribusiness.

This metadata analysis was done to show the impact of pan-genomics in comparative fungal genomics. Furthermore, we found 1567 species related to 12 genera of fungi, and

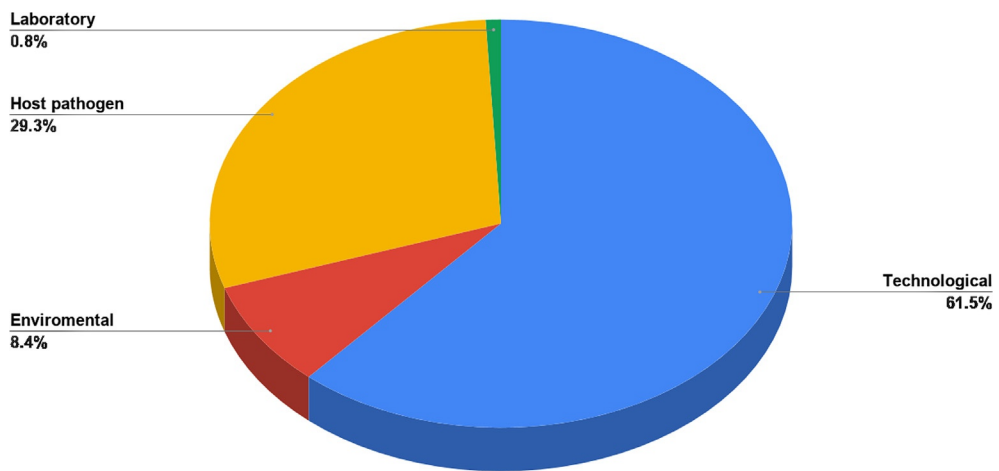


Fig. 1 The pie chart demonstrates the frequency of the studied fungi from each group.

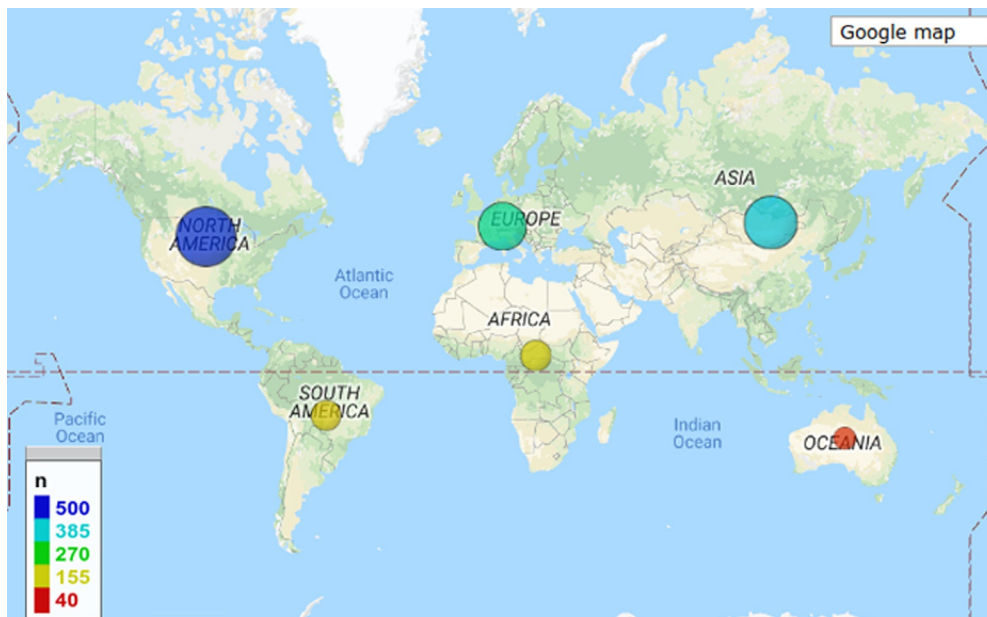


Fig. 2 The image shows the geographical a distribution of 1567 species of 12 different genera of fungal isolates.

they were distributed globally, where n means number of isolates for each country (Fig. 2).

Among these 12 genera we found that *Saccharomyces cerevisiae* is the most frequently (Fig. 3) studied genus. Around 83% of researches have been related to *S. cerevisiae* in the last 2 years of published research article [25].

2.1 Application of pan-genomics on advantageous fungus

The set of complete genes in all the strains of a specific species is known as pan-genome [20]. The genus *Saccharomyces* is among the most important and broadly studied model eukaryotic organisms. The fermented beverages production commonly used *S. cerevisiae* yeasts, dates at least as back 7000 BC, in china [26]. In order to comprehend the significance of selection during domestication and understand the levels of genetic diversity among wine yeasts, a number of pan-genome analyses have been done using commercial wine yeasts and industrial yeasts. A set of 83 strains of *S. cerevisiae* was used for the pan-genome analysis to identify the copy number variations in this yeast distributed in different industrial environments [26].

Another comparative work of 43 strains of *S. cerevisiae* isolated from fermenting grape was used to analyze genome renewal, and they propose that natural wine yeast strains can undergo such modifications and, thereby, change a multiple heterozygote into

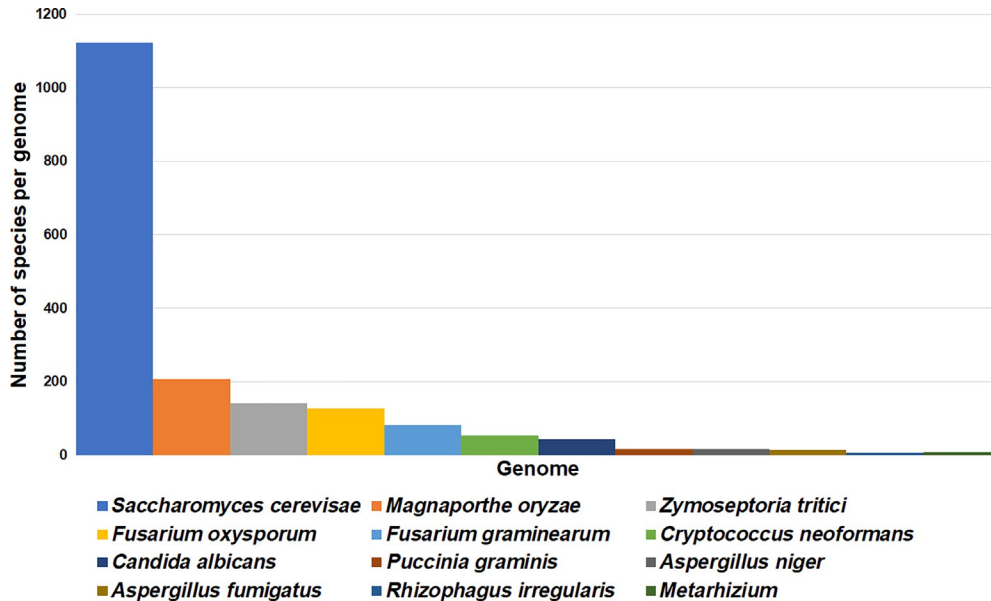


Fig. 3 The figure represents the most frequently studied fungi among the 12 genera.

completely homozygous diploids, some of which may replace the original heterozygous diploid [27]. The pan-genome studies of very polymorphic eukaryotic pathogens utilizing the accessory genome gives a better understanding for adaptive evolution. The genomics study of this yeast has enhanced our understanding of the evolutionary dynamics of natural populations when comparing with the domesticated strains, during infections, and during laboratory experiments [28]. Apart from *S. cerevisiae* [29], population genomic studies have also characterized the metabolic, genetic, and biogeographic diversity of *Saccharomyces paradoxus* [30], *Saccharomyces kudriavzevii* [31], and *Saccharomyces uvarum* [32]. As well as all the organisms, yeast genome sequences largely describe their genetic makeups; however, the comparative genomic studies have given better shape to the historical and genetic processes in their evolution [28].

2.2 Application of pan-genomics in disadvantageous fungus

Effect of chromosomal rearrangements on genes can lead to functional variation between individuals and influenced the expression of phenotypic attributes [33]. The inter and intraspecific structural variation among genomes of fungi has already been reported [34]. This structural variation among the pathogens can affect their host range. For instance, in the fungus *Melanopsichium pennsylvanicum*, gene loss are responsible for the hosts jump from dicotyledonous to monocotyledon plant hosts [33,35]. Unexpected

number of fungal and fungal-like diseases have been recently afflicted animals and plants, and some of them are the most severe die-offs and extinctions ever witnessed in wild species, and are a peril for food security [20]. Emerging infectious diseases (EIDs) brought about by fungi are progressively perceived as exhibiting a danger to food security around the world [36]. Until date, several fungal genomes are responsible to accurate and complete genome assemblies using long-read sequencing technologies [37,38]. Various symbiotic interactions have been recorded between insects and fungi. Although the genomics has already been elucidated in many fungi that expanded our knowledge on this group, there is still much to explore the genomic features of the insect-commensal relationships [39]. *Zymoseptoria tritici* is a pathogen of wheat causing Septoria tritici blotch, and a recently published work using pan-genome analysis of this pathogen identified that host specialization has evolved by gene deletions and chromosomal rearrangements. In this aforementioned study, the authors used five isolates for the pan-genome analysis and 15,749, 9,149, and 6600 nonredundant proteins were identified as pan-genome, core, and accessory genome, respectively [33] (Table 1).

Table 1 Pan-genomics studies on different fungi

Fungi	Importance	Comparative Genomics	Strains/ isolates	Reference
<i>Saccharomyces cerevisiae</i>	Industrial important	1. Pangenome Analysis of <i>Saccharomyces cerevisiae</i>	25	[40]
		2. Report of the whole-genome sequencing and phenotyping of 1011 <i>Saccharomyces cerevisiae</i> isolates	1011	[25]
<i>Rhizophagus irregularis</i>	Plant pathogen	Genome assembly and gene annotation of the model strain <i>Rhizophagus irregularis</i> DAOM197198, and gene comparison with five different isolates of <i>Rhizophagus irregularis</i>	6	[41]
<i>Puccinia graminis</i> <i>f. sp. tritici</i>	Plant pathogen	Comparative genomics of Australian isolates of the wheat stem rust pathogen <i>Puccinia graminis f. sp. tritici</i> and draft genome was built for a founder Australian <i>Pgt</i> isolate	16	[42]

Table 1 Pan-genomics studies on different fungi—cont'd

Fungi	Importance	Comparative Genomics	Strains/ isolates	Reference
<i>Zyloseptoria tritici</i>	Plant pathogen	Pangenome analysis of <i>Zyloseptoria tritici</i>	123	[33, 38]
<i>Metarhizium spp.</i>	Insect pathogen	Pangenome analysis for <i>Metarhizium spp.</i>	7	[43]
<i>Coccidioides posadasii</i> , <i>Coccidioides immitis</i> and other fungus of order Onygenales	Human fungal pathogen	Genome sequencing and comparison of the primary human pathogens <i>C. immitis</i> and <i>C. posadasii</i>	17	[44]
<i>Fusarium graminearum</i>	Cereal pathogen	Sequencing of genomes of 60 diverse <i>F. graminearum</i> isolates from North America, and also the assembly of the first pan-genome for <i>F. graminearum</i> to clarify population-level differences in gene content potentially contributing to pathogen diversity.	70	[45]
<i>Fusarium meridionale</i> / <i>Fusarium Asiaticum</i> / <i>Fusarium graminearum</i>	Plant/Cereal pathogen	Genomic comparison and gene content analysis of six newly isolates from the species complex, including the first available genomes of <i>F. asiaticum</i> and <i>F. meridionale</i> , with four other genomes	10	[46]

3 Conclusions and future prospective

Although people frequently demonstrate a nonmycophilic or even mycophobic relation with fungi, these group of organisms are vital on many aspects of human life, including medicine, food, and farming, and also play key roles in nature, such as in the carbon biogeochemical cycle. The comparative genomics approach based on sequence similarity with statistical analysis helps in identifying the essential genomic content common among all fungal isolates of a same species as well as the subset of genes encoding novel functions as variable genome. Biotechnology consists of the use of organisms for the development

of processes and products of economic or social interest. It is recognized as one of the technologies for the 21st century with higher potential impact on global problems (diseases, nutrition, and environmental pollution) and sustainable industrial development (use of renewable resources, “green technology,” and reduction of global warming). Based on the search and discovery of industrially exploitable biological resources, the scientific and technological advances achieved by fungal pan-genomics studies in recent years have revolutionized traditional approaches to the exploitation of biological resources for biotechnology.

References

- [1] F. Badotti, F.S. de Oliveira, C.F. Garcia, A.B. Vaz, P.L. Fonseca, L.A. Nahum, et al., Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi), *BMC Microbiol.* 17 (1) (2017) 42.
- [2] D. Moore, 21st Century Guidebook to Fungi, *Q. Rev. Biol.* 87 (4) (2012) 396.
- [3] Sarah C. Watkinson NPM, Lynne Boddy. *The Fungi*. San Diego, Elsevier Science Publishing Co Inc.
- [4] G.M. Gadd, *Fungi in Biogeochemical Cycles*, CAB International, 2006.
- [5] N.H. Nguyen, Z. Song, S.T. Bates, S. Branco, L. Tedersoo, J. Menke, et al., FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild, *Fungal Ecol.* 20 (2016) 241–248.
- [6] M. Hofrichter, *The Mycota: A Comprehensive Treatise on Fungi as Experimental Systems for Basic and Applied Research*, second ed., Springer, 2010.
- [7] *Fungal Biomolecules: Sources, Applications and Recent Developments*, Wiley-Blackwell, 2015.
- [8] T.M. Butt, C. Jackson, N. Magan (Eds.), *Fungi as Biocontrol Agents: Progress Problems and Potential*, CABI, 2001.
- [9] H. Singh, *Mycoremediation*, Wiley, 2006.
- [10] C. Girometta, A. Picco, R. Baiguera, D. Dondi, S. Babbini, M. Cartabia, et al., Physico-mechanical and thermodynamic properties of mycelium-based biocomposites: A review, *Sustainability* 11 (1) (2019).
- [11] R. Prasad, *Fungal Nanotechnology*, Springer, 2017.
- [12] R. F. Species 2000 & ITIS Catalogue of life, 20th February 2019. 2019.
- [13] M. Blackwell, The fungi: 1, 2, 3 ... 5.1 million species? *Am. J. Bot.* 98 (3) (2011) 426–438.
- [14] J.W. Spatafora, M.C. Aime, I.V. Grigoriev, F. Martin, J.E. Stajich, M. Blackwell, The fungal tree of life: from molecular systematics to genome-scale phylogenies, *Microbiol. Spectr.* 5 (5) (2017).
- [15] L. Tedersoo, S. Sánchez-Ramírez, U. Kõljalg, M. Bahram, M. Döring, D. Schigel, et al., High-level classification of the fungi and a tool for evolutionary ecological analyses, *Fungal Divers.* 90 (1) (2018) 135–159.
- [16] J. Choi, S.-H. Kim, A genome tree of life for the fungi kingdom, *Proc. Natl. Acad. Sci.* 114 (35) (2017) 9391–9396.
- [17] D.S. Hibbett, M. Blackwell, T.Y. James, J.W. Spatafora, J.W. Taylor, R. Vilgalys, Phylogenetic taxon definitions for Fungi, Dikarya, Ascomycota and Basidiomycota, *IMA Fungus* 9 (2018) 291–298.
- [18] Genome Online Database (GOLD) n.d. [Internet]. [cited March 2019]. Available from: <https://gold.jgi.doe.gov/measurements>.
- [19] D.S. Hibbett, J.E. Stajich, J.W. Spatafora, Toward genome-enabled mycology, *Mycologia* 105 (6) (2013) 1339–1349.
- [20] J.E. Stajich, Fungal genomes and insights into the evolution of the kingdom, *Microbiol. Spectr.* 5 (4) (2017).
- [21] H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: the bacterial pan-genome, *Curr. Opin. Microbiol.* 11 (5) (2008) 472–477.

- [22] R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: <https://www.r-project.org/>.
- [23] Wickham H. ggplot2 2009.
- [24] Schneider, A GPS visualizer. Available from: http://www.gpsvisualizer.com/map_input?form=data.
- [25] J. Peter, M. De Chiara, A. Friedrich, J.-X. Yue, D. Pflieger, A. Bergström, et al., Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates, *Nature* 556 (7701) (2018) 339–344.
- [26] B. Dunn, C. Richter, D.J. Kvitck, T. Pugh, G. Sherlock, Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments, *Genome Res.* 22 (5) (2012) 908–924.
- [27] R.K. Mortimer, P. Romano, G. Suzzi, M. Polsinelli, Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts, *Yeast* 10 (12) (1994) 1543–1552.
- [28] C.T. Hittinger, A. Rokas, F.Y. Bai, T. Boekhout, P. Goncalves, T.W. Jeffries, et al., Genomics and the making of yeast biodiversity, *Curr. Opin. Genet. Dev.* 35 (2015) 100–109.
- [29] P.K. Strobe, D.A. Skelly, S.G. Kozmin, G. Mahadevan, E.A. Stone, P.M. Magwene, et al., The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen, *Genome Res.* 25 (5) (2015) 762–774.
- [30] G. Liti, D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James, et al., Population genomics of domestic and wild yeasts, *Nature* 458 (7236) (2009) 337–341.
- [31] C.T. Hittinger, P. Gonçalves, J.P. Sampaio, J. Dover, M. Johnston, A. Rokas, Remarkably ancient balanced polymorphisms in a multi-locus gene network, *Nature* 464 (7285) (2010) 54–58.
- [32] P. Almeida, C. Gonçalves, S. Teixeira, D. Libkind, M. Bontrager, I. Masneuf-Pomarède, et al., A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*, *Nat. Commun.* 5 (1) (2014).
- [33] C. Plissonneau, F.E. Hartmann, D. Croll, Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome, *BMC Biol.* 16 (1) (2018).
- [34] M.E. Zolan, Chromosome-length polymorphism in fungi, *Microbiol. Rev.* 59 (4) (1995) 686–698.
- [35] R. Sharma, B. Mishra, F. Runge, M. Thines, Gene loss rather than gene gain is associated with a host jump from monocots to dicots in the smut fungus *Melanopsichium pennsylvanicum*, *Genome Biol. Evol.* 6 (8) (2014) 2034–2049.
- [36] M.C. Fisher, D.A. Henk, C.J. Briggs, J.S. Brownstein, L.C. Madoff, S.L. McCraw, et al., Emerging fungal threats to animal, plant and ecosystem health, *Nature* 484 (7393) (2012) 186–194.
- [37] L. Faino, M.F. Seidl, E. Datema, G.C.M. van den Berg, A. Janssen, A.H.J. Wittenberg, et al., Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome, *MBio* 6 (4) (2015).
- [38] C. Plissonneau, A. Stürchler, D. Croll, The evolution of orphan regions in genomes of a fungal pathogen of wheat, *MBio* 7 (5) (2016).
- [39] Y. Wang, M. Stata, W. Wang, J.E. Stajich, M.M. White, J.-M. Moncalvo, et al., Comparative genomics reveals the core gene toolbox for the fungus–insect symbiosis, *MBio* 9 (3) (2018).
- [40] J. Schacherer, G. Song, B.J.A. Dickins, J. Demeter, S. Engel, B. Dunn, et al., AGAPE (automated genome analysis pipeline) for pan-genome analysis of *Saccharomyces cerevisiae*, *PLoS One* 10 (3) (2015).
- [41] E.C.H. Chen, E. Morin, D. Beaudet, J. Noel, G. Yildirim, S. Ndikumana, et al., High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*, *New Phytol.* 220 (4) (2018) 1161–1171.
- [42] N.M. Upadhyaya, D.P. Garnica, H. Karaoglu, J. Sperschneider, A. Nemri, B. Xu, et al., Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes, *Front. Plant Sci.* 5 (2015).
- [43] X. Hu, G. Xiao, P. Zheng, Y. Shang, Y. Su, X. Zhang, et al., Trajectory and genomic determinants of fungal–pathogen speciation and host adaptation, *Proc. Natl. Acad. Sci.* 111 (47) (2014) 16796–16801.
- [44] T.J. Sharpton, J.E. Stajich, S.D. Rounsley, M.J. Gardner, J.R. Wortman, V.S. Jordar, et al., Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives, *Genome Res.* 19 (10) (2009) 1722–1731.

- [45] A.C. Kelly, T.J. Ward, Population genomics of *Fusarium graminearum* reveals signatures of divergent evolution within a major cereal pathogen, *PLoS One* 13 (3) (2018).
- [46] S. Walkowiak, O. Rowland, N. Rodrigue, R. Subramaniam, Whole genome sequencing and comparative genomics of closely related Fusarium Head Blight fungi: *Fusarium graminearum*, *F. meridionale* and *F. asiaticum*, *BMC Genomics* 17 (1) (2016).

CHAPTER 5

Pan-genomics of veterinary pathogens and its applications

Thiago de Jesus Sousa^a, Arun Kumar Jaiswal^{a,b}, Raquel Enma Hurtado^a, Stephane Fraga de Oliveira Tosta^a, Siomar de Castro Soares^b, Anne Cybelle Pinto Gomide^a, Luiz Carlos Junior Alcantara^d, Debmalya Barh^c, Vasco Azevedo^a, Sandeep Tiwari^a

^aPG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

^bDepartment of Immunology, Microbiology and Parasitology, Institute of Biological Science and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Brazil

^cCentre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Purba Medinipur, India

^dLaboratório de Flavivírus, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, Brazil

1 Introduction

Pan-genome is an approach that contributes to the research of bacterial pathogenesis. This terminology was proposed in 2005 in research with the bacterium *Streptococcus agalactiae*, by the researcher Tettelin and collaborators [1]. In this work, they define the pan-genome as a set of genes in a given study group, considering core genome, the genes present in all strains in the group of study; dispensable genes as absent genes in one or more strains; and, genes that are considered unique in each lineage of the study group. Pan-genome can be considered open or closed, depending on the bacterial ability to acquire exogenous regions (DNA) [1] and the lifestyle that will determine this issue [2]. From the sequencing, one can thoroughly study each region of the genome, contributing with unpublished information. Since 2005, with the era of new sequencers, the speed, ease, and reliability of data have been increasing and with them the number of bacterial genomes deposited in public databases [3]. Pan-genome studies can be applied with different goals, such as taxonomy, reverse vaccinology, gene variation, pathogenesis [4], among others. This chapter is focused on the pan-genomics studies carried out on pathogenic bacteria that cause veterinary diseases, including the ones responsible for zoonotic diseases. From the genetic repertoire studies, the key points (genes) supposedly involved in the spread of disease, bacterial resistance, infection, adhesion, can be detected, leading to practical solutions against the disease being studied. An important fact is an identification, from taxonomic studies among the lineages, of horizontal gene transfer, which in addition to contributing to evolutionary information, may be used to infer possibly emerging pathogens, once the previously harmless pathogen may become pathogenic. Horizontal gene transfer causes a considerable impact on genomic plasticity,

***In silico* approaches for prioritizing drug targets in pathogens.**

Mariana Santana^{1‡}, Stephane Fraga de Oliveira Tosta^{1‡}, **Arun Kumar Jaiswal**, Leticia de Castro Oliveira, **Siomar C. Soares**, Anderson Miyoshi, Luiz Carlos Junior Alcantara, Vasco Azevedo^{*}, Sandeep Tiwari^{*}

Submitted: Book: Mitigation of Antimicrobial Resistance, July 2019

Abstract

Antimicrobial resistance is a natural evolutionary process in response to antimicrobial exposure; however, the indiscriminate use of antimicrobials is accelerating this progression. The development of resistance happens when microorganisms (e.g. bacteria, fungi, viruses, and parasites) evolve mechanisms to evade damage (e.g. drug inactivation/alteration, efflux pumps, porin loss, biofilm formation, reduced intracellular drug accumulation, modification of drug binding sites) caused by the contact with antimicrobial drugs, such as antibiotics, antifungals, antivirals, antimalarial, and anthelmintic, which involves genetic changes. The comparative genomics associated to Pan-genomics, subtractive genomics, structural bioinformatics and metabolic pathways analysis approaches are currently applied to reach the development of new antibiotics and fight antimicrobial resistance. Targeted drug development retains major challenges from candidate selection to *in vitro* and *in vivo* experiments and clinical trials. Yet, the advances in scientific knowledge (i.e. disease and pathogen) and research and development R&D, the advent of OMICS approaches (e.g. genomics, transcriptomics and proteomics), and bioinformatics breakthroughs conduct to a ‘big-data era’ that improved identification of putative targets via the application of *in silico* tools that shortened the timeline in a cost-efficient manner. In this chapter we are focusing on different bioinformatics strategies for prioritizing drug targets in pathogens.

IV. Appendix

C. Curriculum Vitae



ARUN KUMAR JAISWAL

Email: arunjaiswal1411@gmail.com | Tel: +55-31973439010

Curriculum vitae: <http://lattes.cnpq.br/5417027899674045>

Skype: arunjaiswal1411@gmail.com

PROFILE SUMMARY

- Currently pursuing PhD. (last year) in Bioinformatics from department of biological science (LGCM- Laboratory of Cellular and Molecular Genetics), at Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. I am a dynamic and highly motivated **Research Personal** in the field of Bioinformatics, Genomics, Proteomics, Bacterial Disease, etc.

CORE COMPETENCIES

- Bioinformatics, Genome, Genomics, Structural and Computational Biology, Comparative Genomics, Molecular Biology and Biochemical techniques, Recombinant DNA Technology.

PERSONAL INFORMATION

BIBLIOGRAPHIC CITATION: JAISWAL; AK, KUMAR JAISWAL; ARUN, JAISWAL; ARUN K,

PROFESSIONAL ADDRESS: INSTITUTO DE CIÊNCIAS BIOLÓGICAS.

UNIVERSIDADE FEDERAL DE MINAS

GERAIS PAMPULHA

31270901 - BELO HORIZONTE, MG - BRASIL

PHONE: (031) 34092554 EXTENSION NUMBER: 031 FAX: (031) 34092554

WWW.ICB.UFMG.BR, <http://www.pgbioinfo.icb.ufmg.br/>

EDUCATIONAL INFORMATION

- M. Sc. in Bioinformatics, School of Biotechnology Devi Ahilya Vishwa Vidyalaya, Indore, India, 69.92%, 2014.
- B. Sc. in Biotechnology, Botany and Chemistry, Veer Kunwar Singh University, Ara, India, 64.13%, 2011.
- Intermediate, Bihar School Examination Board, Patna, Bihar India, 63.55%, 2008.
- High School, Bihar School Examination Board, Patna, Bihar India, 69.0%, 2006.

APPOINTMENTS

CAPES (CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Postgraduate Fellowship Programme (PhD fellowship) UFMG, Brazil February 2016, till February 2020

Project Assistant: School of Biotechnology (DAVV, Indore, India), 17th September 2014 to 31 December 2015

ACHIEVEMENTS AND AWARDS:

- **Selected in: Bioinformatics Industrial Training Programme (BIITP) 2014-2015**
- **CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Postgraduate Fellowship Programme (PhD fellowship) UFMG, Brazil** February 2016 to February 2020.
- Best Poster Honorable Mention Award, X-Meeting 14th, International Conference of AB3C, Brazil.

LANGUAGES

ENGLISH: Comprehends Well, Speaks Well, Reads Well, And Writes Well.

HINDI: Comprehends Well, Speaks Well, Reads Well, And Writes Well.

PORTUGUESE: Comprehends Well, Speaks Well, Reads Well, And Writes Reasonably.

Date of birth: 14th November 1990

1

Permanent Address: C/O Jagdish Prasad Jaiswal, Near Kanha PCO, Funduridhari Bichpara, Ward no-10, Ambikapur, Surguja Chhattisgarh, India, 497001.

Current address: Rua professor Nelson de Sena, 95, Apartment 202, Pampulha, Belo Horizonte – MG Belo Horizonte
MG, Brazil, 31270-660

EXPERTISE

- Homology Modeling, Phylogenetics, Phylogenomics, Structure Biology, Reverse Vaccinology, Docking- Protein-Ligand (Autodock Vina) & Protein-Protein, Pan-genomics, Subtractive genomics, Pan-genome. Computer- Aided Drug Designing, Simulation- Protein & Protein-Drug Complex (GROMACS, VMD-NAMD), Visualizing software- Chimera and Pymol.
- Next Generation Sequence Data Analysis (NGS DATA)- Fastq files, Assembly and Annotation (FastQC, Newbler, SPAdes, Mira and CLC Genomics Workbench v7).

PROGRAMMING SKILLS

- Python and R (Limited), Good experience on Linux

PROFESSIONAL TRAINING AND CERTIFICATES

- Summer Training “ON QUALITY CONTROL OF BEVERAGES AND TREATMENT OF WATER” from (PEPSI) LUMBINI BEVERAGES PVT. LTD. Hajipur , India. 5th October to 30th October 2009.
- Certified Portuguese calls for 40 hours. Federal University of Minas Gerais. held in Belo Horizonte - Brazil between 15th to 26th February, 2016.
- Participated certificated course in the winter course of the post-graduation program in Bioinformatics. Federal University of Minas Gerais. Held in Belo Horizonte - Brazil between 11th to 15th July, 2016.
- Certified participation in an “International course of Bioinformatics in Molecular and Evolutionary Epidemiology of Virus ”, (CIBEMEV), held at Institute of Biological Science, Federal University of Minas Gerais. Belo Horizonte - Brazil between 25th to 29th September, 2017.
- Identifying Gene Regulatory Networks from Gene Expression. FIOCRUZ, Campus Virtual 2019. Federal University of Minas Gerais. Engineering School, Department of Structural Engineering 2019- held in Belo Horizonte - Brazil between 11th to 15th March 2019.

PUBLICATIONS

1. **Arun Kumar Jaiswal**, Sandeep Tiwari¹, Syed Babar Jamal, Leticia de Castro Oliveira, Leandro Gomes Alves, Vasco Azevedo, Preetam Ghosh, Carlo Jose Freira Oliveira, Siomar C. Soares. “**The pan-genome of *Treponema pallidum* reveals differences in genome plasticity between subspecies related to venereal and non-venereal syphilis**”. (*Accepted:- BMC Bioinformatics, 24th December 2019*).
2. Stephane Fraga de Oliveira Tosta, Mariana Santana Passos, Rodrigo Kato, Álvaro Salgado, **Arun Kumar Jaiswal**, Siomar C. Soares, Vasco Azevedo, Marta Giovanetti, Sandeep Tiwari, Luiz Carlos Junior Alcantara. “**Multi-epitope based vaccine against Yellow fever virus applying immunoinformatics analysis**”. (*Accepted:- Journal of Biomolecular Structure & Dynamics*).
3. Thaís Cristina Vilela Rodrigues[†], **Arun Kumar Jaiswal**[†], Alissa de Sarom, Leticia de Castro Oliveira, Carlo José Freire Oliveira, Preetam Ghosh, Sandeep Tiwari, Fábio Malcher Miranda, Leandro de Jesus Benevides, Vasco Ariston de Carvalho Azevedo and Siomar de Castro Soares. “**Reverse vaccinology and subtractive genomics reveal new therapeutic targets against *Mycoplasma pneumoniae*: a causative agent of pneumonia**”. (*Royal Society of Open Science. 2019 ; Volume 6 Issue 7. doi.org/10.1098/rsos.190907*).
4. Mariana Santana, Stephane Tosta, Rodrigo Kato, **Arun Kumar Jaiswal**, Siomar C. Soares, Luiz Carlos Junior Alcantara, Preetam Ghosh, Debmalya Barh, Vasco Azevedo, Anderson Mioshi, Sandeep Tiwari. “**Multi epitope based vaccine against diphtheria: An immunoinformatics approach to cope up with toxigenic and non-toxigenic *Corynebacterium diphtheriae* strains**”. (*Submitted:- Journal of Biomolecular Structure & Dynamics*).
5. Mariana Santana, **Arun Kumar Jaiswal**, Stephane Fraga de Oliveira Tosta, Rodrigo Kato, Siomar C. Soares, Luiz Carlos Junior Alcantara, Anderson Miyoshi, Vasco Azevedo, Sandeep Tiwari. “**A reverse Vaccinology and immunoinformatics based approach for multi-epitope vaccine against *Clostridioides difficile* infection**”. (*Submitted:- PeerJ*).”

6. **Arun Kumar Jaiswal**, Sandeep Tiwari, Syed Babar Jamal, Stephane Fraga de Oliveira Tosta, Rodrigo Profeta, Preetam Ghosh, Vasco Azevedo, Siomar C. Soares. “**Syphilis: Clinical, Epidemiological and Biological features with future perspectives.**” (**Submitted:- PeerJ**)
7. Sandeep Tiwari, Debmalya Barh, M. Imchen, Eswar Rao, Ranjith K. Kumavath, S. Prabu Seenivasan, **Arun Kumar Jaiswal**, Syed B. Jamal, Vanaja Kumar, Preetam Ghosh, Vasco Azevedo.”**Acetate Kinase (AcK) is Essential for Microbial Growth and Betel-derived Compounds Potentially Target AcK, PhoP and MDR Proteins in M. tuberculosis, V. cholerae and Pathogenic E. coli: An in silico and in vitro Study**”. (*Current Topics in Medicinal Chemistry, Curr Top Med Chem. 2018;18(31):2731-2740. doi: 10.2174/1568026619666190121105851*).
8. Alissa de Sarom, **Arun Kumar Jaiswal**, Sandeep Tiwari, Letícia de Castro Oliveira, Debmalya Barh, Vasco Azevedo, Carlo Jose Oliveira and Siomar de Castro Soares“**Putative vaccine candidates and drug targets identified by reverse vaccinology and subtractive genomics approaches to control Haemophilus ducreyi, the causative agent of chancroid**”. (*Journal of The Royal Society Interface, 2018; J. R. Soc. Interface 15: 20180032. doi.org/10.1098/rsif.2018.0032*).
9. **Arun Kumar Jaiswal**, Sandeep Tiwari, Syed Babar Jamal, Debmalya Barh, Vasco Azevedo and Siomar C. Soares “**An In Silico Identification of Common Putative Vaccine Candidates against Treponema pallidum: A Reverse Vaccinology and Subtractive Genomics Based Approach**”. (*International Journal of Molecular Science, (Int. J. Mol. Sci. 2017, 18(2), 402;. DOI:10.3390/ijms18020402)*).
10. Siomar de Castro Soares, Letícia de Castro Oliveira, **Arun Kumar Jaiswal**, Vasco Azevedo "**Genomic Islands: an overview of current software tools and future improvements**". (*Journal of Integrative Bioinformatics, 2016 Dec 22;13(1):301. doi.org/10.1515/jib-2016-301*).
11. Supriya Ratnaparkhe, Milind B. Ratnaparkhe, **Arun Kumar Jaiswal**, Anil Kumar. "**Strain Engineering for Improved Bio-Fuel Production**". (*Current Metabolomics, 2016; 4 ;1, DOI : 10.2174/2213235X03666150818222343*).

BOOK CHAPTER

- **Arun Kumar Jaiswal**✉, Sandeep Tiwari✉, Guilherme Campos Tavares, Wanderson Marques, Letícia de Castro Oliveira, Izabela Coimbra Ibraim, Luis Carlos Guimarães, Anne Cybelle Pinto Gomide, Syed Babar Jamal, Yan Pantoja, Basant K. Tiwary, Andreas Burkovski, Faiza Munir, Hai Ha Pham Thi, Nimat Ullah, Amjad Ali, Marta Giovanetti, Luiz Carlos Junior Alcantara, Jaspreet Kaur, Dipali Dhawan, Madangchanok Imchen, Ravali Krishna Vennapu, Ranjith Kumavath, Mauricio Corredor, Henrique César Pereira Figueiredo, Debmalya Barh, Vasco Azevedo, Siomar C. Soares. **Pan-omics focused to Crick’s Central Dogma. (Publisher- Elsevier, Imprint, Paperback ISBN: 9780128170762, Book: Pan-genomics: Applications, Challenges, and Future Prospects. Published Date: 17th, January 2020.)**.
- Thiago de Jesus Sousa, **Arun Kumar Jaiswal**, Raquel Enma Hurtado, Stephane Fraga de Oliveira Tosta, Siomar de Castro Soares, Anne Cybelle Pinto Gomide, Luiz Carlos Junior Alcantara, Debmalya Barh, Vasco Azevedo, Sandeep Tiwari. **Pan-genomics of veterinary bacteria and its applications. (Publisher- Elsevier, Imprint, Paperback ISBN: 9780128170762, Book: Pan-genomics: Applications, Challenges, and Future Prospects. Published Date: 17th, January 2020.)**.
- Rodrigo Bentes Kato[†], **Arun Kumar Jaiswal**[†], Sandeep Tiwari, Vasco Azevedo, Aristóteles Góes-Neto. **Pan-genomics of Fungi and its applications. (Publisher- Elsevier, Imprint, Paperback ISBN: 9780128170762, Book: Pan-genomics: Applications, Challenges, and Future Prospects. Published Date: 17th, January 2020.)**.
- Mariana Santana, Stephane Fraga De Oliveira Tosta, **Arun Kumar Jaiswal**, Letícia de Castro Oliveira, Siomar de Castro Soares, Anderson Miyoshi, Luiz Carlos Junior Alcantara, Vasco Azevedo, Sandeep Tiwari. In silico approaches for prioritizing drug targets in pathogens. Publisher- Springer, MITIGATION OF ANTIMICROBIAL RESISTANCE (**Under review**).

NATIONAL AND INTERNATIONAL CONFERENCE, WORKSHOP & SEMINARS ATTENDED

- Attended Seminar in “100th Indian Science Congress” at Kolkata, India 2013.
- “Role of Biotechnology in Human Welfare”, organized by School of Biotechnology, Devi Ahilya University, held in Indore – India in January, 2014.
- International Conference on “Emerging Challenges in Biotechnology Human Health and Environment” Under Golden Jubilee Celebration of the University organized by School of Biotechnology, Devi Ahilya University, held in Indore, India – between 18th to 20th December, 2014.
- Participated in “International Associated Laboratory – LIA/ 2018- Bact-Inflam” (INRA-UFGM), held at Federal University of Minas Gerais. Belo Horizonte - Brazil between 22nd to 23rd May, 2018.
- Computer Forensic Lecture, an activity promoted by the Forensic Science Academic League (LACF-UFTM) Federal University of Triângulo Mineiro, Uberaba-MG- Brazil, 20th March, 2018.
- Participated in “The Nanopore Based Genetic Sequencing Technology Course for Temporal Investigation and Dengue Outbreak Epidemiology: Training, Research, Surveillance, and Scientific Dissemination”. BeloHorizonte, Brazil, 19th to 30th August, 2019.
- Participated in “15th, 14th, 13th and 12th X-meeting, International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), In year 2019 (Campos do Jordão-Brazil), 2018 (São Pedro-Brazil), 2017 (São Pedro-Brazil) and 2016 (Belo Horizonte-Brazil)”

LECTURE

- 1^o Central-west Congress on Clinical Immunology on “An Overview of Subtractive genomics and Reverse Vaccinology approach to predict therapeutic targets in Bacteria”. Held at Mineiros-Goias, Brazil, 9th June, 2017.
- Discipline of Molecular Modeling of Biological Systems, Post-graduate program in physiological sciences on “Molecular Docking”, Federal University of Triângulo Mineiro, Uberaba-MG- Brazil, 29th June, 2017.
- V Summer course in Immunoparasitology on “*In silico* Drug Target Identification in Bacterial Pathogen”, Federal University of Triângulo Mineiro, Uberaba-MG- Brazil, 29th January to 2nd February, 2018.
- Lecture entitled “Comparative Genomics approaches for candidate drug and vaccine targets identification in pathogenic Bacteria”. Postgraduate program in Bioinformatics. Federal University of Minas Gerais. Belo Horizonte – Brazil, 29th June, 2018.

ABSTRACTS SELECTED FOR PRESENTATION

- A Structural bioinformatics approach for Functional characterization of *Treponema pallidum* subspecies hypothetical proteins. **X-Meeting 2019- 15th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at Campos do Jordão – Brazil, between 30th October to 01 November, 2019.**
- An In Silico approach for the identification of vaccine and drug targets against Mycoplasma genitalium, causative agent of sexually transmitted pelvic inflammatory disease (PID). **X-Meeting 2019- 15th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at Campos do Jordão – Brazil, between 30th October to 01 November, 2019.**
- IMMUNOINFORMATICS-AIDED DESIGN OF POTENTIAL VACCINE CANDIDATES AND DRUGS’ TARGETS AGAINST Trypanosoma CRUZI. **XLIV Congress of the Brazilian Society of Immunology-Immunotherapy: recent advances and future for therapeutic interventions, IMMUNO 2019, held at Florianopolis – SC.- Brazil, between 29th september to 2nd October, 2019.**

- New vaccine and drug targets of *Mycoplasma pneumoniae* revealed by reverse vaccinology and subtractive genomics. *Associated International Laboratory Meeting, held at BeloHorizonte between 21th and 23th August 2019 (Presented by Thais Cristina).*
- A Reverse Vaccinology and Subtractive Genomics Approach for the Common Therapeutics Identification against *Mycobacterium leprae* and *Mycobacterium lepromatosis*. *X-Meeting 2018- 14th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 24th and 26th October of 2018. (Best Poster Honorable Mention Award)*
- In silico identification of drug and vaccine targets in *Corynebacterium ulcerans*. *X-Meeting 2018- 14th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 24th and 26th October of 2018.*
- A Reverse Vaccinology and Subtractive Genomics Based Approach for the Identification of Common Putative Vaccine and drug Candidates against *Clostridioides difficile* infection. *X-Meeting 2018- 14th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 24th and 26th October of 2018.*
- Comparative genomics analysis of *Bartonella henselae* discloses different patterns of adaptation to host. *X-Meeting 2018- 14th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 24th and 26th October of 2018.*
- Comparative genomic analysis with a multidrug resistant *Klebsiella pneumoniae* strain recently isolated from a patient in Brazil. *Gene Time Conference (Post-graduation in genetics-UFMG) 2018- held in Belo Horizonte - Brazil between 23rd and 24th August of 2018.*
- The Pan-Genome of *Treponema pallidum* Reveals Differences in Genome Plasticity between the subspecies. *X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 4th and 6th October of 2017.*
- Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Salmonella serovers*. *X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 4th and 6th October of 2017.*
- Functional Analysis of hypothetical proteins unveils putative metabolic pathways, essential genes and therapeutic drug and vaccine target in *Trypanosoma cruzi*: A bioinformatics based approach. *X-Meeting 2017 -13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at São Pedro - Brazil between 4th and 6th October of 2017.*
- In silico Identification of common putative vaccine candidates against *Treponema pallidum*: A reverse vaccinology based approach. *X-Meeting 2016- 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at UFMG, Belo Horizonte - Brazil between 16th to 18th November of 2016. X-meeting 2016.*
- In-silico analyses for the discovery of drug and vaccine targets in *Corynebacterium camporealensis*: A Novel Hierarchical Approach. *X-Meeting 2016- 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held at UFMG, Belo Horizonte - Brazil between 16th to 18th November of 2016. X-meeting 2016.*

CLASSES AND OTHER RESEARCH RELATED ACTIVITIES

- **Certified** for giving 40 hours class on “Bioinformatics: Comparative Genomics and Molecular Modeling”, at the III Winter Course in Physiological Sciences , (UFTM) Federal University of the Triângulo Mineiro, on the 31st July to 4th August, 2017.

REFERENCES

Supervisor

- **Professor Dr. Siomar de Castro Soares, MSc., PhD., Adjunct Professor:** Institute of Biological and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Rua Getúlio Guaritá, S/N, Uberaba-MG, Brazil. CEP-38025-180. **Email:** siomars@gmail.com, siomar@icbn.uftm.edu.br.

Co-Supervisor

- **Professor Dr. Vasco Ariston de Carvalho Azevedo:** Deputy Head of the Department of Genetics, Ecology and Evolution, Department of General Biology, Institute of Biological Science (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil. CP 486, CEP 31270-901. **Email:** vascoariston@gmail.com.

Co-Supervisor

- **Dr. Sandeep Tiwari, M.Sc, PhD,** Postdoctoral Research Fellow: Laboratory of Cellular and Molecular Genetics (LGCM), Institute of Biological Science, Federal University of Minas Gerais Av. Antonio Carlos 6627, Pampulha, Belo Horizonte Minas Gerais, Brazil CO 486, CEP 31270-901. **Email:** sandip_sbtbi@yahoo.com

V. Bibliography

- Akalin, P.K. (2006). Introduction to bioinformatics. *Mol Nutr Food Res* 50(7), 610-619. doi: 10.1002/mnfr.200500273.
- Altman, R.B., and Klein, T.E. (2007). Biomedical informatics training at Stanford in the 21st century. *J Biomed Inform* 40(1), 55-58. doi: 10.1016/j.jbi.2006.02.005.
- Ariel, N., Zvi, A., Grosfeld, H., Gat, O., Inbar, Y., Velan, B., et al. (2002). Search for Potential Vaccine Candidate Open Reading Frames in the *Bacillus anthracis* Virulence Plasmid pXO1: In Silico and In Vitro Screening. *Infection and Immunity* 70(12), 6817-6827. doi: 10.1128/iai.70.12.6817-6827.2002.
- Bagnoli, F., Fontana, M.R., Soldaini, E., Mishra, R.P., Fiaschi, L., Cartocci, E., et al. (2015). Vaccine composition formulated with a novel TLR7-dependent adjuvant induces high and broad protection against *Staphylococcus aureus*. *Proc Natl Acad Sci U S A* 112(12), 3680-3685. doi: 10.1073/pnas.1424924112.
- Baker, B.J., and Armelagos, G.J. (1988). The origin and antiquity of syphilis: paleopathological diagnosis and interpretation. *Curr Anthropol* 29(5), 703-738. doi: 10.1086/203691.
- Barh, D., Tiwari, S., Jain, N., Ali, A., Santos, A.R., Misra, A.N., et al. (2011). In silico subtractive genomics for target identification in human bacterial pathogens. *Drug Development Research* 72(2), 162-177. doi: 10.1002/ddr.20413.
- Bhatia, B., Ponia, S.S., Solanki, A.K., Dixit, A., and Garg, L.C. (2014). Identification of glutamate ABC-Transporter component in *Clostridium perfringens* as a putative drug target. *Bioinformation* 10(7), 401-405. doi: 10.6026/97320630010401.
- Bowie, J.U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253(5016), 164-170. doi: 10.1126/science.1853201.
- Cherneskie, T. (2006). <An Update and Review of the Diagnosis and Management of Syphilis.>
- Cox, S.S., van der Giezen, M., Tarr, S.J., Crompton, M.R., and Tovar, J. (2006). Evidence from bioinformatics, expression and inhibition studies of phosphoinositide-3 kinase signalling in *Giardia intestinalis*. *BMC Microbiol* 6, 45. doi: 10.1186/1471-2180-6-45.
- Cule, J. (1992). The Great Maritime Discoveries and World Health. *Journal of the History of Medicine and Allied Sciences* 47(4), 513.
- De Maayer, P., Chan, W.Y., Rubagotti, E., Venter, S.N., Toth, I.K., Birch, P.R., et al. (2014). Analysis of the *Pantoea ananatis* pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. *BMC Genomics* 15, 404. doi: 10.1186/1471-2164-15-404.
- de Melo, F.L., de Mello, J.C., Fraga, A.M., Nunes, K., and Eggers, S. (2010). Syphilis at the crossroad of phylogenetics and paleopathology. *PLoS Negl Trop Dis* 4(1), e575. doi: 10.1371/journal.pntd.0000575.
- de Sarom, A., Kumar Jaiswal, A., Tiwari, S., de Castro Oliveira, L., Barh, D., Azevedo, V., et al. (2018). Putative vaccine candidates and drug targets identified by reverse vaccinology and subtractive genomics approaches to control *Haemophilus ducreyi*, the causative agent of chancroid. *J R Soc Interface* 15(142). doi: 10.1098/rsif.2018.0032.
- Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., et al. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* 11(10), R107. doi: 10.1186/gb-2010-11-10-r107.
- Etz, H., Minh, D.B., Henics, T., Dryla, A., Winkler, B., Triska, C., et al. (2002). Identification of in vivo expressed vaccine candidate antigens from *Staphylococcus aureus*. *Proc Natl Acad Sci U S A* 99(10), 6573-6578. doi: 10.1073/pnas.092569199.
- EXPLORING LIFE, I.I. (2014). *Syphilis: Then and Now*, *The Scientist* [Online]. Available: <https://www.the-scientist.com/features/syphilis-then-and-now-38045> [Accessed].
- Fernandez-Pinar, R., Lo Sciuto, A., Rossi, A., Ranucci, S., Bragonzi, A., and Imperi, F. (2015). In vitro and in vivo screening for novel essential cell-envelope proteins in *Pseudomonas aeruginosa*. *Sci Rep* 5, 17593. doi: 10.1038/srep17593.

- Ferreira, R.S., Simeonov, A., Jadhav, A., Eidam, O., Mott, B.T., Keiser, M.J., et al. (2010). Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. *J Med Chem* 53(13), 4891-4905. doi: 10.1021/jm100488w.
- Foa, A. (1990). "The new and the old: The spread of syphilis (1494-1530)," in *Sex and gender in historical perspective*, eds. E. Muir, G. Ruggiero, M.A. Gallucci, M.M. Gallucci & C.C. Gallucci. (Baltimore, Md London: Johns Hopkins University Press), 26-45.
- Forrai, J. (2011). <History of Different Therapeutics of Venereal Disease Before the Discovery of Penicillin.pdf>.
- Forsea, D., Popescu, R., and Popescu, C.M. (1997). *Compendiu de dermatologie si Venerologie*. Editura Tehnica.
- Gal-Mor, O., and Finlay, B.B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol* 8(11), 1707-1719. doi: 10.1111/j.1462-5822.2006.00794.x.
- Gupta, C.L., Akhtar, S., and Bajpai, P. (2014). In silico protein modeling: possibilities and limitations. *EXCLI J* 13, 513-515.
- Hansen, E.E., Lozupone, C.A., Rey, F.E., Wu, M., Guruge, J.L., Narra, A., et al. (2011). Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc Natl Acad Sci U S A* 108 Suppl 1, 4599-4606. doi: 10.1073/pnas.1000071108.
- Hofer, A., Steverding, D., Chabes, A., Brun, R., and Thelander, L. (2001). Trypanosoma brucei CTP synthetase: a target for the treatment of African sleeping sickness. *Proc Natl Acad Sci U S A* 98(11), 6412-6416. doi: 10.1073/pnas.111139498.
- Hosen, M.I., Tanmoy, A.M., Mahbuba, D.A., Salma, U., Nazim, M., Islam, M.T., et al. (2014). Application of a subtractive genomics approach for in silico identification and characterization of novel drug targets in *Mycobacterium tuberculosis* F11. *Interdiscip Sci* 6(1), 48-56. doi: 10.1007/s12539-014-0188-y.
- Jamal, S.B., Hassan, S.S., Tiwari, S., Viana, M.V., Benevides, L.J., Ullah, A., et al. (2017). An integrative in-silico approach for therapeutic target identification in the human pathogen *Corynebacterium diphtheriae*. *PLoS One* 12(10), e0186401. doi: 10.1371/journal.pone.0186401.
- Jauch, R., Yeo, H.C., Kolatkar, P.R., and Clarke, N.D. (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins* 69 Suppl 8, 57-67. doi: 10.1002/prot.21771.
- John, B., and Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res* 31(14), 3982-3992. doi: 10.1093/nar/gkg460.
- Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature* 358(6381), 86-89. doi: 10.1038/358086a0.
- Katz, K.A., and Klausner, J.D. (2008). Azithromycin resistance in *Treponema pallidum*. *Curr Opin Infect Dis* 21(1), 83-91. doi: 10.1097/QCO.0b013e3282f44772.
- Kelly, D.F., and Rappuoli, R. (2005). Reverse vaccinology and vaccines for serogroup B *Neisseria meningitidis*. *Adv Exp Med Biol* 568, 217-223. doi: 10.1007/0-387-25342-4_15.
- Khan, F.I., Wei, D.Q., Gu, K.R., Hassan, M.I., and Tabrez, S. (2016). Current updates on computer aided protein modeling and designing. *Int J Biol Macromol* 85, 48-62. doi: 10.1016/j.ijbiomac.2015.12.072.
- Klebe, G. (2006). Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 11(13-14), 580-594. doi: 10.1016/j.drudis.2006.05.012.
- Kojima, N., and Klausner, J.D. (2018). An Update on the Global Epidemiology of Syphilis. *Curr Epidemiol Rep* 5(1), 24-38. doi: 10.1007/s40471-018-0138-z.
- Liao, C., Peach, M.L., Yao, R., and Nicklaus, M.C. (2013). "Molecular docking and structure-based virtual screening," in *In Silico Drug Discovery and Design.*, 6-20.

- Liwo, A., Lee, J., Ripoll, D.R., Pillardy, J., and Scheraga, H.A. (1999). Protein structure prediction by global optimization of a potential energy function. *Proc Natl Acad Sci U S A* 96(10), 5482-5485. doi: 10.1073/pnas.96.10.5482.
- Luger, A. (1991). [The significance of Karl Landsteiner's works for syphilis research]. *Wien Klin Wochenschr* 103(5), 146-151.
- Mahoney, J.F., Arnold, R.C., and Harris, A. (1943). Penicillin Treatment of Early Syphilis-A Preliminary Report. *Am J Public Health Nations Health* 33(12), 1387-1391. doi: 10.2105/ajph.33.12.1387.
- Medini, D., Donati, C., Tettelin, H., Massignani, V., and Rappuoli, R. (2005). The microbial pangenome. *Curr Opin Genet Dev* 15(6), 589-594. doi: 10.1016/j.gde.2005.09.006.
- Montigiani, S., Falugi, F., Scarselli, M., Finco, O., Petracca, R., Galli, G., et al. (2002). Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infect Immun* 70(1), 368-379. doi: 10.1128/iai.70.1.368-379.2002.
- Morton, R.S. (1968). Another look at the Morbus Gallicus. Postscript to the meeting of the Medical Society for the Study of Venereal Diseases, Geneva, May 26-28, 1967. *Br J Vener Dis* 44(2), 174-177. doi: 10.1136/sti.44.2.174.
- Nikolaev, D.M., Shtyrov, A.A., Panov, M.S., Jamal, A., Chakchir, O.B., Kochemirovsky, V.A., et al. (2018). A Comparative Study of Modern Homology Modeling Algorithms for Rhodopsin Structure Prediction. *ACS Omega* 3(7), 7555-7566. doi: 10.1021/acsomega.8b00721.
- Norris, S.J., Paster, B.J., Moter, A., and Göbel, U.B. (2006). "The Genus *Treponema*," in *The Prokaryotes*., 211-234.
- Peeling, R.W., Mabey, D., Kamb, M.L., Chen, X.S., Radolf, J.D., and Benzaken, A.S. (2017). Syphilis. *Nat Rev Dis Primers* 3, 17073. doi: 10.1038/nrdp.2017.73.
- Peng, J., and Xu, J. (2010). Low-homology protein threading. *Bioinformatics* 26(12), i294-300. doi: 10.1093/bioinformatics/btq192.
- Quétel, C., Braddock, J., and Pike, B. (1990). *History of syphilis*. Polity Press Cambridge.
- Qvist, G. (1977). John Hunter's alleged syphilis. *Ann R Coll Surg Engl* 59(3), 206-209.
- Rapid, p.-o.-c.t.f.S.t.f.o.d. (2013). *McGill University Health Centre* [Online]. Available: <https://publications.mcgill.ca/medenews/2013/02/27/rapid-point-of-care-tests-for-syphilis-the-future-of-diagnosis/> [Accessed].
- Rappuoli, R. (2001). Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 19(17-19), 2688-2691. doi: 10.1016/s0264-410x(00)00554-5.
- Rasko, D.A., Altherr, M.R., Han, C.S., and Ravel, J. (2005). Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol Rev* 29(2), 303-329. doi: 10.1016/j.femsre.2004.12.005.
- Rasko, D.A., Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F., Gajer, P., et al. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190(20), 6881-6893. doi: 10.1128/JB.00619-08.
- Ripphausen, P., Nisius, B., Peltason, L., and Bajorath, J. (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J Med Chem* 53(24), 8461-8467. doi: 10.1021/jm101020z.
- Ross, B.C., Czajkowski, L., Hocking, D., Margetts, M., Webb, E., Rothel, L., et al. (2001). Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* 19(30), 4135-4142. doi: 10.1016/s0264-410x(01)00173-6.
- Rothschild, B.M. (2005). History of syphilis. *Clin Infect Dis* 40(10), 1454-1463. doi: 10.1086/429626.
- Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4), 725-738. doi: 10.1038/nprot.2010.5.
- Salado-Rasmussen, K. (2015). Syphilis and HIV co-infection. Epidemiology, treatment and molecular typing of *Treponema pallidum*. *Dan Med J* 62(12), B5176.
- Sefton, A.M. (2001). The Great Pox that was...syphilis. *J Appl Microbiol* 91(4), 592-596. doi: 10.1046/j.1365-2672.2001.01494.x.

- Sette, A., and Rappuoli, R. (2010). Reverse vaccinology: developing vaccines in the era of genomics. *Immunity* 33(4), 530-541. doi: 10.1016/j.immuni.2010.09.017.
- Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268(1), 209-225. doi: 10.1006/jmbi.1997.0959.
- Singh, A.E., and Romanowski, B. (1999). Syphilis: review with emphasis on clinical, epidemiologic, and some biologic features. *Clin Microbiol Rev* 12(2), 187-209.
- Sivashankari, S., and Shanmughavel, P. (2007). Comparative genomics - a perspective. *Bioinformatics* 1(9), 376-378. doi: 10.6026/97320630001376.
- Soares, S.C., Silva, A., Trost, E., Blom, J., Ramos, R., Carneiro, A., et al. (2013). The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. *PLoS One* 8(1), e53818. doi: 10.1371/journal.pone.0053818.
- Sousa, S.F., Cerqueira, N.M., Fernandes, P.A., and Ramos, M.J. (2010). Virtual screening in drug design and development. *Comb Chem High Throughput Screen* 13(5), 442-453. doi: 10.2174/138620710791293001.
- Souza, E.M.d. (2005). A hundred years ago, the discovery of *Treponema pallidum*. *Anais Brasileiros de Dermatologia* 80(5), 547-548.
- Stamm, L.V. (2010). Global challenge of antibiotic-resistant *Treponema pallidum*. *Antimicrob Agents Chemother* 54(2), 583-589. doi: 10.1128/AAC.01095-09.
- Stamm, L.V. (2016). Syphilis: Re-emergence of an old foe. *Microb Cell* 3(9), 363-370. doi: 10.15698/mic2016.09.523.
- statistics, U.S.-P.p.k.r. (2019). *UniProtKB/Swiss-Prot* [Online]. Available: <https://web.expasy.org/docs/relnotes/relstat.html> [Accessed].
- Stephens, D.S., Greenwood, B., and Brandtzaeg, P. (2007). Epidemic meningitis, meningococcaemia, and *Neisseria meningitidis*. *Lancet* 369(9580), 2196-2210. doi: 10.1016/S0140-6736(07)61016-2.
- Syphilis: causes, s., diagnosis, treatment, and prevention (2019). *Treponema pallidum* [Online]. Available: <https://medcubic.com/node/141> [Accessed].
- Tampa, M., Sarbu, I., Matei, C., Benea, V., and Georgescu, S.R. (2014). Brief history of syphilis. *J Med Life* 7(1), 4-10.
- Tettelin, H., Maignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* 102(39), 13950-13955. doi: 10.1073/pnas.0506758102.
- UniProt, C. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 37(Database issue), D169-174. doi: 10.1093/nar/gkn664.
- Varma, R., Estcourt, C., and Mindel, A. (2013). "Syphilis," in *Sexually Transmitted Diseases.*, 427-462.
- Wadapurkar, R.M., and Vyas, R. (2018). Computational analysis of next generation sequencing data and its applications in clinical oncology. *Informatics in Medicine Unlocked* 11, 75-82. doi: 10.1016/j.imu.2018.05.003.
- Wizemann, T.M., Heinrichs, J.H., Adamou, J.E., Erwin, A.L., Kunsch, C., Choi, G.H., et al. (2001). Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect Immun* 69(3), 1593-1598. doi: 10.1128/IAI.69.3.1593-1598.2001.
- Workowski, K.A. (2015). <Sexually Transmitted Diseases Treatment Guidelines, 2015>.
- Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 5, 17. doi: 10.1186/1741-7007-5-17.
- Xavier, E.R., Capanema, B.P., Ruiz, J.C., Oliveira, G., Meyer, R., D'Afonseca, V., et al. (2008). Brazilian genome sequencing projects: state of the art. *Recent Pat DNA Gene Seq* 2(2), 111-132. doi: 10.2174/187221508784534203.

- Xia, X. (2017). Bioinformatics and Drug Discovery. *Curr Top Med Chem* 17(15), 1709-1726. doi: 10.2174/1568026617666161116143440.
- Yang, J., and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res* 43(W1), W174-181. doi: 10.1093/nar/gkv342.
- Zetola, N.M., and Klausner, J.D. (2007). Syphilis and HIV infection: an update. *Clin Infect Dis* 44(9), 1222-1228. doi: 10.1086/513427.
- Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 18(3), 342-348. doi: 10.1016/j.sbi.2008.02.004.