UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS

DEPARTAMENTO DE BIOQUÍMICA E IMUNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIONFORMÁTICA

DIVERGENOME: UMA PLATAFORMA BIOINFORMÁTICA PARA O ESTUDO DA DIVERSIDADE GENÉTICA HUMANA E APLICAÇÕES NA IDENTIFICAÇÃO DE EPISÓDIOS DE SELEÇÃO NATURAL NA EVOLUÇÃO HUMANA

Orientador: Prof.Dr. Eduardo Martín Tarazona Santos

Co-Orientadora: Dra. Alessandra Aparecida Campos

Universidade Federal de Minas Gerais
Belo Horizonte – Janeiro de 2011

Wagner Carlos Santos Magalhães

DIVERGENOME: UMA PLATAFORMA BIOINFORMÁTICA PARA O ESTUDO DA DIVERSIDADE GENÉTICA HUMANA E APLICAÇÕES NA IDENTIFICAÇÃO DE EPISÓDIOS DE SELEÇÃO NATURAL NA EVOLUÇÃO HUMANA

Tese apresentada ao Programa de Bioinformática do Instituto de Ciências Biológicas da UFMG como requisito para a obtenção do título de Doutor em Bioinformática

Orientador: Eduardo Martín Tarazona Santos

Co-Orientadora: Dra. Alessandra Aparecida Campos

Universidade Federal de Minas Gerais
Intituto de Ciências Biológicas
Departamento de Biologia Geral

Belo Horizonte – Janeiro de 2011

"A única coisa que interfere com meu aprendizado é a minha educação."

(Albert Eisntein)

ÍNDICE

Agradecimentos

Manifesto aqui minha apreciação a todos os que de alguma maneira me ajudaram. Mesmo temendo esquecer-me de alguém (e desde já pedindo desculpas) aqui humildemente deixo meus agradecimentos.

Ao Professor e meu orientador Eduardo Tarazona, pelo seu entusiasmo e esforço em desenvolver pesquisa de ponta e empenho para fazer as coisas acontecerem, além dos ensinamentos científicos sempre direcionados.

A minha co-orientadora Alessandra Campos.

Ao Dr. Stephen Chanock e a Dra. Meredith Yeager pela orientação e ajuda no estágio de doutoramento no National Cancer Institute, Laboratory of Translational Genomics and Core Genotype Facility que gentilmente abriram seus laboratórios, e me receberam de forma impecável.

Aos meus pais, Carlos e Vânia, e a minha irmã Juliana por todo apoio nesses longos anos de estudo. Tive apoio incondicional e insentivo de todas as formas, sem esse suporte familiar nunca poderia ter chegado até aqui.

A Ana Paula, minha noiva, que esteve comigo sempre ao meu lado mesmo quando estava distante. Sem seu apoio e carinho não sei onde estaria.

Aos colegas do LDGH, Maria Clara, Luciana Werneck,Maíra, Giordano, Marilia, Moara, Laélia, Lívia, Juliana, Márcia, Camila, Maíra, Latife, Roxan, Fernandas pela amizade e confiança.

Ao Gustavo Cerqueira, Flávia Azeredo e ao meu roommate em Bethesda no estágio sanduíche "Simon", que me ajudaram muito nos meus primeiros dias nos U.S

Aos meus amigos de doutorado, Leandro, Deive, Eduardo, Sergio, Rodrigo, Bernardo, Rômulo pelas várias conversas e ajudas.

A CAPES pelo suporte financeiro e concessão de bolsas no Brasil e nos EUA.

Lista de Abreviaturas e símbolos

SNP – single nucleotide polymorphism – (polimorfismo de base única)

INDEL – polimorfismo de inserção-deleção

CNV – polimorfismos de número de cópias

DNAsp- DNA sequence polimorphism

Lista de Figuras e Tabelas

Resumo

Neste trabalho desenvolvemos uma plataforma de gerenciamento de dados e projetos, provenientes de estudos de genética de populações e epidemiologia genética, a DIVERGENOME. A plataforma apresenta dois componentes funcionais: A) uma base de dados relacional, o DIVERGENOMEdb, desenvolvida com o objetivo de armazenar os dados de forma segura e organizada e integrar diferentes fontes de informação (disponíveis em repositórios públicos e gerados localmente), dados genéticos (genótipos e haplótipos, provenientes de diferentes tipos de polimorfismos) e informações epidemiológicas (fenótipos, constituídos de variáveis qualitativas e quantitativas); e B) um conjunto de scripts para manipulação de formatos de arquivos, o DIVERGENOMEtools, com o objetivo de otimizar a tarefa de conversão de formatos para análises em diferentes software, tarefa comprovadamente árdua e fonte de grande número de erros evidenciados nos resultados finais das análises. Nossa plataforma apresenta uma nova metodologia para a integração de diferentes scripts permitindo maior número possível de conversões e fácilitando sua extensão. Uma primeira versão da ferramenta pode ser acessada em (www.cebio.org/pipelineldgh/). Os diferentes componentes da plataforma foram utilizados na condução dos trabalhos sobre a ação dos fatores evolutivos que moldam a diversidade genética apresentados na dissertação, mostrando-se eficientes às suas propostas. Para garantir a o acesso de forma rápida e ampla utilização de nossa plataforma pela comunidade científica desenvolvemos ainda uma interface web dessa forma não exigindo do usuário conhecimentos prévios de programação e gerenciamento de bancos de dados.

Abstract

At this work we developed a management platform for data and projects from population genetics and genetic epidemiology fields called DIVERGENOME. It is composed of two functional components: A) a relational database, which aims to safely store, organize and integrate different sources of information and datasets (available at public repositories and locally produced), as well as genetic data (genotypes and haplotypes inferred using different types of polymorphisms) and epidemiologic information (phenotypes, characterized by quantitative and qualitative variables); and B) a set of scripts written using the programming language Perl, called DIVERGENOMEtools, that enables users to handle and change file formats according to their and software required for data analysis, a step which bears several basic but error-prone tasks. Our conversion tool outlines a new strategy, graph based, that integrates the scripts available by creating dynamic conversion pipelines attempting to maximize the number of formats available and to easy incorporate new scripts to the system. A first version of DIVERGENOMEtools may be accessed at (www.cebio.org/pipelineldgh/). The different modules of our platform demonstrated to be efficient to their objective. Finally, we developed a web interface to make easy the access of all functionalities of our system.

PREFÁCIO

Um dos maiores desafios da genética moderna é entender a diversidade observada atualmente no genoma humano. Caracterizando a variação entre indivíduos e populações, espera-se ser possível entender problemas como respostas diferenciais a agentes farmacológicos, susceptibilidade diferencial a doenças e a complexa interação entre fatores genéticos e ambientais na produção de fenótipos.

A Bioinformática é uma área relativamente nova de pesquisa que apresenta nos últimos anos um grande crescimento e pode ser considerada como uma linha de pesquisa, que envolve aspectos multidisciplinares e que surgiu a partir do momento em que se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos, bioquímicos e de biologia molecular. Esta nova disciplina utiliza grande disponibilidade de dados gerados de diferentes fontes, na tentativa de integrá-las e explorar de forma mais robusta esse conjunto de informações disponíveis.

A bioinformática envolve a união de diversas linhas de conhecimento – a ciência da computação, a engenharia de softwares, a matemática, a estatística e a biologia molecular – e tem como finalidade principal desvendar a grande quantidade de dados, que vem sendo obtidos através de sequências de DNA e proteínas. Para o desenvolvimento de genomas completos, a informática é imprescindível assim como a biologia molecular moderna não estaria tão avançada hoje, se não fossem os recursos computacionais existentes.

As bases de dados em biologia molecular são importantes, principalmente para proporcionar à comunidade científica uma forma de tornar os dados (produzidos mundialmente) acessíveis de forma fácil, rápida e inteligente. Os bancos de dados constituem

uma das melhores maneiras para armazenar e recuperar de forma eficiente subconjuntos do universo dos dados disponíveis. A crescente quantidade de informação disponível em genética humana, principalmente atribuída nos últimos anos ao desenvolvimento de novas tecnologias de genotipagem em paralelo e novas tecnologias de sequenciamento (*Next Generation Sequencing* - NGS), tem levado a um crescente número de banco de dados biológicos e, consequentemente, intensiva busca e mineração em bases referentes a doenças genéticas e demais interesses.

Embora possa ter frustrado alguns pesquisadores mais tradicionais, essa grande proliferação de dados apresenta um novo paradigma para aqueles envolvidos em aréas popularmente chamadas "*omicas*" e que precisam automatizar análises de grandes escalas de dados para melhor explorá-los. Nesta dissertação, aborda-se a importância do estudo do padrão de variação genética observado no genoma humano, enfatizando a ação da seleção natural sobre esta variação moldando o padrão de variabilidade observado atualmente e como a bioinformática e suas ferramentas interagem com esta questão.

No primeiro capítulo aborda-se o que seja talvez uma das primeiras áreas à qual foi atribuído o nome Bioinformática: a criação de bancos de dados para armazenamento de dados biológicos e o desenvolvimento de ferramentas para automatização das análises de grandes conjuntos de dados de forma eficiente e em menor tempo. Apresenta-se um primeiro trabalho, publicado no periódico *Investigative Genetics*, um pipeline para auxiliar os estudos que envolvam análises de dados de ressequenciamento, com o objetivo de diminuir erros inerentes ao processo de manipulação e criação de diferentes formatos de arquivos utilizados nas diferentes etapas envolvidas nesse processo.

Ainda no primeiro capítulo, apresento um segundo trabalho em preparação, uma plataforma bioinformática denominada *DIVERGENOME,* que é composta por dois módulos. No primeiro módulo, um banco de dados relacional, *DIVERGENOMEdb*, enquanto no segundo apresento uma extensão às ferramentas desenvolvidas no primeiro artigo *DIVERGENOMEtools*, que permite a manipulação de mais formatos de dados, em relação ao pipeline de ressequenciamento, para *software* de genética de populações e epidemiologia genética.

Há interesse na detecção de genes e, ou regiões genômicas que foram alvo da seleção natural, com o intuito de desvendar os processos evolutivos que atuaram na nossa espécie e em outras. No segundo capítulo, abordo a utilização de ferramentas bioinformáticas (*scripts*) públicas ou desenvolvidas pelo nosso grupo de pesquisa no estudo da seleção natural. São apresentados dois artigos, publicados em coautoria, e um manuscrito com a aplicação, onde foram aplicados diferentes testes para o desvio da neutralidade intra- e inter-específicos em genes de interesse biomédico.

No terceiro capítulo, através de uma revisão realizada durante o período de estágio de doutoramento no National Cancer Institute e publicada no periódico ***Carcinogenesis***, apresento um dos tópicos mais contemplados nos últimos anos, os estudos de varredura genômica (Genome-wide association studies – GWAS) e que tem se mostrado como uma das melhores metodologias na elucidação de associações entre variantes genéticas e desfechos patológicos ou fenótipos de interesse.

Ao longo desta dissertação, procuro ainda ressaltar a necessidade de futuros estudos voltados para o entendimento de processos evolutivos que moldam a variabilidade genética observada no presente, além do modo como a Bioinformática pode auxiliar nesse caminho

através do desenvolvimento de metodologias mais sofisticadas e robustas de análise e interpretação, auxiliando a exploração mais eficiente dos dados gerados.

1. DESENVOLVIMENTO DE FERRAMENTAS BIOINFORMÁTICAS PARA ESTUDOS DE GENÉTICA DE POPULAÇÕES E EPIDEMIOLOGIA GENÉTICA

Introdução

1.1 Variação Genômica e Bioinformática

O sequenciamento do genoma humano (Lander *et al*., 2001; Venter *et al*., 2001) tem possibilitado novos enfoques para a compreensão da origem dos padrões de diversidade genética nas populações humanas (Sachidanandam *et al*., 2001; Frazer *et al*., 2007). Os padrões de diversidade, observados ao longo do genoma, permitem inferir os eventos evolutivos que os geraram (Fagundes *et al*., 2007). Em particular nos últimos anos, com o grande volume de informações sobre a diversidade genética humana, foi possível: (1) utilizar dados de diferentes regiões do genoma para fazer inferências mais robustas sobre a história demográfica humana (Voight *et al*., 2005; Fagundes *et al*., 2007); (2) determinar a estrutura genética das populações humanas a partir de um grande número de polimorfismos representativos do genoma (Rosenberg *et al*., 2003; Li *et al*., 2008); (3) estudar os padrões de diversidade de centenas de genes de interesse biomédico (Packer *et al*., 2006); e (4) inferir a ação da seleção natural e estudar as adaptações aos diferentes ambientes experimentados pela população humana ao longo de sua história evolutiva (Jakobsson *et al*., 2008). Dados biológicos advindos do conhecimento genômico são relativamente complexos dada sua diversidade e seu inter-relacionamento. Assim, toda essa informação disponibilizada pela genômica só é possível de ser organizada, analisada e interpretada com o auxílio da Bioinformática.

Nos últimos anos, diversas iniciativas públicas foram desenvolvidas com o objetivo de organizar e disponibilizar recursos bioinformáticos para armazenagem e manipulação dos dados gerados (Tabela 1). Dentre os projetos desenvolvidos, destaca-se o catálogo da diversidade genética humana, HapMap. O projeto HapMap (The International HapMap Project) é uma iniciativa multicêntrica para a identificação e catalogação de polimorfismos genéticos compartilhados e específicos entre diferentes populações. O projeto teve início com quatro populações: Europeus, Japoneses, Chineses e Africanos, sendo que atualmente conta com uma cobertura de 11 populações (http://hapmap.ncbi.nlm.nih.gov/).

Vários grupos utilizaram os dados disponibilizados pelo projeto HapMap em diferentes estudos. Dentre esses estudos, destacam-se: a utilização dos dados para inferir a ação da seleção natural nas diferentes populações cobertas pelo projeto (Nielsen *et al*., 2005; Sabeti *et al*., 2007); a influência do viés de averiguação, introduzido pela metodologia utilizada no projeto HapMap (Clark *et al*., 2005); os padrões de estruturação populacional (Clayton *et al*., 2005); a eficiência e potência dos estudos de associação (De Bakker *et al*., 2005; Bhangale *et al*., 2008), além de outros projetos que estão sendo desenvolvidos no presente momento, principalmente devido a ampliação do número de populações, 4 populações até a segunda fase e 11 populações na terceira fase.

Outro grande conjunto de informações sobre a variabilidade humana vem sendo produzido com o projeto SNP500Cancer (http:www.*snp500cancer.nci.nih.gov*), especialmente orientado para validar SNPs em genes envolvidos em carcinogênese. O projeto SNP500Cancer é uma das várias iniciativas desenvolvidas para caracterização da variação genética, com o objetivo de entender a etiologia de diferentes tipos de câncer. O banco de dados faz parte do Projeto de Anatomia Genômica do Câncer do *National Cancer Institute* – NCI, dos EUA. O projeto SNP500Cancer estuda amostras de DNA de 102 indivíduos dos

repositórios celulares no *Coriell Institute of Medical Research*, sendo esses sujeitos representantes de quatro etnicidades: Afro-americanos, Caucasianos, Hispânicos e Asiáticos (Packer *et al*., 2006).

Outra importante ferramenta no desenho de estudos farmacogenômicos baseados em distâncias genéticas interpopulacionais é a base de dados PharmGKB: *Pharmacogenetics Knowledge Base* (www.pharmgkb.org). A base de dados PharmGKB contém informações que incluem genes, proteínas, sequências referências, regiões de interesse, haplótipos e populações dos indivíduos. Além disso, há informações sobre os fenótipos celulares, farmacocinética, cinética enzimática, descrição de fármacos, informações sobre estudos clínicos, administração e metabolismo de fármacos (Hewett *et al*., 2002).

No entanto, como pode ser observado, a maior parte das iniciativas dos grupos de pesquisa mundiais (incluído o Projeto Internacional HapMap) tem-se limitado ao estudo de populações de origem européia, africana ou asiática. Por este motivo, é importante que pesquisadores do Brasil e América Latina realizem estudos de descoberta de SNPs e determinação da estrutura haplotípica em genes de interesse biomédico e evolutivo em populações de nativos americanos e latinos americanos. Esses estudos permitirão o desenvolvimento de estudos genético-epidemiológicos mais robustos, que considerem as particularidades das populações nativas e miscigenadas de América Latina, bem como a realização de inferências evolutivas mais consistentes sobre a história dessas populações.

Além da disponibilidade de informação gerada e dos vários projetos desenvolvidos com a utilização desses dados durante a implementação das iniciativas públicas, diversas ferramentas bioinformáticas (sistemas para visualização dos dados, pacotes em ambientes de programação, formatos de arquivos) foram desenvolvidas.

Entre as ferramentas desenvolvidas, merece destaque o BioMart. O BiorMart é um sistema de código aberto para visualização e administração, que permite formular *queries* com diferentes critérios para recuperação eficiente de dados genômicos. O Biormart também é integrado com outros recursos bioinformáticos externos ao sistema, tais como: Galaxy (Giardine *et al*., 2005); BioConductor (Gentleman *et al*., 2004); the Distributed Annotation System (DAS) (Barrio *et al*., 2009; Messina e Sonnhammer, 2009); Cytoscape (Cline *et al*., 2007); e Taverna (Hull *et al*., 2006). Esta característica torna possível a integração dos dados, utilizando-se este sistema com diferentes bancos de dados. O BioMart é também parte do projeto GMOD (*Generic Model Organism Database*) http://www.gmod.org. Atualmente, Biomart tem suporte para as plataformas MySQL, Oracle e Postgres.

O modelo genérico de banco de dados para organismos modelos (*Generic Model Organism Database (GMOD) Project*) é outro conjunto de *scripts* de código aberto, desenvolvidos para visualizar e administrar banco de dados para diferentes organismos modelos. Com a utilização do sistema, é possível visualisar informações genômicas e outras informações importantes de diferentes organismos (Stein *et al*., 2002; O'connor *et al*., 2008).

Entre as plataformas, R (*R Development Core Team*, 2008; http://www.r-project.org) é um ambiente e uma poderosa linguagem de programação voltada para a manipulação de dados estatísticos, modelagem e visualização de gráficos. R tem sido cada vez mais usada em análises de dados biológicos (Schmid *et al*., 2006; Stranger *et al*., 2007; Todd *et al*., 2007; Aldrich *et al*., 2008). Diversos pacotes em R foram desenvolvidos paralelamente aos diferentes tipos de dados e análises requeridos. Esses pacotes encontram-se disponíveis no repositório do projeto, nos sítios http://www.bioconductor.org/ e http://cran.r-project.org/web/views/Genetics.html.

O formato SAM (*Sequence Alignment/Map*) e o pacote de scripts SAM*tools* foram desenvolvidos para tratar os dados gerados pelo projeto *1000genomes*. Este projeto apresenta um enfoque diferente (sequenciamento) das estratégias anteriores (genotipagem). O projeto 1000 Genomes é o primeiro projeto com o objetivo de sequenciar o genoma completo de 1000 indivíduos, bem como fornecer um mapa da variação genética (polimorfismos raros) em 12 populações.

Com base no exposto nos parágrafos anteriores, fica claro o papel fundamental da Bioinformática no desenvolvimento de ferramentas que permitam a análise e integração dos grandes volumes de dados gerados sobre a diversidade humana e de outros organismos. Atualmente, a Bioinformática é imprescindível para a manipulação de dados biológicos.

1.2 Bancos de Dados

A comunicação de dados é comprovadamente uma das atividades indispensáveis ao avanço do conhecimento. No entanto, quando a quantidade de informação é demasiadamente abundante, ela se torna de difícil preservação e, ou manipulação. Surge, então, a necessidade de ser coletada, estruturada e armazenada de forma eficiente e permanente.

Um dos principais desafios enfrentados na interface entre estudos de variabilidade genética e bioinformática é o armazenamento inteligente e eficiente dos dados biológicos gerados (Excoffier e Heckel, 2006; Cirulli e Goldstein, 2010). Dessa forma, fica sob a responsabilidade da Bioinformática possibilitar o acesso, manutenção e análise dessas informações. Os dados, por si só, não apresentam valor antes das análises e seu presente volume torna praticamente impossível, mesmo para pesquisadores experientes, interpretá-los

manualmente. Por esta razão, a criação de bancos para armazenamento de dados biológicos torna-se importante. Além disso, o enriquecimento dos dados com informações públicas complementares é importante para o entendimento dos processos biológicos aos quais eles estão relacionados (Manolio *et al*., 2009). Exemplos de informações relevantes são dados de SNPs já anotados, contidos em bancos de referência como o dbSNP (Sherry *et al*., 1999a; Smigielski *et al*., 2000; Sherry *et al*., 2001) e o HapMap (Frazer *et al*., 2007); e dados de vias metabólicas, armazenados em bancos como o KEGG (Ogata *et al*., 1999; Kanehisa *et al*., 2010).

Tabela 1 - Principais bancos de dados sobre variabilidade genética, uma breve descrição de suas principais características e o endereço (*Web-site*) nos quais os recursos podem ser acessados

Database	Descrição	Web-site
DbSNP	dbSNP database é o repositório central para SNPs e polimorfimos de inserção/deleção. Os dados no dbSNP são integrados com outros dados genômicos disponíveis no NCBI.	http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap	O projeto HapMap é uma iniciativa para catalogar SNPs em diferentes populações humanas e construir um mapa de desequilíbrio de ligação do genoma dessas populações. Esta informação é a base dos chips comerciais utilizados atualmente nos estudos de varredura genômica orientados a encontrar genes responsáveis por doenças e diferentes respostas individuais a medicamentos e fatores ambientais.	http://hapmap.ncbi.nlm.nih.gov/
1000 Genomes	O projeto 1000 Genomes é uma colaboração entre grupos de pesquisa dos US, UK, China e Alemanha e uma extensão do HapMap para sequenciar 1000 genomas humanos, usando tecnologias de nova geração para sequenciamento.	http://www.1000genomes.org/
JSNPs DATABASE	O Banco de dados de SNPs japonês tem como meta identificar cerca de 150.000 SNPs em regiões gênicas, distribuídos no genoma humano e disponibilizar essas informações publicamente para o desenvolvimento de novas ferramentas para análise desse tipo de variação genômica.	http://snp.ims.u-tokyo.ac.jp/
SNP500Cancer	SNP500cancer é parte do projeto Cancer Genome Anatomy e foi desenhado especificamente para validar SNPs em genes envolvidos em câncer, que sejam comuns nos principais grupos étnicos que formam a população dos Estados Unidos: Africanos, europeus, asiáticos e hispânicos.	http://snp500cancer.nci.nih.gov/
PharmGkb	Pharmacogenomics Knowledge base, coleta informação do impacto de variações genômicas na resposta de fármacos. O banco anota relações entre variantes gênicas e relações gene-fármaco-doença via revisão de literatura, sumarizando importantes vias de genes e fármacos.	http://www.pharmgkb.org/
KEGG PATHWAY	KEGG: Kyoto Encyclopedia of Genes and Genomes é um conjunto de bancos de dados integrados, consistindo de 16 databases com informações genômicas, vias metabólicas e informações químicas. Este banco de dados tem sido amplamente utilizado para interpretação de grandes conjuntos de dados gerados por novas tecnologias de sequenciamento.	http://www.genome.jp/kegg/

SeattleSNPs	**SeattleSNPs** é um projeto parte do National Heart Lung e Blood Institute's (NHLBI) dos Estados Unidos. Este projeto é direcionado para a identificação, genotipagem e modelagem de estudos de associação entre single nucleotide polymorphisms (SNPs) em genes candidatos e vias metabólicas relacionadas a respostas inflamatórias.	http://pga.gs.washington.edu/

Um sistema de bancos de dados é um sistema computadorizado de manutenção de registros. Os bancos de dados podem ser considerados como o equivalente eletrônico de um armário de arquivamento; ou seja, ele é um repositório ou recipiente para uma coleção de arquivos de dados computadorizados, permitindo acesso aos diferentes níveis dos dados (Date, 2003), conforme apresentado na Figura 2.

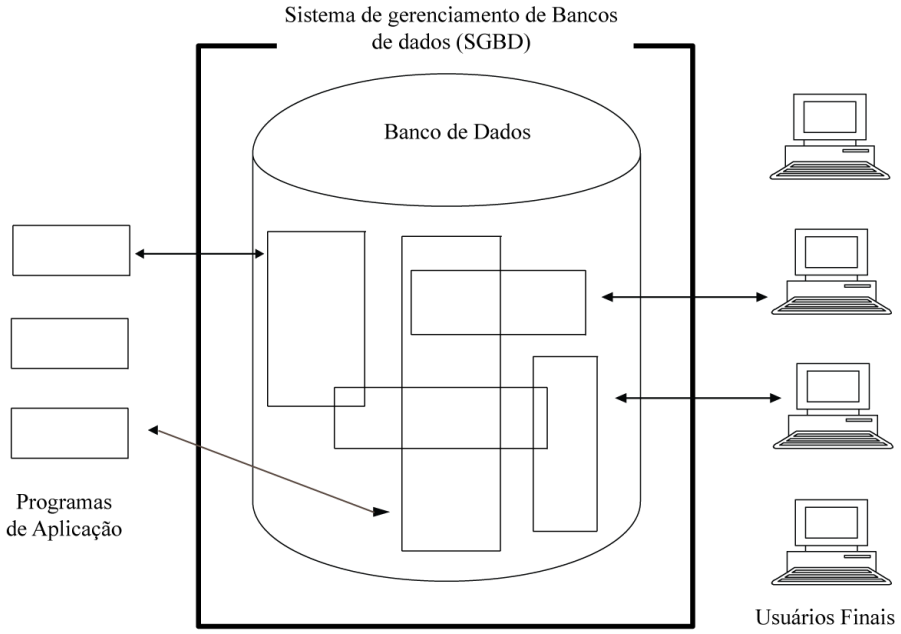


Figura 1 - Representação simplificada de um sistema de banco de dados. Modificado a partir de C J Date.

Os bancos de dados são ferramentas de extrema importância na bioinformática, pois permitem tanto o armazenamento quanto a recuperação de forma eficiente dos dados armazenados, provenientes de diferentes fontes e estudos biológicos. Existem dois tipos

principais de sistemas de gerenciamento de bancos de dados em uso atualmente: sistemas de indexação de arquivos simples e relacionais. Há um terceiro tipo, orientado a objetos, cuja popularidade está começando a aumentar (Gibas e Jambeck, 2001).

Bancos de dados de arquivos simples são a forma mais simples de bancos de dados. Estes bancos constituem uma coleção ordenada de arquivos, geralmente em conformidade com um formato padrão de conteúdo. Este modelo de banco de dados é análogo a um grande arquivo, em que a informação é recuperada por meio de ordenação e indexação dos dados nele contidos. Um índice extrai um atributo específico de um arquivo e alinha o valor do atributo no índice com um nome do arquivo e uma localização. Vários bancos de dados começaram como banco de dados de arquivos simples, também conhecidos como *flat files*, sendo que um exemplo clássico é o banco de dados de proteínas, PDB – Protein Data Bank (http://www.pdb.org/pdb/home/home.do) (Bernstein *et al*., 1977; 1978).

Um segundo tipo, o banco de dados relacionais armazena dados em tabelas separadas, cada uma contendo um conjunto de informações que pode ser combinado entre diferentes tabelas. Os dados nas tabelas são organizados em linhas, sendo que cada linha representa um registro no banco de dados. Uma linha pode conter várias informações separadas (campos) e cada campo pode conter uma informação distinta. A função dos sistemas de gerenciamento de bancos de dados consiste em fazer conexões entre diferentes tabelas relacionadas do banco, localizando rapidamente os elementos comuns entre estas que estabelecem relacionamento. A rede de tabelas e relacionamentos que compõe um banco de dados é denominada como esquema de entidade-relacionamentos.

Entre o banco de dados físico e o usuário existe uma camada de software, conhecida como sistema de gerenciamento de banco de dados (SGBD). Todas as requisições de acesso ao banco de dados são tratadas pelo SGBD. O MySQL é um dos sistemas de gerenciamento

de banco de dados relacional com código aberto mais utilizado em aplicações biológicas e está disponível para sistemas operacionais Unix e Windows, que utiliza a linguagem SQL (Linguagem de Consulta Estruturada, do inglês *Structured Query Language*) como interface. O MySQL é um SGBD estritamente relacional. No MySQL, a estrutura que mantém os blocos (ou registros) de informações são as tabelas. Por sua vez, os registros são constituídos de objetos menores, que podem ser manipulados pelos usuários, conhecidos por tipos de dados (*datatypes*). Juntos, um ou mais *datatypes* formam um registro (*record*). Uma hierarquia de banco de dados pode ser considerada como: Banco de dados > Tabela > Registro > Tipo de dados. Os tipos de dados possuem diversas formas e tamanhos, permitindo ao programador criar tabelas específicas de acordo com suas necessidades, (http://www.mysql.com/).

Um exemplo cássico de banco de dados relacional com importância capital para as ciências genômicas é o dbSNP (Sherry *et al*., 1999a; Sherry *et al*., 1999b; Sherry *et al*., 2000; Smigielski *et al*., 2000; Sherry *et al*., 2001), o maior banco de dados de variações nucleotídicas. O dbSNP é parte do National Center for Biotechnology Information (NCBI) dos Estados Unidos, sendo um dos repositórios mais importantes de polimorfismos. O dbSNP é constituído por um grande conjunto de databases espécie específicos, que contém cerca de 12 milhões de polimorfismos não redundantes. O dbSNP é um banco de dados relacional com 100 tabelas e implementado em um servidor SQL.

1.3 Ferramentas Bioinformáticas e Pipelines para estudos de genética de populações

Uma vez que os dados biológicos estão armazenados de forma consistente e estão disponíveis aos pesquisadores, há a necessidade de desenvolver métodos para extração destes a partir das bases de dados. É essencial que essas bases de dados sejam facilmente acessíveis e que as buscas sejam intuitivas, permitindo ao pesquisador recuperar dados específicos de acordo com suas necessidades. Além disso, os dados deveriam ser disponibilizados de forma clara, consistente e, quando possível, em formatos já utilizados por programas de análise, facilitando assim suas interpretações.

Ferramentas bioinformáticas são *scripts* e *software* desenvolvidos para conduzir as análises. Para o desenvolvimento dessas ferramentas, alguns aspectos devem ser observados: 1) devem ser de uso simples, não exigindo conhecimentos computacionais avançados ou prévios do usuário final; 2) devem ser acessíveis e preferencialmente de código aberto, permitindo assim um desenvolvimento contínuo; 3) deve haver confiabilidade na manipulação de dados, para evitar perda ou alteração dos mesmos; 4) robustez de execução, para tolerância a falhas na execução e prevenção de perda de resultados (Gibas e Jambeck, 2001).

Pipeline, que em português significa encadeamentos de funções, pode ser definido como um processo pelo qual dois ou mais programas podem ser executados de maneira coordenada, de forma que o output de cada um é redirecionado como input do próximo. Assim, o conjunto dos programas que são executados desta forma passa a se comportar como um novo programa, com o input direcionado ao primeiro programa e o output vindo do último. Esta automação de passos consecutivos para a realização de uma análise possibilita

aos pesquisadores analisarem eficientemente processos com múltiplos passos e grandes quantidades de dados. Este ambiente permite também o controle sistemático de erros envolvidos nas análises.

Ferramentas bioinformáticas e pipelines influenciaram a descrição e o avanço de várias áreas da biologia, desde análises de sequências, aquisição de literatura e o desenvolvimento de hipóteses da evolução de diferentes organismos (Giardine *et al*., 2005). A habilidade em processar e interpretrar grandes volumes de dados é essencial com o desenvolvimento de novas tecnologias de geração de dados.

1.4 Publicações

1.4.1 Artigo I

From Phred-Phrap-Consed-Polyphred to DNAsp: a pipeline to facilitate population genetics re-sequencing studies

O ressequenciamento de regiões-alvo é umas das estratégias mais utilizadas nos trabalhos em genética de populações, permitindo a análise da variação sem o viés de averiguação próprio de outras estratégias como a análise de SNPs ou INDELs. Dentre os vários estudos em que utilizam esses tipos de dados, podem ser citados como exemplos: inferências evolutivas em humanos (Fagundes *et al*., 2007), animais (Vargas *et al*., 2008), plantas (Novaes *et al*., 2010), microrganismos (Grynberg *et al*., 2008), estudos epidemiológicos desenhados para capturar polimorfismos raros (Parikh *et al*., 2010; Petersen *et al*., 2010), responsáveis por fenótipos complexos e estudos em famílias ou populações restritas com alta incidência de doenças genéticas específicas (Souza *et al*., 2008).

Com o objetivo de facilitar as várias etapas presentes em estudos evolutivos e genéticos que envolvem dados de ressequenciamento, foi desenvolvido um sistema online "*web-based tool*" que transforma arquivos em diferentes formatos compatíveis com vários software de genética de populações. Usando o nosso pipeline de ressequenciamento, é possível utilizar o arquivo de saída dos conjuntos de análises de sequências Phred-Phrap-Polyphred-Consed e transformá-los em arquivos de entrada para os programas PHASE e fastPHASE, bem como em formatos amplamente usados em análises genéticas, como por exemplo: SDAT e Prettybase. Com o uso do pipeline, ainda é possível utilizar as informações contidas no arquivo de saída do Phase, para gerar o arquivo de entrada do software DNAsp, com o qual podem ser calculadas diferentes estatísticas usadas em genética de populações.

Accepted to be published in Investigative Genetics Journal

Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies

Moara Machado[1,*], Wagner CS Magalhães[1,*], Allan Sene[1], Bruno Araújo[1], Alessandra C Faria-Campos[2], Stephen J Chanock[3], Leandro Scott[4], Guilherme Corrêa-Oliveira[4], Eduardo Tarazona-Santos[1], Maira R Rodrigues[1]

*These authors contributed equally to this paper

[1] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486, Belo Horizonte, MG, CEP 31270-910, Brazil.

[2] Departamento de Ciências da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha, Belo Horizonte, MG, CEP 31270-910, Brazil.

[3] Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, USA. 8717 Grovemont Circle Advanced Technology Center, Room 127, Gaithersburg, MD, 20877, USA.

[4] Genomics and Computational Biology Group and Center for Excellence in Bioinformatics, René Rachou Institute, Fundação Oswaldo Cruz, Av. Augusto de Lima 1715, Belo Horizonte, MG, 30190-002, Brazil

Corresponding Author:

Dr. Eduardo Tarazona-Santos

Departamento de Biologia Geral

Instituto de Ciências Biológicas

Universidade Federal de Minas Gerais.

Av. Antonio Carlos 6627, Pampulha.

Belo Horizonte, MG, CEP 31270-910, Brazil.

Telephone: +55 31 3409-2597

Fax: +55 31 3409-2567

E-mail: edutars@icb.ufmg.br

Authors' e-mail addresses:

Moara Machado: mmoara@terra.com.br

Wagner CS Magalhães: wcsmagalhaes@gmail.com

Allan Sene: allansenne@gmail.com

Bruno Araújo: brunoar@dcc.ufmg.br

Alessandra Faria-Campos: alessa@dcc.ufmg.br

Stephen J Chanock: chanocks@mail.nih.gov

Leandro Scott: leandroscott@cebio.org

Guilherme Corrêa Oliveira: oliveira@cebio.org

Maira Rodrigues: maira.r.rodrigues@googlemail.com

Abstract

Background

Targeted re-sequencing is one of the most powerful and widely used strategies for population genetics studies because it allows an unbiased screening for variation that is suitable for a wide variety of organisms. Examples of studies that require re-sequencing data are evolutionary inferences, epidemiological studies designed to capture rare polymorphisms responsible for complex traits and screenings for mutations in families and small populations with high incidences of specific genetic diseases. Despite the advent of Next-Generation Sequencing technologies, Sanger sequencing is still the most popular approach in population genetics studies because of the widespread availability of automatic sequencers based on capillary electrophoresis and because it is still less prone to sequencing errors, which is critical in population genetics studies. Two popular software applications for re-sequencing studies are Phred-Phrap-Consed-Polyphred, which performs base calling, alignment, graphical edition and genotype calling, and DNAsp, which performs a set of population genetics analyses. These independent tools are the start and end points of basic analyses. In between the use of these tools, there is a set of basic but error-prone tasks to be performed with re-sequencing data.

Results

To assist with these intermediate tasks, we developed a pipeline that facilitates data handling typical of re-sequencing studies. Our pipeline (1) consolidates different outputs produced by distinct Phred-Phrap-Consed contigs sharing a reference sequence; (2) checks for genotyping inconsistencies; (3) reformats genotyping data produced by Polyphred into a matrix of genotypes with individuals as rows and segregating sites as columns; (4) prepares input files for haplotype inferences using the popular software PHASE; and (5) handles PHASE output files that contain only polymorphic sites to

reconstruct the inferred haplotypes including polymorphic and monomorphic sites as required by population genetics software for re-sequencing data such as DNAsp.

Conclusion

We tested the pipeline in re-sequencing studies of haploid and diploid data in humans, plants, animals and microorganisms and observed that it allowed a substantial decrease in the time required for sequencing analyses, as well as being a more controlled process that eliminates several classes of error that may occur when handling datasets. The pipeline is also useful for investigators using other tools for sequencing and population genetics analyses.

Background

Targeted re-sequencing is one of the most powerful and widely used strategies for population genetics studies because it allows screening of variation in a way that is unbiased in respect to the allele frequency spectrum and because it is suitable for a wide variety of living organisms. Although there is a plethora of new opportunities from Next-Generation Sequencing (NGS) technologies (Mardis e Wilson, 2009), re-sequencing studies are traditionally performed using Sanger DNA sequencing. This is due in part to the widespread availability of automatic sequencers based on capillary electrophoresis and also to the fact that Sanger sequencing is still less prone to base-calling errors (Harismendy *et al.*, 2009), which is critical in population genetics studies, in which accurate identification of substitutions carried by unique chromosomes (singletons) is highly informative (Gutenkunst *et al.*, 2009). Examples of studies in different areas of genetics that require re-sequencing data are: (a) inferences of past demographic parameters of populations of humans (Fagundes *et al.*, 2007; Nielsen *et al.*, 2009), animals (Vargas *et al.*, 2008), plants (Novaes *et al.*, 2010) and microorganisms (Grynberg *et al.*, 2008), and of the action of natural selection based on ascertainment-bias-free allelic spectra (Nielsen *et al.*, 2005; Andres *et al.*, 2009; Eduardo Tarazona-Santos, 2010; Fuselli *et al.*, 2010); (b) epidemiological studies designed to capture rare polymorphisms responsible for complex traits (Bhangale *et al.*, 2008; Parikh *et al.*, 2010; Petersen *et al.*, 2010); (c) screening for variation in populations that are not included in public databases such as HapMap, to optimally select informative SNPs (tag-SNPs) for association studies (Carlson *et al.*, 2004); (d) forensic studies or analyses based on mt-DNA data (Budowle *et al.*, 2009; Budowle e Van Daal, 2009) and (e) screenings for mutations in families or small populations with high incidences of specific genetic diseases (Souza *et al.*, 2008). Two of the most popular, powerful and freely available tools for re-sequencing studies are (1) the software package Phred-Phrap-Consed-Polyphred (PPCP)

(Nickerson *et al.*, 1997; Ewing e Green, 1998; Ewing *et al.*, 1998; Gordon *et al.*, 1998; Montgomery Kt, 2008) that performs base calling, alignment, graphical edition and polymorphism identification; and (2) DNAsp (Rozas *et al.*, 2003), which performs a wide set of population genetics analyses through a user-friendly Windows interface. Because these tools were created by different groups, they are not integrated, notwithstanding their wide combined use. Frequently, they are the start and end points of basic analyses for many population genetics re-sequencing studies. In between the use of these tools, there are a set of basic but error-prone tasks to be performed with re-sequencing data. To facilitate these tasks, we developed and tested a pipeline that improves the handling of sequencing data. Although our pipeline was created with the wide community of investigators using PPCP and DNAsp in mind, it is also useful for investigators who use other sequence analysis tools, such as Sequencher [Gene Codes Corporation, US] and SeqScape [Applied Biosystems, US], or other population genetics packages, such as VariScan (Vilella *et al.*, 2005), the command-line-based version of DNAsp that is designed for large-scale datasets. Forthcoming versions of our pipeline will be integrated with forthcoming Phred-Phrap functions to analyze NGS data and with other computationally robust population genetics tools, such as the libsequence library (http://molpopgen.org/software/libsequence.html, (Thornton, 2003)).

We assume the case of an investigator who is partially or totally re-sequencing a specific genomic region in a set of individuals and that a reference sequence is available for this targeted region (Figure 1). After experimentally obtaining the re-sequencing data (usually with a minimal individual coverage of 2X using forward and reverse primers), the sequencing analyses are performed with software such as Phred-Phrap-Consed. For our purposes (i.e., population genetics studies), we define a contig as set of aligned sequences obtained from a set of individuals using the same sequencing primer or a pair of forward/reverse sequencing primers (Figure 1) with a minimum individual coverage of 2X for each sequenced base. In conjunction with Phred-Phrap-Consed, Polyphred is

frequently used to automatically identify polymorphic sites and to call genotypes for each read, but in our experience (Tarazona-Santos e Tishkoff, 2005; Fuselli *et al.*, 2007; Tarazona-Santos *et al.*, 2008; Tarazona-Santos *et al.*, 2010), visual inspection of peaks is necessary to ensure high quality data. After data production and application of Quality Control (QC) filters (e.g. based on Phred scores), the following information should be available for entry into the pipeline: (1) the sequenced regions defined by their coordinates with respect to the reference sequence and (2) for these regions, the coordinates of the observed segregating sites and their observed genotypes for each read. The pipeline assumes that this information is available in the output format of Polyphred (i.e., the Polyphred output file generated for each contig).

Implementation

Design and building

The pipeline was developed as an online system using the Perl programming language for handling dynamic scripting. The current version runs on a Linux/Apache Web server. To guarantee portability and accessibility, the system was fully tested in different operating systems and web browsers (see Section Availability and requirements).

An overview of the web-based system's architecture is shown in Figure 2. The arrows represent the flow of data and controls across the system's modules (boxes in Figure 2) and are labeled according to their order of execution. The system starts by receiving the user's choice of start and end points for the pipeline, which represent, respectively, the type of input file that the user has and the format into which the user wants to transform the original file. In accordance with the combination of these start and end points, the system determines the input files (module "Determine Input") that the user needs to provide in order to complete the chosen path through the pipeline. The required input files are presented to the user as a Web page tailored by the "Generate HTML" module. The user can then upload the input files that he or she wants to convert to the format required for a specific

population genetics program. These files are received by the system's “Coordination module,” which controls the execution of all required steps through the pipeline, including a verification step (the "Verification module”) for checking whether the provided input files are in their correct formats. Depending on the combination chosen by the user for start and end points, different scripts are invoked by the “Coordination module” (as illustrated in Figure 2). Each script has a specific functionality that is related to a determined file transformation procedure. It is important to note that the modular design of the system’s architecture is intended to facilitate future extensions of the pipeline to include other functionalities.

Web Interface

The system's external shell, behind which lies the described architecture, is the web interface illustrated in Figure 3. The gray rectangles in Figure 3 represent the steps of the pipeline that are not automated, such as PHASE and DNAsp execution. The light colored rectangles represent the modules or functionalities provided by the pipeline, which can be combined to reach the desired output. The user-friendly interface allows the user to select the desired start and end points of the pipeline by clicking within the rectangles (or modules) composing the pipeline. Whenever the user clicks on one of the rectangles, a brief explanation of the type of input file that it accepts and the output file that it generates is shown. After that, the system indicates the input files that need to be provided by the user in order to run the chosen path through the pipeline. This is performed dynamically depending on the user's choice of start and end points. After the selection of start and end points, no user intervention is needed until the final output is presented.

Results

The web interface of the pipeline is shown in Figure 3. The pipeline allows the procedures described below to be performed using a web page with a graphical and user-friendly interface

(http://www.cebio.org/pipelineldgh). Step 1 integrates different outputs produced by different Phred-Phrap-Consed-Polyphred contigs that share a reference sequence (Figure 1 B). For instance, this step can combine different exons of a gene that were independently amplified and re-sequenced, so that they might be analyzed using a shared reference genomic sequence (Figure 1 C). Step 2 reformats genotypes from reads in Polyphred output file format into a user-friendly rectangular matrix of genotypes with individuals as rows and segregating sites as columns (i.e., SDAT format) (Figure 1 D). In this step, the pipeline consolidates reads from the same individuals (sharing the same identifier) by checking for genotype inconsistencies among different reads of the same individual (e.g., forward and reverse reads of the same amplicon). In the case of diploid data, if the investigator prefers to infer haplotypes using the popular software PHASE (Stephens *et al.*, 2001), which requires multiple runs with specific parameters, the pipeline prepares the input files for PHASE (Step 3) (Figure 1 D). PHASE output files contain the inferred haplotypes for each individual but only include the segregating sites. For some population genetics analyses using re-sequencing data (e.g., DNAsp), it is necessary to reconstruct the entire sequence, including both monomorphic and polymorphic sites. Step 4 of the pipeline uses the reference sequence and the information from PHASE output files (positions of segregating sites in relation to the reference sequences and inferred haplotypes for each individual) to reconstruct for the targeted region the two DNA sequences corresponding to the two inferred haplotypes of each individual. The pipeline generates a FASTA file that may be used as input for DNAsp or other population genetics tools (Figure 1 D).

Discussion

The following features of the pipeline deserve additional commentary.

1. Data production and the use of the pipeline. There are different experimental approaches to generate data for a re-sequencing population genetics project. It is possible to continuously re-

sequence an entire region or to target specific discontinuous subregions, such as exons (Figure 1 A). To achieve these goals, different strategies that combine PCR and re-sequencing are available. For instance, it is possible to amplify regions of ~400–600 bps that will be independently re-sequenced (Packer *et al.*, 2006). It is also possible to amplify larger regions consisting of a few kilobases by long-PCR (Tarazona-Santos e Tishkoff, 2005) and to perform more than one re-sequencing reaction on each amplicon. In our experience, independent of the wet-lab strategy, two procedures are advisable to analyze the sequencing data. First, we recommend the use of a unique reference sequence for the entire genomic region, which allows unambiguous determination of the position of variable sites independently of their position on each read. Second, each set of reads that is re-sequenced using the same sequencing primers (or with forward and reverse primers) should be aligned separately (i.e., in different Phrap-Consed contigs). These procedures minimize the mix of good and bad quality calls for a specific position in the same contig, which facilitates both automatic and visual genotype calls.

When using PPCP to analyze reads in small- to medium-scale re-sequencing studies, we perform visual verification of the chromatograms. Although Polyphred genotype calls are very useful, the process is prone to mistakes, particularly for heterozygous genotypes. We observed that this miscalling happens in around 2.5% of genotype calls (in 15% of the inferred SNPs), considering good quality reads (phred scores > 30) and data generated with *Applied Biosystems* BigDye v.3.1 and run in a 3730 or 3100 *Applied Biosystems* sequencer (calculated from unpublished data from ETS and SJC on the basis of ~7Mb re-sequenced in a population genetics study). For this reason, we visually check all *Consed* chromatogram peaks that are both monomorphic (called by Phred) and polymorphic (called by Polyphred).

2. Haploid data. Our pipeline was developed keeping in mind the more general case of diploid data. However, it may be easily used with haploid data. We recommend that users interested in analyzing

haploid data follow the same procedures specified for the analysis of diploid data, assuming that all genotypes are homozygous. They should run the PHASE software to generate its output file, which is necessary to create the FASTA file required by DNAsp. Considering that, the current version of the pipeline assumes diploidy, two identical sequences will be generated for each haploid individual, and one of them should be discarded.

3. Haplotype inferences using PHASE. Although the latest version of DNAsp (v. 5.0) incorporates the algorithm implemented in the PHASE software (Stephens *et al.*, 2001), investigators may prefer running PHASE separately for several reasons: the need to use different parameters for burn-in and length of the runs; the possibility of performing the computationally demanding haplotype inferences in a more powerful computer, or the preference for the PHASE for Linux/Unix platforms, which bypasses the limitations of the Windows version. We developed the pipeline with the user who prefers to run PHASE separately in mind. However, for large datasets, inferences using PHASE may be computationally prohibitive. In this case, a faster, although less accurate method was implemented using the software fastPHASE (Scheet e Stephens, 2006). Because input files for fastPHASE and PHASE are the same, our pipeline is compatible with both programs.

4. QC procedures of the pipeline. In addition to saving time preparing input files, our pipeline has a set of QC procedures that are executed before any of the file formatting steps is performed. This includes the identification of inconsistent genotype calls for different reads of the same individuals and the verification of the different input/output files' formats.

5. Future developments. We will continue to expand the functions of our pipeline, so that it will include (a) reformatting of SDAT files to generate a Haploview input file for linkage disequilibrium analysis; (b) the option of reformatting files in both directions (e.g., being able to generate the Polyphred output from the SDAT format; (c) the option to specify if the data to be analyzed are

haploid or diploid and conveniently adapting outputs to this information; (d) the possibility of generating either the SDAT file format or the DNAsp FASTA file for diploid organisms using the IUPAC ambiguity nomenclature for heterozygous genotypes.

Conclusions

Our pipeline is designed to handle re-sequencing data and is complementary to resources such as FORMATOMATIC (Manoukis, 2007) and CONVERT (http://www.agriculture.purdue.edu/fnr/html/faculty/rhodes/students%20and%20staff/glaubitz/software.htm), which are useful for analyzing microsatellites and SNPs, but not for sequencing data. We tested our pipeline with several users who were performing re-sequencing studies of haploid and diploid loci in humans, plants, animals and microorganisms. We verified that our pipeline is robust and substantially decreases the time required for re-sequencing data analyses. Also, our pipeline allows for a more controlled process that eliminates several classes of error that may occur in population genetics, epidemiological, clinical and forensic studies when handling such data.

Author's contributions

ETS conceived the project. WCSM, AS, ETS and BA developed the scripts used in this work. MM tested different versions and parts of the pipeline and interacted with several investigators and research groups and wrote the Web service documentation. AS, BA, WCSM and MR designed and integrated the pipeline modules and developed the web interface. ETS and MR supervised the project. SJC provided resources and participated during the early parts of the project. LS provided resources for hosting and maintaining the web interface under the supervision of GCO. All the authors read and approved the final manuscript. ETS, MR and WCSM wrote the manuscript.

Availability and requirements

The Sequencing Pipeline is available at http://www.cebio.org/pipelineldgh.

The web-based system will be freely available for academic purposes.

Operating systems: Windows, 32-bit Linux, 64-bit Linux, MAC-OS.

Programming languages: Perl, HTML and JavaScript.

Browsers: Internet Explorer (Windows), Firefox (Linux, Windows), Safari (MAC-OS)

Acknowledgments

We are grateful to Flavia Siqueira, Rodrigo Redondo, Renata Acacio, Sharon Savage and Charles Chung for helping us test the pipeline and to the Bioinformatics group of the Core Genotyping Facilities of NCI for their participation in discussions about the pipeline. This work is supported by the National Institutes of Health – Fogarty International Center (1R01TW007894-01 to ETS), Brazilian National Research Council (CNPq), Brazilian Ministry of Education (CAPES Agency) and Minas Gerais State Foundation in Aid of Research (FAPEMIG).

References

1. Mardis ER, Wilson RK: **Cancer genome sequencing: a review.** *Human Molecular Genetics* 2009, **18:**R163-R168.

2. Harismendy O, Ng PC, Strausberg RL, Wang XY, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biology* 2009, **10:**-.

3. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD: **Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data.** *Plos Genetics* 2009, **5:**-.

4. Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L: **Statistical evaluation of alternative models of human evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**17614-17619.

5. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, et al: **Darwinian and demographic forces affecting human protein coding genes.** *Genome Research* 2009, **19:**838-849.

6. Vargas SM, Araujo FCF, Monteiro DS, Estima SC, Almeida AP, Soares LS, Santos FR: **Genetic diversity and origin of leatherback turtles (Dermochelys coriacea) from the Brazilian coast.** *Journal of Heredity* 2008, **99:**215-220.

7. Novaes RML, De Lemos JP, Ribeiro RA, Lovato MB: **Phylogeography of Plathymenia reticulata (Leguminosae) reveals patterns of recent range expansion towards northeastern Brazil and southern Cerrados in Eastern Tropical South America.** *Molecular Ecology* 2010, **19:**985-998.

8. Grynberg P, Fontes CJF, Hughes AL, Braga EM: **Polymorphism at the apical membrane antigen 1 locus reflects the world population history of Plasmodium vivax.** *Bmc Evolutionary Biology* 2008, **8:**-.

9. Eduardo Tarazona-Santos, Cristina Fabbri, Meredith Yeager, Wagner C.S. Magalhaes, Laurie Burdett, Andrew Crenshaw, Davide Pettener, Stephen J. Chanock: **Diversity in the Glucose Transporter-4 Gene (SLC2A4) in Humans Reflects the Action of Natural Selection along the Old-World Primates Evolution.** *PloS One* 2010, **5:**e9827.

10. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data.** *Genome Res* 2005, **15:**1566-1575.

11. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, et al: **Targets of Balancing Selection in the Human Genome.** *Molecular Biology and Evolution* 2009, **26:**2755-2764.

12. Fuselli S, de Filippo C, Mona S, Sistonen J, Fariselli P, Destro-Bisol G, Barbujani G, Bertorelle G, Sajantila A: **Evolution of detoxifying systems: the role of environment and population history in shaping genetic diversity at human CYP2D6 locus.** *Pharmacogenetics and Genomics* 2010, **20:**485-499.

13. Parikh H, Deng ZM, Yeager M, Boland J, Matthews C, Jia JP, Collins I, White A, Burdett L, Hutchinson A, et al: **A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33.** *Human Genetics* 2010, **127:**91-99.

14. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, Arslan AA, Bueno-de-Mesquita HB, Gallinger S, Gross M, et al: **A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33.** *Nature Genetics* 2010, **42:**224-U229.

15. Bhangale TR, Rieder MJ, Nickerson DA: **Estimating coverage and power for genetic association studies using near-complete variation data.** *Nature Genetics* 2008, **40:**841-843.

16. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *American Journal of Human Genetics* 2004, **74:**106-120.

17. Budowle B, Ge JY, Aranda XG, Planz JV, Eisenberg AJ, Chakraborty R: **Texas Population Substructure and Its Impact on Estimating the Rarity of Y STR Haplotypes from DNA Evidence*.** *Journal of Forensic Sciences* 2009, **54:**1016-1021.

18. Budowle B, van Daal A: **Extracting evidence from forensic DNA analyses: future molecular biology directions.** *Biotechniques* 2009, **46:**339-+.

19. Souza CP, Valadares ER, Trindade ALC, Rocha VL, Oliveira LR, Godard ALB: **Mutation in intron 5 of GTP cyclohydrolase 1 gene causes dopa-responsive dystonia (Segawa syndrome) in a Brazilian family.** *Genetics and Molecular Research* 2008, **7:**687-694.

20. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Research* 1998, **8:**175-185.

21. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Research* 1998, **8:**186-194.

22. Gordon D, Abajian C, Green P: **Consed: A graphical tool for sequence finishing.** *Genome Research* 1998, **8:**195-202.

23. Nickerson DA, Tobe VO, Taylor SL: **PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing.** *Nucleic Acids Research* 1997, **25:**2745-2751.

24. Montgomery KT IO, Li L, Loomis S, Obourn V, Kucherlapati R.: **PolyPhred analysis software for mutation detection from fluorescence-based sequence data.** *Current Protocol in Human Genetics* 2008, **Oct:**Chapter 7:Unit 7.16.

25. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods.** *Bioinformatics* 2003, **19:**2496-2497.

26. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J: **VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data.** *Bioinformatics* 2005, **21:**2791-2793.

27. Thornton K: **libsequence: a C++ class library for evolutionary genetic analysis.** *Bioinformatics* 2003, **19:**2325-2327.

28. Tarazona-Santos E, Tishkoff SA: **Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus.** *Genes and Immunity* 2005, **6:**53-65.

29. Tarazona-Santos E, Bernig T, Burdett L, Magalhaes WCS, Fabbri C, Liao J, Redondo RA, Welch R, Yeager M, Chanock SJ: **CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains.** *Hum Mutat* 2008, **29:**623-632.

30. Fuselli S, Gilman RH, Chanock SJ, Bonatto SL, De Stefano G, Evans CA, Labuda D, Luiselli D, Salzano FM, Soto G, et al: **Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity.** *Pharmacogenomics Journal* 2007, **7:**144-152.

31. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *American Journal of Human Genetics* 2001, **68:**978-989.

32. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi LQ, Sicotte H, Staats B, Acharya M, Crenshaw A, et al: **SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes.** *Nucleic Acids Research* 2006, **34:**D617-D621.

33. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase.** *American Journal of Human Genetics* 2006, **78:**629-644.

34. Manoukis NC: **FORMATOMATIC: a program for converting diploid allelic data between common formats for population genetic analysis.** *Molecular Ecology Notes* 2007, **7:**592-593.

Figure legends

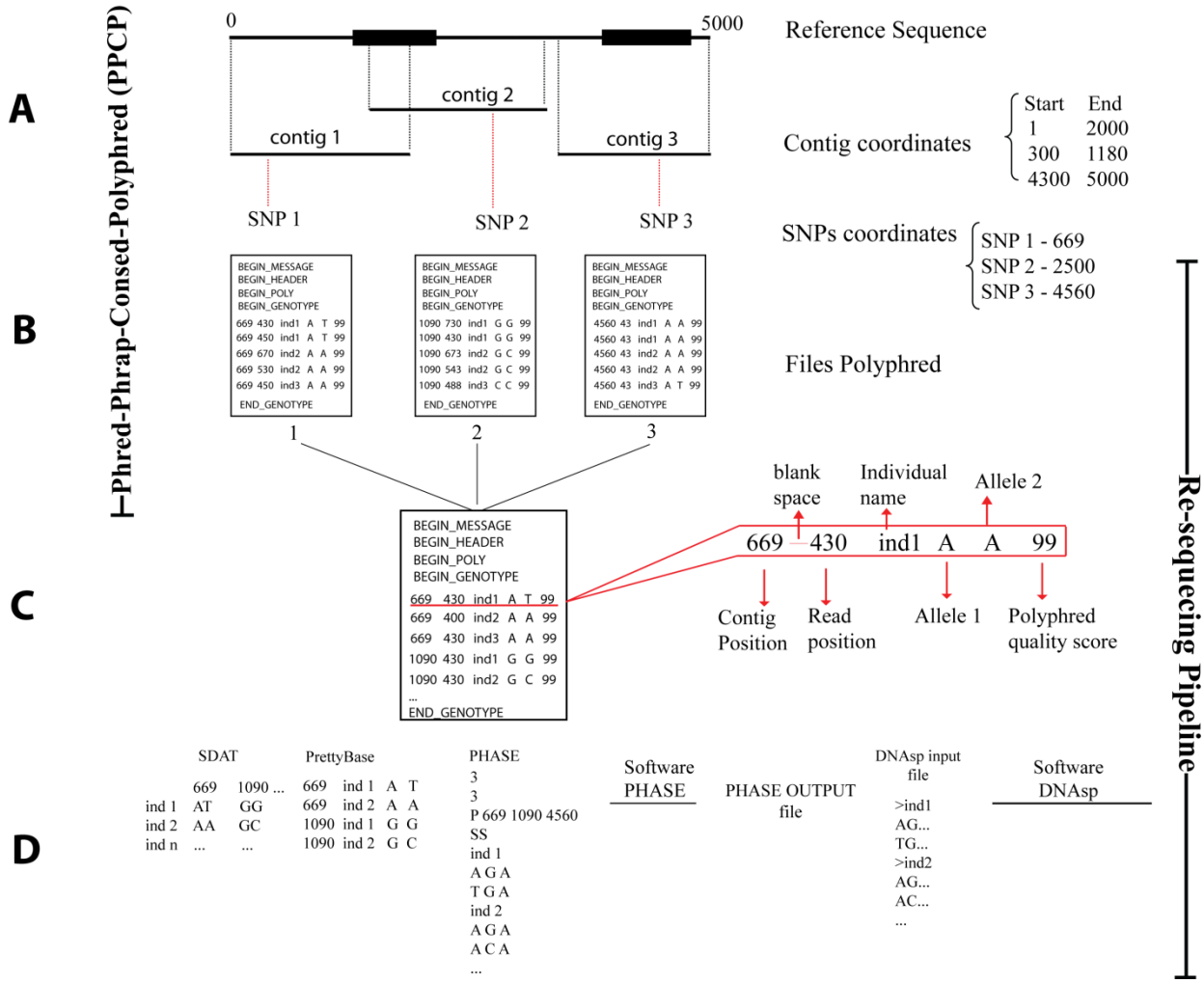


Figure 1 - Overview of the re-sequencing data analysis process integrated with the pipeline. Example of an entire process of re-sequencing analysis for a specific genomic region from a set of individuals. (A) Re-sequencing steps, base calling, alignment and heterozygous site identification for an entire region sharing a reference sequence; (B) PolyPhred output files of discontinuous sub-regions, such as exons, re-sequenced independently; (C) Consolidation of different exons of a gene that were independently amplified and re-sequenced using the same reference genomic sequence; (D) Files formats that can be handled by the pipeline either as input or output files.

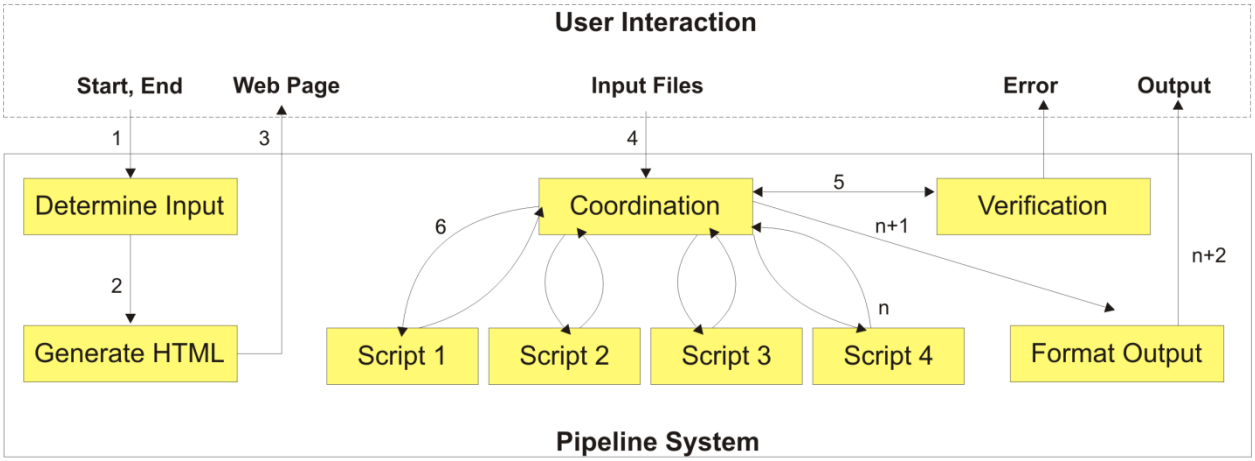


Figure 2 - System's background operation and user interaction. The arrows represent the flow of data and controls across the system's modules (boxes) and are labeled according to their order of execution. The system starts by receiving the user's choice of start and end points for the pipeline, which represent, respectively, the type of input file that the user has and the format into which the user wants to transform the original file. In accordance with the combination of these start and end points, the system determines the input files (module "Determine Input") that the user needs to provide in order to complete the chosen path through the pipeline. The required input files are presented to the user as a Web page tailored by the "Generate HTML" module. The user can then upload the input files that he or she wants to convert to the format required for a specific population genetics program. These files are received by the system's "Coordination module," which controls the execution of all required steps through the pipeline, including a verification step (the "Verification module") for checking whether the provided input files are in their correct formats. Depending on the combination chosen by the user for start and end points, different scripts are invoked by the "Coordination module". These scripts generate outputs that are presented to the user through the "Format output module".

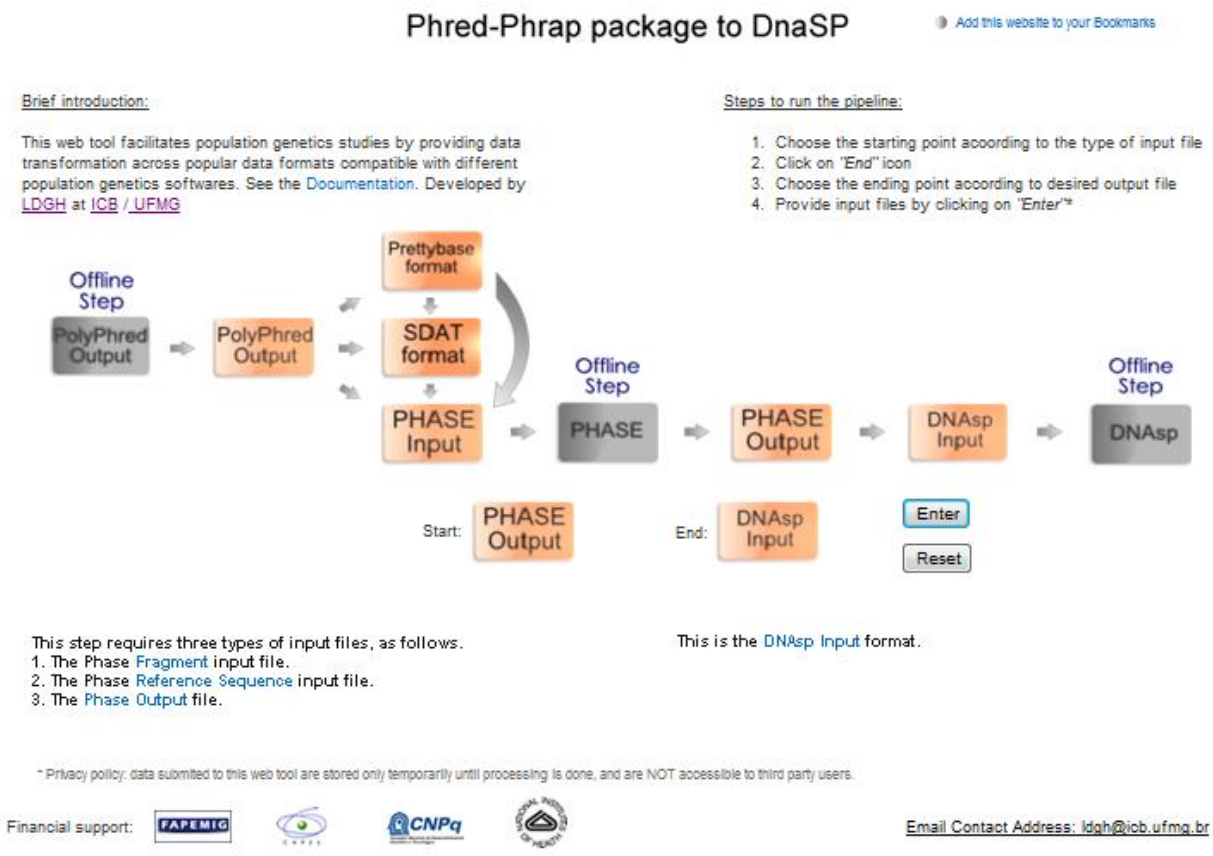


Figure 3 - Overview of the main pipeline's web interface. The user-friendly interface It allows the user to select the desired input and output file formats by clicking within the rectangles (modules) composing the pipeline. The web page includes a brief introduction to the pipeline with access to end-user documentation, and a description of the basic steps needed to run the pipeline. Contact email address is provided to guarantee permanent user support.

1.4.2 Manuscrito I

DIVERGENOME: a bioinformatics tool to assist the analysis of genetic variation

A plataforma bioinformática DIVERGENOME foi desenvolvida com o objetivo de facilitar o armazenamento, a recuperação e análise de dados provenientes de estudos de genética de populações e epidemiologia genética. A plataforma é dividida em dois componentes: um banco de dados relacional, o DIVERGENOMEdb; e um conjunto de ferramentas para facilitar a análise dos dados, o DIVERGENOMEtools. Os objetivos específicos da proposta são: (1) desenvolver um banco de dados, DIVERGENOMEdb, que organize, reúna e relacione uma série de informações genotípicas e fenotípicas de indivíduos participantes em estudos de genética de populações e epidemiologia genética; (2) desenvolver ferramentas de compatibilidade de dados, o DIVERGENOMEtools, que permitam a utilização dos dados armazenados no DIVERGENOMEdb pelos programas que compõem os procedimentos de análise de dados nos estudos-alvo; (3) aplicar técnicas de integração de dados para enriquecimento do banco DIVERGENOMEdb com informações relevantes de outros bancos de dados biológicos. Por exemplo: para estudos de associação, estudos epidemiológicos com informações complementares para o entendimento dos processos biológicos aos quais eles estão relacionados; atualmente estamos implementando (4) um método para combinar as funcionalidades das ferramentas desenvolvidas, de forma a permitir a composição de procedimentos mais complexos de análise de dados, criando dessa forma um *pipeline dinâmico*. Ainda com o objetivo de facilitar a recuperação dos dados e tornar sua manipulação mais intuitiva, também estamos desenvolvendo uma interface web para todo o sistema DIVERGENOME. DIVERGENOMEdb, que tem também as funções: (a) servir como repositório de dados genéticos para um laboratório de médio porte, de tal forma que os

dados produzidos pelo grupo se encontrem sempre disponíveis mesmo depois que estudantes e posdocs deixaram o grupo; (b) a totalidade dos dados de cada projeto, incluindo dados produzidos por um grupo e dados de comparação, pode ficar armazenada em DIVERGENOMEdb, podendo ser disponibilizada como material suplementar das publicações, o que facilita a reprodutibilidade das análises realizadas.

(*To be submitted to NAR, Bioinformatics or BMC Bioinformatics*)

DIVERGENOME: a bioinformatics platform to assist population genetics and genetic epidemiology studies

Wagner C. S. Magalhães[1*], Maíra Rodrigues[1*], Donnys Silva[1], Márcia L. Iannini[1], Gustavo C. Cerqueira[3], Alessandra A. Faria-Campos[2], Eduardo Tarazona-Santos[1#]

*These authors contributed equally to this paper

[1]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486, Belo Horizonte, MG, CEP 31270-910, Brazil.

[2]Departamento de Ciência da Computação, Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, Pampulha, Belo Horizonte, MG, CEP 31270-910, Brazil.

[3]Institute of Genome Sciences, University of Maryland, Baltimore Street BioPark II, 6th floor Baltimore, MD, 21201 US.

[#] Corresponding author:

Eduardo Tarazona-Santos

Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486, Belo Horizonte, MG, CEP 31270-910, Brazil. Tel: ++55 31 3409-2572

Fax: ++55 31 3409-2567

E-mail: edutars@icb.ufmg.br

ABSTRACT

DIVERGENOME is a web accessible open-source platform (http://localhost/divergenome) developed to help investigators in data storage and analysis for population genetics and genetic epidemiology studies. The platform contains two components. The first component, DIVERGENOMEdb, is a relational database developed using MySQL. It allows to safely storing individual genotypes from different types of data such as contigs (resulted from re-sequencing projects), SNPs/INDELs and microsatellites. Genotype data can be linked to a description of the protocols used to generate them. Individuals can be linked to populations, as well as to individual phenotypic information that are collected in biomedical studies, allowing using different kinds of variables. The database structure permits easy integration with other data types, including public databases such as the HapMap project, opening prospects for future implementations. The second component, DIVERGENOMEtools, is a dynamic pipeline composed of a set of scripts, developed using a graph-based coordination algorithm and implemented in the programming language Perl. It enables the conversion of either queries submitted to the database as well as independent files to many popular file formats required by popular population genetics and genetic epidemiology software.

Key words: Databases, Genetic Epidemiology, Population Genetics, tools

INTRODUCTION

The production of biological data by high-throughput technologies has revolutionized Biology. In genetics, classical and emerging scientific questions are being approached using SNPs and CNVs genotyping and Next Generation Sequencing (NGS) platforms (1-3). Today, the body of investigators in biology is composed by few big research groups that produce high-throughput data, and thousands of small- and medium-size groups that, in addition to produce smaller amounts of data, use and integrate the data produced by the former to resolve relevant scientific questions. While large-scale genomics initiatives such as the HapMap project, CGEMs and the 1000-genomes rely on powerful computational and bioinformatics support to assist in the production and analyses of data (4), there are very few bioinformatics platforms oriented to small-medium groups to storage, handle and integrate data from different sources, as well as to assist in efficiently performing different kinds of analyses. As a consequence, these tasks are frequently performed sub-optimally, frequently handling data files manually, which is an error prone task that is seldom coupled with adequate quality control procedures. Here we developed a bioinformatics platform, DIVERGENOME, to assist population genetics and genetic epidemiology studies performed by small-medium scale research groups. The platform is composed by two components: 1) **DIVERGENOMEdb**, a relational database developed using MySQL, and 2) **DIVERGENOMEtools,** a set of data conversion tools for many popular file formats required by population genetics and genetic epidemiology. These tools are organized into a *dynamic pipeline*. DIVERGENOMEdb allows to safely storing individual genotypes from different types of polymorphisms: contigs (resulted from re-sequencing projects), SNPs/INDELs, and microsatellites. Genotypes can be linked to a description of the laboratory protocols used to generate them. Individuals can be linked to populations, as well as phenotypes that may be

collected in genetic epidemiology studies, allowing for different kinds of variables. In DIVERGENOMEtools**,** each tool is an independent module that receives an input file with format A, performs some conversion task on the input file and returns an output file with format B. Different tools are combined in a dynamic conversion pipeline that increase the number of data format conversions available to the user. We use a dynamic implementation for the pipeline to cover a major drawback in currently available pipelines designed in a static way (with the execution steps hardcoded into programs and scripts): the inclusion of new tools is costly in terms of manual and error prone tasks. In such cases, it needs an experienced programmer to change the hardcoded steps to include new tools in a static pipeline, while guaranteeing its well functioning. This is a big concern if we want to develop pipelines that are continuously updated with new software developments. The dynamic pipeline approach makes DIVERGENOMEtools an easily extendable system that can keep up with the constant developments in the bioinformatics field. Because DIVERGENOMEdb and DIVERGENOMEtools are integrated, data extracted from the database may be analyzed using the tools. Moreover, DIVERGENOME is open-source, freely available software, and can be accessed from the command line or through a web interface.

Implementation

Design and building

DIVERGENOMEdb stores and links information on genotypes, polymorphisms, individuals, populations, and individual phenotypes. The design of our relational database which entity-relationship diagram is shown in Supplementary Figure 1 which may be divided in three parts: (A) the first one is responsible to manage data from populations, individuals,

quantitative and qualitative variables as well information of biological samples. (B) The second part allows defining Projects, that are a set of individuals (from the first part) screened for a set of polymorphisms or a genomic region. The access to the data occurs through Projects defined by users to manage their data which can be set as public (may be visualized to unregistered users) or private (may accessed only by users which permission was given by the coordinator of the project); and (C) The third part stores the individual genotypes and polymorphisms information, as well as their annotations (e.g. dbSNP code (rs#) when available, gene, a reference sequence, the dbSNPs links). Genetic variation information stored on DIVERGENOMEdb can be retrieved and used to run several population genetics and genetic epidemiology software with the assistance of DIVERGENOMEtools. The design adopted for DIVERGENOMEdb enables to easily incorporate new instances to the database, which may be accommodated into the graphical interface. **DIVERGENOMEdb** has been hosted using the MySQL version 5.1.45 (http://www.mysql.org/) database management system. The software DBDesigner 4.0.5.6 (http://www.fabforce.net/dbdesigner4) was used to develop the data model project. The whole system is hosted in a Unix-based server running the Apache Web server and can be downloaded and hosted locally.

Registration and Data Entry

In DIVERGENOME, data entry and modification are possible only for registered users. There are three levels of registered users, as outlined below in hierarchical order:

(i) Administrators have full access to all database functionalities and contents.

(ii) Project Coordinators have data entry and modification rights and can register and create accounts for project members within Projects.

(iii) Project members can download and search public data as well as those data from their respective projects (on which the coordinator had given access rights).

Additional information can be accessed on the platform documentation.

Tools

DIVERGENOMEtools is a dynamic pipeline composed of a set of conversion tools (modules) for popular population genetics and genetic epidemiology software. These tools were developed using the Perl programming language. We designed the pipeline to have two properties. First, it is easily extensible, so that new tools can be incorporated to the platform at any time. Second, we maximized conversion functionalities offered to the user, so that simple tools can be combined to provide a bigger variety of possible conversions. To achieve these properties, we designed each conversion tool as an independent module that simply receives an input file in format A, performs some processing on the input file and returns an output file in format B. In addition, we use a dynamic pipeline to combine these tools functionalities in a coordinated mode, by passing the output of one module as the input of the next module and so forth. Dynamism is achieved through a graph-based approach in pipeline implementation (Rodrigues et al. in preparation). The idea is to represent the connectivity of tools with a directed graph (5) in which data or file formats are the graph vertexes and programs or scripts that process them (via format conversion) are the graph edges. Therefore, if there is an edge (E) connecting two vertexes (A) and (B), being (E) the incoming edge of (B) and the outgoing edge of (A), it means that script (E) receives data or file format (A) as input and generates format (B) as output. The actual implementation of this graph-based approach comprises four elements: (i) a tool Registry containing the list of conversion tools and their accepted input and output data formats, (ii) a graph representation of the Registry,

(iii) a graph-traversing algorithm that finds a path between two points in the graph, and (iv) the dynamic pipeline algorithm that coordinates all previous elements. The later algorithm works generally as follows: (1) receives as input the Registry file and the start (original file format) and end (desired file format) points of the pipeline chosen by the user; (2) builds a graph based on the Registry file; (3) applies the graph-traversing algorithm to find a path through the graph connecting formats A and B received as input; (4) executes the path returned in step 3 (FIGURE 1). The path through the graph is actually the sequence of tools that need to be executed to generate the user's desired output file format. With this approach, to incorporate a new tool into the pipeline, we need only to update the tool Registry, and the dynamic pipeline algorithm is responsible for generating the new pipeline "on-the-fly", during execution. We are currently using Dijkstra's algorithm as the graph-traversing algorithm in step (3) above. Dijkstra's algorithm implements a solution for the "*travelling salesman problem (TSP)*", one of the most intensively studied problems in computational mathematics (6). One analogy with the *travelling salesman problem* may be done with each tool (module) representing a city: the first input file represents the present position (starting point), the desired output file format represents the final destination (ending point), and the best combination of tools to convert one to another represents the shortest pathway between the cities (FIGURE 1). With the combined conversion tools provided by the pipeline, investigators will be able to visualize their data in different formats and as input files for different population genetics and statistical software, thus facilitating its analyses. At the moment, the following population genetics packages are covered: *PHASE (7)*, *FastPHASE* (8), *DNAsp (9)*, *Haploview (10)*, Haplopainter (11), *STRUCTURE* (12), SWEEP (http://www.broadinstitute.org/mpg/sweep/index.html), and common file formats (SDAT, Prettybase and Pedigree) handled by genetic epidemiology software as GLU (http://code.google.com/p/glu-genetics/) and PLINK

(http://pngu.mgh.harvard.edu/~purcell/plink/). It is important to note that the modular and dynamic design of the pipeline system's architecture are intended to facilitate future extensions of the pipeline to include other functionalities.

Web Interface

DIVERGENOME is accessed through a web-based interface offering users a simple interaction and friendly navigation. The Web interface implements scripts that perform requests to the MySQL server and the Apache web server (http://www.apache.org), thus connecting DIVERGENOMEdb and DIVERGENOMEtools.

To guarantee portability and accessibility, the system was tested in different operating system's and web browsers.

RESULTS and DISCUSSION

Tools options, files and diagnostics

DIVERGENOME currently supports 9 different target programs, including many commonly used programs, such as *PHASE (7)*, *FastPHASE* (8), *DNAsp (9)*, *Haploview (10)*, *STRUCTURE* (17), Haplopainter (11), SWEEP, GLU and PLINK.. It also accepts 11 different file formats. Each conversion tool has its own internal control that validates the input file and only after that converts it to the desired file format, otherwise an error message is returned.

Study of case:

Diversity in the Glucose Transporter-4 Gene (*SLC2A4*) in Humans Reflects the Action of Natural Selection along the Old-World Primates Evolution

Glucose is an important source of energy for living organisms. In vertebrates, it can be ingested with the diet and transported into the cells by conserved mechanisms and molecules, such as the trans-membrane Glucose Transporters (GLUTs) protein family. Members of this family have tissue specific expression, biochemical properties and physiologic functions that together, contribute to the regulation of blood sugar levels as well as its distribution. GLUT4 –coded by *SLC2A4* (chromosome 17p13), is an insulin sensitive glucose transporter with a critical role in glucose homeostasis (15-16). All data handling for population genetics analyses (i.e. haplotype phasing inference, extended-haplotype-homozygosity statistic) for this work were performed using a set of scripts from the platform DIVERGENOME.

The integration between phenotypic and genotypic data achieved using our platform allows an efficient use of many qualitative and quantitative traits commonly collected in epidemiological studies that now may be incorporated as co-variants in analysis of genome-wide association studies. The inferred cross-link between genomic and phenotypic information allows access to a large body of information to find answers to several biological questions. The database structure also permits easy integration with other data types and opens up prospects for future implementations.

In particular, our database will be storing data producing by different genome-wide association studies in Latin America populations, for instance, the EPIGEN/Brazil project,

which aims to genotyping ~7000 individuals from three Brazilian cohorts with at least 10 years of study for different clinical outcomes.

Availability

DIVERGENOME can be accessed freely at http://hosted/divergenome

Author's contributions

ETS conceived the project. WCSM, DS, ETS and MR developed the project. ETS and MR supervised the project. All the authors read and approved the final manuscript. ETS, MR and WCSM wrote the manuscript.

FUNDING

This work is supported by the National Institutes of Health – Fogarty International Center (1R01TW007894-01 to ETS), Brazilian National Research Council (CNPq), Brazilian Ministry of Education (CAPES Agency) and Minas Gerais State Foundation in Aid of Research (FAPEMIG).

ACKNOWLEDGEMENTS

We thank ….

REFERENCES

1. Gilad, Y., Pritchard, J.K. and Thornton, K. (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends Genet*, **25**, 463-471.

2. Mardis, E.R. and Wilson, R.K. (2009) Cancer genome sequencing: a review. *Human Molecular Genetics*, **18**, R163-R168.

3. Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X.Y., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. *et*

al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, -.

4. Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J. *et al.* (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451-1455.

5. Cormen, T.H.L., Charles E.; Rivest, Ronald L.; Stein, Clifford. (2001) *Introduction to Algorithms*. Second Edition ed. MIT Press and McGraw-Hill, Cambridge.

6. Cormen, T.H.L., Charles E.; Rivest, Ronald L.; Stein, Clifford. (2001), *Introduction to Algorithms*. Second ed. ed. MIT Press and McGraw-Hill, Cambridge, pp. 595–601.

7. Stephens, M., Smith, N.J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978-989.

8. Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629-644.

9. Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496-2497.

10. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263-265.

11. Thiele, H. and Nurnberg, P. (2005) HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics*, **21**, 1730-1732.

12. Falush, D., Stephens, M. and Pritchard, J.K. (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567-1587.

13. Packer, B.R., Yeager, M., Burdett, L., Welch, R., Beerman, M., Qi, L.Q., Sicotte, H., Staats, B., Acharya, M., Crenshaw, A. *et al.* (2006) SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Research*, **34**, D617-D621.

14. Staats, B., Qi, L.Q., Beerman, M., Sicotte, H., Burdett, L.A., Packer, B., Chanock, S.J. and Yeager, M. (2005) Genewindow: an interactive tool for visualization of genomic variation. *Nature Genetics*, **37**, 109-110.

15. Olson, A.L. and Pessin, J.E. (1996) Structure, function, and regulation of the mammalian facilitative glucose transporter gene family. *Annu Rev Nutr*, **16**, 235-256.

16. Huang, S. and Czech, M.P. (2007) The GLUT4 glucose transporter. *Cell Metab*, **5**, 237-252.

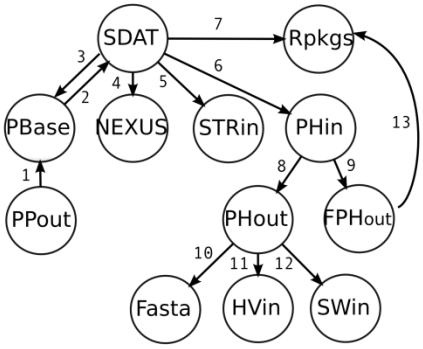
17. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

Figure 1-

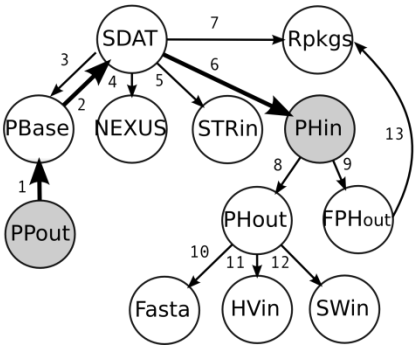
1) Tool Registry

Input	Output	Tool	Tool_ID
PPout	PBase	PP2PB	1
PBase	SDAT	PB2SDAT	2
SDAT	PBase	SDAT2PB	3
SDAT	NEXUS	SDAT2NX	4
SDAT	STRin	SDAT2STR	5
SDAT	PHin	SDAT2PH	6
SDATA	Rpkgs	SDAT2RP	7
PHin	PHout	PHASE	8
PHin	FPHout	fastPHASE	9
PHout	Fasta	PH2Fasta	10
PHout	HVin	PH2HV	11
PHout	SWin	PH2SW	12
FPHout	Rpkgs	FPH2RP	13

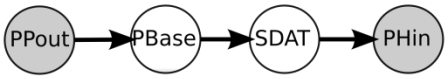
2) Tool Graph



3) Graph Traversing Algorithm (from PPout, to PHin)



4) Resulting Dynamic Pipeline




```
1  input: PPout, PHin, File_PPout
2
3  File_PBase <-- PP2PB(File_PPout)
4  File_SDAT  <-- PB2SDAT(File_PBase)
5  File_PHin  <-- SDAT2PH(File_SDAT)
6
7  return File_PHin
```

Overview of the DIVERGENOMEtools. 1) Tool Registry it shows the table which contains the list of scripts (tools), input and outputs available. 2)Tool Graph – describes the relationship between the formats and the scripts. 3) Graph Traversing Algorithm – it shows the input and output selected and the path (bold). 4) Resulting Dynamic Pipeline, linear representation of the scripts and the command line which will be executed.

Supplementary figure – 1

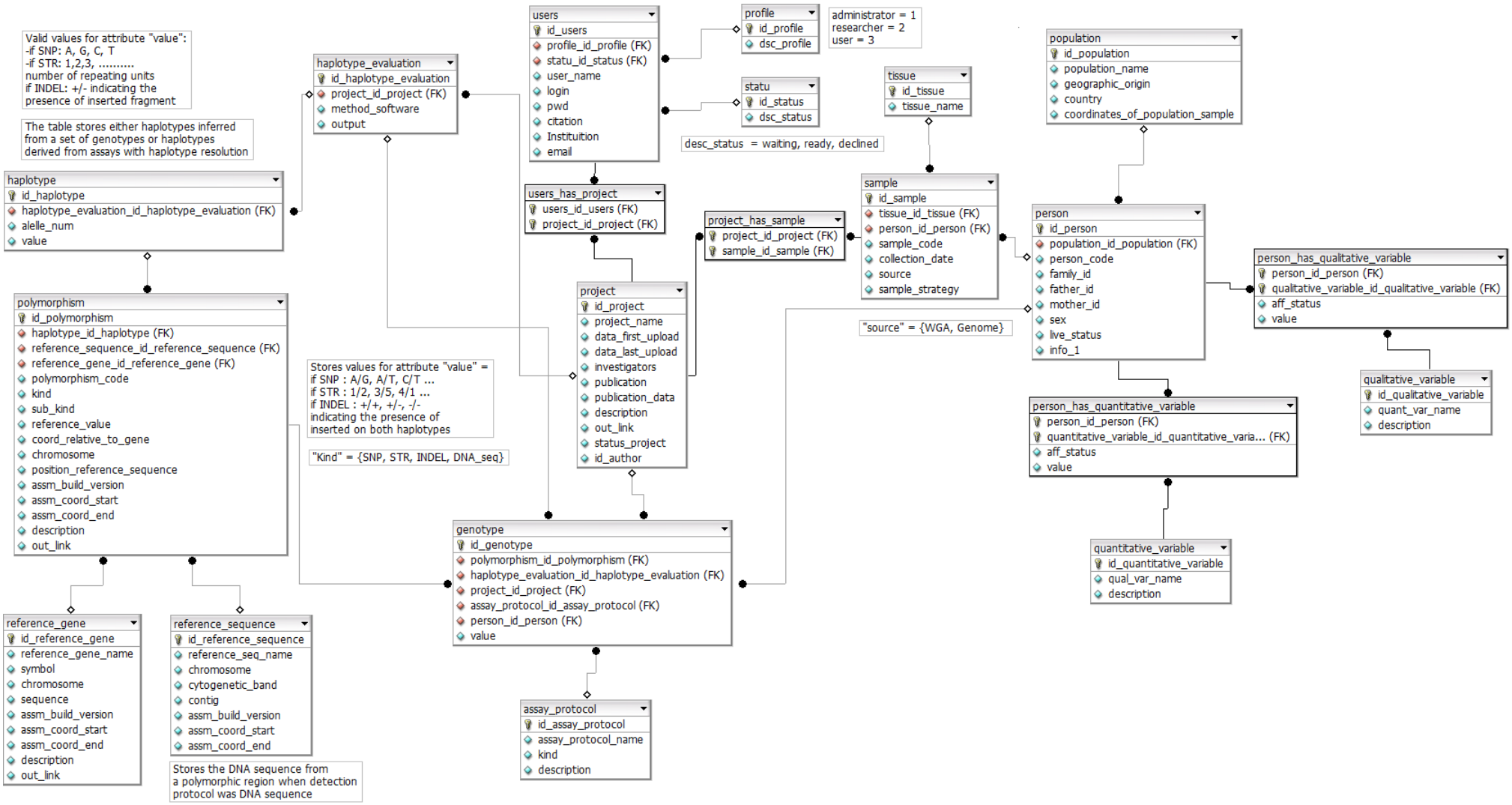
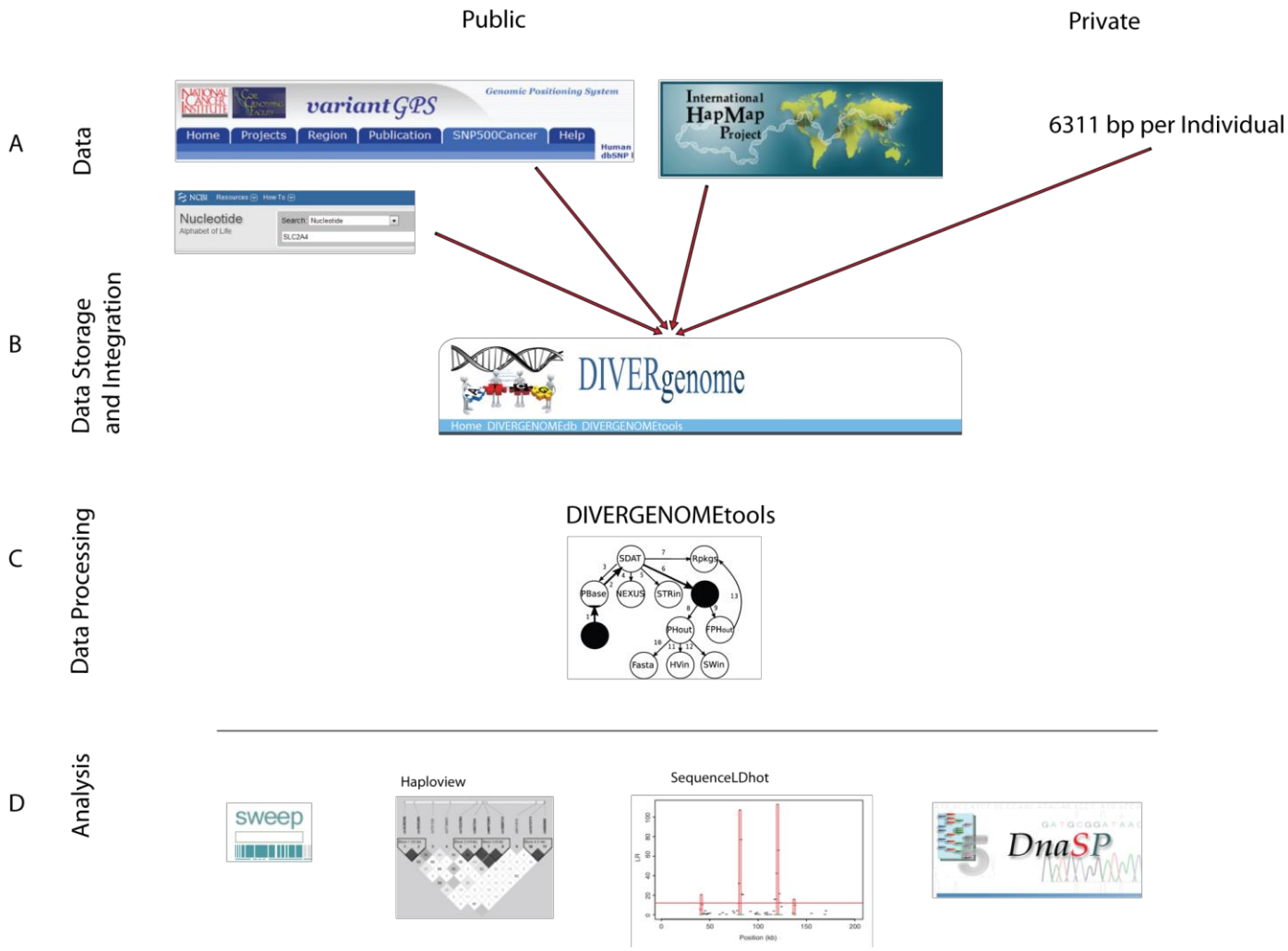


Diagram of Entitity- Relationship model (DER).



Study of case outlining the platform functionalities. A) Data source -shows data integration from different sources (public and private). B) Data Storage and integration – Using DIVERGENOMEdb, data might be manipulated and combined allowing users recovery specific data subsets according to their biological question. C) Data processing - DIVERGENOMEtools, a set of scripts which allows convert data files formats to be used in different program analysis. Analysis – an example of some software commonly used for population genetics.

1.5 Referências

ALDRICH, M. C. *et al.* Comparison of statistical methods for estimating genetic admixture in a lung cancer study of African Americans and Latinos. *Am J Epidemiol* [S.I.], v. 168, n. 9, p. 1035-46, Nov 1 2008.

ANDRES, A. M. *et al.* Targets of Balancing Selection in the Human Genome. *Molecular Biology and Evolution* [S.I.], v. 26, n. 12, p. 2755-2764, Dec 2009.

BARRIO, A. M. *et al.* Annotation and visualization of endogenous retroviral sequences using the Distributed Annotation System (DAS) and eBioX. *Bmc Bioinformatics* [S.I.], v. 10, p. -, 2009.

BERNSTEIN, F. C. *et al.* Protein Data Bank - Computer-Based Archival File for Macromolecular Structures. *European Journal of Biochemistry* [S.I.], v. 80, n. 2, p. 319-324, 1977.

_____. Protein Data Bank - Computer-Based Archival File for Macromolecular Structures. *Archives of Biochemistry and Biophysics* [S.I.], v. 185, n. 2, p. 584-591, 1978.

BHANGALE, T. R. *et al.* Estimating coverage and power for genetic association studies using near-complete variation data. *Nature Genetics* [S.I.], v. 40, n. 7, p. 841-843, Jul 2008.

BUDOWLE, B. *et al.* Texas Population Substructure and Its Impact on Estimating the Rarity of Y STR Haplotypes from DNA Evidence*. *Journal of Forensic Sciences* [S.I.], v. 54, n. 5, p. 1016-1021, Sep 2009.

BUDOWLE, B.; VAN DAAL, A. Extracting evidence from forensic DNA analyses: future molecular biology directions. *Biotechniques* [S.I.], v. 46, n. 5, p. 339-+, 2009.

CARLSON, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* [S.I.], v. 74, n. 1, p. 106-120, Jan 2004.

CHUNG, C. C. *et al.* Genome-wide association studies in cancer-current and future directions. *Carcinogenesis* [S.I.], v. 31, n. 1, p. 111-120, Jan 2010.

CIRULLI, E. T.; GOLDSTEIN, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* [S.I.], v. 11, n. 6, p. 415-425, Jun 2010.

CLARK, A. G. *et al.* Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* [S.I.], v. 15, n. 11, p. 1496-1502, Nov 2005.

CLAYTON, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* [S.I.], v. 37, n. 11, p. 1243-1246, Nov 2005.

CLINE, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* [S.I.], v. 2, n. 10, p. 2366-2382, 2007.

DATE, C. J. *Introdução a Sistemas de Bancos de Dados*. Tradução de VIEIRA, D. 8th. ed. Rio de Janeiro: Elsevier, 2003.

DE BAKKER, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nature Genetics* [S.I.], v. 37, n. 11, p. 1217-1223, Nov 2005.

DONNELLY, P. Progress and challenges in genome-wide association studies in humans. *Nature* [S.I.], v. 456, n. 7223, p. 728-731, Dec 11 2008.

DUERR, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* [S.I.], v. 314, n. 5804, p. 1461-3, Dec 1 2006.

EASTON, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* [S.I.], v. 447, n. 7148, p. 1087-93, Jun 28 2007.

EDUARDO TARAZONA-SANTOS, C. F., MEREDITH YEAGER, WAGNER C.S. MAGALHAES, LAURIE BURDETT, ANDREW CRENSHAW, DAVIDE PETTENER, STEPHEN J. CHANOCK. Diversity in the Glucose Transporter-4 Gene (SLC2A4) in Humans Reflects the Action of Natural Selection along the Old-World Primates Evolution. *PloS One* [S.I.], v. 5, n. 3, p. e9827, 2010.

ELBERS, C. C. *et al.* Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol* [S.I.], v. 33, n. 5, p. 419-31, Jul 2009.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* [S.I.], v. 8, n. 3, p. 186-194, Mar 1998.

EWING, B. *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* [S.I.], v. 8, n. 3, p. 175-185, Mar 1998.

EXCOFFIER, L.; HECKEL, G. Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* [S.I.], v. 7, n. 10, p. 745-758, Oct 2006.

FAGUNDES, N. J. R. *et al.* Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* [S.I.], v. 104, n. 45, p. 17614-17619, Nov 6 2007.

FRAZER, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* [S.I.], v. 449, n. 7164, p. 851-U3, Oct 18 2007.

FUSELLI, S. *et al.* Evolution of detoxifying systems: the role of environment and population history in shaping genetic diversity at human CYP2D6 locus. *Pharmacogenetics and Genomics* [S.I.], v. 20, n. 8, p. 485-499, Aug 2010.

_____. Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. *Pharmacogenomics Journal* [S.I.], v. 7, n. 2, p. 144-152, Apr 2007.

GENTLEMAN, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* [S.I.], v. 5, n. 10, p. -, 2004.

GIARDINE, B. *et al.* Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* [S.I.], v. 15, n. 10, p. 1451-1455, Oct 2005.

GIBAS, C.; JAMBECK, P. *Developing Bioinformatics Computer Skills*. 1st edition. ed.: O'Reilly Media, 2001.

GORDON, D. *et al.* Consed: A graphical tool for sequence finishing. *Genome Research* [S.I.], v. 8, n. 3, p. 195-202, Mar 1998.

GRYNBERG, P. *et al.* Polymorphism at the apical membrane antigen 1 locus reflects the world population history of Plasmodium vivax. *Bmc Evolutionary Biology* [S.I.], v. 8, p. -, Apr 29 2008.

GUDMUNDSSON, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* [S.I.], v. 39, n. 8, p. 977-83, Aug 2007.

GUTENKUNST, R. N. *et al.* Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *Plos Genetics* [S.I.], v. 5, n. 10, p. -, Oct 2009.

HARISMENDY, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* [S.I.], v. 10, n. 3, p. -, 2009.

HEWETT, M. *et al.* PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* [S.I.], v. 30, n. 1, p. 163-5, Jan 1 2002.

HULL, D. *et al.* Taverna: a tool for building and running workflows of services. *Nucleic Acids Research* [S.I.], v. 34, p. W729-W732, Jul 1 2006.

JAKOBSSON, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* [S.I.], v. 451, n. 7181, p. 998-1003, Feb 21 2008.

KANEHISA, M. *et al.* KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* [S.I.], v. 38, p. D355-D360, Jan 2010.

LANDER, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* [S.I.], v. 409, n. 6822, p. 860-921, Feb 15 2001.

LI, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* [S.I.], v. 319, n. 5866, p. 1100-1104, Feb 22 2008.

MANOLIO, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* [S.I.], v. 461, n. 7265, p. 747-753, Oct 8 2009.

MANOUKIS, N. C. FORMATOMATIC: a program for converting diploid allelic data between common formats for population genetic analysis. *Molecular Ecology Notes* [S.I.], v. 7, n. 4, p. 592-593, Jul 2007.

MARDIS, E. R.; WILSON, R. K. Cancer genome sequencing: a review. *Human Molecular Genetics* [S.I.], v. 18, p. R163-R168, Oct 15 2009.

MESSINA, D. N.; SONNHAMMER, E. L. L. DASher: a stand-alone protein sequence client for DAS, the Distributed Annotation System. *Bioinformatics* [S.I.], v. 25, n. 10, p. 1333-1334, May 15 2009.

MONTGOMERY KT, I. O., LI L, LOOMIS S, OBOURN V, KUCHERLAPATI R. PolyPhred analysis software for mutation detection from fluorescence-based sequence data. *Current Protocol in Human Genetics* [S.I.], v. Oct, n. Oct, p. Chapter 7:Unit 7.16, 2008.

NICKERSON, D. A. *et al.* PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research* [S.I.], v. 25, n. 14, p. 2745-2751, Jul 15 1997.

NIELSEN, R. *et al.* Darwinian and demographic forces affecting human protein coding genes. *Genome Research* [S.I.], v. 19, n. 5, p. 838-849, May 2009.

_____. Genomic scans for selective sweeps using SNP data. *Genome Res* [S.I.], v. 15, n. 11, p. 1566-75, Nov 2005.

NOVAES, R. M. L. *et al.* Phylogeography of Plathymenia reticulata (Leguminosae) reveals patterns of recent range expansion towards northeastern Brazil and southern Cerrados in Eastern Tropical South America. *Molecular Ecology* [S.I.], v. 19, n. 5, p. 985-998, Mar 2010.

O'CONNOR, B. D. *et al.* GMODWeb: a web framework for the Generic Model Organism Database. *Genome Biol* [S.I.], v. 9, n. 6, p. R102, 2008.

OGATA, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* [S.I.], v. 27, n. 1, p. 29-34, Jan 1 1999.

PACKER, B. R. *et al.* SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Research* [S.I.], v. 34, p. D617-D621, Jan 1 2006.

PARIKH, H. *et al.* A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33. *Human Genetics* [S.I.], v. 127, n. 1, p. 91-99, Jan 2010.

PETERSEN, G. M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics* [S.I.], v. 42, n. 3, p. 224-U29, Mar 2010.

ROSENBERG, N. A. *et al.* Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* [S.I.], v. 73, n. 6, p. 1402-1422, Dec 2003.

ROZAS, J. *et al.* DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* [S.I.], v. 19, n. 18, p. 2496-2497, Dec 12 2003.

SABETI, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* [S.I.], v. 449, n. 7164, p. 913-U12, Oct 18 2007.

SACHIDANANDAM, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* [S.I.], v. 409, n. 6822, p. 928-933, Feb 15 2001.

SCHEET, P.; STEPHENS, M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* [S.I.], v. 78, n. 4, p. 629-644, Apr 2006.

SCHMID, K. J. *et al.* Evidence for a large-scale population structure of Arabidopsis thaliana from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet* [S.I.], v. 112, n. 6, p. 1104-14, Apr 2006.

SHERRY, S. T. *et al.* The NCBI dbSNP database for Single Nucleotide Polymorphisms and other classes of minor genetic variation. *American Journal of Human Genetics* [S.I.], v. 65, n. 4, p. A101-A101, Oct 1999a.

_____. Use of molecular variation in the NCBI dbSNP database. *Human Mutation* [S.I.], v. 15, n. 1, p. 68-75, 2000.

_____. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* [S.I.], v. 29, n. 1, p. 308-311, Jan 1 2001.

_____. dbSNP - Database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research* [S.I.], v. 9, n. 8, p. 677-679, Aug 1999b.

SLADEK, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* [S.I.], v. 445, n. 7130, p. 881-5, Feb 22 2007.

SMIGIELSKI, E. M. *et al.* dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research* [S.I.], v. 28, n. 1, p. 352-355, Jan 1 2000.

SOUZA, C. P. *et al.* Mutation in intron 5 of GTP cyclohydrolase 1 gene causes dopa-responsive dystonia (Segawa syndrome) in a Brazilian family. *Genetics and Molecular Research* [S.I.], v. 7, n. 3, p. 687-694, 2008.

STEIN, L. D. *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res* [S.I.], v. 12, n. 10, p. 1599-610, Oct 2002.

STEPHENS, M. *et al.* A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* [S.I.], v. 68, n. 4, p. 978-989, Apr 2001.

STRANGER, B. E. *et al.* Population genomics of human gene expression. *Nat Genet* [S.I.], v. 39, n. 10, p. 1217-24, Oct 2007.

TARAZONA-SANTOS, E. *et al.* CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains. *Hum Mutat* [S.I.], v. 29, n. 5, p. 623-32, May 2008.

TARAZONA-SANTOS, E.; TISHKOFF, S. A. Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. *Genes and Immunity* [S.I.], v. 6, n. 1, p. 53-65, Feb 2005.

THORNTON, K. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* [S.I.], v. 19, n. 17, p. 2325-2327, Nov 22 2003.

TODD, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* [S.I.], v. 39, n. 7, p. 857-64, Jul 2007.

VARGAS, S. M. *et al.* Genetic diversity and origin of leatherback turtles (Dermochelys coriacea) from the Brazilian coast. *Journal of Heredity* [S.I.], v. 99, n. 2, p. 215-220, Mar-Apr 2008.

VENTER, J. C. *et al.* The sequence of the human genome. *Science* [S.I.], v. 291, n. 5507, p. 1304-+, Feb 16 2001.

VILELLA, A. J. *et al.* VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* [S.I.], v. 21, n. 11, p. 2791-2793, Jun 1 2005.

VOIGHT, B. F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America* [S.I.], v. 102, n. 51, p. 18508-18513, Dec 20 2005.

2. UTILIZAÇÃO DE FERRAMENTAS BIOINFORMÁTICAS PARA O ESTUDO DA VARIAÇÃO GENÔMICA HUMANA

2.1 Introdução

A genética de populações inclui o estudo de diversas forças que resultam em mudanças evolutivas nas espécies ao longo do tempo. O padrão de diversidade genética das populações humanas é o resultado da combinação de diversos fatores evolutivos: (1) fatores que atuam sobre regiões genômicas específicas, como mutações, recombinação e seleção natural (Jorde *et al*., 2001; Tishkoff e Verrelli, 2003); (2) fatores que atuam sobre todo o genoma, como aqueles relacionados com a história demográfica das populações: ex. tamanho populacional e as suas flutuações (deriva genética), sub-estruturação, fluxo gênico e padrões de acasalamento (Excoffier e Ray, 2008; Charlesworth, 2009; Novembre e Di Rienzo, 2009).

Vários processos podem criar variação genética nas populações ou ainda promover a reorganização da variação preexistente seja esta dentro da população, seja entre subpopulações. No entanto, a mutaçãos é a fonte primordial de inovações genéticas. O processo de mutação pode ocorrer em diferentes níveis de organização do genoma, podendo causar desde a mudança de um único nucleotídeo, a até mesmo deleções e inserções de cromossomos inteiros ou ainda grandes partes genômicas. Mutações são responsáveis pela criação de novas variantes; enquanto a seleção natural, migração e deriva genética agem sobre essa variabilidade criada fazendo com que se altere ou não ao longo do tempo (Bamshad *et al*., 2003).

Mutações ocorrem através de diferentes mecanismos envolvidos na replicação e reparo do DNA, criando diferentes tipos de polimorfismos (Figura 1). Um desses consiste na substituição de uma única base por outra na mesma posição (*Single nucleotide polymorphisms*-

SNPs), sendo este o polimorfismo mais comum presente no genoma humano (Frazer *et al*., 2007). Mutações nucleotídicas ocorrem a uma taxa entre ~ 2.5 $x10^{-8}$ e ~1.1 $x10^{-10}$ por sítio por geração, dependendo do genoma e da região genômica em estudo (Nachman e Crowell, 2000; Roach *et al*., 2010). Regiões específicas, conhecidas como ilhas de CpG, podem apresentar taxas de mutação diferenciadas ao valor médio observado no genoma. (Fryxell e Moon, 2005). Outro tipo de polimorfismo, os microssatélites (também denominados *short tandem repeats* - *STRs*), são sequências de 1 a 6 bases de comprimento repetidas em *tandem*. O polimorfismo existe porque cada cromossomo pode ter um número diferente de repetições. Os microssatélites apresentam uma taxa de mutação de 10^{-3} a 10^{-4} mutações (aumento ou redução de uma ou mais unidades repetitivas) por locus por geração(Brinkmann *et al*., 1998; Rosenberg *et al*., 2003).

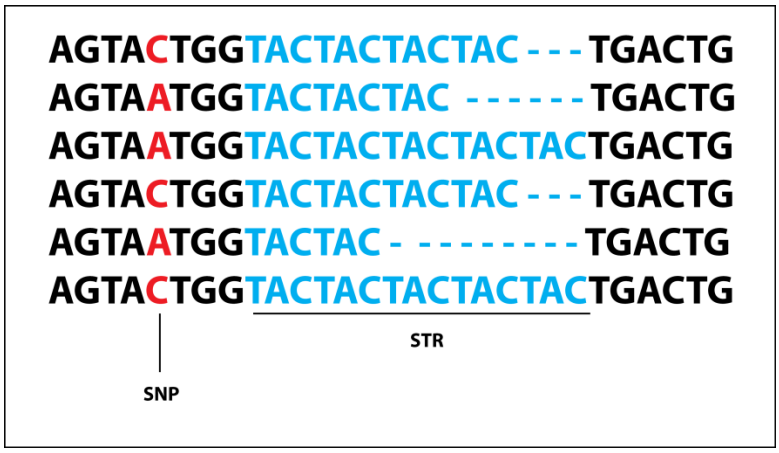


Figura 2 – Representação de polimorfismos de base única – SNP (em vermelho) e microssatélites – STR (em azul)

Recombinação é uma das principais forças evolutivas que afetam os padrões de diversidade genética (Wang e Rannala, 2009). A influência da recombinação sob a diversidade genética humana tem causado grande debate nos últimos anos (Stumpf e Mcvean, 2003; Myers *et al*., 2005; Wang e Rannala, 2009). A recombinação atua sobre a diversidade rearranjando os

alelos criados pela mutação, aumentando a quantidade de combinações genéticas, e portanto, aumentando a gama de alelos que podem sofrer a ação da seleção natural. Esse mecanismo ajudaria a seleção a ser mais eficiente (Myers *et al*., 2005). Observações demonstram que regiões do genoma que apresentam altos valores de diversidade também apresentam altas taxas de recombinação, evidenciando o caráter mutagênico da recombinação como demosntrado por (Hellmann *et al*., 2003).Em humanos, a taxa de recombinação é positivamente associada aos níveis de diversidade (quando medida em megabases)(Hellmann *et al*., 2005). No entanto, não é claro se este padrão observado é devido ao acaso (deriva genética), por *footprints* de pressões seletivas ao longo do genoma ou ainda pela correlação espúria de fatores neutros, como, por exemplo, composição de bases (Spencer *et al*., 2006).

A seleção natural enunciada por Darwin em "A origem das espécies" (1859), tem sido parte do pensamento evolucionista e muitas vezes até mesmo confundida com o termo evolução. Seleção natural é o termo que relaciona um tipo de influência do meio ambiente na seleção de variantes alélicas (polimorfismos) em uma população: definido pela probabilidade diferencial dos genótipos de um locus serem passados a geração seguinte. Pressões seletivas afetam regiões específicas do genoma (Sabeti *et al*., 2002). A forma e a intensidade com que estes eventos de pressões seletivas afetam a diversidade são dependentes de outros fatores evolutivos, tais como as taxas de mutação e de recombinação (Nielsen *et al*., 2005; Nielsen *et al*., 2009).

Fatores evolutivos dependentes da história demográfica da população como mudanças de tamanho da população ao longo do tempo devido ao acaso (deriva genética), ou movimentos populacionais também apresentam efeitos no padrão de diversidade observado ao longo do genoma (Excoffier, 2002; Goldstein e Chikhi, 2002; Balaresque *et al*., 2007; Campbell e Tishkoff, 2008). Por exemplo, grandes reduções no tamanho populacional (*bottlenecks)* são

responsáveis por diminuições da diversidade observada, enquanto expansões populacionais agem de forma inversa aumentando a diversidade observada através do rápido aumento populacional, levando a um aumento do espaço amostral para surgimento de novas variantes, como observado em (Kimmel *et al*., 1998; Novembre e Di Rienzo, 2009).

No entanto, os mecanismos pelos quais os fatores evolutivos moldam o padrão de diversidade genética são distintos. Variações nos níveis de diversidade também são observados devido a subdivisões populacionais. Subdivisões de populações aumentam a variabilidade entre elas, dado que cada população poderá acumular diferenças ao longo do tempo (Goldstein e Chikhi, 2002).

A história demográfica de diferentes populações humanas pode ser em princípio inferida a partir da análise dos padrões de variação genética. Em humanos, SNPs tem sido usados como principais marcadores destes estudos (Sachidanandam *et al*., 2001; Altshuler *et al*., 2005; Savage *et al*., 2005). O estudo da diversidade genômica fornece importantes informações sobre os processos evolutivos que incidiram sobre o genoma e produziram o padrão de diversidade observado, tanto ao nível de todo o genoma (fatores demográficos), quanto em nível de regiões específicas do genoma (pressões seletivas exercidas por um determinado ambiente).

2.2 Modelo de Wrigth-Fisher

O modelo de Wright-Fisher descreve como uma população ideal transmite seus genes aos seus descendentes e com estes evoluem ao longo do tempo. O modelo segue as seguintes premissas:

1. Tamanho infinito e constante, isto é, o número de indivíduos na população não muda ao longo das gerações.
2. Panmixia: essa premissa postula que todos os indivíduos tem a mesma probabilidade de cruzar entre si, sem uma divisão interna (estratificação).
3. Não existe sobreposição de gerações, ou seja, todos os indivíduos que pertencem a uma única geração se reproduzem e morrem ao mesmo tempo.
4. Ausência de recombinação. Os genes envolvidos nesse modelo não sofrem ação da recombinação, isso implica que esse modelo só deve ser usado com regiões que não estão sujeitas a essa força evolutiva, como por exemplo, segmentos do cromossomo Y, X e DNA mitocondrial.
5. Ausência de seleção. Todos os indivíduos têm a mesma probabilidade de sobreviver e produzir prole, transmitindo dessa forma seus genótipos.

Como pode ser observado, o modelo de Wright-Fisher não é baseado em uma população real o que fica claro pelo número de premissas do presente modelo e que não podem ser observadas em populações naturais.O modelo de Wright-Fisher é particularmente importante como hipótese nula (Hudson, 2002).

2.3 Seleção Natural e Neutralidade

A evolução adaptativa de genes e genomas tem sido ultimamente apontada como responsável por parte da evolução morfológica, comportamentamental, fisiológica assim como pela divergência das espécies. No entanto, em genética de populações, diferentes modelos evolutivos têm procurado esclarecer os fatores que levam à diferenciação genética entre as populações, por exemplo, a Teoria Neutra, proposta por (Kimura, 1968; Crow, 1987).

Antes dos anos 60, vários geneticistas assumiam que a maioria dos polimorfismos eram mantidos na população devido à ação da seleção balanceadora. No entanto, em 1968 o geneticista japonês Motoo Kimura propôs que, em nível molecular, mutações neutras seriam mais freqüentes que os demais tipos de mutação e que sua fixação ocorreria por efeitos puramente estocásticos, mediados pelos fatores evolutivos de mutação e deriva genética. A formulação original da teoria Neutra estava focada nas mutações que são, a rigor, seletivamente neutras, sendo seu destino determinado pela deriva genética, onde em muitos casos, a seleção natural não é necessária para explicar o nível de polimorfismos observado em uma população. A razão pela qual o acaso tem tamanha importância nas mudanças genéticas, argumentava Kimura, residia no fato de que a maioria das variantes genéticas são evolutivamente equivalentes.

Esta nova idéia sugere que as mutações responsáveis pelo surgimento de características adaptativas vantajosas, contribuiriam pouco para a variabilidade genética das populações por serem extremamente raras e se fixarem muito rapidamente (por seleção natural). Em suas considerações sobre a teoria neutra, Kimura excluiu as mutações prejudiciais já que estas não contribuiriam nem para a variabilidade genética nem para a evolução molecular, uma vez que são rapidamente eliminadas por meio da "seleção negativa" ou purificadora (Kimura, 1976; 1977a; b; Crow, 1987).

Entretanto, devido às limitações dessa teoria em explicar taxas médias de evolução diferenciais e tipos de mutação (por exemplo, sinônimas e não sinônimas), a Teoria Neutra foi sendo substituída pela Teoria Quase-Neutra. Sob essa teoria, grande parte da variabilidade entre populações ocorre devido à deriva genética. Dessa forma, a adaptação ocorre devido a pressões seletivas fracas em variantes comuns, ao invés de se dar através de fortes pressões seletivas em variantes raras. A Teoria Quase-Neutra admite três classes de mutações quanto à pressão seletiva: neutras, quase-neutras (*slightly deleterious*) e deletérias (ou vantajosas). Além disso, assume que as taxas de evolução estão relacionadas ao tamanho efetivo das populacional. Mesmo assim, ainda hoje, há um intenso debate entre defensores da teoria neutra e da seleção adaptativa (Hurst, 2009).

É importante ressaltar que a teoria neutra da evolução, ainda que tenha causado muita controvérsia no ambíto científico, não nega a existência da seleção natural nem sua importância para a evolução. E que, ao contrário de Darwin, que não dispunha dos conhecimentos de biologia molecular, Kimura com a teoria neutra lida essencialmente com variações à nível molecular.

Estudos de genética de populações tem sido utilizados tanto para explicar os padrões de diversidade genética humana em termos de história populacional (deriva), quanto para entender as bases genéticas das adaptações fenotípicas (a ação da seleção natural). Entretanto, um dos principais obstáculos referentes às inferências evolutivas repousa justamente em discriminar entre a ação da deriva genética e a seleção natural (Balaresque *et al*., 2007), dado que estes dois fatores evolutivos podem produzir padrões de diversidade genética semelhantes.

A seleção natural pode ser dividida em classes: direcional, estabilizadora, disruptiva e balanceadora. A seleção direcional, ou seleção Darwiniana, está relacionada ao incremento da frequência de um alelo que aumente o *fitness* do indivíduo (Hurst, 2009). A seleção

estabilizadora, mantêm a frequência dos alelos em um valor ótimo. A seleção disruptiva aumenta a frequência de valores nos dois extremos da distribuição da característica, enquanto diminui a frequência de valores intermediários. A seleção balanceadora aumenta a frequência de alelos de diferentes características que sofrem pressão do ambiente. Dessa forma, fenótipos que apresentam grande diferenciação entre populações, possivelmente, estão relacionados a polimorfismos apresentando grandes diferenças nas frequências alélicas (Myles *et al*., 2008). O Desequilíbrio de ligação pode ser definido como a associação estatística entre dois alelos localizados em diferentes loci e indica se um alelo está associado ao outro em uma frequência maior que o esperado sob a hipótese de neutralidade (Slatkin, 2008). Regiões sob seleção positiva têm alto desequilíbrio de ligação, fato que pode ser atribuído ao aumento da frequência do alelo selecionado ser mais rápido do que a ação da recombinação no local onde o alelo está situado (Sabeti *et al*., 2002). A seleção purificadora, ou negativa, elimina mutações deletérias. De acordo com a premissa que indivíduos bem adaptados têm maior valor adaptativo, provavelmente esse tipo de seleção é a mais comum. A seleção balanceadora atua no sentido de favorecer a diversidade através da co-dominância, seleção dependente da frequência ou coevolução parasita-hospedeiro cíclica. Neste caso, os alelos não são fixados e não podem ser ditos como deletérios (Hurst, 2009).

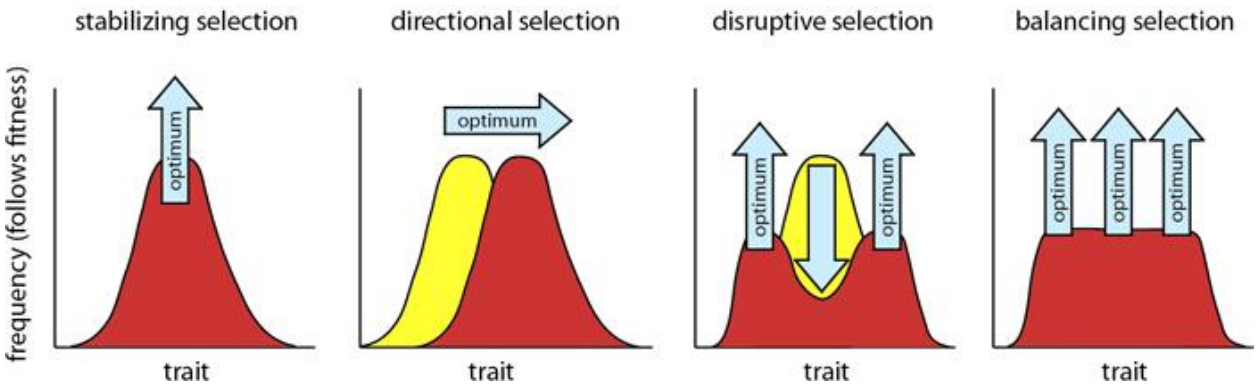


Figura 3 - A área vermelha denota a distribuição de frequência atual de uma determinada característica observada em indivíduos de uma população. A área amarela denota a mesma distribuição na segunda geração. Se houver tempo suficiente, a frequência seguirá as mudanças nos valores da aptidão associados com um valor particular caracterísctica em questão, assim que estas figuras descrevem também característica de partes relevantes da paisagem adaptável. As setas indicam o ponto ótimo (ou para onde está movendo) e assim os valores da característica que serão selecionados positivamente. Todos valores restantes da característica estão sob seleção negativa.

2.4 Testes para a hipótese de evolução sobre neutralidade

Um dos objetos de estudo da genética de populações molecular é como inferir a ação da seleção natural em regiões genômicas específicas em populações (Hudson *et al*., 1987; Tajima, 1989). Nas últimas quatro décadas foi observado o desenvolvimento de um grande número de testes estatísticos para o desvio das condições esperadas sob neutralidade (Ewens e Feldman, 1974; Watterson, 1978b; a; Fu e Li, 1993; Fay e Wu, 2000; Nielsen, 2005; Bird *et al*., 2007; Andres *et al*., 2009). Estes testes para a hipotése de evolução neutra junto com a teoria do coalescente, (Hudson, 1983; Kingman, 2000; Hudson, 2002) tem se tornado nos últimos anos uma importante ferramenta nos estudos de genética de populações dada a possibilidade de realizar simulações de amostras de genes extraídas de populações, condicionadas ao padrão de variabilidade observado na amostra. Estimando parâmetros a partir dos dados observados e comparando essas estatísticas com suas distribuições neutras obtidas sob simulações é possível

inferir eventos demográficos e evolutivos experimentados pelas popualções no passado (Hein *et al*., 2005).

2.4.1 Teste de Ewens-Watterson

O teste para a hipótese de desvio da neutralidade de Ewens-Watterson, descrito por (Ewens, 1972) e (Watterson, 1978b), pode ser considerado um dos primeiros e mais elegantes testes de para o desvio da hipótese de evolução neutra. Baseado nas premissas do modelo de alelos infinitos proposto por (Kimura, 1968), que considera que cada nova mutação gera um novo alelo, e que a população está em equilíbrio de mutação-deriva, a taxa de homozigose observada (F_{obs}) é comparada com a esperada (F_{esp}).

O valor da taxa de homozigose esperada é derivado atualmente de amostras simuladas com o mesmo número de alelos da amostra observada, a simulação é baseada na teoria de amostragem de alelos sob neutralidade de (Ewens, 1972). Amostras com $F_{obs} > F_{esp}$, apresentam poucos alelos em altas frequências e vários alelos em baixas frequências, situação compatível com seleção direcional ou expansão populacional. Amostras com $F_{obs} < F_{esp}$, apresentam alelos em frequencias intermediárias, situação compatível com seleção balanceadora, situação comumente observada em loci do sistema HLA, gargalo de garrafa ou estruturação populacional (Watterson, 1978b; Nielsen, 2005; Harris e Meyer, 2006). O valores de significância do teste são dados pela proporção de F_{esp} inferiores ou iguais ao valor observado.

2.4.2 Teste D de Tajima

A estatística D de Tajima (Tajima, 1989) é um dos testes de neutralidade mais utilizados para dados de sequenciamentos e ressequenciamento. Baseada na diferença entre dois estimadores do parâmetro θ, símbolo usado para denotar o produto $4N\mu$ para os loci nucleares, onde N representa o tamanho da população e μ a taxa de mutação. Um de seus estimadores é o π que corresponde ao número médio de diferenças nucleotídicas entre pares de sequencias (θ_π). O outro estimador é o theta de Waterson (θ_w), que é o número de sítios segregantes presentes na amostra, corrigido para o tamanho da amostra. Sob a premissa de evolução neutra os dois estimadores θ são equivalentes e ambos estimam o valor correto de θ. Essa é a razão pela qual dentro da hipótese de neutralidade os dois valores são iguais e a diferencia é próxima a zero, variância igual a 1, embora não estejam normalmente distribuídos, esses valores seguem a distribuição beta.

Tajima's D (1989).

$$D = \frac{\Pi_n - {}^{K}/_{a_n}}{\sqrt{Var(\Pi_n - {}^{K}/_{a_n})}}$$

onde

$$a_n = \sum_{k=1}^{n-1} \frac{1}{k}$$

sendo n o número de cromossomos na amostra e k o número de sítios segregantes.

Em cenários de seleção positiva, purificadora ou expansão populacional um excesso de *singletons* e variantes apresentando baixas frequências são observados. Se isso acontece, o

número de sítios segregantes *k* é grande quando comparado com π e então θw é maior que $\theta\pi$. Consequentemente, valores de *D* vão ser negativos e quanto mais negativos maior é o desvio da hipótese de evolução sob neutralidade. Em outro cenário, presença de seleção balanceadora ou estruturação populacional um excesso de variantes com frequências intermediárias é observado e *K* é pequeno comparado com aos valores de π levando a valores positivos de *D*. A significância do teste *D* de Tajima é calculada pela proporção de valores de *D* de amostras obtidas por simulações utilizando a teoria do coalescente com o mesmo número de sítios segregantes que a amostra observada.

2.4.3 Testes de Fu e Li

Os testes de Fu e Li (Fu e Li, 1993) pertencem à classe de testes baseados na comparação entre um estimador de θ e o número de mutações únicas derivadas (*singletons*) presentes nos ramos externos da genealogia. Estes testes são D, F, D* e F*, onde a principal diferença entre eles é que os dois primeiros precisam de um grupo externo (*outgroup*). Da mesma maneira que no teste de Tajima (Tajima, 1989), esses testes são desenhados para diferenciar dois estimadores de θ sob a hipótese de evolução neutra.

Os testes de Fu e Li se diferenciam do teste de (Tajima, 1989), pela diferente interpretação das mutações presentes na filogenia, essas mutações são classificadas como mutações internas e externas, mutações que ocorrem nos ramos internos (mutações antigas) daquelas que ocorrem nos ramos externos da filogenia (mutações recentes), respectivamente, onde:

Fu and Li's D (1993)

$$D = \frac{\eta - a_n \eta_e}{\sqrt{u_D \eta + \upsilon_D \eta^2}}$$

onde

$$\eta = \sum_i^m S_i$$

$$\eta_e = \sum_i^m e_i$$

$$\upsilon_D = 1 + \frac{a_n^2}{b_n - a_n^2}\left(c_n - \frac{n+1}{n-1}\right)$$

$$u_D = a_n - 1 - \upsilon_D$$

Entretanto, quando não há a presença de um grupo externo (*outgroup*) é difícil inferir o número de *singletons*, e, considerando todos os *singletons* como derivados, claramente há uma superestimação do número de singletons derivados. Para resolver isso Fu e Li desenvolveram dois testes que corrigem para essa superestimação.

Fu and Li's D*

$$D^* = \frac{\left(\frac{n}{n-1}\right)\eta - a_n \eta_S}{\sqrt{u_{D*}\eta + \upsilon_{D*}\eta^2}}$$

onde

$$\upsilon_{D*} = \left[\left(\frac{n}{n-1}\right)^2 b_n + a_n^2 d_n - 2\frac{n a_n (a_n+1)}{(n-1)^2}\right] / (a_n^2 + b_n)$$

e

$$u_{D*} = \frac{n}{n-1}\left(a_n - \frac{n}{n-1}\right) - \upsilon_{D*}$$

Outros dois testes ainda apresentados no mesmo trabalho são: a diferença normalizada entre $\Pi_n \; and \; \eta_e$ como apresentado, pela estatística F.

Fu and Li's F (1993)

$$F = \frac{\Pi_n - \eta_e}{\sqrt{u_F \eta + \upsilon_F \eta^2}}$$

onde

$$\upsilon_F = \left[c_n + \frac{2(n^2+n+3)}{9n(n-1)} - \frac{2}{n-1}\right] / (a_n^2 + b_n)$$

e

$$u_F = \left[1 + \frac{n+1}{3(n-1)} - 4\,\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)\right] / a_n - \upsilon_F)$$

E sua variação com a ausência de uma grupo externo.

Fu and Li's F* (1993)

$$F^* = \frac{\Pi_n - \frac{n-1}{n}\,\eta_S}{\sqrt{u_{F*} \eta + \upsilon_{F*} \eta^2}}$$

onde

$$\upsilon_{F*} = \left[d_n + \frac{2(n^2+n+3)}{9n(n-1)} - \frac{2}{n-1}\left(4b_n - 6 + \frac{8}{n}\right)\right] / (a_n^2 + b_n)$$

e

$$u_{F*} = \left[\frac{n}{n-1} + \frac{n+1}{3(n-1)} - 2\,\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)\right] / a_n - \upsilon_{F*})$$

Em ambos casos, a estatística é baseada na mesma idéia, que sob a hipótese de neutralidade o número esperado de mutações externas E[ηE] = θw = $\theta\pi = 4N\mu$.

A significância para os testes de Fu e Li é calculada pela proporção de valores das estatísticas calculadas nas amostras obtidas por simulações utilizando a teoria do coalescente.

2.4.4 H de Fay e Wu

O teste de Fay e Wu (Fay e Wu, 2003), compara polimorfismos com frequências intermediárias com polimorfismos em frequências elevadas. Sob o cenário de seleção neutra, a distribuição dos polimorfismos apresenta uma distribuição em forma de uma curva em L, com um alto número de polimorfismos comuns (frequências intermediárias) e poucos polimorfismos raros (frequências baixas). Sob a hipótese de seleção direcional, o cenário se inverte, os polimorfismos comuns próximos à variante sob seleção também aumentam suas frequências, sofrendo assim o chamado "efeito carona" (*Hitcking effect*). (Fay e Wu, 2000) demonstraram que um excesso de polimorfismos apresentando altas frequências, indicados pela significâncias dos valores de H, é compatível com o efeito carona. O teste também requer a utilização de um grupo externo para a inferências dos polimorfismos derivados. Este teste utiliza uma idéia similar aos outros descritos anteriormente além de desenvolver um novo estimador para θ:

$$\theta_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)}$$

onde S_i é o número de variantes derivadas encontradas i vezes na amostra. θ_H é um novo estimador de θ que atribui mais peso às variantes que apresentam frequências maiores.

A par com com esse novo estimador de θ eles então desenvolveram a estatística H:

$$H = \frac{\theta_\pi - \theta_H}{\sqrt{Var(\theta_\pi - \theta_H)}}$$

Como nos testes desenvolvidos por Tajima (Tajima, 1989) e (Fu e Li, 1993), sob a hipótese de Neutralidade é esperado que os dois estimadores de θ sejam 4Nμ. O teste de Fay e Wu é especialmente indicado para detectar o efeito carona, uma vez que ele aumenta a frequência de variantes derivadas.

2.4.5 Testes de padrão de divergência e polimorfismos

Dentre os testes que consideram dados de padrões de divergência e de polimorfismos (mutações inter e intra-específicas), dois testes merecem ser destacados: Hudson-Kreitman-Aguade (HKA) e McDonald-Kreitman (MK).

2.4.5.1 Teste de Hudson-Kreitman-Aguadé

O teste de Hudson-Kreitman-Aguadé, parte da premissa que, sob a hipótese de neutralidade, polimorfismos dentro da mesma espécie e a divergência entre as espécies são resultado do mesmo processo (Hudson *et al*., 1987). O teste HKA assume que mudanças

populacionais afetam igualmente todo o genoma enquanto a seleção natural afeta regiões específicas, apenas alguns loci, e, com isso, em condições de neutralidade, a razão entre polimorfismos inter e intra-específicos deve ser constante em vários loci independentes. O teste usa dados de múltiplas sequências de loci não ligados de pelo menos duas espécies relacionadas para testar se os polimorfismos e a divergência destes loci são compatíveis. De acordo com o teste, se um locus tem uma alta taxa de mutação, ambos polimorfismos e divergência devem ser altos, enquanto se um locus tem baixa taxa de mutação, ambos polimorfismos e divergência devem ser baixos.

Tabela 2 – Estimativas da quantidade de variação intra (within species) e inter-específica (between species) entre espécies de *Drosophila melanogaster* e *Drosophila sechelia* para o gene Adh e a região flanqueadora 5'. (Tabela modificada de Hudson et al., 1987).

	Adh	Flanking region	Ratio (Adh/flanking)
Within species	0.101	0.022	4.59
Between specie	0.056	0.052	1.08
Ratio (within/between)	1.80	0.42	

2.4.5.2 Teste de McDonald-Kreitman

O teste MK (Mcdonald e Kreitman, 1991) assume que, sob neutralidade, a razão entre divergências não sinônimas, caracterizadas pelas substituições nucleotídicas que alteram o aminoácido na proteína, e substituições sinônimas, onde a modificação nucleotídica não altera a sequência protéica, deve ser semelhante à razão entre polimorfismos não sinônimos e sinônimos, uma vez que são resultados da deriva genética e da fixação de mutações neutras. Dessa forma se ambos tipos de mutações, sinônimas e não sinônimas são evolutivamente neutras, a proporção de mutações sinônimas e não sinônimas intra-espécificas e a proporção inter-específca deve ser a mesma. O teste de McDonald-Kreitman examina essa predição.

Polimorfismos em regiões codificantes (exons) de espécies relacionadas são classificados em 4 categorias em uma tabela de contigência de 2x2, dependendo se o sítios tem um polimorfismo ou uma diferença fixada e se esse polimorfismo é sinônimo ou não sinônimo, (tabela 3).

A hipótese nula (evolução neutra) é a independência entre as linhas e as colunas da tabela e pode ser testada aplicando-se um teste de X^2 (Qui-quadrado) ou teste exato de Fisher se os valores são pequenos.

(Mcdonald e Kreitman, 1991) sequenciaram o gene da enzima Alcool desidrogenase de três espécies dentro do subgrupo de *Drosophila melanogaster* e obtiveram os valores apresentados na tabela 3. Com um p valor de 0.006 sugerindo um desvio da hipótese nula, evolução neutra. Isso é devido a um número maior de diferenças entre as espécies que dentro da espécie. Dessa forma sugeriram que a seleção positiva é a força evolutiva agindo nas diferenças encontradas.

Devido ao fato de que modelos de variação no mesmo gene são mais possíveis de serem comparados, o teste MK é considerado estatisticamente mais robusto que o teste HKA.

Tabela 3 – Número de polimorfismos não sinônimos (Nonsynonymous) e sinônimos (Synonymous) para substituições fixadas (fixed) entre espécies e polimorfismos intra-específicos (polymorphic). (a) Visão geral do cálculo para o teste; (b) aplicação do teste para o locus Adh em três espécies de Drosophila (McDonald and Kreitman, 1991) e (c) para o locus G6pd para as espécies *D*. *Melanogaster* e *D*. *Simulans* (Eanes et al., 1993).

	(a) General		(b) Adh		(c) G6pd	
	Fixed	Polymorphic	Fixed	Polymorphic	Fixed	Polymorphic
Nonsynonymous	NF	NP	7	2	21	2
Synonymous	SF	SP	17	42	26	36
Ratio	NF/SF	NP/SP	0.41	0.05	0.81	0.06

2.4.6 Teste da Extensão da Homozigosidade Haplótipica (EHH)

A homozigosidade haplotípica (HH) é uma medida do desequilíbrio de ligação, para 2 ou mais marcadores, e pode ser calculado como apresentado por (Sabeti *et al*., 2002). A Homozigosidade Haplotípica Extendida (*Extended Haplotype Homozigosity*, EHH) é a distância x de um conjunto específico de SNPs "*core*" como definido por (Sabeti et al. 2002). EHH estima o nível de novos haplótipos formados pela ação da recombinação e mutações em regiões adjacentes de ambos os lados do "core".

Diferente das outras estatísticas o teste EHH (Sabeti *et al*., 2002; Sabeti *et al*., 2006) é um teste heurístico. Dessa forma, seu valor de significância não pode ser acessado por simulações usando o modelo de evolução neutra, sendo para tanto necessário grandes conjuntos de dados empíricos. EHH, foi concebido sob a premissa de que dentro de um cenário de evolução neutra, mutações recentes (novas) são encontradas em baixas frequências e estão presentes em regiões que apresentam altos valores de desequilíbrio de ligação, enquanto mutação antigas podem ser encontras em frequências altas ou baixas em regiões com baixos valores de desequilíbrio de ligação. Esse fato pode ser atribuído ao maior tempo de exposição à ação da recombinação, diminuindo dessa forma a extensão dos haplótipos. No entanto, se um alelo aumenta em frequência, em um pequeno espaço de tempo, por seleção diferencial (seleção positiva), não há tempo suficiente para a ação da recombinação e com isso é possível observar alelos (mutações recentes) em regiões com altos valores de desequilíbrio de ligação (Figura 4).

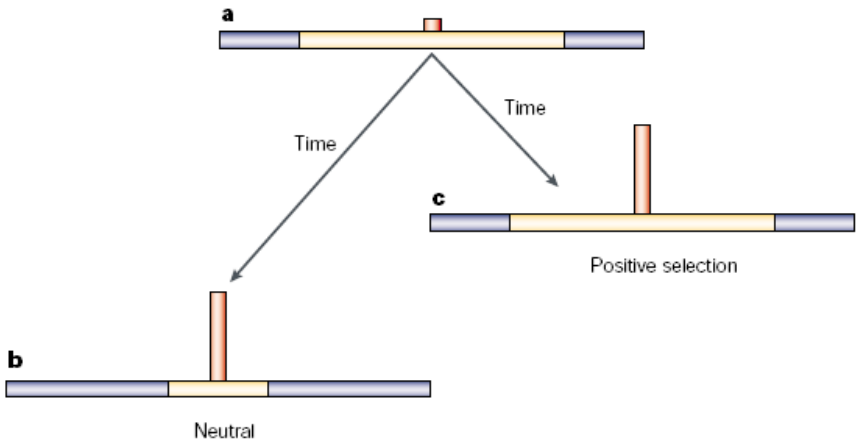


Figura 4 - Detecção de Seleção Natural positiva recente utilizando desequilíbrio de ligação. a) Um novo alelo (vermelho) em uma frequência relativamente baixa (indicado pela barra vermelha) em um haplótipo (azul) que é caracterizado por uma grande região em forte desequilíbrio de ligação (amarelo) entre o alelo de interesse e marcadores ligados. b) Ao longo do tempo, a frequência do alelo aumenta como resultado de deriva genética e a recombinação local reduz o desequilíbrio de ligação entre o alelo e os marcadores. c) O alelo influenciado pela seleção positiva recente aumenta sua frequência mais rápido que a recombinação local pode reduzir a região em desequilíbrio de ligação entre o alelo e os marcadores.

(Sabeti *et al*., 2002) avaliaram o desequilíbrio de ligação associado a alelos cuja frequência é alta devido à seleção natural, utilizando como estimador a EHH ou haplótipo candidato, e testam a hipótese nula (neutra) que os diferentes alelos (ou haplótipos) de um locus tem níveis de EHH similares, uma vez que a deriva genética atua homogeneamente sobre os diferentes alelos (haplótipos). A hipótese alternativa é que um dos alelos (haplótipos) sob efeito de seleção positiva recente tem associado um nível maior de EHH. O método foi inicialmente aplicado aos genes G6PD e TNFSF5 (Sabeti *et al*., 2002).

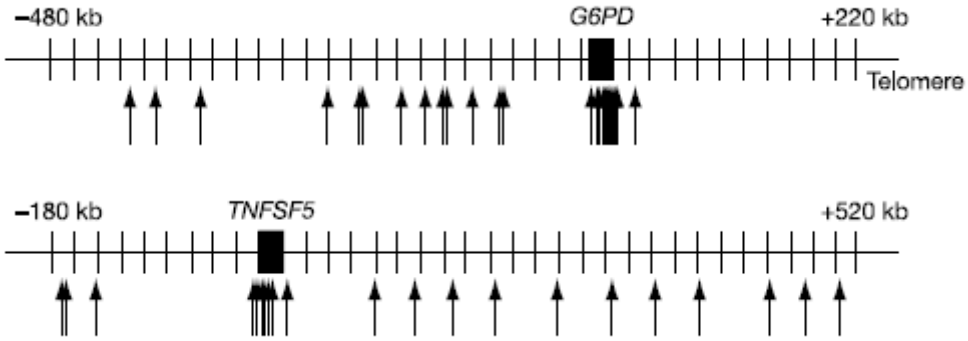


Figura 5 - Desenho experimental do *core* e região em desequilíbrio de ligação para os genes G6PD e TNFSF5. A região do *core* é marcada por uma alta densidade de SNPs (setas) dentro da região codificante do gene. Adicionalmente, SNPs em regiões genômicas, usados para examinar o decaimento do desequilíbrio de ligação para cada *core*, também são mostrados.

Sabeti e colaboradores verificaram a ação da seleção positiva no gene CCR5-Δ32 usando este método. CCR5 é uma quimiocina que participa da entrada do vírus HIV, e apresenta várias mutações não sinônimas na população humana (Sabeti *et al*., 2005).

2.4.7 Teste iHS

O teste iHS é aplicado a SNPs individuais e é calculado o partir do EHH(Voight *et al*., 2006), que pode ser definida como a integral da diminuição observada da Homozisosidade Haplotípica Extendida (a área dentro da curva de EHH) *versus* a distância a partir de um alelo central até o valor de EHH atingir 0.05 (Voight *et al*., 2006). A razão do log de EHH para o alelo ancestral e derivado é então normalizado para que tenha média igual a zero e variância igual a 1. Ambos valores altos positivos e negativos de iHS são indicativos de haplótipos mais longos que o esperado sob neutralidade.

(Voight *et al*., 2006) desenvolveram e aplicaram o teste iHS para dados do projeto HapMap e demonstraram que a maioria dos sinais de seleção encontrados pela metodologia são regiões específicos. Esses resultados são contrários a estudos prévios com menor resolução (Pluzhnikov *et al*., 2002).

2.5 Viés de Averiguação

Testes de neutralidade, principalmente aqueles baseados no espectro de frequência de mutações, estão apoiados em uma descrição acurada da frequência. No entanto, isso só pode ser obtido através de ressequenciamento de todos os cromossomos da amostra para a região genômica candidata. Entretanto, muitos pesquisadores por motivos de custo ou praticidade usam genotipagem. Essa técnica implica na seleção previa dos SNPs a serem genotipados, o que já deixa claro que essa informação não vai ser obtida para todos os sítios segregantes. Esse é o viés de averiguação.

Viés de averiguação pode ser produzido por dois mecanismos, embora eles não sejam exclusivos: 1) não detectando todos os SNPs na amostra, 2)selecionando somente alguns SNPs. O único jeito de não introduzir o viés de averiguação seria realizar o resequenciamento de toda a amostra (Picoult-Newberg *et al*., 1999; Altshuler *et al*., 2000). Usando esse procedimento é mais provável detectar SNPs em frequências intermediárias e altas, portanto SNPs mais fáceis de serem genotipados que os SNPs raros. Além disso, vem sendo demonstrado que o espectro de frequências (a distribuição dos alelos em classes diferentes de frequência) difere dependendo da estratégia de seleção de SNPs (Nielsen e Signorovitch, 2003; Nielsen *et al*., 2004).

Por outro lado, o viés de averiguação pode ser causado pela seleção de SNPs que serão genotipados provenites de outras fontes ao invés de serem selecionados na própria amostra.

Usualmente esses SNPs são selecionados a partir de painéis mais conhecidos, tais como HapMap (http://www.hapmap.org/; (Frazer *et al*., 2007) ou Perlegen (http://www.perlegen.com/; (Hinds *et al*., 2005), os quais, por si só, já apresentam um viés de averiguação.

Protocolos para a seleção de SNPs podem variar, mais de forma geral eles envolvem um ou a combinação dos seguintes critérios: (a) Selecionar SNPs com alelo de menor frequência alélica superior a 5-10 %, (b) seleção por distância, um SNPs a cada determinado número de bases. (c) selecionar por distância mais não uniformemente, por exemplo maior densidade em regiões gênicas, (d) selecionar SNPs polimórficos em todas as populações de interesse. (e) selecionando SNPs que são polimórficos em somente uma população (Moreno-Estrada et al. 2008). A influência da seleção de SNPs no espectro de frequência depende do critério adotado, mas em alguns casos eles podem variar consideravelmente. Como observado, independentemente de como o viés de averiguação foi produzido seu efeito final é sempre a distorção do real estado do espectro de frequências. Como consequência, os dados obtidos pela genotipagem não podem ser analisados pelos testes de neutralidade, sem considerações prévias (Pickrell *et al*., 2009). Esse problema tem sido previamente reportado por (Kreitman e Di Rienzo, 2004). (Soldevila *et al*., 2005), demonstraram que os efeitos locais da ação da seleção balanceadora mostrados para o gene *PRPN* por (Mead *et al*., 2003) foram devidos ao viés de averiguação. De fato eles usaram a metodologia de descoberta que leva à perda de alelos de baixa frequência (Soldevila *et al*., 2005).

Embora nenhum teste de neutralidade possa ser propriamente usado devido ao viés de averiguação, muitos trabalhos vêm sendo desenvolvidos na tentativa de detectar esses casos. O principal esforço para resolver esse problema tem sido direcionado no desenvolvimento de testes como os novos métodos baseados na extensão haplotípica EHH (Sabeti et al. 2002), obter valores

críticos de intervalos de confiança para testes de neutralidade construídos com base em simulações que incorporam o viés de averiguação como aqueles trabalhados por (Voight *et al*., 2006) ou (Carlson *et al*., 2004), e diretamente corrigidos por estimadores estatísticos (Nielsen, 2000; Wakeley *et al*., 2001; Nielsen *et al*., 2004).

2.6 Publicações

2.6.1 Artigo II

CYBB, and NADPH-Oxidase Gene: Restricted Diversity in Humans and Evidence for differential Long-Term Purifying Selection on Transmenbrane and Cytosolic Domain

O complexo NAPH oxidase, é um complexo enzimático que cataliza a redução do oxigênio molecular para O_{2-} gerando espécies reativas de oxigênio (reactive oxygen species – (ROS)), uma reação crítica para a atividade anti-microbiana dos fagócitos, (Chanock *et al*., 1994; Heyworth *et al*., 2003). O complexo inclui duas proteínas transmembrana, as sub-unidades gp91-phox e gp22-phox (expressos pelos genes *CYBA* e *CYBB*), e três proteínas citoplasmáticas, as sub-unidades, p40-phox, p47-phox, and p67-phox (expressas pelos genes *NCF4, NCF1*e *NCF2*). Mutações em qualquer um dos quatro genes, *CYBB*, *CYBA NCF1* e *NCF2*, pode resultar na manifestação da doença granulomatose crônica (CGD;OMIM #306400).

Participei neste trabalho na discussão e análises envolvendo os testes de neutralidade inter-específicos. Dentre os testes de neutralidade, fiz as análises que utilizaram o Software MLHKA (Wright e Charlesworth, 2004), que implementa o método de verossimilhança para o teste de Hudson-Kreitman-Aguade (HKA). O teste usa o número de polimorfismos fixados entre humanos e um *outgroup*, (ex. chimpanzé, no caso), para testar se a taxa de evolução para duas regiões genômicas independentes depende somente das taxas de mutação, às quais estão sujeitas essas regiões como seria esperado em uma hipótese de neutralidade, (Hudson *et al*., 1987). Complementarmente neste trabalho, também fui responsável pela análise da associação entre a severidade de mudanças aminoacídicas ao longo da filogenia de cinco mamíferos utilizados no estudo, usando a matriz de Grantham. Esta matriz de distâncias classifica as alterações físico-

químicas entre aminoácidos (Grantham, 1974). Como resultado foi verificado que as mudanças causadoras da doença são, em média, mais severas que as mudanças inter-específicas.

HUMAN MUTATION 29(5), 623–632, 2008

RESEARCH ARTICLE

CYBB, an NADPH-Oxidase Gene: Restricted Diversity in Humans and Evidence for Differential Long-Term Purifying Selection on Transmembrane and Cytosolic Domains

Eduardo Tarazona-Santos,[1,2] Toralf Bernig,[1] Laurie Burdett,[3] Wagner C.S. Magalhaes,[2] Cristina Fabbri,[4] Jason Liao,[1] Rodrigo A.F. Redondo,[2] Robert Welch,[3] Meredith Yeager,[3] and Stephen J. Chanock[1]*

[1]*Section of Genomic Variation, Pediatric Oncology Branch, National Cancer Institute (NCI), National Institutes of Health (NIH), Gaithersburg, Maryland;* [2]*Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil;* [3]*Intramural Research Support Program, Scientific Applications International Corporation (SAIC) Frederick, NCI-Frederick, Maryland, and Core Genotype Facility, NCI, NIH, Gaithersburg, Maryland;* [4]*Area di Antropologia, Dipartimento di Biologia Evoluzionistica e Sperimentale, Università di Bologna, Bologna, Italia*

Communicated by Pui-Yan Kwok

CYBB encodes the gp91-phox protein of the phagocytic NADPH oxidase; the innate immunity-related enzymatic complex responsible for the respiratory burst. Mutations in CYBB can cause chronic granulomatous disease (CGD), a primary immunodeficiency characterized by ineffective microbicidal activity, for which over 150 family-specific mutations have been described. It is also plausible that common SNPs in CYBB alter the expression or function of gp91-phox, determining differences in susceptibility to complex disorders such as autoimmune or infectious diseases. We have resequenced the exons, UTRs, and intronic regions of CYBB in 102 ethnically diverse individuals and genotyped nine tag-SNPs in 942 individuals from 52 worldwide populations. The 28 observed SNPs (none of which nonsynonymous) reside on 28 haplotypes that can be collapsed into five clades. CYBB shows lower diversity than other X-chromosome genes and most of the between-population genetic variance was observed among Africans and non-Africans. The African population shows the highest diversity and the lowest linkage disequilibrium (LD). Because there is extensive shared LD among non-Africans, tag-SNPs can be effectively employed in gene-centric association studies and are portable across Eurasian and Native American populations. Comparison of CYBB coding sequences among mammals evidences the action of long-term purifying selection, which is stronger on the C-terminal cytosolic domain than on the N-terminal transmembrane domain of gp91-phox. Hum Mutat 29(5), 623–632, 2008. Published 2008 Wiley-Liss, Inc.†

KEY WORDS: population genetics; haplotypes; respiratory burst; tag-SNPs; linkage disequilibrium; innate immunity; *CYBB*

INTRODUCTION

The phagocyte NADPH oxidase, also known as the "respiratory burst oxidase," is an enzymatic complex that catalyzes the reduction of oxygen to O_2^- and generates reactive oxygen species (ROS), a critical reaction for the microbicidal activity of phagocytes [Chanock et al., 1994; Heyworth et al., 2003]. The NADPH oxidase includes two membrane-spanning polypeptide subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*), which comprise a flavocytochrome b_{558}, and three cytoplasmic polypeptide subunits, p40-phox, p47-phox, and p67-phox (encoded by *NCF4*, *NCF1*, and *NCF2*). Mutations in any one of four genes (*CYBB*, *CYBA*, *NCF1*, and *NCF2*) can result in chronic granulomatous disease (CGD; OMIM#306400), a primary immunodeficiency. Most CGD patients have no measurable respiratory burst and less than 5% generate a very low level of ROS [Heyworth et al., 2003]. Nearly 70% of CGD cases are X-linked, due to mutations in *CYBB* (Xp21.1; OMIM#306481; Fig. 1) [Winkelstein et al., 2000; Heyworth et al., 2003] and over 500 family-specific mutations have been described in this gene (*CYBB* browser database; http://bioinf.uta.fi/CYBBbase).

The Supplementary Material referred to in this article can be accessed at http://www.interscience.wiley.com/jpages/1059-7794/suppmat

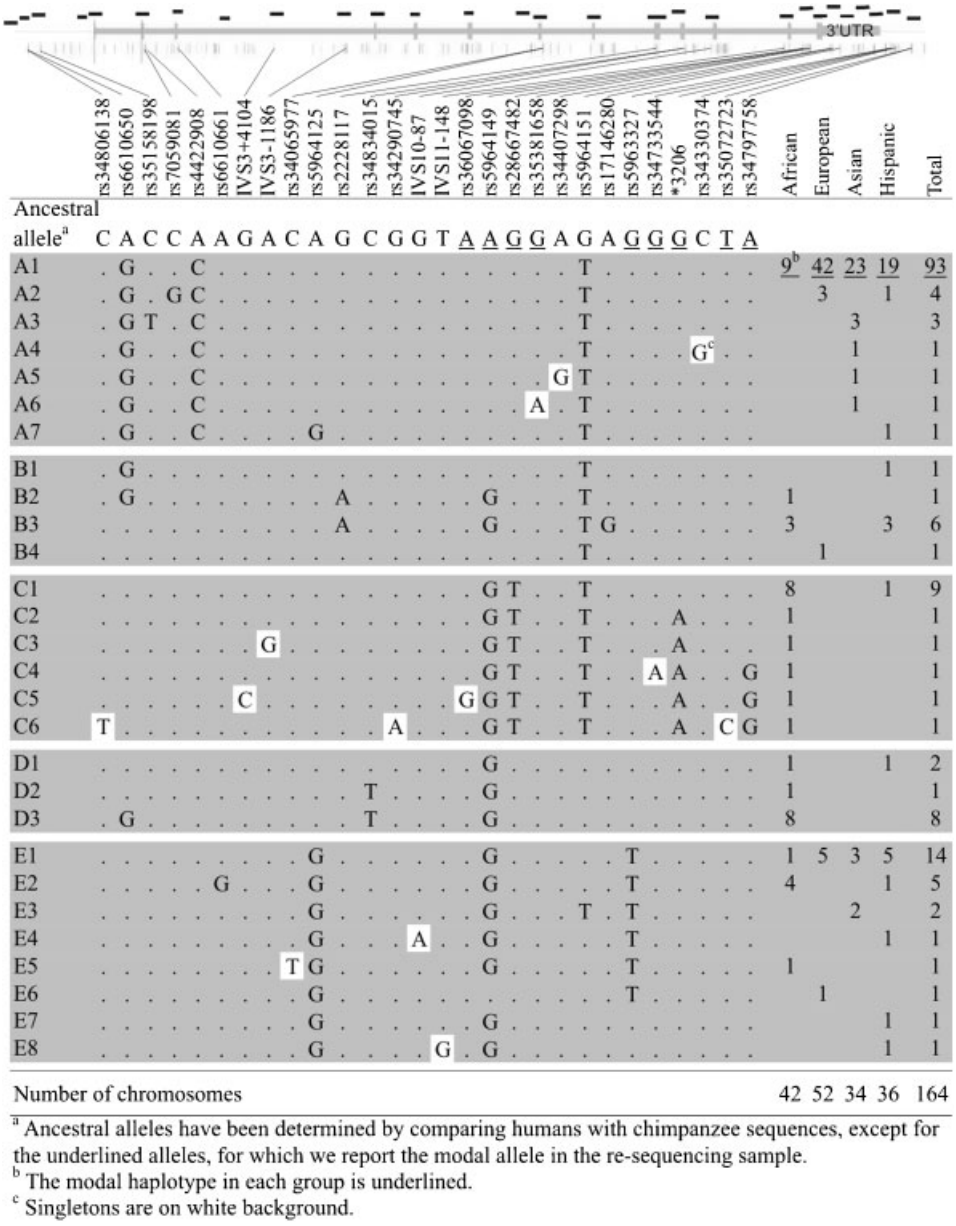
Received 28 June 2007; accepted revised manuscript 20 September 2007.

*Correspondence to: Stephen J. Chanock, Pediatric Oncology Branch, NCI, Advanced Technology Center, 8717 Grovemont Circle, Bethesda, MD 20892-4605. E-mail: chanocks@mail.nih.gov

Grant sponsors: Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute (NCI), Center for Cancer Research, University of Bologna, Conselho Nacional de Desenvolvimento Científico eTecnológico (CNPq) (Brazil), Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) (Brazil), Coordenação de Aperfeçoamento de Pessoal (CAPES) (Brazil).

DOI 10.1002/humu.20667

Published online 15 February 2008 in Wiley InterScience (www.interscience.wiley.com).



[a] Ancestral alleles have been determined by comparing humans with chimpanzee sequences, except for the underlined alleles, for which we report the modal allele in the re-sequencing sample.
[b] The modal haplotype in each group is underlined.
[c] Singletons are on white background.

FIGURE 1. **Genomic structure of *CYBB*, SNPs, and haplotype frequencies in the resequencing panel. Resequenced regions are denoted by horizontal bars above the gene. Exons are represented by vertical bars.**

Complementary work in animal models and in vitro has confirmed the significance of the NADPH oxidase for immunity against bacterial and fungal pathogens [Buckley, 2004]. Thus, it is plausible that subtler variation in its expression or function contributes to autoimmune or infectious diseases such as tuberculosis, malaria, and other parasitic infections [Wang et al., 2003]. For instance, Uhlemann et al. [2004] have reported an association between microsatellite alleles in the promoter region of *CYBB* and severity of malaria in Gabon populations, and correlative in vitro experiments suggest that these alleles could be associated with differences in NADPH oxidase activity. Skewed lyonization in female carriers of *CYBB* mutations is known to be associated with higher susceptibility to autoimmune diseases [Anderson-Cohen et al., 2003]. Moreover, Roos et al. [2003] have postulated that an imbalance in products of the respiratory burst could produce tissue damage in a range of diseases, such as gout, chronic obstructive pulmonary disease, rheumatoid arthritis, and also may be involved in the pathogenesis of cardiovascular diseases [Brandes and Kreuzer, 2005].

Despite the involvement of the NADPH oxidase in the pathogenesis of Mendelian and complex diseases, our knowledge of genetic variation in *CYBB* has been mostly derived from the analysis of X-linked CGD patients. Here, we have resequenced 164 chromosomes and complemented this analysis by selecting nine common tag-SNPs that can be used as surrogates for untested SNPs due to the pattern of linkage disequilibrium (LD) in 942 individuals from globally diverse populations [Cann et al., 2002]. We addressed the following issues: 1) the genetic variation within the coding region of *CYBB*; 2) the pattern of genetic diversity of *CYBB* across worldwide populations; 3) the portability of tag-SNPs in genetic epidemiology studies across populations—both within- and between-continents. Moreover, we tested the hypothesis that the pattern of variation of *CYBB* fits the neutral model across human populations and at a larger evolutionary scale.

MATERIALS AND METHODS

Samples

A total of two datasets of anonymous samples were used. The first one (i.e., the resequencing panel) was composed by 102 unrelated individuals of the SNP500Cancer project (http://snp500cancer.nci.nih.gov), which includes [Packer, et al. [2006]: Africans (19 females and four males from the United States and Pigmies), Europeans (22 females and eight males from the Utah pedigree from the Centre d'Etude du Polymorphisme Humaine (CEPH) and from the National Institute of Environmental Health Sciences (NIEHS) Environmental Genome Project), Asians (11 females and 12 males from Pakistan, China, Cambodia, Japan, Taiwan, and Melanesia), and Hispanics (13 females and 10 males from Mexico, Puerto Rico, and South America). The second dataset (i.e., the SNPs panel) derives from the Human Genome Diversity Panel (HGDP-CEPH), and includes 942 individuals (621 males and 321 females) from 52 global populations [Cann et al., 2002] from Sub-Saharan Africa (seven populations), North Africa (one), the Middle East (three), Europe (eight); Central-South Asia (nine), East Asia (17) Oceania (two), and the Americas (five). This version of the HGDP excludes one individual from each pairs of first- and second-degree relatives, matching as much as possible the H952 panel of Rosenberg [2006].

PCR Amplification, Sequencing, and Genotyping of SNPs

Oligonucloetide with M13 tails were designed to amplify and sequence critical regions of the gene. In total, 29 primer pairs were designed to include the 13 exons, the 5′-, the 3′-UTRs and intronic regions of *CYBB* (Reference sequence: *CYBB* transcript in the human genome build 35 v1; Fig. 1). Bidirectional sequence analysis was performed on 10,384 bp following the protocol described by Packer et al. [2006]. We genotyped nine common tag-SNPs discovered in the resequencing panel on the SNPs panel (see below for details about tag-SNPs selection). Genotyping were performed using *Taqman* assays (Applied Biosystems, Foster City, CA), as described in the SNP500Cancer website. SNPs without rs number are reported following rules of the SNP500Cancer database (https://snp500cancer.nci.nih.gov/terms_SNP.cfm). SNPs information is deposited in the public database SNP500Cancer and can be visualized in the public browser Genewindow (http://genewindow.nci.nih.gov) [Staats et al., 2005].

Haplotype Inferences and Population Genetic Analyses

Because *CYBB* resides on the X chromosome, male haplotypes were observed. This information was used to improve haplotype inferences on females, by using the method implemented in the software Phase v.2.1 [Stephens and Donnelly, 2003].

To estimate within-population diversity in the re-sequencing panel, two estimators of the parameter $\theta = 3N_e\mu$ were calculated: π, which is the per-site mean number of pairwise differences between sequences [Tajima, 1983], and θs, based on the number of segregating sites (S) [Watterson, 1975]. We also computed the haplotype diversity in both panels [Nei, 1987]. For the resequencing panel, we assessed the departure from the allelic spectra expected under neutrality using the D statistics of Tajima [1989], and the D and F statistics of Fu and Li [1993]. These analyses were performed using DNAsp 4.00 [Rozas et al., 2003]. Phylogenetic relationships were explored by calculating a Median Joining Network [Bandelt et al., 1999]. Differentiation between populations (F_{ST}) was calculated assuming the Tamura and Nei [1993] model of nucleotide substitution for resequencing data, and the pairwise number of differences for the SNPs panel. We used the analysis of molecular variance (AMOVA) [Excoffier et al., 1992] to assess the genetic structure of populations. Under the island model of population structure, the expected value of F_{ST} at equilibrium is $1/(1+4N_em)$ [Cavalli-Sforza and Bodmer, 1971], where N_e is the effective population size and m the migration rate. Based on this expectation, we corrected the F_{ST} values for X chromosome loci for comparison with autosomal loci by applying the formula $F_{STau}/(1–F_{STau}) = 0.75\ F_{STX}/(1–F_{STX})$. For the SNPs panel, the matrix of pairwise F_{ST} among the 52 populations was represented by a nonmetric multidimensional scaling calculated using Statistica v.4.0 (Statsoft Inc, Tulsa, OK). We used the software Arlequin 2.0 [Schneider et al., 2000] for F_{ST} and AMOVA calculations.

Patterns of LD and Tag-SNPs

We calculated the recombination parameter $\rho = 4N_er$ (r is the recombination rate between adjacent sites per generation) using the method developed by Li and Stephens [2003] and implemented in the software Phase 2.1. Pairwise LD was measured by r^2 [Hill and Robertson, 1968] and its significance was assessed calculating logarithm of the odds (LOD) scores [Gabriel et al., 2002]. We used the approach of Carlson et al. [2004] to identify "bins of linkage disequilibrium" (i.e., sets of *CYBB* SNPs that are in LD). For each "bin," we selected one representative "tag-SNP." For this procedure, we imposed an $r^2 \geq 0.80$ among tag- and tagged-SNPs. The software Haploview 3.2 [Barrett et al., 2005] was used for calculations of LD and tag-SNPs.

Evolutionary Inferences Based on Interspecific Variation

To test the fitness of the data to the neutral model of evolution at an interspecific level, two analyses were conducted: 1) The Hudson-Kreitman-Aguade (HKA) test [Hudson et al., 1987] uses fixed differences between humans and one chimpanzee to test if the rate of evolution for two genomic regions depends only on their mutation rate, as expected under neutrality. We used the maximum-likelihood version of this test implemented in the software MLHKA [Wright and Charlesworth 2004]. The following loci (for which there is no evidence of natural selection) were used for comparison with *CYBB*: *APOE* [Fullerton et al., 2000], *LPL* [Clark et al., 1998], *IL13* [Tarazona-Santos and Tishkoff, 2005], intron 44 of *DMD* [Nachman and Crowell 2000], and 10 X-chromosome loci studied by Kitano et al. [2003]. 2) We estimated the ratio of nonsynonymous (dN) to nonsynonymous substitutions (dS), $\omega = dN/dS$ (for which the expected value(s) under neutrality would be equal to 1), using the maximum likelihood approach implemented in the software PALM [Yang, 1997]. We performed comparisons for the *CYBB* coding sequences between *H. sapiens* (BC032720.1), *P. troglodytes*, and *P. pygmaeus* (sequenced by us), *B. taurus* (NM_174035.2), *M. musculus* (NM_007807.2), and *R. rattus* (NM_023965.1). To compare the spectra of nonsynonymous changes for CGD patients and at an interspecific level, we used the matrix of chemical distances among amino acids of Grantham [1974].

RESULTS

Patterns of Diversity

We did not find in the resequencing panel carriers of any reported CGD mutations (Fig. 1; CYBB browser database in January 2007 or those summarized by Heyworth et al. [2001], Jirapongsananuruk et al. [2002], and Stasia et al. [2005]). The absence of nonsynonymous SNPs suggests that this type of

TABLE 1. Analysis of Intrapopulation Diversity, Recombination and Test of Neutrality Based on Resequence Analysis of the Four SNP500Cancer Populations

Populations	African	European	Asian	Hispanic	Eurasian	World
Number of chromosomes	42	52	34	36	86	164
Segregating sites	21	7	10	13	11	28
Singletons	8	0	3	5	3	13
Haplotype structure						
Number of inferred haplotypes	14	5	7	12	10	28
Number of common haplotypes (frequency > 0.05)	12	5	4	9	5	12
Haplotype diversity ± SD	0.88 ± 0.03	0.34 ± 0.08	0.53 ± 0.10	0.70 ± 0.08	0.42 ± 0.07	0.66 ± 0.04
R_{min}[a]	1	0	0	2	1	4
ρ	8×10^{-5}	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	2×10^{-5}	8×10^{-5}
θ estimators						
$\pi \pm$ SD ($\times 10^3$)	0.36 ± 0.03	0.12 ± 0.03	0.15 ± 0.04	0.28 ± 0.03	0.13 ± 0.03	0.26 ± 0.02
$\theta_W \pm$ SD ($\times 10^3$) (per site)	0.42 ± 0.15	0.13 ± 0.06	0.21 ± 0.09	0.27 ± 0.11	0.19 ± 0.07	0.42 ± 0.12
Neutrality tests						
Tajima's D	−0.473	−0.274	−0.813	0.084	−0.789	−1.106
Fu and Li's D	−1.050	1.110	0.395	−0.977	0.345	−2.627[b]
Fu and Li's F	−0.980	0.811	0.114	−0.814	0.042	−2.399[b]

[a]Minimum number of recombination events.
[b]$P < 0.05$.

substitution in *CYBB* is usually deleterious in humans. As expected, we observed variants throughout the 5′flanking region, the 3′UTR, and in introns. Both the resequencing and SNPs panels show the highest diversity in African populations (Tables 1 and 2). The higher geographic resolution of the SNPs panel revealed that within the Asia-Oceania region, East Asian and Oceania are the least and the most diverse populations, respectively. Moreover, in an analysis of the set of 164 chromosomes, the allelic spectra (i.e., how many substitutions are observed across different classes of allele frequency) are those expected for a locus that has evolved under neutrality, as evidenced by the nonsignificance of Tajima's D and Fu-Li's D and F tests of neutrality (Table 1).

Haplotype Structure and Distribution

In an analysis of the resequencing panel, we calculated a median joining network [Bandelt et al. 1999], that shows the relationships between the 28 observed/inferred haplotypes and their distribution across the four self-described ethnic groups. Results in Figures 1 and 2 reveal that: 1) Haplogroups (i.e., groups of closely related haplotypes) A and E are the most differentiated, and include respectively, the first and second most common and ubiquitous haplotypes A1 and E1. 2) Haplogroups C and D are predominately observed in African populations, but are also present in Hispanics, probably due to Post-Columbian admixture. These results are consistent with the distribution of the haplotypes defined by the nine common SNPs genotyped in the SNPs-panel (Table 2). The portion of genetic variance allocated among populations (F_{ST}) is 0.24 ($P < 0.0001$) when Africans, Europeans, and Asians of the resequencing panel are considered (excluding Hispanics because showing a wide variation in admixture levels), and 0.21 ($P < 0.0001$; Fig. 3) on the basis of the 52 worldwide populations of the SNPs-panel. When these values are corrected for effective population size, to allow comparisons with autosomal loci, they correspond to 0.19 and 0.16, respectively, which are slightly higher than the average F_{ST} calculated among human populations (0.10–0.12) [Barbujani et al., 1997]. The analysis of the genetic structure for both panels (Table 3 and Fig. 3, respectively) shows that: first, the largest differentiation is between African and non-African populations; second, Eurasian populations are homogenous ($F_{ST} = 0.01$, $P = 0.10$), although small differences in the haplotype distribution are observed (see Supplementary Tables S1 and S2, available online at http://www.interscience.wiley.com/jpages/1059-7794/suppmat); third, only a small portion of genetic variance is observed among populations within the geographic groups defined in Table 2 ($F_{SC} = 0.03$, $P < 0.001$). Furthermore, the analysis of the SNPs-panel shows that North-African and Oceania populations are differentiated from Europeans and Asians (Fig. 3; Table 2).

LD and Tag-SNPs

As expected based on its larger effective population size [Tishkoff and Verrelli, 2003], the African population shows a larger recombination parameter ρ (Table 1 for the resequencing panel) and lower LD than non-Africans (Fig. 4).

We have identified tag-SNPs in the resequencing panel using the "tagger" approach, which is based on the analysis of pairwise LD [Carlson et al., 2004]. As expected, the number of tag-SNPs required to capture the haplotype structure of *CYBB* in Europeans (two tag-SNPs), Asians (four), and Hispanics (five) is smaller than that for Africans (11 tag-SNPs; see Fig. 4A). We tested the robustness of LD estimates and the portability of tag-SNPs by comparisons with the SNPs panel, in which we performed LD analysis and verified tag-SNPs in the following groups of homogeneous populations (Fig. 4): Sub-Saharan Africa, Europe, the Middle East, Central Asia, East Asia, and Oceania. In general, for a specific population, the pattern of LD is similar between the two sets analyzed in this study, and the tag-SNPs selection is also comparable. Moreover, because there is substantial shared LD across non-African populations, tag-SNPs are portable across these groups. An exception is observed among the resequenced Hispanics and Native Americans from the SNPs panel, due to the large European ancestry (as high as 70%) of the former (data not shown).

Pattern of Interspecific Divergence and the Spectra of Nonsynonymous Changes

We combined data from the resequencing panel and data for divergence from chimpanzee to investigate if natural selection has acted on human–chimpanzee *CYBB* lineages. The HKA test revealed that both for the entire human sample as well as for each of the four studied populations, the neutral model of evolution accounts for the small differences among the ratio of polymorphisms to fixed human–chimpanzee differences calculated for *CYBB*

TABLE 2. *CYBB* Haplotype Frequencies in the SNPs Panel in Eight Worldwide Regional Population Groups*

	rs7059081	rs4422908	rs6610661	rs5964125	rs5964149	rs28667482	rs34407298	rs5964151	rs5963327	Africa	North Africa	Middle East	Europe	Central South Asia	East Asia	Oceania	America	World
Ancestral allele[a]	C	A	A	A	A	G	A	G	G									
A1,3,4,6	.	C	.	.	.	.	.	T	.	19(0.16)	22(0.58)	178(0.84)	179(0.80)	191(0.83)	274(0.92)	14(0.40)	79(0.75)	956
A2	G	C	.	.	.	.	.	T	.		3(0.08)	8(0.04)	13(0.06)	4(0.02)		1(0.01)		29
A5	.	C	.	.	.	.	G	T	.							3(0.09)		3
A7	.	C	.	G	.	.	.	T	.				1(0.00)					1
A8	.	C	.	.	.	.	.	T	T				1(0.00)					1
A9	.	C	.	.	.	T	.	T	.	1(0.01)								1
B1,4	.	.	.	.	.	.	.	T	.					2(0.01)				2
B2,3	.	.	.	.	G	.	.	T	.	22(0.18)		3(0.01)		1(0.00)			1(0.01)	27
B5	.	.	.	G	G	.	.	T	.	1(0.01)								1
C1-6	.	.	.	.	G	T	.	T	.	35(0.29)	2(0.05)	2(0.01)		2(0.01)		5(0.14)		46
D1-3	.	.	.	.	G	.	.	.	.	25(0.21)	4(0.11)	1(0.00)						30
E1	.	.	.	G	G	.	.	.	T	9(0.07)	6(0.16)	21(0.10)	29(0.13)	28(0.12)	23(0.08)	13(0.37)	23(0.22)	152
E2	.	.	G	G	G	.	.	.	T	9(0.07)				1(0.00)				10
E9	.	C	.	G	G	.	.	.	T				2(0.01)					2
G1	.	C	.	.	G	.	.	.	T		1(0.03)						1(0.01)	2
Number of chromosomes										121	38	213	225	229	297	35	105	1263
Number of different haplotypes										8	6	6	6	7	2	4	5	15
Number of polymorphic sites										7	7	7	6	8	5	7	6	9
Haplotype diversity										0.81	0.64	0.29	0.35	0.29	0.14	0.69	0.39	0.41
S.D. of haplotype diversity										0.01	0.08	0.04	0.04	0.04	0.03	0.04	0.05	0.02

*Since haplotypes are based on genotyping a subset of 9 SNPs from those of Figure 1, some haplotypes of Figure 1 are collapsed in this table. Haplotypes A8, A9, B5, E9, and G1 have not been identified in the resequencing panel, and have been inferred for the first time in the SNPs-panel. Relative frequencies are showed in parentheses. In this table, populations have been collapsed in regional groups.Haplotype frequencies for each of the 52 populations are in supplementary material (available online).

[a]Ancestral alleles have been determined by comparing humans with chimpanzee sequences, except for the underlined alleles, for which we report the modal allele in the resequencing panel.

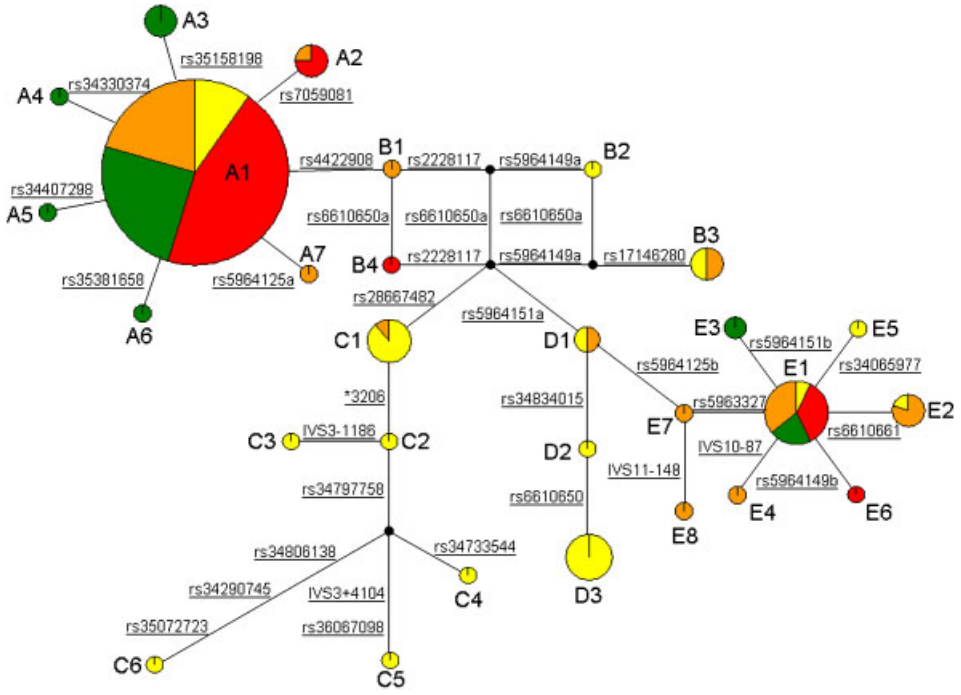


FIGURE 2. Median joining network of the CYBB haplotypes observed in the resequencing panel. The network was constructed assuming ε = 0. The resulting network is unrooted and assumes that 32 substitutions occurred. The data are compatible with a minimum of four recombination events or four substitutions that occurred twice: rs6610650, rs5964125, rs5964149, and rs5964151. The sizes of the circles are proportional to haplotype frequencies in the worldwide sample. For each haplotype, its presence in each population is denoted by colors: yellow for African, red for European, green for Asian, and orange for the admixed Hispanic, and the areas for each of the color is proportional to the fraction of that haplotypes presents on each of the four populations.

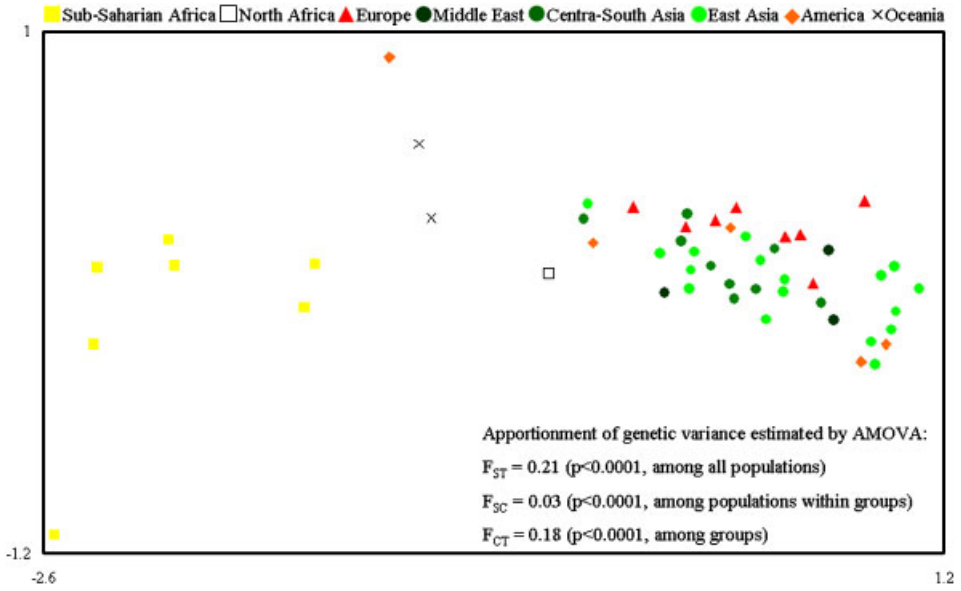


FIGURE 3. Nonmetric MDS representation of pairwise F_{ST} values among the 52 populations of the SNPs panel (stress value = 0.07). The result of the AMOVA, based on the 52 populations grouped in eight regional groups (in different colors), is reported. The genetic distance matrix used to perform the MDS is available online as Supplementary Table S2.

TABLE 3. Analysis of Differentiation Between Populations for the Resequencing panel

	Africa	Europe	Asia	Hispanic
Africa		**0.316**	**0.264**	**0.092**
Europe	**0.257**		0.000	**0.107**
Asia	**0.211**	0.000		**0.073**
Hispanic	**0.070**	**0.082**	**0.056**	

Estimated F_{ST} values are above the diagonal. Below the diagonal are the F_{ST} values corrected as if the effective population sizes of X chromosome genes were equal to autosomal ones. Values in bold denote F_{ST} values significantly different from 0.

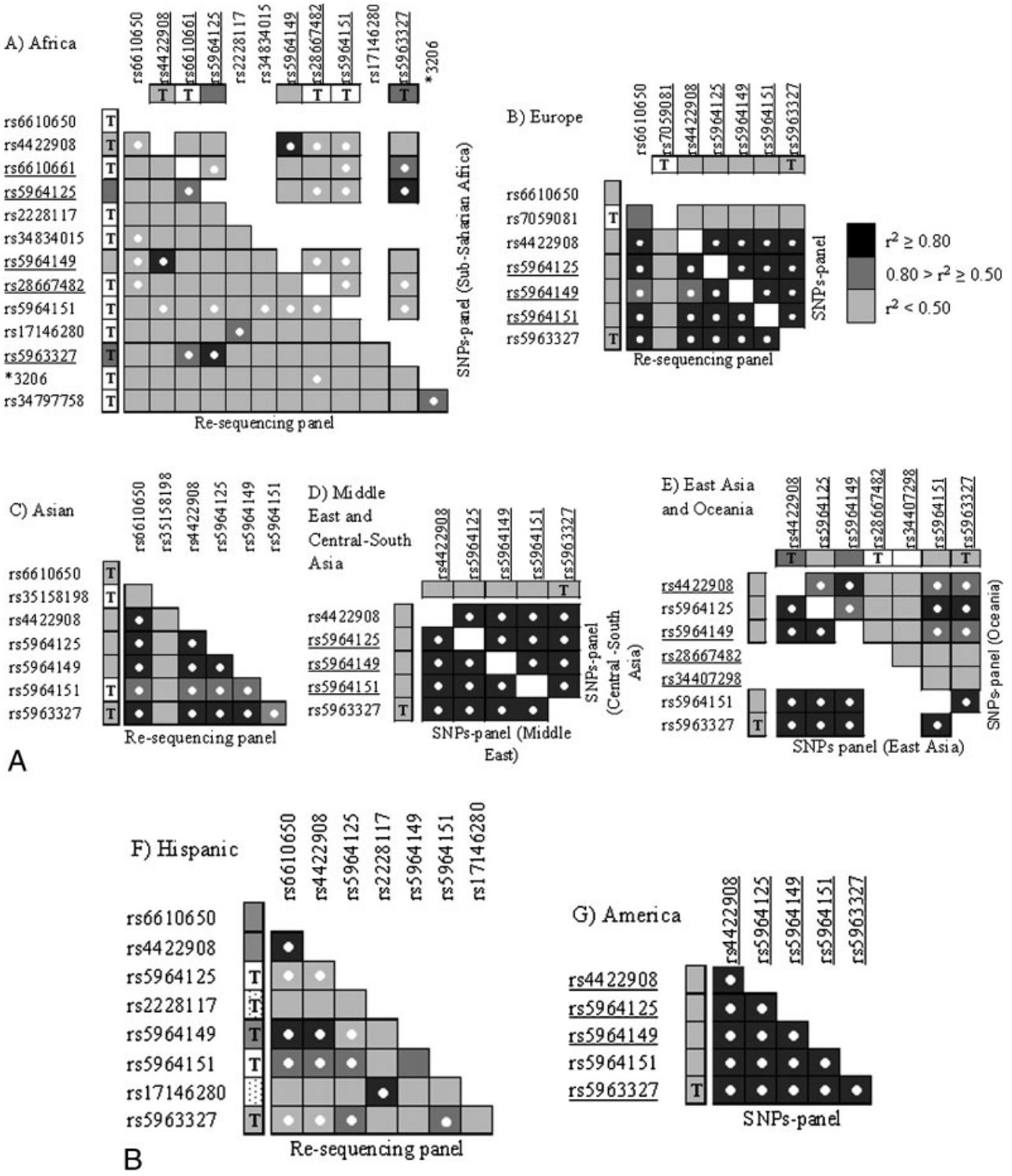


FIGURE 4. Pairwise LD across *CYBB* and tag-SNPs. LD is assessed by r^2 for common SNPs (minor allele frequency [MAF] $>5\%$) in the resequencing and SNPs panels. SNPs genotyped in the SNPs panel are underlined. Significant r^2 values (LOD >2) are denoted by white circles. Tag-SNPs identified by the "tagger" algorithm [Carlson et al., 2004] are denoted by T and are shown for the resequencing panel and for the SNPs panel for eight regional groups of populations: Sub-Saharan Africa (A, upper triangle), Europe (B, upper triangle), Middle East (D, lower triangle), Central Asia (D, upper triangle), East Asia (E, lower triangle), and Oceania (E, upper triangle). Each tag-SNP is a surrogate for SNPs associated through an $r^2 > 0.80$. On the vertical and horizontal bars we use the same non-white backgrounds to represent a tag-SNP and the corresponding set of tagged SNPs. For instance, for Africans, rs5963327 is a tag-SNP of rs5964125 and therefore, is represented on the vertical and horizontal bars on the same dark gray background. White backgrounds on the horizontal and vertical bars correspond to *single* tag-SNPs (i.e., that do not have tagged SNPs).

and the same ratio calculated for other loci assumed to be neutral (see Supplementary Table S3 for results).

Considering the absence of nonsynonymous changes in the human–chimpanzee lineage, we expanded our analysis to gain statistical power. By using one coding sequences from human, chimpanzee, orangutan, cow, mouse, and rat, we estimated that for *CYBB*, the rate of nonsynonymous substitutions (dN) is only ~9% the rate of synonymous ones (dS) ($\omega = dN/dS = 0.11$; $P < 0.0001$). The ω value is almost the same as that calculated for the lineages of great apes, by assuming a model that allows variation of ω across different branches of the phylogeny [Yang, 1997] (data not shown).

We further compared the spectra of nonsynonymous X-linked CGD mutations (from the *CYBB* Mutation Browser) with amino acid changes across the phylogeny of the five analyzed mammals and verified that disease mutations are on average more radical than interspecific substitutions (average Grantham values: 92.06 vs. 68.14, Mann-Whitney U test = 2739.5, $P < 0.01$) [Grantham, 1974]. Furthermore, we conducted separate analyses of the two parts of the gp91-phox peptide, the N-terminal half, which includes six transmembrane domains, the heme moieties, and interacts with gp22-phox; and the C-terminal half, which is the cytosolic component containing NADPH- and FAD-binding sites. Our interspecific analysis shows that the transmembrane N-terminal half is more variable (71 amino acid changes in 277 residuals) and show a larger $\omega = 0.19$ than the cytosolic C-terminal half (30 amino acid changes in 293 residuals, $\omega = 0.05$). Thus, the latter has evolved under a stronger purifying selection at least during mammalian evolutionary history. On the other hand, although no significant differences were observed in the number of CGD mutations among the N-terminal half of gp91-phox (44 mutations) and the C-terminal half (42 mutations), CGD amino acid changes in the former appear to be on average more radical (average Grantham values: 106.7) than in the latter (average Grantham values: 88.4), although this difference is not significant (Mann-Whitney U test = 788.5; $P = 0.24$).

DISCUSSION

To understand the pattern of diversity at *CYBB* in humans, we first resequenced ~30% of the *CYBB* locus (including the entire coding region) in 102 individuals from five continents and then, to achieve higher geographic resolution, we genotyped a set of nine informative SNPs (tag-SNPs) on a set of 942 individuals from 52 globally diverse populations (the Human Genome Diversity Panel). This combined approach has permitted, for instance, to perceive small differences in the pattern of within-population diversities within Asia, and to determine the *CYBB* haplotype

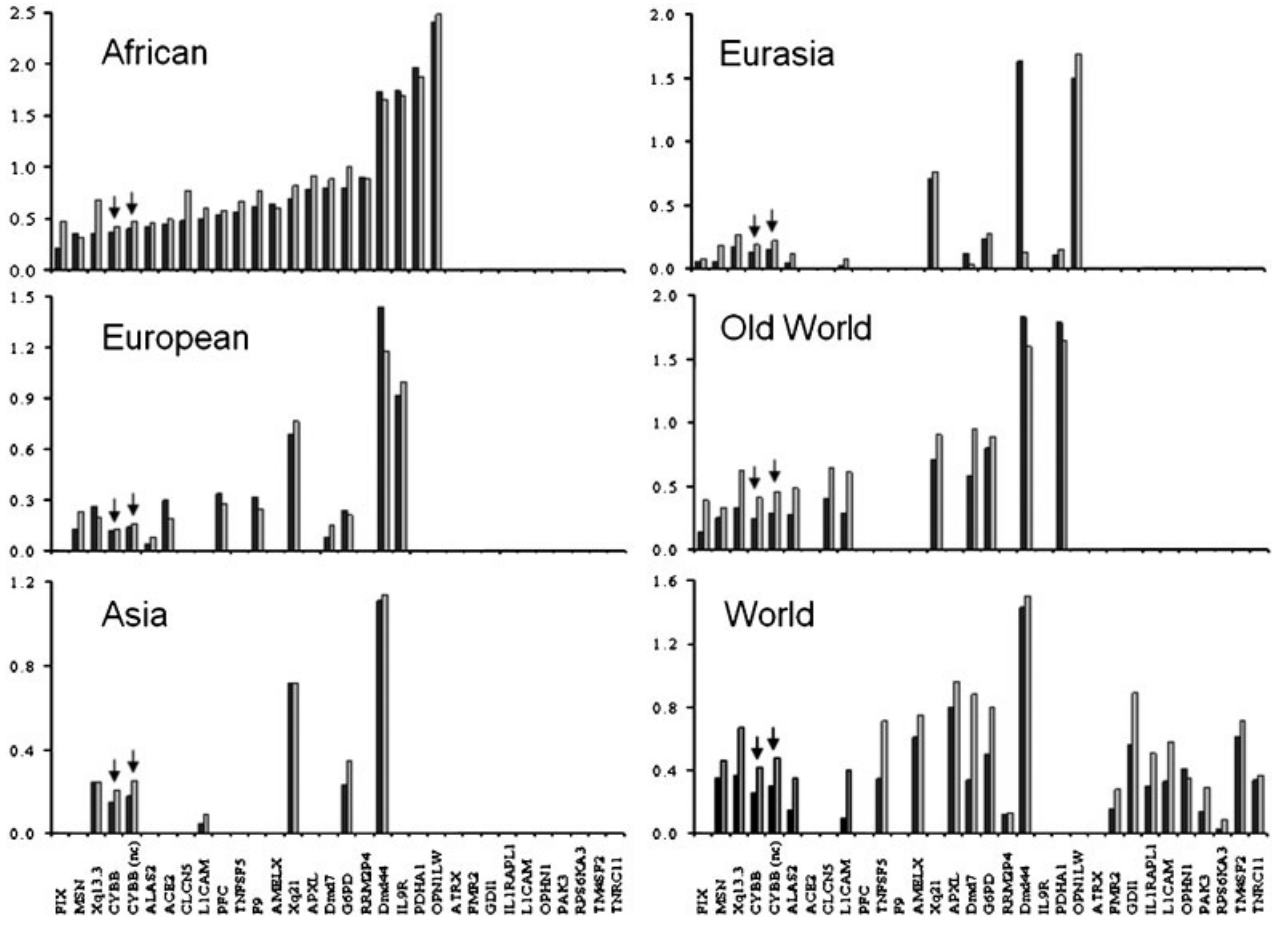


FIGURE 5. **Nucleotide diversity (π, dark grey bars) and θ_w (light grey bars) of *CYBB* and published X chromosome genes based on resequencing analysis. Comparative data are extracted from eleven publications: (Akey *et al.*, 2004) (*ACE2, PFC, F9, IL9R*), (Harris and Hey, 2001) (*FIX, PDHAI*), (Hammer *et al.*, 2004) (*APXL, AMELX, TNFSF5, RRM2P4*), (Nachman *et al.*, 2004) (*ALAS2, MSN*), (Nachman and Crowell, 2000) (introns 44 and 7 of *DMD*), (Nachman and Crowell, 2000) (*L1CAM, G6PD*), (Verrelli and Tishkoff, 2004) (*OPN1LW*), (Kitano *et al.*, 2003) (*ATRX, FMR2, GDI1, IL1RAPL1, L1CAM, OPHNI, PAK3, RPS6KA3,TM4SF2,TNRC11*), (Alonso and Armour, 2004) (*CLCN5*), (Kaessmann *et al.*, 1999) (Xq13.3) and (Yu *et al.*, 2002) (Xq21.1-21.33). CYBB(nc): non coding region of *CYBB*.**

structure in autochthonous Native American and Oceania populations, which are underrepresented in genetic studies.

Heyworth et al. [2001] and the *CYBB* Mutation Browser summarize more than 500 mutations in CGD patients, which are scattered across all *CYBB* exons and include missense and nonsense mutations, insertion and deletions producing frameshifts, and mutations in the promoter and splice sites. The contrast among this wide spectrum of disease mutations and the absence of nonsynonymous SNPs in healthy individuals from different populations suggests that mutations of this type are deleterious and that the coding region is highly constrained.

Extending the evolutionary scale of our analyses, we found that these contrasting spectra are consistent with the observed low rate of nonsynonymous substitutions calculated from sequences of five mammals, which is only 9% that of synonymous changes. This is evidence of strong purifying selection across mammalian *CYBB* lineages. However, this pattern observed across mammals does not exclude episodes of evolution of *CYBB* in which different evolutionary forces have been predominant. For instance, the analysis of the rhesus macaque genome provides evidence for positive natural selection, acting probably on the *CYBB* rhesus lineage [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007]. We also observed that disease mutations are on average more radical than interspecific amino acid substitutions, consistent with previous observations by Miller and Kumar [2001] for other genes involved in Mendelian diseases. Based on an analysis of interspecific comparison, the cytosolic C-terminus of the gene is less variable than the trans-membrane N-terminus, and has probably been under stronger purifying selection. Altogether, our analyses suggest that conservative amino acid substitutions such as those observed at an interspecific level are not well tolerated in humans. Their absence in our human sample indicates that if they exist in human populations, they are very rare.

Our analysis of the *CYBB* haplotype structure on globally diverse populations may be summarized as follows: 1) *CYBB* nucleotide diversity (π) is low when compared to other X-chromosome genes, in particular for European and African populations (Fig. 5). 2) Intrapopulation genetic diversity for *CYBB* is higher in African than in non-African populations, which is consistent with the larger effective population size of the former [Tishkoff and Verrelli, 2003]. 3) Most of the genetic variance among populations is due to comparisons between African and non-Africans, while Eurasian populations are quite similar. 4) Due to the large extent of shared LD among populations (Fig. 4), for genetic epidemiology studies, tag-SNPs are portable among samples from the same continent and across Eurasian populations. Also, *CYBB* tag-SNPs ascertained in Eurasia appear to be portable to Native Americans.

In conclusion, we have characterized the genetic diversity of *CYBB* and shown that this gene is a target of ancient purifying selection. However, natural selection appears to be weaker in the transmembrane terminus compared to the cytosolic terminus. The observed intra- and interpopulation diversity in non-African populations are consistent with the extensive, shared LD and therefore, for the design of genetic epidemiological studies of complex disease involving *CYBB*, tag-SNPs are portable among populations with little or no African ancestry.

ACKNOWLEDGMENTS

We thank Renee Chen, Maureen Kiley, Andrew Eckert, Shafaq Presswala, and the Sequencing Group of the Core Genotyping Facility for their technical assistance; and to Gilles Thomas, Sharon Savage, James Taylor VI, and Silvia Fuselli for discussions. C.F. was a supported by the University of Bologna, E.T.S. by CNPq (Brazil) and FAPEMIG (Brazil); R.R. by CNPq, and W.C.S.M. by CAPES (Brazil).

REFERENCES

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286.

Alonso S, Armour JA. 2004. Compound haplotypes at Xp11.23 and human population growth in Eurasia. Ann Hum Genet 68:428–437.

Anderson-Cohen M, Holland SM, Kuhns DB, Fleisher TA, Ding L, Brenner S, Malech HL, Roesler J. 2003. Severe phenotype of chronic granulomatous disease presenting in a female with a de novo mutation in gp91-phox and a non familial, extremely skewed X chromosome inactivation. Clin Immunol 109:308–317.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48.

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. Proc Natl Acad Sci USA 94:4516–4519.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Brandes RP, Kreuzer J. 2005. Vascular NADPH oxidases: molecular mechanisms of activation. Cardiovasc Res 6511:16–27.

Buckley RH. 2004. Pulmonary complications of primary immunodeficiencies. Paediatr Respir Rev 5(Suppl A):S225–S233.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. 2002. A human genome diversity cell line panel. Science 296:261–262.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 7411:106–120.

Cavalli-Sforza LL, Bodmer W. 1971. The genetics of human population. New York: Dover Publications.

Chanock SJ, el Benna J, Smith RM, Babior BM. 1994. The respiratory burst oxidase. J Biol Chem 269:24519–24522.

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595–612.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881–900.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Hammer MF, Garrigan D, Wood E, Wilder JA, Mobasher Z, Bigham A, Krenz JG, Nachman MW. 2004. Heterogeneous patterns of variation

among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. Genetics 167:1841–1853.

Harris EE, Hey J. 2001. Human populations show reduced DNA sequence variation at the factor IX locus. Curr Biol 11:774–778.

Heyworth PG, Curnutte JT, Rae J, Noack D, Roos D, van Koppen E, Cross AR. 2001. Hematologically important mutations: X-linked chronic granulomatous disease (second update). Blood Cells Mol Dis 2711:16–26.

Heyworth PG, Cross AR, Curnutte JT. 2003. Chronic granulomatous disease. Curr Opin Immunol 15:578–584.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 1161:153–159.

Jirapongsananuruk O, Niemela JE, Malech HL, Fleisher TA. 2002. CYBB mutation analysis in X-linked chronic granulomatous disease. Clin Immunol 104:73–76.

Kaessmann H, Heissig F, von Haeseler A, Paabo S. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22:78–81.

Kitano T, Schwarz C, Nickel B, Paabo S. 2003. Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. Mol Biol Evol 20:1281–1289.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233.

Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328.

Nachman MW, Crowell SL. 2000. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, Dmd, in humans. Genetics 155:1855–1864.

Nachman MW, D'Agostino SL, Tillquist CR, Mobasher Z, Hammer MF. 2004. Nucleotide variation at Msn and Alas2, two genes flanking the centromere of the X chromosome in humans. Genetics 167:423–437.

Nei M. 1987. Molecular evolutionary genetics. New York: Princeton University Press.

Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, Sicotte H, Staats B, Acharya M, Crenshaw A, Eckert A, Puri V, Gerhard DS, Chanock SJ. 2006. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res 34(Database issue):D617–D621.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

Roos D, van Bruggen R, Meischl C. 2003. Oxidative killing of microbes by neutrophils. Microbes Infect 5:1307–1315.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70:841–847.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin v. 2.0: a software for population genetics data analysis. Genetics and Biometrics Laboratory. University of Geneva, Switzerland. p 111.

Staats B, Qi L, Beerman M, Sicotte H, Burdett LA, Packer B, Chanock SJ, Yeager M. 2005. Genewindow: an interactive tool for visualization of genomic variation. Nat Genet 37:109–110.

Stasia MJ, Bordigoni P, Floret D, Brion JP, Bost-Bru C, Michel G, Gatel P, Durant-Vital D, Voelckel MA, Li XJ, Guillot M, Maquet E, Martel C, Morel F. 2005. Characterization of six novel mutations in the CYBB gene leading to different sub-types of X-linked chronic granulomatous disease. Hum Genet 116:72–82.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526.

Tarazona-Santos E, Tishkoff SA. 2005. Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. Genes Immun 6:53–65.

Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340.

Uhlemann AC, Szlezak NA, Vonthein R, Tomiuk J, Emmer SA, Lell B, Kremsner PG, Kun JF. 2004. DNA phasing by TA dinucleotide microsatellite length determines in vitro and in vivo expression of the gp91phox subunit of NADPH oxidase and mediates protection against severe malaria. J Infect Dis 189:2227–2234.

Verrelli BC, Tishkoff SA. 2004. Signatures of selection and gene conversion associated with human color vision variation. Am J Hum Genet 75:363–375.

Wang JP, Rought SE, Corbeil J, Guiney DG. 2003. Gene expression profiling detects patterns of human macrophage responses following *Mycobacterium tuberculosis* infection. FEMS Immunol Med Microbiol 39:163–172.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276.

Winkelstein JA, Marino MC, Johnston RB, Jr, Boyle J, Curnutte J, Gallin JI, Malech HL, Holland SM, Ochs H, Quie P, Buckley RH, Foster CB, Chanock SJ, Dickler H. 2000. Chronic granulomatous disease. Report on a national registry of 368 patients. Medicine (Baltimore) 79: 155–169.

Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071–1076.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556.

Yu N, Fu YX, Li WH. 2002. DNA polymorphism in a worldwide sample of human X chromosomes. Mol Biol Evol 19:2131–2141.

2.6.2 Artigo III

Diversity in the Glucose Transporter-4 Gene (*SLC2A4*) in Humans Reflects the Action of Natural Selection along the Old-World Primates Evolution

A glicose é uma das principais fontes de energia para quase todos os organismos. Em organismos vertebrados, pode ser ingerida através da dieta e transportada para dentro das células por diferentes mecanismos e moléculas, como por exemplo uma famíla de proteína de transportadores transmenbrana de glicose (Glucose Transporters- GLUTs). Membros dessa família possuem uma distribuição tecido específica, propriedades bioquímicas e fisiológicas que juntas regulam os níveis de acúcar no sangue e sua distribuição. GLUT4 – expresso pelo gene *SLC2A4*, é um transportador de glicose regulado pelos níveis de insulina no sangue com um papel crítico na homeostase da glicose.

Este trabalho teve como objetivo analisar o papel da seleção natural sobre o gene *SLC2A4*. Dentro desta proposta participei na discussão e construção dos cenários adicionais ao cenário padrão de evolução sobre neutralidade com tamanho populacional constante (Hudson, 2002). Estes cenários foram utilizados para gerar simulações e acessar a significância dos testes de neutralidade D de Tajima (Tajima, 1989) e F de Fu e Li (Fu e Li, 1993). Todos os cenários foram desenvolvidos utilizando o software ms (Hudson, 2002). Para testar para a presença de regiões com altas taxas de recombinação utilizamos o software SequenceLDhot (Fearnhead, 2006). Apliquei o teste de homozigosidade haplotípica extendida (*Extended Haplotype Homozigosity* - EHH), o qual mede se um alelo ou haplótipo específico que está sobre ação da seleção natural apresenta valores maiores de desequilibrio de ligação que o esperado sobre neutralidade, com regiões genômicas adjacentes, (Sabeti *et al*., 2002). Para testar a influência da seleção natural sobre o gene *SLC2A4* de uma forma mais ampla também foram aplicados os

testes implementados no pacote PAML (Yang, Z. H., 2007). O pacote PAML implenta a metodologia de maxíma verossimilhança (Yang, Z. G., 2007) para estimar as razões entre substituições não sinônimas (dN) e sinônimas (dS), representado por omega (ω), ω = dN/dS, para os códons de *SLC2A4* sobre a hipótese de vários modelos evolutívos. Esses modelos permitem inferir sobre a evolução dos códons ao longo da filogenia e discriminar entre os códons aqueles que experimentaram eventos mais acentuados ou mais brandos de seleção purificadora, neutralidade e seleção adaptativa.

OPEN ACCESS Freely available online

PLoS one

Diversity in the Glucose Transporter-4 Gene (*SLC2A4*) in Humans Reflects the Action of Natural Selection along the Old-World Primates Evolution

Eduardo Tarazona-Santos[1,2☯]*, Cristina Fabbri[1,3☯], Meredith Yeager[4,5], Wagner C. Magalhaes[1,2], Laurie Burdett[4,5], Andrew Crenshaw[4,5], Davide Pettener[3], Stephen J. Chanock[1]

1 Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Bologna, Italy, **4** Intramural Research Support Program, SAIC Frederick, National Cancer Institute - Frederick Cancer Research and Development Center (NCI-FCRDC), Frederick, Maryland, United States of America, **5** Core Genotype Facility, National Cancer Institute, National Institutes of Health, Gaithersburg, Maryland, United States of America

Abstract

Background: Glucose is an important source of energy for living organisms. In vertebrates it is ingested with the diet and transported into the cells by conserved mechanisms and molecules, such as the trans-membrane Glucose Transporters (GLUTs). Members of this family have tissue specific expression, biochemical properties and physiologic functions that together regulate glucose levels and distribution. GLUT4 –coded by *SLC2A4* (17p13) is an insulin-sensitive transporter with a critical role in glucose homeostasis and diabetes pathogenesis, preferentially expressed in the adipose tissue, heart muscle and skeletal muscle. We tested the hypothesis that natural selection acted on *SLC2A4*.

Methodology/Principal Findings: We re-sequenced *SLC2A4* and genotyped 104 SNPs along a ~1 Mb region flanking this gene in 102 ethnically diverse individuals. Across the studied populations (African, European, Asian and Latin-American), all the eight common SNPs are concentrated in the N-terminal region upstream of exon 7 (~3700 bp), while the C-terminal region downstream of intron 6 (~2600 bp) harbors only 6 singletons, a pattern that is not compatible with neutrality for this part of the gene. Tests of neutrality based on comparative genomics suggest that: (1) episodes of natural selection (likely a selective sweep) predating the coalescent of human lineages, within the last 25 million years, account for the observed reduced diversity downstream of intron 6 and, (2) the target of natural selection may not be in the *SLC2A4* coding sequence.

Conclusions: We propose that the contrast in the pattern of genetic variation between the N-terminal and C-terminal regions are signatures of the action of natural selection and thus follow-up studies should investigate the functional importance of differnet regions of the *SLC2A4* gene.

Citation: Tarazona-Santos E, Fabbri C, Yeager M, Magalhaes WC, Burdett L, et al. (2010) Diversity in the Glucose Transporter-4 Gene (*SLC2A4*) in Humans Reflects the Action of Natural Selection along the Old-World Primates Evolution. PLoS ONE 5(3): e9827. doi:10.1371/journal.pone.0009827

Editor: Anita Brandstaetter, Innsbruck Medical University, Austria

Received December 27, 2009; **Accepted** March 1, 2010; **Published** March 23, 2010

Funding: This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute, Center for Cancer Research. CF and DP were supported by the University of Bologna, ET-S by NIH, Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil) and Fundação de Amparo a Pesquisa de Minas Gerais (Brazil) and WCSM by Brazilian Ministry of Education (Agency for the Development of Graduate Education-CAPES). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: edutars@icb.ufmg.br

☯ These authors contributed equally to this work.

Introduction

Glucose is an important source of energy for living organisms. In vertebrates, it can be ingested with the diet and transported into the cells by conserved mechanisms and molecules, such as the trans-membrane Glucose Transporters (GLUTs) protein family. Members of this family have tissue specific expression, biochemical properties and physiologic functions that together, contribute to the regulation of blood sugar levels as well as its distribution. GLUT4 –coded by *SLC2A4* (chromosome 17p13), is an insulin-sensitive glucose transporter with a critical role in glucose homeostasis. In absence of insulin, GLUT4 is maintained sequestered in intracellular vesicles in tissues where it is preferentially expressed: adipose tissue, heart muscle and skeletal muscle [1,2]. Within minutes of insulin stimulation, GLUT4 molecules move to the cell surface to transport glucose into the cell, reducing blood glucose and allowing the intracellular synthesis of glycogen and triglycerides. GLUT4 also plays a role during prolonged exercise [3], when demand for glucose by contracting muscles is associated with its translocation from intracellular vesicles to the cell membrane to favor glucose uptake. Based on the critical role of GLUT4 in glucose homeostasis, and the association of hyperglycemia with metabolic disorders such as insulin resistance, type-2 diabetes, dyslipidaemia, hypertension

and obesity [4,5], structural and functional studies of GLUT4 have received great attention: a Pubmed search using the query "GLUT4 and glucose transporter" reports 250 publications in 2008 and 940 during the 2004–2008 quinquennium. On a structural basis, the GLUT4 protein has 12 membrane-spanning domains, with both the amino and carboxyl termini intracellularly oriented. Moreover, the human GLUT4 promoter region has been identified within 895 bp upstream of the transcription initiation site, containing *cis* regulatory domains for the Myocyte Enhancer Factor 2 and the Domain I Binding Protein, both required for regulation of transcription [6].

Forty-six years ago, James Neel posited the "thrifty" genotype hypothesis, suggesting that variants that increase diabetes type II susceptibility under modern life were advantageous in past environments characterized by food shortage [7]. He noticed that in human populations, diabetic offspring tend to be weightier than non-diabetics offspring, and that "the diabetic genotype" was a "thrifty genotype, in the sense of being exceptionally efficient in the intake and/or utilization of food". Recently, Anna Di Rienzo and colleagues have tested and discussed this hypothesis in a modern population genetics framework [8,9] and have shown that, consistent with the Neel hypothesis, the pattern of diversity of *Calpain-10* (*CAPN10*), a candidate gene with polymorphisms associated with diabetes type II, suggests evidence of balancing natural selection. In this context, it is important to test if the diversity of other genes playing a role in glucose metabolism, such as *SLC2A4*, also bears the signature of natural selection. Moreover, because glucose metabolism is critical for energy availability across all living organisms, it is important to infer if a signature of natural selection is recent or if, alternatively, it predates the coalescent of human lineages. Indeed, genes involved in glucose metabolism are overrepresented among genes that have experienced positive selection in its promoter region during human evolution [10]. To address these issues, we re-sequenced the *SLC2A4* locus in 102 ethnically diverse individuals and described its pattern of diversity in different populations. We compared the pattern of human polymorphisms with divergence from other mammals and tested the hypothesis that natural selection has shaped *SLC2A4* diversity.

Materials and Methods

Samples

Two datasets of anonymous samples were used. The first one (i.e. the re-sequencing panel) was composed by 102 unrelated individuals of the SNP500Cancer project (http://snp500cancer.nci.nih.gov/) [11], which includes: 24 African ancestry (15 African Americans from the United States and 9 Pygmies), 23 admixed Latin American (from Mexico, Puerto Rico and South America), 31 Europeans (from the CEPH/UTAH pedigree and the NIEHS Environmental Genome Project) and 24 Asians-Oceanians (from Melanesia, Pakistan, China, Cambodia, Japan and Taiwan). The second dataset (i.e. the SNPs-panel) includes a subset of 280 individuals from the HGDP-CEPH Panel [12], belonging to the following 13 populations: (http://snp500cancer.nci.nih.gov/terms_ethnic_hdp.cfm): San, Bantu, Mandenka and Yoruba from Sub-Saharian Africa; Sindhi, Pathan and Han from Asia; French, North-Italian, Tuscan and Orcadian from Europe; and Pima and Maya from the Americas.

PCR amplification, sequencing and SNPs genotyping

In the re-sequencing panel, we performed bi-directional sequencing of 6311 bp per individual, encompassing the most of the *SLC2A4* gene and ~1 kb upstream the gene (Reference sequence: chromosome 17, positions 7124832-7131142 of the NCBI human genome build 36.3). A fragment of 949 bp at the end of the 3'UTR could not be reliably sequenced because of a high density of A/T bases. For PCR amplification and sequencing we followed the protocol described by Packer et al. [11]. The orthologous chimpanzee and rhesus genomic sequences were used to determine ancestral states of polymorphisms. For analysis of long range linkage disequilibrium, we used data from 56 and 48 SNPs mapped ~0.5 Mb upstream and downstream of *SLC2A4* from the *Affymetrix SNP Array 5.0*, genotyped in the SNP500Cancer individuals ([13], see supplementary File S1 for the list of SNPs).

In the SNPs-panel we genotyped 5 common and representative *SLC2A4* SNPs (i.e. tag-SNPs *in sensu* Carlson et al. [14], see below for the criteria used for tag-SNPs selection) identified in the re-sequencing panel: rs5418, rs16956647, rs5435, rs5436, and rs5417. For this genotyping, we used Taqman assays (Applied Biosystems, Foster City, CA, US) following the protocols described in http://snp500cancer.nci.nih.gov/.

Evolutionary and population genetics analyses

We tested the Hardy-Weinberg equilibrium using the test of Guo and Thompson [15], implemented in the software Arlequin 3.0 [16]. Insertion-deletions (INDELs) were excluded from further population genetics analyses. We assessed intra-population variability in the following way: For the re-sequencing data we used estimators of the θ parameter based on the infinite-site-model of mutations: π, the per-site mean number of pair-wise differences between sequences [17], and by θw, based on the number of segregating sites (S) [18]. Instead, for the SNPs-panel, we calculated from haplotyes the gene diversity *in sensu* Nei et al. [19]. We measured pair-wise between-populations diversity measuring its percentage of the total genetic variance present in both populations (F_{ST}), and we also performed the Analysis of Molecular Variance (AMOVA) to measure the apportionment of genetic variance within and among populations or groups of populations [20], using the software Arlequin 3.0.

We inferred haplotypes considering SNPs with a Minor Allele Frequency (MAF) ≥ 0.05 in at least one population, using the method by Stephens and Sheet [21], that takes into account decay of linkage disequilibrium with distance among SNPs. The recombination parameter ρ was also calculated for each population from the re-sequencing panel by using the method of Li and Stephens [22]. These inferences were performed by the software Phase v.2.1.1 (see supplementary File S1 for additional specifications). Graphical relationships between haplotypes of the re-sequencing panel were explored by a Reduced Median Network, as implemented in the software Network 4.1.1.2 [23].

To investigate if the observed patterns of variability in human population is consistent with the neutral model, we used the tests of Tajima's D [24], Fu and Li's D* and Fu and Li 's F* [25] on the re-sequencing panel. In addition to the standard null hypothesis of neutrality under constant population size, we tested for the African population the significance of these statistics against a family of null hypotheses that consider scenarios of exponential demographic growth, which is consistent with its demographic history, in particular since the Pleistocene-Holocene [26]. We constructed the distribution of the statistics to be tested under these null hypotheses using the software ms [27] (see supplementary File S1 for details).

Linkage disequilibrium (LD) was estimated by r^2 [28] for SNPs with MAF≥ 0.05 in at least one population and its significance assessed by LOD scores, using software Haploview v.3.2 [29,30]. Based on the pattern of intragenic LD that emerged from the re-sequencing panel, we identified *SLC2A4* multi-population tag-SNPs (that may be used as surrogates for untyped SNPs [13]), with

a threshold $r^2>0.64$. For analyses of long range LD using the 104 *Affymetrix* SNPs covering ~1 Mb region, we first inferred long-range haplotypes using the algorithm by Scheet and Stephens [31], implemented in the software fastPHASE.v130.beta (details in supplementary File S1). We tested for the presence of recombination hotspots along the ~1 Mb using the approximate marginal likelihood method by Fearnhead [32] implemented in the software SequenceLDhot. For the long-range phased data, we applied the test for positive natural selection of Sabeti et al. [33], based on the Extended-Haplotype-Homozygosity statistic, which measures if a specific allele/haplotype under selection shows a higher LD with the sorrounding genomic region. We applied this test using haplotypes of the 8 common *SLC2A4* SNPs. Data handling for population genetics analyses were perfomed using a set of scripts from the platform DIVERGENOME (developed by Magalhães WCS and Tarazona-Santos ET).

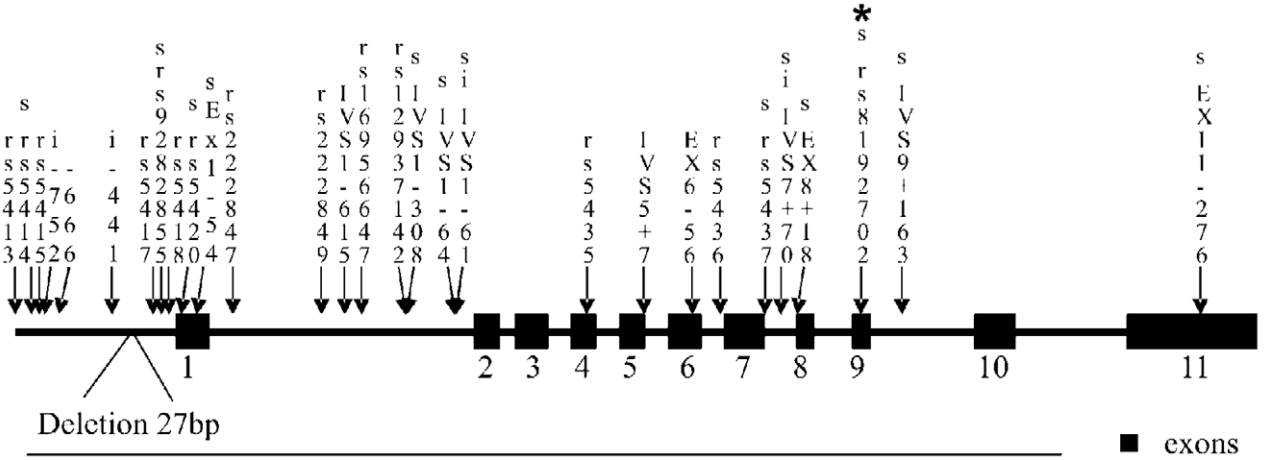
To explore evolutionary conservation across different species, we measured for each polymorphic position the conservation score of the Genome Browser website (assembly March 2006, http://genome.ucsc.edu/), based on multiple alignment of 17 vertebrate species [34]. To test the fitness of the data to the neutral model including inter-specific comparisons, we performed neutrality tests based on the comparison of polymorphisms and divergence rates from chimpanzee and rhesus: the McDonald and Kreitman test [35] that compares synonymous (assumed to be neutral) and nonsynonymous sites; and the adaptation of the Kolmogorov-Smirnov statistic (D_{KS}) by McDonald [36], developed to test the hypothesis that the ratio of polymorphisms to divergence is homogeneous along a genomic region. This statistic is based on the maximum absolute difference between the observed and expected cumulative numbers of polymorphisms. These tests were performed by DNAsp 4.10 and Slider softwares, respectively. To gain insights into the evolutionary history of *SLC2A4* at a larger evolutionary scale, we identified regions in the coding sequence associated to different kinds of selection through the evolutionary history of mammals. We compared *SLC2A4* coding sequences among the following mammals for which information is publicly available: *H. sapiens* (NM_001042.2), *P. troglodytes* (XM_001155036.1), *M. mulatta* (XM_001107391.1), *B. taurus* (NM_174604.1), *M. musculus* (NM_009204.2), *R. norvegicus* (NM_012751.1), *S. scrofa* (NM_001128433.1), *E. caballus* (NM_001081866.1). We used the maximum likelihood approach developed by Yang [37] to estimate ratios of non-synonymous (dN) to synonymous (dS) substitutions (ω = dN/dS) for *SLC2A4* codons under a variety of evolutionary models (see supplementary File S1). This method allows inferences about the evolution of a coding region along a phylogeny and to discriminate among codons that have evolved under strong or weak purifying selection, neutrality or adaptive positive selection. After fitting the data to an appropriate evolutionary model, a Bayes Empirical Bayes approach was used to infer the ω parameter for each codon. We performed this analysis using the software PAML [38].

Results

By re-sequencing the *SLC2A4* gene and ~1 kb upstream it, we detected 29 polymorphisms, including one non-synonymous singleton in exon 9 (Figure 1). All SNPs/INDELs fit Hardy-Weinberg proportions in the studied populations, both in the re-sequenced and the follow-up SNP genotyping. Two features of the observed pattern of diversity are interesting. First, across the four studied populations, all the eight common SNPs are concentrated upstream of exon 7 (on the first ~3700 bp of the gene), while the region downstream of intron 6 (~2600 bp) only harbors 6 singletons in Europeans/Africans, and no variation in Asians and Latin Americans. This lack of common variation in the C-terminal part of the gene is even more surprising after verifying trough the UCSC Genome Browser that among mammals, the genomic region downstream of intron 6 is as much variable as the region upstream of exon 7 (data not shown). Second, the African set shows a larger Watterson's θ (which depends on the number of segregating sites), but unexpectedly, they show a lower nucleotide diversity (which mostly depends on common variants, $\pi_{SLC2A4} = 0.00038$) than non-Africans (Table 1, [39,40,41]. For most of the human genome, African populations show larger π values than non-Africans, which is likely due to the bottleneck occurred approximately 40–50 thousand years ago during the migration of humans "Out of Africa" [42]. The observed π_{SLC2A4} in the African population is also the twenty-second lowest value when compared with 329 re-sequenced genes (seventh percentil of the distribution, december 2009) analyzed in an African-American sample by the Seattle SNPs initiative (see http://pga.gs.washington.edu/summary_stats.html and [43]). Therefore, in addition to the lack of common variation downstream of intron 6 in humans, *SLC2A4* has an uncommon pattern of variation in Africans, characterized by a high number of segregating sites and singletons but low nucleotide diversity.

Based on the 8 common polymorphisms with a MAF≥0.05 in at least one population (all located upstream of exon 7) we inferred 11 haplotypes (Figure 1). The Reduced Median Network in Figure 2 illustrates the phylogenetic relationships among haplotypes and their distribution in human populations. The differentiation between human populations (F_{ST}) observed in the re-sequencing panel for *SLC2A4* is 3.8% ($P = 0.013$), which is lower than the 10–12% observed on average among human populations [44]. This result reflects the fact that only the African population is differentiated from the homogeneous non-African ones, which is mainly due to differences in frequencies of haplotypes A2 and A7 (Figure 2). The analysis of the SNPs-panel produced results that were consistent with those of the re-sequencing panel (see details see the supplementary File S1).

Based on the observed pattern of diversity of *SLC2A4*, we tested the hypothesis that it was shaped by natural selection. We interrogated the evolutionary basis of the low nucleotide diversity observed in Africans by analyzing the re-sequencing panel with tests of natural selection that are based on the proportions of rare and common polymorphisms (i.e. the allelic spectrum) expected under neutrality. First, we assumed a null hypothesis of neutrality and constant population size (Table 1). While the allelic spectra of non-African populations are consistent with the null hypothesis, Africans show more rare alleles than expected, which is evidenced by negative and significant values ($P<0.02$) of the Fu-Li's D* and F* statistics. The Tajima's D statistics for the African sample also corresponds to the low fifth-percentile when compared with the 329 genes sequenced in an African-American sample by the Seattle-SNPs initiative (http://pga.gs.washington.edu/summary_stats.html). Based on the contrasting pattern of diversity along *SLC2A4*, we compared the allelic spectra of the regions upstream of exon 7 and downstream of intron 6 and observed that, while Africans show an excess of rare alleles (measured by $D^*_{Fu\text{-}Li}$ and $F^*_{Fu\text{-}Li}$) in both regions (data not shown), the presence of 3 singleton and no common variation downstream of intron 6 in the European population is not compatible with the null hypothesis of neutrality ($D^*_{Fu\text{-}Li} = -3.131$ and $F_{Fu\text{-}Li} = -3.134$, $P<0.05$). This comparison was not applied to Asians and Hispanic population because they show no variation downstream of intron 6. These results suggest that under the assumption of constant population size, an observed excess of rare alleles is compatible with a selective sweep or with background selection against deleterious mutations



■ exons

✱ non synonymous substitution

s singleton

i INDELs

Haplotypes	rs5415	rs5417	rs5418	rs222847	rs222849	rs16956647	rs5435	rs5436	African	European	Asian	Hispanic	World
A1	**C**	**C**	**G**	**T**	**C**	**C**	**C**	**C**	18	4		4	26
A2	.	.	.	.	.	.	.	**T**	7	1			8
A3	.	.	.	.	.	**T**	.	.	1	1	6	2	10
A4	.	.	.	.	**T**	.	**T**	.		1	1	3	5
A5	**T**	.	.	.	.	.	.	.	1		3		4
A6	**T**	.	.	.	**T**	.	.	.	1		2	1	4
A7	**T**	.	.	.	**T**	.	**T**	.	5	18	14	11	48
B1	.	**A**	**A**	.	.	.	.	.	15	32	22	25	94
B2	.	**A**	**A**	.	.	.	**T**	.		2			2
B3	.	**A**	**A**	**A**	.	.	.	.		2			2
B4	.	**A**	**A**	**A**	.	.	**T**	.		1			1
Number of chromosomes									48	62	48	46	204
Number of common haplotypes									7	9	6	6	11
Number of common polymorphic sites									7	8	6	6	8
Haplotype diversity									0.744	0.653	0.698	0.647	0.714
S.D. of haplotype diversity									0.036	0.048	0.045	0.061	0.024

Figure 1. Genomic structure of SLC2A4, substitutions found, inferred haplotypes and their frequencies. Substitutions are represented by arrows and when no dbSNP name is available, named as in the SNP500Cancer database. A total of 29 polymorphisms (25 SNPs and 4 INDELs) were detected in the 204 worldwide re-sequenced chromosomes. Forty five percent of the substitutions were singletons and only 8 reached a MAF>0.05 in at least one studied population. Comparison with the homologous chimpanzee sequence suggests that for all SNPs the ancestral allele is modal in humans. In the human genome, there is a 27 bp fixed deletion 348 bp upstream of the transcription initiation site. Three non-coding SNPs are in evolutionarily conserved positions (UCSC Genome Browser, [33]): rs5415 (conservation score: 0.96), within the promoter region, as well as rs222847 and rs222849, both with conservation score of 0.99 and within the first intron. Only one of the 4 coding-SNPs is non-synonymous (rs8192702, Ala358Val, a conservative substitution in exon 9, in the ninth trans-membrane domain), observed in a European. Haplotypes are inferred using only the 8 common SNPs.
doi:10.1371/journal.pone.0009827.g001

affecting the variation of *SLC2A4* in Africans and Europeans. We also assumed a set of null hypotheses for human populations based on scenarios of demographic expansion. In this case, the excess of rare alleles in Africans is compatible with neutrality under the following scenarios: (a) an exponential growth that started at least 2400 generations (~60000 years) ago from the 0.001% of the current population size and (b) with a very recent expansion (~200 generations, ~5000 years) from the 0.0001% of the current

Table 1. Summary of intra-population diversity indexes and tests of neutrality based on re-sequencing analysis of the four SNP500Cancer populations.

Populations	African	European	Asian	Hispanic	World
N. of chromosomes	48	62	48	46	204
Segregating sites	20	13	8	9	25
Singletons	13	5	1	3	11
Common SNPs (MAF[a]>0.05)	6	6	6	5	5
ρ (per gene)	1.70	0.48	0.45	0.63	7.34
θ estimators					
$\pi \pm$ SD ($\times 10^{-3}$)	0.38±0.04	0.43±0.03	0.44±0.02	0.40±0.04	0.43±0.02
$\theta_W \pm$ SD ($\times 10^{-3}$) (per site)	0.71±0.25	0.44±0.16	0.29±0.13	0.32±0.14	0.67±0.19
Neutrality tests					
Tajima's D	−1.483	−0.064	1.453	0.587	−1.016
Fu and Li's D*	−3.069 [b]	−1.176	0.594	−0.656	−2.986 [b]
Fu and Li's F*	−2.992 [b]	−0.941	1.023	−0.226	−2.630 [c]
P of McDonald-Kreitman test	0.544	1.000	1.000	1.000	1.000

[a]Minor Allele Frequency.
[b]$P<0.02$.
[c]$P<0.05$.
doi:10.1371/journal.pone.0009827.t001

population size. Therefore, *SLC2A4* African allelic spectrum is compatible with an evolutionary history that may involve a combination of population expansion and/or natural selection (selective sweep or background selection).

For *SLC2A4*, Africans show the highest recombination parameter ρ and the lowest LD, consistent with studies on other genomic regions and with the human evolutionary history ([41], Table 1 and Figure 3), although substantial intragenic LD is shared across human populations. We performed an analysis of long range LD on the genomic region of ~1 Mb containing *SLC2A4* at its center (see supplementary File S1), to gain information about possible recent events of natural selection. Based on the information from

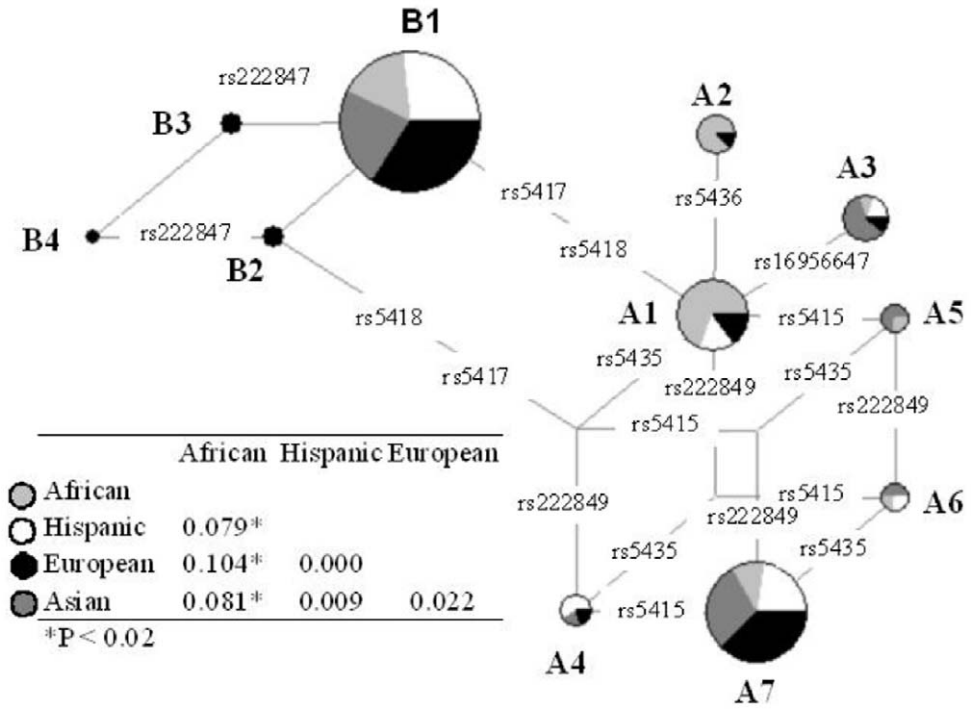


Figure 2. Reduced Median Network of SLC2A4 haplotypes inferred in the re-sequencing panel and matrix of pairwise F_{ST}. Haplotypes were inferred from the 8 polymorphisms with a MAF <0.05 in at least one population. Each circle represents a different haplotype, its size is proportional to its relative frequency and the presence in each population is indicated with different gray tonalities. Base substitutions are indicated along branches. The reticulated network reflects the action of recombination or recurrent substitution.
doi:10.1371/journal.pone.0009827.g002

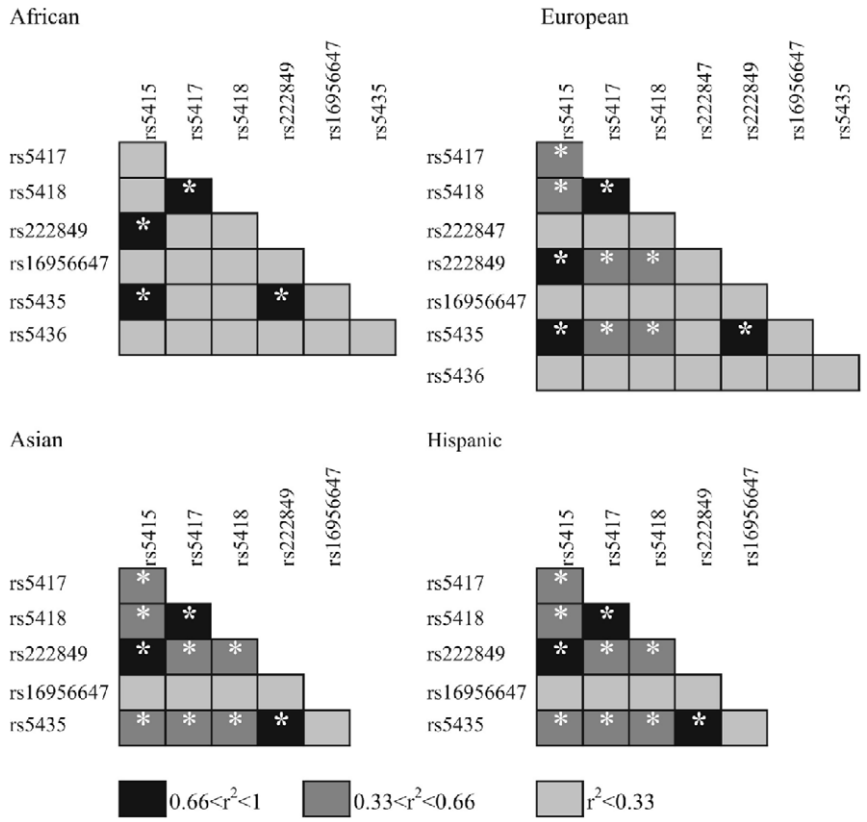


Figure 3. Pairwise linkage disequilibrium in *SLC2A4* in human populations as ascertained in the re-sequencing panel. Significant r^2 values (LOD >2) are denoted by white asterisks.
doi:10.1371/journal.pone.0009827.g003

~50 SNPs mapped on ~0.5 Mb at each side of *SLC2A4*, we first verified that there is no statistical evidence of recombination hotspots near *SLC2A4* [32]. Then we determined that this gene is not located within a block of LD in any of the four studied populations. Also, none of the *SLC2A4* common haplotypes is associated with increased measurements of LD, when measured by the Extended-Haplotype-Homozygosity statistic [45]. Thus, we have no evidence of ongoing positive selection associated with this gene.

To further assess if the lack of common variants downstream of intron 6 may be due to natural selection at inter-specific level, we applied the Kolmogorov-Smirnov statistic (KS), which belongs to a family of statistics that test if the ratio of polymorphism to divergence along a gene is homogenous, as expected under neutrality [36]. Among these tests, the KS statistic has the highest power to detect patterns in which one end of a gene has high polymorphism and the other end has low polymorphism, as in the case of *SLC2A4*. Moreover, it does not require an arbitrary division of the *SLC2A4* in two parts to be compared (e.g. upstream of exon 7 and downstream of intron 6), a procedure that would be necessary if the classical Hudson-Kreitman-Aguade test (HKA [46]) were applied (but see the supplementary File S1 for results of this classical test). We used two outgroups: chimpanzee (diverged from humans 5–6 millions of years-MY ago) and rhesus monkey (diverged from humans 20–25 MY ago). When we used the chimpanzee as outgroup, we did not reject the null neutral expectation that the ratio of polymorphisms to divergence is homogeneous across *SLC2A4* (supplementary File S1). However, when we used rhesus monkey as outgroup, this pattern changed, and there is significantly less human polymorphisms in Africans, Asians and Latin Americans in the second part of the gene than expected based on the divergence among humans and rhesus (Figure 4). This is even more evident when we consider that all polymorphisms observed downstream of intron 6 are singletons (see also the supplementary File S1 for HKA results). Therefore, if natural selection contributed to reduce the diversity in the second part of *SLC2A4*, this may not be an event restricted to the human evolutionary history, since the comparison with chimpanzee shows that a lower rate of accumulation of substitutions downstream of intron 6 was already evident along the lineages of 5–6 MY that separate humans and chimpanzees. However, divergence downstream of intron 6 accumulated faster in the timeframe between human-rhesus and human-chimpanzee divergences, at rates comparable to the region upstream of exon 7. These results are consistent with an episode of natural selection occurred after the divergence between lineages leading to humans and rhesus (20–25 MY), but predating the divergence between humans and chimpanzee (5–6 MY). Alternatively, the absence of significance observed when the chimpanzee was used as the outgroup may be due to a reduced statistical power determined by few fixed differences between humans and chimpanzees. In this case, natural selection would have not predated the divergence among humans and chimpanzees.

To determine if the observed pattern of diversity is due to the action of natural selection on *SLC2A4* coding region, we obtained maximum likelihood estimations [37] of the ratios of non-synonymous (dN) to synonymous (dS) substitutions (ω = dN/dS) for *SLC2A4* codons under a variety of evolutionary models. The ω parameter is expected to be 1 under neutrality, <1 (dN < dS) under purifying selection and >1 (dN > dS) under positive

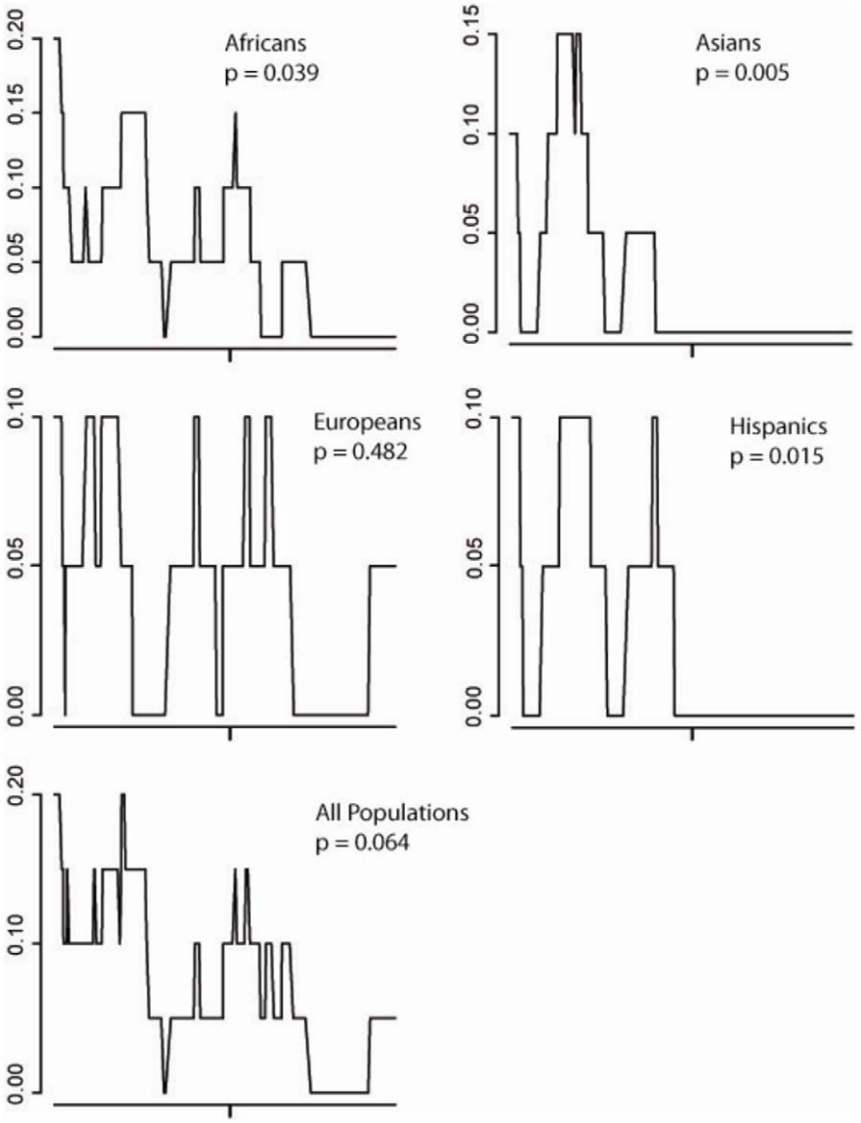


Figure 4. Proportions of polymorphisms to fixed substitutions among humans and rhesus (P/K), calculated by a sliding window approach. Each window includes 20 substitutions. The P value for the Kolmogorov-Smirnoff statistic by McDonald [35] was used to test if the P/K ratio was homogeneous along the gene (see Supplementary File S1 for results using the chimpanzee as outgroup). To be conservative, we evidence the highest P value among those obtained assuming values of recombination parameter r equal to 0, 2, 4 and 6. In the horizontal axes, the vertical tick mark indicates the intron 6- exon 7 boundary. The pattern of significance is the same when *Mus musculus* or *Rattus norvegicus* are used as outgroups. Excluding chimpanzee and rhesus; *M. musculus* and *R. norvegicus* are the mammals most closely related to humans for which *SLC2A4* genomic sequences are available in NCBI databases.
doi:10.1371/journal.pone.0009827.g004

selection. The best fit of our data is obtained for models that (see the supplementary File S1 for detailed results): (1) allow for values of $\omega \leq 1$ to vary across *SLC2A4* coding region, (2) do not show strong evidence of relaxation of purifying selection along the primate lineages and, (3) do not show evidence of positive selection. In particular, the discrete Model 3 of Yang [37], that allow for $K = 2$ different classes of ω (without restrictions for the value of this parameter) best fit the data, and suggests that ~85% of *SLC2A4* codons evolved under strong purifying selection ($\omega \approx 0.007$) and ~15% under a weaker purifying selection ($\omega \approx 0.506$, Figure 5). There is no association among the distribution of these two classes of codons and their location in the transmembrane domains of GLUT4. Also, codons that evolved under strong purified selection are not associated (Fisher exact test $P = 0.41$) with the region encompassing exons 7–11, where no common polymorphisms are present in humans and a reduced rate of accumulation of substitutions is observed along the chimpanzee-human genomic lineage. This result suggests that our results for the Kolmogorov-Smirnov test, possibly attributed to the action of natural selection, do not depend on variation in the *SLC2A4* coding region.

Discussion

Considering the evolutionary timeframe of mammals, we observed no evidence of positive natural selection for the *SLC2A4* coding sequence, although inferences about ω using the Yang [37] approach has sufficient power for a protein with more than 500 codons, such as GLUT4 [47,48]. While most codons (~85%) are under strong purifying selection, for sixty of them (15%) purifying

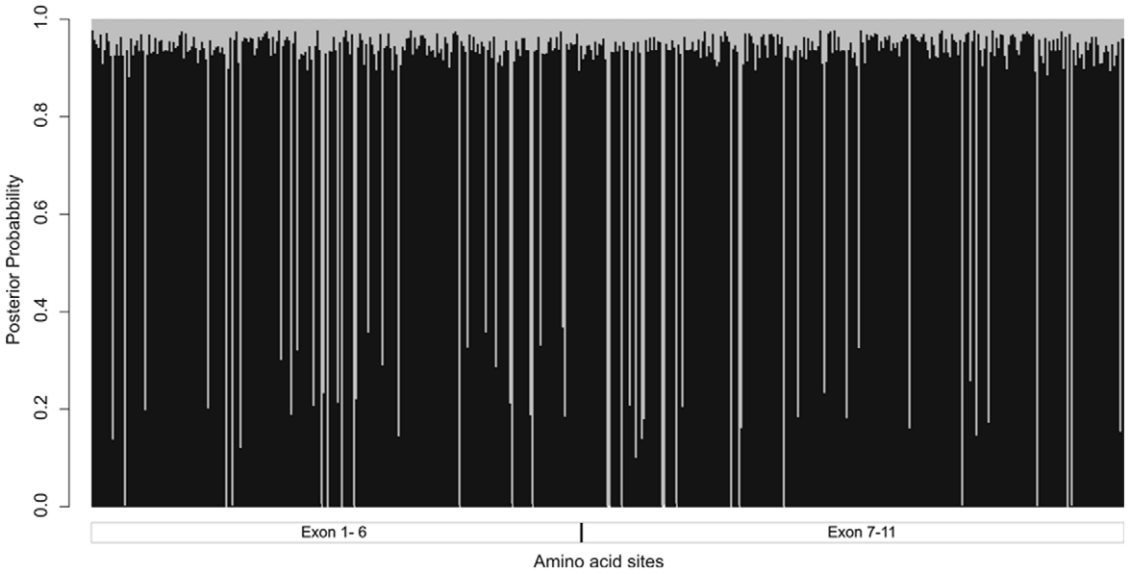


Figure 5. Probability of evolving under strong ($\omega_s = 0.007$, in black) or weak ($\omega_w = 0.506$, in gray) purifying selection for each of the *SLC2A4* codons (in the horizontal axis).
doi:10.1371/journal.pone.0009827.g005

selection was weaker. In fact, codons of the latter category present non-synonymous substitutions (19 of them more than one at the same codon) along the mammal phylogeny. Classifying *SLC2A4* codons in two classes of purifying selection is a simplification, but we think this is a reasonable assignment that derives from the evolutionary model that best fit our data (Model 3 of Yang [37], supplementary File S1). In any case, this simplification allowed us to verify that these classes of codons are not associated with portions of *SLC2A4* upstream of exon 7 or downstream of intron 6. Therefore, the pattern of substitution across the phylogeny of mammals coding region does not explain the lack of common variation in humans nor the lower divergence along the human-chimpanzee lineages for the second part of the gene.

We observed that when we used the rhesus monkey (that diverged from humans 20–25 MY ago) as outgroup and applied the Kolmogorov-Smirnov neutrality test, we do not observe along the human-rhesus lineages the paucity of variation downstream of intron 6 that is observed for human polymorphisms. We interpret this result as evidence that natural selection reduced the variability downstream of *SLC2A4* intron 6 during the last 25 MY, and the current pattern of diversity observed in modern humans reflects this event. However, an alternative explanation is that comparisons with the chimpanzee - an evolutionarily closed outlier; have less statistical power than comparisons with the rhesus monkey and therefore, our data may be also compatible with a more recent action of natural selection, though not recent enough to be detected using neutrality tests based on linkage disequilibrium [33]. Because we did not observe relevant changes in ω along the primate phylogeny of *SLC2A4* coding sequence, we hypothesize that natural selection acted on a non-coding region of *SLC2A4*. In fact, only neutrality tests such as the KS statistic, which application is not limited to coding regions, are able to capture a pattern like this. Two kinds of selection may reduce genetic diversity: background purifying selection and a selective sweep leading to a hitchhiking event [49]. However, it is unlikely that background purifying selection started to act on a large non-coding region only at a certain point during the last 20–25 MY, after the divergence of humans and rhesus lineages. Instead, a selective sweep is consistent with the lack of variation along a genomic region (such as the second part of *SLC2A4*), with the low nucleotide diversity observed in African populations and with the excess of rare alleles and negative values of the Tajima statistics for the region downstream of intron 6 in Africans and Europeans (although this may be due in part to the demographic history of these populations as suggested by coalescent simulations). What is not inconsistent with a selective sweep scenario, but makes it less likely, is the fact that the observed lack of variation is mainly restricted to the region downstream of intron 6, and we did not find evidence for the existence of a recombination hotspot within the *SLC2A4* locus that prevents the propagation of the signature of natural selection along a larger genomic region. In favor of consistency with a selective sweep scenario, we may also mention that *SLC2A4* is within a genomic region where LD is in general low (supplementary File S1), and therefore, the signature of natural selection determined by a selective sweep would be necessarily restricted to a small region. If a complete selective sweep occurred during the last 20–25 MY along the rhesus-human lineage, this may be compatible with a "transpecies" version of the "thrifty" genotype hypothesis (see Introduction of [8]). In this hypothetical scenario, we may not see association between diabetes susceptibility and *SLC2A4* variants [50] because a selective sweep lead to the existence of a small genomic region with no common variants, and the fixed haplotype may be "thrifty". By examining the pattern of long-range LD, we did not find evidence of an ongoing selective sweep within a temporal frame of ~25000 years (the timescale at which a selective sweep left a signature in the pattern of LD, [33]). In fact, none of the common *SLC2A4* haplotypes (defined by SNPs upstream of exon 7) is associated to a large surrounding region of LD - a pattern expected under a recent selective sweep.

Because population samples included in this study (as in most human population genetics studies) are not optimal for the population genetics inferences to be addressed, it is important to consider the limitations of our results. By genotyping five SNPs in an additional worldwide samples from the HGDP-CEPH Panel, we found a haplotype structure that was consistent with that observed in the re-sequencing panel. Although African and Asian/Oceanian samples include individuals with diverse origin and

therefore, are structured, we would not expect the paucity of variation observed downstream of intron 6, or the excess of rare alleles in the African sample to be an artifact of our sample composition. Instead, the population structure observed in the African and Asian samples is expected to generate a deficit of rare alleles (and an excess of common alleles), and therefore, our results reporting an excess of rare alleles (or the lack of common variants) are conservative in light of our sampling strategy [25].

In conclusion, after performing extensive sequencing of *SLC2A4*, we determined that it has a peculiar pattern of genetic variation, with the first part of the gene showing common and rare variants in a fashion compatible with neutral evolution. However the second part of the gene shows no common variants as well as a pattern of diversity that is not compatible with neutrality, but compatible with an event of natural selection that reduced the level of substitution in this region during the last 20–25 MY. Although the natural selection scenario is compatible with the observed data, we recommend caution since claims of natural selection should require replication on larger samples to be accepted, and if possible, understanding of its biological/functional basis.

Supporting Information

File S1

Found at: doi:10.1371/journal.pone.0009827.s001 (3.13 MB DOC)

Acknowledgments

The authors are grateful to Silvia Fuselli and Rodrigo Redondo for discussions of our results, to Renee Chen and the Sequencing Group of the Core Genotyping Facility (National Cancer Institute) for their technical assistance.

Author Contributions

Conceived and designed the experiments: SJC. Performed the experiments: CF. Analyzed the data: ETS CF WCM. Contributed reagents/materials/analysis tools: MY WCM LB AC DP SJC. Wrote the paper: ETS CF SJC.

References

1. Olson AL, Pessin JE (1996) Structure, function, and regulation of the mammalian facilitative glucose transporter gene family. Annu Rev Nutr 116: 235–56.
2. Huang S, Czech MP (2007) The GLUT4 glucose transporter. Cell Metab 5: 237–52.
3. Suh SH, Paik IY, Jacobs K (2007) Regulation of blood glucose homeostasis during prolonged exercise. Mol Cells 23: 272–9.
4. Brand-Miller J, Dickinson S, Barclay A, Celermajer D (2007) The glycemic index and cardiovascular disease risk. Curr Atheroscler Rep 9: 479–85.
5. Teran-Garcia M, Rankinen T, Bouchard C (2008) Genes, exercise, growth, and the sedentary, obese child. J Appl Physiol 105: 988–1001.
6. Oshel KM, Knight JB, Cao KT, Thai MV, Olson AL (2000) Identification of a 30-base pair regulatory element and novel DNA binding protein that regulates the human GLUT4 promoter in transgenic mice. J Biol Chem 275: 23666–73.
7. Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14: 353–62.
8. Vander Molen J, Frisse LM, Fullerton SM, Qian Y, Del Bosque-Plata L, et al. (2005) Population genetics of CAPN10 and GPR35: implications for the evolution of type 2 diabetes variants. Am J Hum Genet 76: 548–60.
9. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21: 596–601.
10. Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet 39: 1140–1144.
11. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, et al. (2006) SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res 34(Database issue): D617–21.
12. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–2.
13. Hughes AL, Welch R, Puri V, Matthews C, Haque K, et al. (2008) Genome-wide SNP typing reveals signatures of population history. Genomics 92: 1–8.
14. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 7411: 106–20.
15. Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48: 361–72.
16. Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47–50.
17. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–60.
18. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256–76.
19. Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press: New York, .
20. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–91.
21. Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76: 449–62.
22. Li N, Stephens M (2003) Modelling linkage disequilibrium and identifying recombination hotspots using SNP data. Genetics 165: 2213–2233.
23. Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141: 743–53.
24. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–95.
25. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709.
26. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, et al. (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102: 18508–13.
27. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–8.
28. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226–231.
29. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–9.
30. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–5.
31. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–44.
32. Fearnhead P (2006) SequenceLDhot: detecting recombination hotspots. Bioinformatics 22: 3061–6.
33. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614–20.
34. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucl Acids Res 32 (Suppl 1): D493–D496.
35. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.
36. McDonald JH (1998) Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol Biol Evol 15: 377–384.
37. Yang Z (2007a) Adaptive Molecular Evolution. In Handbook of Statistical Genetics, Edited by Balding DJ, Bishop M and Cannings C. Third Edition, John Wiley & Sons, Ltd, Susex, UK, Volume 1. pp 377–406.
38. Yang Z (2007b) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24: 1586–91.
39. Tarazona-Santos E, Tishkoff SA (2005) Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. Genes Immun 6: 53–65.
40. Tarazona-Santos E, Bernig T, Burdett L, Magalhaes WC, Fabbri C, et al. (2008) CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains. Hum Mutat 29: 623–32.
41. Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9: 403–33.
42. Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, et al. (2007) Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA 104: 17614–9.
43. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, et al. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2: e286.

44. Barbujani G, Goldstein D (2004) Africans and Asians abroad: genetic diversity in Europe Annu Rev Genomics Hum Genet 5: 119–50.
45. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–7.
46. Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 1161: 153–9.
47. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170.
48. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, et al. (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4: e1000144.
49. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857–68.
50. Frayling TM (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet 8: 657–62.

2.6.3 Manuscrito II

The Complex Evolutionary History of Human NADPH Oxidase Genes (*CYBB, CYBA, NCF2 and NCF4*): Inferences about the action of Natural Selection

O complexo enzimático NADPH é um complexo enzimático que catalisa a redução do oxigênio para O^{2-} gerando espécies reativas de oxigênio, uma reação crítica para a atividade microbicida dos fagócitos. Em células não fagocíticas, NADPH oxidase produz baixas quantidades de O_{2-}, e em alguns casos, alterações nesta taxa de produção, podem estar associados com doenças degenerativas e insuficiência cardíaca. NADPH oxidase inclui duas proteínas transmembrana, as sub-unidades gp91-phox e gp22-phox (expressos pelos genes CYBA e CYBB), e três proteínas citoplasmáticas, as sub-unidades, p40-phox, p47-phox, and p67-phox (expressas pelos genes *NCF4*, *NCF1*e *NCF2*). Mutações nos genes *CYBB*, *CYBA*, *NCF1* e *NCF2* podem resultar no desenvolvimento de granulomatose crônica, uma imunodeficiencia primária. Neste trabalho testamos a hipótese de que a seleção tem moldado a diversidade presente nos genes que compoe o complexo NADPH em duas escalas temporais: evolução dos mamíferos e evolução humana recente. Durante a evolução dos mamíferos, *CYBA*, *NCF2* e *NCF4* tem predominantemente evoluído sobre influência de seleção purificadora. Para isso participei nas análises das regiões codificantes em mamíferos, dados públicos. As análise foram realizadas utilizando o pacote PAML (Yang, Z. H., 2007). O pacote PAML implenta a metodologia de máxima verossimilhança (Yang, Z. G., 2007) para estimar as razões entre substituições não sinônimas (dN) e sinônimas (dS), representado por omega (ω), ω = dN/dS, para os códons de *NCF4*, *NCF1*e NCF2 sobre a hipótese de vários modelos evolutivos. Aplicamos testes de neutralidade baseados no espectro de frequência alélico, D de Tajima and F e D de Fu e Li, (Tajima, 1989; Fu e Li, 1993). Para estes testes usamos como hipótese nula, o modelo clássico de

Wright-Fisher, e um modelo mais realista para populacões humanas inferidos com base em dados genéticos provenientes de marcadores multi-alélicos (Voight *et al*., 2005). Todos os cenários evolutivos foram gerados usando o programa ms (Hudson, 2002).

(To be submitted to …)

THE COMPLEX EVOLUTIONARY HISTORY OF HUMAN NADPH OXIDASE GENES (*CYBB*, *CYBA*, *NCF2* and *NCF4*): INFERENCES ABOUT THE ACTION OF NATURAL SELECTION

Eduardo Tarazona-Santos[1,2], Moara Machado[2], Wagner CS Magalhães[2], Fernanda Lyon[2], Laurie Burdett[3], Renee Chen[1], Andrew Crenshaw[3], Cristina Fabbri[4], Laelia Pinto[2], Rodrigo Redondo[2], Ben Sestanovich[1], Meredith Yeager[3], Stephen J Chanock[1]

[1] Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, USA. 8717 Grovemont Circle, Advanced Technology Center, Room 127, Gaithersburg, MD, 20877, USA.

[2] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486, Belo Horizonte, MG, CEP 31270-910, Brazil.

[3] Intramural Research Support Program, SAIC Frederick, NCI-FCRDC, Frederick, MD, 21702, USA and Core Genotype Facility, National Cancer Institute, NIH, Gaithersburg, Maryland, USA.

[4] Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Via Selmi 3, 40126, Bologna, Italy.

CORRESPONDING AUTHORS:

Eduardo Tarazona-Santos

Departamento de Biologia Geral

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais.

Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486,

Belo Horizonte, MG, CEP 31270-910, Brazil.

Tel: 55 31 34092597

Fax: 55 31 34092567

E-mail: edutars@icb.ufmg.br

Stephen J Chanock

Laboratory of Translational Genomics

Division of Cancer Epidemiology and Genetics, National Cancer Institute

Advanced Technology Center

8717 Grovemont Circle, Bethesda, MD 20892-4605, US

Tel: 1 301-435-7559

Fax: 1 301-402-3134

E-mail: chanocks@mail.nih.gov

ABSTRACT

The phagocyte NADPH oxidase is an enzymatic complex that catalyzes the reduction of oxygen to O^{2-} and generates reactive oxygen species, a critical reaction for the microbicidal activity of phagocytes. In non-phagocyte cells, NADPH oxidase produces a substantially lower amount of O^{2-}, and in some cases, alterations in production can be associated with neurodegenerative disorders and cardiovascular impairment. NADPH oxidase includes two membrane-spanning polypeptide subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*) and three cytoplasmic polypeptide subunits, p40-phox, p47-phox and p67-phox (encoded by *NCF4*, *NCF1* and *NCF2*). Mutations in *CYBB*, *CYBA*, *NCF1* or *NCF2* can result in Chronic Granulomatous Disease, a primary immunodeficiency. We have tested the hypothesis that natural selection has shaped the diversity of NADPH genes at two temporal scales: the mammalian evolution and recent human evolution. During mammalian evolution, *CYBA*, *NCF2* and *NCF4* coding regions have predominantly evolved driven by purifying natural selection. Conversely, episodes of adaptive natural selection have driven the evolution of *CYBB*, and almost all of these events are concentrated on the extracellular part of this protein, suggesting a currently unknown functional relevance for these inter-specific variants. To infer recent episodes of natural selection, we have re-sequenced 35524bp including the exons, UTRs, promoters and intronic regions of *CYBB*, *CYBA*, *NCF2* and *NCF4* in 102 ethnically diverse healthy individuals. For the four studied genes, diversity and the recombination parameter are higher in Africans than in non-Africans, consistently with the demographic history of human populations. Moreover: (1) *CYBA* shows a pattern of non-synonymous substitution, very high variation in Europeans and an excess of common polymorphisms that is compatible with the action of balancing natural selection. (2) *NCF2* in Asia evidences a particularly differentiated haplotype structure with a modal haplotype

that is rare elsewhere, low diversity and an excess of rare segregating sites, a pattern that is compatible with the action of positive natural selection acting on *NCF2* in Asian populations or with an increase in frequency of rare alleles surfing at the front of an spatial population expansion.

The phagocyte NADPH oxidase, also known as the 'respiratory burst oxidase', is an enzymatic complex with a critical role in innate immunity. It catalyzes the reduction of oxygen to O^{2-}, generating reactive oxygen species (ROS) that are responsible for the microbicidal activity of phagocytes (Chanock et al. 1994; Heyworth et al. 2003). The phagocyte NADPH oxidase includes two membrane-spanning polypeptide subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*) and four cytoplasmatic polypeptide subunits; p40-phox, p47-phox, p67-phox and a GTPase Rac1 or Rac2 (encoded by *NCF4*, *NCF1*, *NCF2*, and *RAC1* or *RAC2*, respectively). When the phagocytosis is induced by invading pathogens, the cytoplasmatic units bind the transmembrane components and activate the enzymatic complex (i.e. producing microbicidal ROS), in a process that is dependent on specific interactions among domains of the NADPH oxidase components (Figure 1, Sumimoto et al. 2005). The relevance of NADPH oxidase in the defense against pathogens is evidenced by the fact that mutations in five NADPH genes (*CYBB*, *CYBA*, *NCF1*, *NCF2*, *NCF4*) can result in Chronic Granulomatous Disease (CGD), a Mendelian recessive heterogeneous immunodeficiency. Indeed, most CGD patients have no measurable respiratory burst and less than 5% generate very low levels of ROS (Heyworth et al. 2003). Nearly 70% of CGD cases are X-linked, due to mutations in *CYBB* (OMIM#306400, Heyworth et al. 2003) and there is high degree of allelic heterogeneity in X-linked and autosomal CGD (see the Immunodeficiency Mutations Database: http://bioinf.uta.fi/base_root/mutation_databases_list.php). Several studies in animal models and in vitro have confirmed the role of the NADPH oxidase in immunity against catalase-positive bacteria and fungi and other pathogens (Buckley 2004), and in addition to the CGD mutations, common variants may determine subtler variation in the expression or function of NADPH genes, contributing to infectious diseases such as tuberculosis and malaria (Wang et al. 2003,

Uhlemann et al. 2004), as well as inflammatory phenotypes such as Crohn disease (Rioux et al. 2007). NADPH oxidases are also expressed with different functions in non-phagocyte cells. Although p22phox (encoded by *CYBA*) is shared by several of these NADPH oxidases (also called Nox), the other components may be different peptides encoded by different Nox genes homologous to the components of the phagocyte enzymatic complex (Sumimoto et al. 2005, San José et al. 2008). Though these non-phagocyte NADPH oxidases produce less O^{2-}, imbalances on ROS levels may cause tissue damage due to oxidative stress, which is correlated with the pathogenesis of gout, chronic obstructive pulmonary disease, rheumatoid arthritis and cardiovascular diseases (Ross et al. 2003, Brandes and Kreuzer 2005). Therefore, variation in NADPH oxidase is pleiotropic. On one hand, it accounts for immunity phenotypes ranging from Mendelian diseases such as CGD, to complex traits such as infectious and autoimmune diseases. On the other hand, these variants seem also to be responsible for pathogenesis of cardiovascular diseases, through endothelium oxidative damage.

Despite the involvement of the NADPH oxidase in the pathogenesis of Mendelian and complex diseases, our knowledge of the sequence diversity of NADPH genes mostly derives from CGD patients. Although targeted SNPs genotyping has been performed in the context of association studies for *CYBA* (Bedard et al. 2009) and *NCF4* (Olsson et al. 2007), none of the large scale re-sequencing efforts such as Seattle SNPs (http://pga.gs.washington.edu/), Innate Immunity PGA (http://www.pharmgat.org/IIPGA2/index_html) or the Cornell-Celera initiative (Bustamante et al. 2005) have included the NADPH oxidase genes. In this study, we extensively studied the pattern of sequence diversity of four of the NADPH genes (*CYBB*, *CYBA*, *NCF2* and *NCF4*) in human populations, and interpreted our results in terms of their evolutionary history (in

particular addressing the potential action of natural selection). We focused on two temporal scales: mammalian evolution and recent human evolution. Several studies have shown the importance of natural selection on the evolution of immunity genes, both at inter-specific (Kosiol et al. 2008) and population levels (Sabeti et al. 2006, Fumagalli et al. 2009, Ferrer-Admetlla et al. 2008, Barreiro et al. 2009, Barreiro and Quintana-Murci 2009). Inferences about the action of natural selection have two implications: First, variants on genes inferred to be under selection have contributed to determine phenotype variability and perhaps, differential susceptibility to rare or common diseases. Second and by definition of natural selection, these variants have been associated with relatively different reproductive efficiencies (i.e. *fitness*) of their carriers, and therefore, they may have biomedical relevance. In this study, our goals are: (1) To infer if the pattern of diversity of human phagocyte NADPH genes reflects the action of different types of natural selection. (2) To understand the relationships among the observed patterns of diversity at the temporal scales of mammals and humans in an evolutionary context, and (3) to understand the biomedical implications of this evolutionary process in human populations. Specifically, we first analyzed the coding sequences of NADPH oxidase genes from different mammals and inferred how purifying natural selection has acted with different intensities, and in particular, if there is evidence of positive natural selection (that rise the frequency of a beneficial variant) at the time scale of mammalian evolution. Then, we re-sequenced *CYBB*, *CYBA*, *NCF2* and *NCF4* for a total of 35524bp for each of 102 ethnically diverse healthy individuals. We excluded *NCF1* from our study because it resides on a region of chromosome 7q11 near a pseudogene that prevents PCR amplification (Chanock et al. 2000). By re-sequencing, we have improved the typical resolution of genotyping studies, screening in an unbiased fashion all common and rare variants of the targeted genomic regions of our sample, and properly inferring the allelic

arquitecture of the studied genes (Ewens 1972). Although genome-wide association studies are successfully contributing to elucidate the influence of common variants on complex phenotypes, it is becoming clear that a component of missing heredability due to rare variants should still emerge (Pritchard 2001, Chang et al. 2009, Manolio et al. 2009) and that a better catalog of the spectrum of rare alleles across the human genome is necessary.

Molecular evolution of NADPH genes along the mammalian phylogeny

To infer how natural selection has acted on NADPH genes through the mammalian evolutionary history, we analyzed the coding regions of NADPH genes from all the mammalian species that were available in the Entrez database in June 2009 (one sequence for each species, see Supplementary Material for details). The most common approach to detect different types of natural selection on a coding region takes advantage of the fact that substitutions come in two classes: nonsynonymous (that change the resulting amino acid sequence of the protein) and synonymous substitutions (which do not change the encoded protein) (Nielsen et al. 2005). When comparing a set of homologous sequences from different species, most if not all of the observed differences are *fixed*: they are monomorphic within species because it has passed enough time for the observed variant to appear, increase in frequency in an ancestral population and reach frequency one (Kimura 1974). We compared the number of fixed synonymous substitution (dS, assumed to be neutral) and fixed non-synonymous substitutions (dN, for which we test the hypothesis of natural selection) using the parameter $\omega = dN/dS$, that is informative about the action of natural selection at inter-specific level (Yang 2007a, Kryazhimskiy and Plotkin 2008). Under neutral evolution of non-synonymous substitutions, these fix at the same

rate than synonymous substitutions, and therefore dN ≈ dS and ω ≈ 1. If non-synonymous substitutions tend to be deleterious, purifying selection maintains them at low frequencies, preventing its fixation at the same rate than synonymous substitutions, determining that dN < dS and therefore ω < 1. On the other hand, if episodes of positive (adaptive) selection are frequent, non-synonymous substitutions increase in frequency and fix more rapidly than neutral synonymous substitutions, and thus, dN > dS and ω > 1. We used the maximum likelihood framework developed by Yang (2007a) to estimate ω for NADPH genes. This approach (implemented in the software PAML, Yang 2007b) allows inferences about the evolution of a coding region along an inter-specific phylogeny, mapping which codons have evolved under strong/weak purifying selection, neutrality or adaptive positive selection (see Supplementary Material for details). The results of this analysis for *CYBB*, *CYBA*, *NCF2* and *NCF4* are shown in Figure 2, which shows for each codon and for known protein domains, the type of natural selection (i.e. the ω estimation) that most likely predominated during the mammalian evolutionary history. For this temporal scale, *CYBA*, *NCF2* and *NCF4* coding regions seem to have evolved driven by a combination of different levels of purifying natural selection, with few codons/domains under nearly neutral evolution in *NCF2* (ω = 0.809) and *NCF4* (ω = 1.159). Exceptions to this pattern are two *CYBA* codons (75 and 180) that respectively, show evidence of positive selection (ω = 3.721) on the maturation domain and the relatively less conserved C-terminal region of *CYBA*.

NADPH oxidase activation depends on the interaction among specific domains of its components. In general, we did not observe association among the interacting domains

evidenced in Figure 1b and specific types of natural selection. Only the case of the first SH3 domain in p67phox *(NCF2)* is noteworthy because it is associated to strong purifying selection (Figure 1b and 2). However, our most striking result regards the evolution of *CYBB*, the most critical component for the integrity of the respiratory burst, as evidenced by the >70% of CGD patients that have mutations in this gene. This fact and the predominant purifying selection on genes involved in Mendelian diseases (Blekhman et al. 2008) would lead to expect for *CYBB* a similar pattern respect to the other NADPH components. Conversely, during the evolution of mammals, episodes of adaptive natural selection have driven the evolution of *CYBB*, and even more important, almost all of these events are concentrated on the small extracellular part of this protein (Figure 3, Taylor et al. 2006). Intriguingly, there is no evident functional explanation for this observation, which should foster structural and functional studies to understand the biological basis of this evolutionary inference. The proximity of these inferred episodes of positive natural selection to glycosylation sites of gp91 is noteworthy, considering the importance of the glycome in immunity (Marth et al. 2008). Although episodes of positive selection had been reported for *CYBB* by genomewide surveys (The International Rhesus Consortium 2007, Koisol et al. 2008), their interesting relations with gp91 structure had not been analyzed.

Population genetics of NADPH genes

We re-sequenced exons, promoters and intronic regions of *CYBB*, *CYBA*, *NCF2* and *NCF4*, for a total of 35524bp for each of 102 healthy individuals of the SNP500Cancer project (http://snp500cancer.nci.nih.gov/, Packer et al. 2006, see Supplementary Material for details),

which includes: 24 African ancestry (15 African Americans from the United States and 9 Pygmies), 31 Europeans (from the CEPH/UTAH pedigree and the NIEHS Environmental Genome Project), 24 Asians-Oceanians (from Melanesia, Pakistan, China, Cambodia, Japan and Taiwan) and 23 admixed Latin American (i.e. Hispanics from Mexico, Puerto Rico and South America). Although this is a suboptimal representation of the worldwide population, this limitation is common to most human genomic diversity projects focused on SNPs genotyping or re-sequencing data. However, based on how human genetic diversity is apportioned within (> 85%) and between populations (<15%, Lewontin 1972, The International HapMap Consortium 2005), even studies based on suboptimal sampling are informative about the genetic structure of human populations as well as to carefully infer the role of different evolutionary factors on its determination (The International HapMap Consortium 2005, Nielsen et al. 2005, Rieder et al. 2008).

The pattern of genetic diversity on a specific genomic region depends both on the demographic history of populations, as well as on locus specific factors such as mutation, recombination and natural selection. Our goal is to infer which combination of evolutionary factors has shaped the pattern of diversity of NADPH genes, and considering the role of NADPH oxidase in defense against pathogens, we focus on the action of natural selection. We assessed intra- and between-population diversity for NADPH genes, and tested the null hypothesis of neutrality: that patterns of diversity may be explained considering only the demographic history of human populations and the mutation and recombination rates of each locus. We applied neutrality tests based on: (1) the allelic spectrum, which is the distribution of polymorphic sites across different classes of allele frequencies (Tajima's D and Fu-Li's D and F statistics [Tajima 1989, Fu and Li 1993])

and, (2) comparisons between the amount of polymorphisms and fived differences with an inter-specific outgroup (McDonald and Kreitman 1987) test and the adapted Kolmogorov-Smirnoff test by McDonald (1998). For the first set of tests, we used as null hypotheses both the classic Wright-Fisher model of neutrality with constant population size, and the more realistic evolutionary scenarios for human populations inferred on the basis of multilocus genetic data by Voight et al. (2005). Null distributions of the neutrality statistics under these evolutionary scenarios were generated using coalescent simulations (Hudson 2002) (See Supplementary Material for methodological details).

We previously analyzed the pattern of nucleotide diversity of *CYBB* on the same samples used in this study (Tarazona-Santos et al. 2008) and observed that this gene shows no common non-synonymous variants. This result is consistent with the fact that most CGD mutations are on *CYBB*, suggesting that substitutions on the coding region of this gene are deleterious and therefore, very rare in human populations. Interestingly, this lack of non-synonymous polymorphisms in humans contrasts with the recurrent episodes of positive selection inferred along the evolution of mammals. In general, non-synonymous substitutions are rare on the human genome (Crawford et al. 2005, Boyko et al. 2008), and when present, they usually show low frequencies, reflecting the action of purifying natural selection (Barreiro et al. 2008). For the NADPH oxidase components, this is also evident for *NCF4*, a gene that shows two rare and conservative (*in sensu* Polyphen, Ramensky et al. 2002) non-synonymous substitutions (T85N and A304E). Conversely, for *NCF2* we observed a combination that seldom occurs in human genes: 9 non-synonymous substitutions, three of them common (see table of haplotypes on the Supplementary Material for details) and six rare. On the other hand and interestingly, the two

non-synonymous substitutions observed in *CYBA* are common and ubiquitous in human populations: Y72H (rs4673) and V174A (rs1049254, in a position where variation among mammalian species is also observed).

In general, for the four studied genes, diversity and levels of recombination are higher in Africans than in non-Africans (see Table 1 for summary statistics and Supplementary Material for haplotype description and frequencies). These results are consistent with the pattern of diversity observed at most of the human genome as a result of the demographic history of the human species (The International HapMap Consortium 2005). In particular, these results reflect the African origin of modern humans and the "out of Africa" migration that, after a bottleneck that occurred 40-80 ky ago, led to the peopling of other continents (Voight et al. 2005, Garrigan and Hammer 2006). This evolutionary history implies that the first divergence between continental human populations was between Africans and the ancestral of non-African populations, and therefore, the highest between-population differentiation is observed between these two groups (Table 2, although *NCF2*, as described below, does not match this feature).

From the four studied genes, *NCF4*, that encodes for p40-phox - a regulatory component of the NADPH oxidase; shows a pattern of diversity that is typical for a gene that have evolved under the influence of the human demographic history, without the action of any form of positive natural selection. In addition to the features described in the previous paragraph, the allelic spectra of *NCF4* in the four studied populations are consistent with the neutral model of evolution (Tables 1 and 2). A model of neutral evolution for *NCF4* is not ad odds with the fact

that among the NADPH genes, *NCF4* is the only for which a genomewide association study (October 2009) have evidenced a replicated association in European populations with a complex trait: Crohn disease, an idiopathic inflammatory bowel disease that predominantly involves the ileum and colon (rs4821544 in intron 1, Rioux et al. 2007, Roberts et al. 2008).

While at inter-specific level *CYBB* evidenced the most interesting evolutionary history, with repeated episodes of positive natural selection, for human populations *CYBA* and *NCF2* show interesting patterns of variation. *CYBA* encodes p22-phox, which is shared as trans-membrane protein by different NADPH oxidases. In addition to harboring two common non-synonymous polymorphisms, the *CYBA* pattern of diversity shows the following characteristics (Tables 1 and 2): (1) it evidences the highest haplotype and nucleotide diversities and recombination levels (ρ) when compared with the other three studied genes, which is consistent with a relatively longer time until the existence of its most recent common ancestor of human lineages. (2) Diversity in *CYBA* is very high in non-African populations, particularly in Europeans. If compared with 329 genes of the SeattleSNPs project re-sequenced in a European sample, π_{CYBA} ranks eleventh (i.e. the 97th-percentil, http://pga.gs.washington.edu/summary_stats.html). (3) The level of differentiation (F_{ST}) between the four human populations is low (Table 2), compared with the average of 0.12 observed at genomic level (The International HapMap Consortium 2005). (4) There are contrasting proportions of polymorphisms/singletons between Africans (a high proportion) and non-Africans (a very low proportion). In particular, in the European population there is an excess of the proportion of common polymorphisms respect to expectation of the neutral model of evolution, as evidenced by the D test of Fu and Li (2003, see Table 1 and Supplementary Table 2). This excess of common variants in the European population is

significant even when we conservatively tested it against a scenario of human evolution that incorporate the "Out of Africa" bottleneck (Voight et al. 2005) and levels of recombination lower to those estimated for *CYBA* in Europeans (ρ_{CYBA} = 8.07 for the sequenced region). Indeed, the "Out of Africa" bottleneck and a lower ρ than the observed value increase the variance of the neutrality statistics, making more difficult to obtain significant results. Altogether, our results suggest that balancing natural selection (that acts maintaining different alleles at high frequency in a population) has contributed to shape the diversity of *CYBA* at least in the European population, since demographic forces alone do not suffice to explain the very high *CYBA* diversity as well as its proportional excess of common polymorphisms. Moreover, a comparative genomic analysis confirms this inference: the ratio of polymorphisms to fixed differences with the chimpanzee is not homogeneous across the gene in the different human populations (Mc Donald 1998, see Supplementary Material for results), as would be expected under neutral evolution. The inference of balancing natural selection is consistent with the association of *CYBA* variants with levels of ROS production (Bedard et al. 2009), as well as with cardiovascular diseases (San José et al. 2008). Advantage of the heterozygous is one of the mechanisms of balancing selection; therefore, we can speculate that the biological basis for heterozygous advantages may be the following: considering that p22-phox is not exclusive of phagocyte NADPH oxidase, but is also part of Nox expressed in other tissues, the dependence of ROS production on *CYBA* variants has to be finely regulated. If *CYBA* variants induce very high levels of ROS, it may favor a phagocyte-dependent efficient response to pathogens but may damage other endothelial tissues. On the other hand, tissue oxidative damage may not occur if the level of ROS production is low, but this may be associated with a less robust phagocyte

respiratory burst against pathogens. In this context, heterozygous individuals with a *CYBA*-dependent intermediate level of ROS production may have been favored by natural selection.

p67, encoded by *NCF2*, is a necessary cytosolic NADPH component for phagocyte ROS production. In the Asian population, *NCF2* diversity also shows a pattern that does not seem compatible with a neutral model of evolution (Table 1). In addition to its low nucleotide diversity, it has a highly differentiated haplotype structure (see frequencies of haplotypes NCF2-D11 and NCF2-E10 in Table SX). In fact, the highest F_{ST} values are observed among Asians and non-Asians (Table 2) and not among Africans and non-Africans, as usually observed for most of the human genome. Moreover, this differentiated haplotype structure is associated with a dramatic excess of rare polymorphisms, a feature that is not observed in the other populations (Table 1) and that is unexpected in Asian populations (Voight et al. 2005). An excess of rare polymorphisms is the pattern of diversity typically expected under a selective sweep scenario: a beneficial substitution (and its associated haplotype) rapidly becomes common (i.e. incomplete sweep) and eventually fixed (i.e. complete sweep), reducing the nucleotide diversity in its surrounding region, rendering rare other standing substitutions. During this process, new rare substitutions appear in the expanding positively selected haplotype, creating a region of high linkage disequilibrium, which is evident for *NCF2* in Asia when compared with the other studied populations (Figure 5). Alternatively, this peculiar pattern of diversity of *NCF2* in Asia may have been generated without the intervention of natural selection during the first colonization of Asia by modern humans. In a process of geographic population expansion, specifically in the front wave of an expansion, some rare alleles/haplotypes (i.e. surfing alleles) may become common by chance, mimicking the pattern of diversity generated by a selective sweep (Excoffier and Ray

2008). Currently, there are no population genetics tools that allow us to discriminate among these non-excluding scenarios.

In this study, we have analyzed the pattern of genetic diversity for four genes that encode for the components of the phagocyte NADPH oxidase complex, and act in concert as part of one of the first line defenses of mammals against pathogens: the respiratory burst. We confirmed that different types of natural selection have been particularly important to shape the pattern of genetic variation of immune-related genes (Ferrer-Admetlla et al. 2008, Barreiro and Quintana-Murci 2009), and added a new dimension to this observation: Tough our statistical power is not the same to infer natural selection events at the different evolutionary scales and populations, it seems clear that even for a set of genes that act in concert, encoding for a unique protein complex, the signatures of the action of different types of natural selection may be different across time, loci and populations. While the signatures of natural selection on the patterns of genetic diversity of the studied genes are reasonable clear, we are not able to specify which pathogens are the selective factors in specific populations.

Our previous analysis of *CYBB* genetic diversity in human populations, of the spectrum of CGD mutations (Tarazona-Santos et al. 2008), and the fact that more than 70% of CGD mutations are due to mutations in this gene, seemed to suggest that this gene was a typical example of strong purifying natural selection. However, when we extended our analyses to the mammalian evolutionary history, we verified that there have been successful evolutionary experiments (i.e. episodes of positive natural selection), and even more important, that these experiments were concentrated on the small extracellular region of gp91, and therefore, variation in this region may

be functionally relevant. It is also interesting that when we changed our time scale to the modern human evolutionary history, we continued to have evidences of the action of natural selection, although acting on different ways. A neutral pattern of variation was inferred for *NCF4*, a cytosolic component traditionally considered as having a regulatory and not essential role for the respiratory burst (although Matute et al. 2009 have reported the first CGD patient with a *NCF4* mutation). Conversely, natural selection seems to have contributed to shape the diversity of *CYBA* and *NCF2*, two essential components for the integrity of the respiratory burst. In particular, a signature of balancing selection (we speculated that likely due heterozygous advantage) was observed for *CYBA*, a gene that also encodes for components of non-phagocytic NADPH oxidases.

Acknowledgements

The authors are grateful to Silvia Fuselli for discussions of our results, and the Sequencing Group of the Core Genotyping Facility (National Cancer Institute) for their technical assistance. This research was supported by the Intramural Research Program of the NIH, NCI, Center for Cancer Research. CF was supported by the University of Bologna, ET-S, MM, WCSM and FL were supported by NIH, Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil) and Fundação de Amparo a Pesquisa de Minas Gerais (Brazil) and WCSM by Brazilian Ministry of Education (CAPES Agency).

Author contributions

SJC conceived the study; ETS, MM, FL, RC, AC, CF, LP and BS generated the data/analyzed the sequences; LB, AC, WCSM and MY provided tools for data generation and analyses; ETS, MM, FL, WCSM and RR performed population genetics analyses; ETS, MM and WCSM prepared tables and figures, ETS and SJC wrote the manuscript. All the authors contributed with discussions about the results.

References

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet. 40:340-5.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, Kidd JR, Kidd KK, Alcaïs A, Ragimbeau J, Pellegrini S, Abel L, Casanova JL, Quintana-Murci L (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 5:e1000562.

Barreiro LB, Quintana-Murci L (2009) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet. Dec 1. [Epub ahead of print].

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M (2008) Natural selection on genes that underlie human disease susceptibility. Curr Biol. 18:883-9.

Brandes RP, Kreuzer J. 2005. Vascular NADPH oxidases: molecular mechanisms of activation. Cardiovasc Res. 65:16-27.

Bedard K, Attar H, Bonnefont J, Jaquet V, Borel C, Plastre O, Stasia MJ, Antonarakis SE, Krause KH. 2009. Three common polymorphisms in the CYBA gene form a haplotype associated with decreased ROS generation. Hum Mutat. 30:1123-33.

Buckley RH. 2004. Pulmonary complications of primary immunodeficiencies. Paediatr Respir Rev 5 Suppl A: S225-33.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG. 2005. Natural selection on protein-coding genes in the human genome. Nature. 437:1153-7.

Chanock SJ, el Benna J, Smith RM, Babior BM. 1994. The respiratory burst oxidase. J Biol Chem 269: 24519-22.

Chanock SJ, Roesler J, Zhan S, Hopkins P, Lee P, Barrett DT, Christensen BL, Curnutte JT, Görlach A (2000) Genomic structure of the human p47-phox (NCF1) gene. Blood Cells Mol Dis. 26:37-46.

Chung CC, Magalhaes W, Gonzalez-Bosquet J, Chanock SJ (2009) Genome-wide Association Studies in cancer - Current and future directions. Carcinogenesis 31: 111-120.

Crawford DC, Akey DT, Nickerson DA (2005) The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet. 6:287-312.

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol. 3:87-112.

Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F (2008) Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol. 181:1315-22.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M (2009) Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. 19:199-212.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nat Rev Genet. 7:669-80.

Heyworth PG, Cross AR, Curnutte JT. 2003. Chronic granulomatous disease. Curr Opin Immunol 15: 578-84.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-8.

Kimura M (1974) The neutral theory of molecular evolution. Cambridge University Press. 384 p.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4(8):e1000144.

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genet. 4:e1000304.

Lewontin, R (1972) The apportionment of human diversity. Evolutionary Biology 6: 391-398.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747-53.

Marth JD, Grewal PK. 2008. Mammalian glycosylation in immunity. Nat Rev Immunol. 8:874-87.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

McDonald JH (1998) Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol Biol Evol 15:377–384.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J Sninsky J, Adams MD, Cargill M (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, Sicotte H, Staats B, Acharya M, Crenshaw A, Eckert A, Puri V, Gerhard DS, Chanock SJ. 2006. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res. 34(Database issue): D617-21.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 69:124-37.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30:3894-900.

Rieder MJ, Livingston RJ, Stanaway IB, Nickerson DA (2008) The environmental genome project: reference polymorphisms for drug metabolism genes and genome-wide association studies. Drug Metab Rev. 40:241-61.

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, RotterJI, Duerr RH, Cho JH, Daly MJ, Brant SR. 2007.

Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet. 39:596-604.

Roberts RL, Hollis-Moffatt JE, Gearry RB, Kennedy MA, Barclay ML, Merriman TR. 2008. Confirmation of association of IRGM and NCF4 with ileal Crohn's disease in a population-based cohort. Genes Immun. 9:561-5.

Rozas J (2009) DNA sequence polymorphism analysis using DnaSP. Methods Mol Biol. 537:337-50.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614-20.

San José G, Fortuño A, Beloqui O, Díez J, Zalba G. 2008. NADPH oxidase CYBA polymorphisms, oxidative stress and cardiovascular diseases. Clin Sci (Lond). 114:173-82.

Sumimoto H, Miyano K, Takeya R. 2005. Molecular composition and regulation of the Nox family NAD(P)H oxidases. Biochem Biophys Res Commun. 338:677-86.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-95.

Tarazona-Santos E, Bernig T, Burdett L, Magalhaes WC, Fabbri C, Liao J, Redondo RA, Welch R, Yeager M, Chanock SJ. 2008. CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains. Hum Mutat. 29:623-32.

Taylor RM, Burritt JB, Baniulis D, Foubert TR, Lord CI, Dinauer MC, Parkos CA, Jesaitis AJ. 2004. Site-specific inhibitors of NADPH oxidase activity and structural probes of flavocytochrome b: characterization of six monoclonal antibodies to the p22phox subunit. J Immunol. 173:7349-57.

Taylor RM, Baniulis D, Burritt JB, Gripentrog JM, Lord CI, Riesselman MH, Maaty WS, Bothner BP, Angel TE, Dratz EA, Linton GF, Malech HL, Jesaitis AJ. 2006. Analysis of human phagocyte flavocytochrome b(558) by mass spectrometry. J Biol Chem. 281:37045-56.

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-320.

Uhlemann AC, Szlezák NA, Vonthein R, Tomiuk J, Emmer SA, Lell B, Kremsner PG, Kun JF. 2004. DNA phasing by TA dinucleotide microsatellite length determines in vitro and in vivo expression of the gp91phox subunit of NADPH oxidase and mediates protection against severe malaria. J Infect Dis. 189:2227-34.

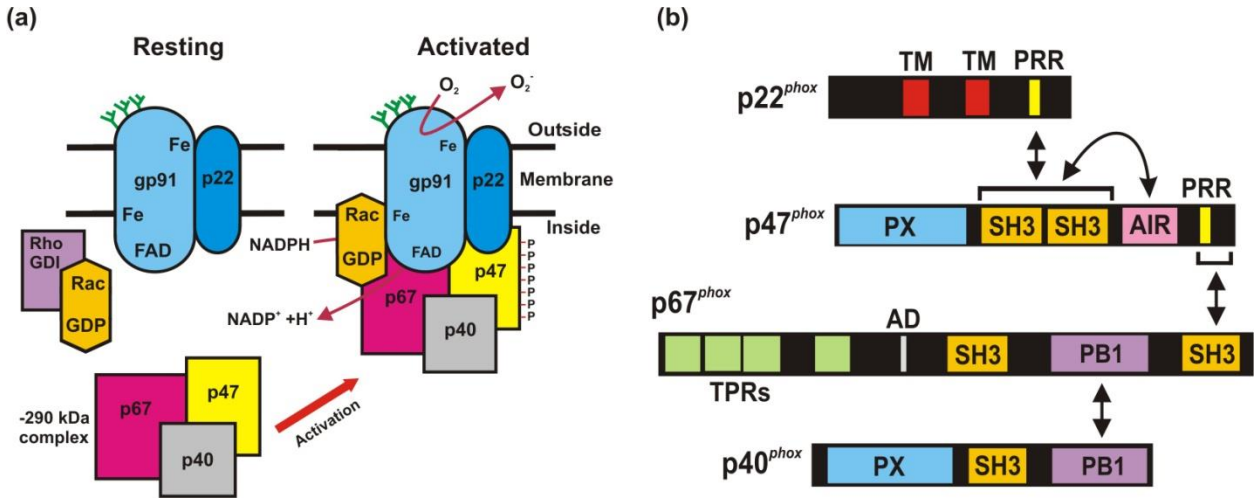
Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A. 102:18508-13.

Wang JP, Rought SE, Corbeil J, Guiney DG. 2003. Gene expression profiling detects patterns of human macrophage responses following Mycobacterium tuberculosis infection. FEMS Immunol Med Microbiol 39:163-72.

Yang Z. 2007a. Adaptive Molecular Evolution. In Handbook of Statistical Genetics, Edited by Balding DJ, Bishop M and Cannings C. Third Edition, John Wiley & Sons, Ltd, Susex, UK. Volume 1, Pp 377-406.

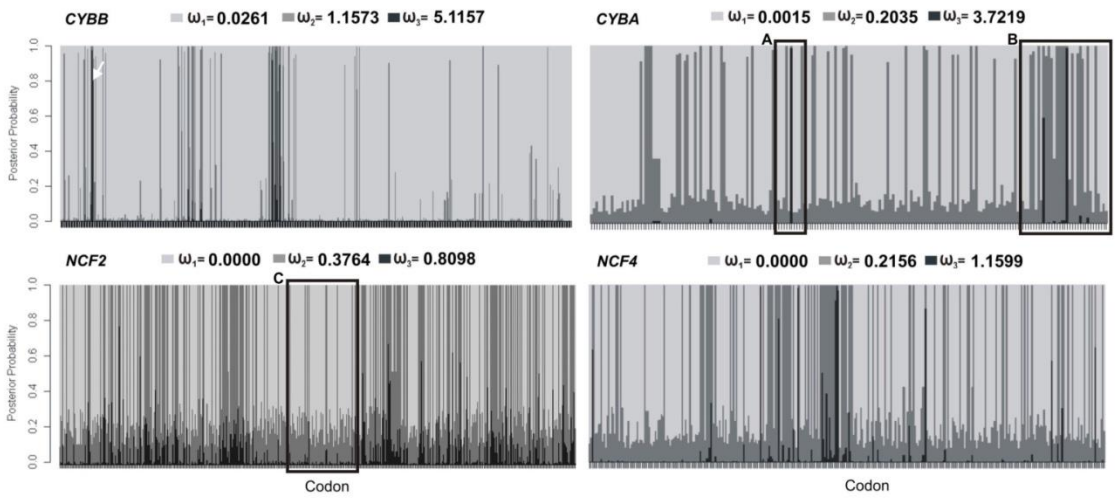
Yang Z. 2007b. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 24:1586-91.

Figure 1



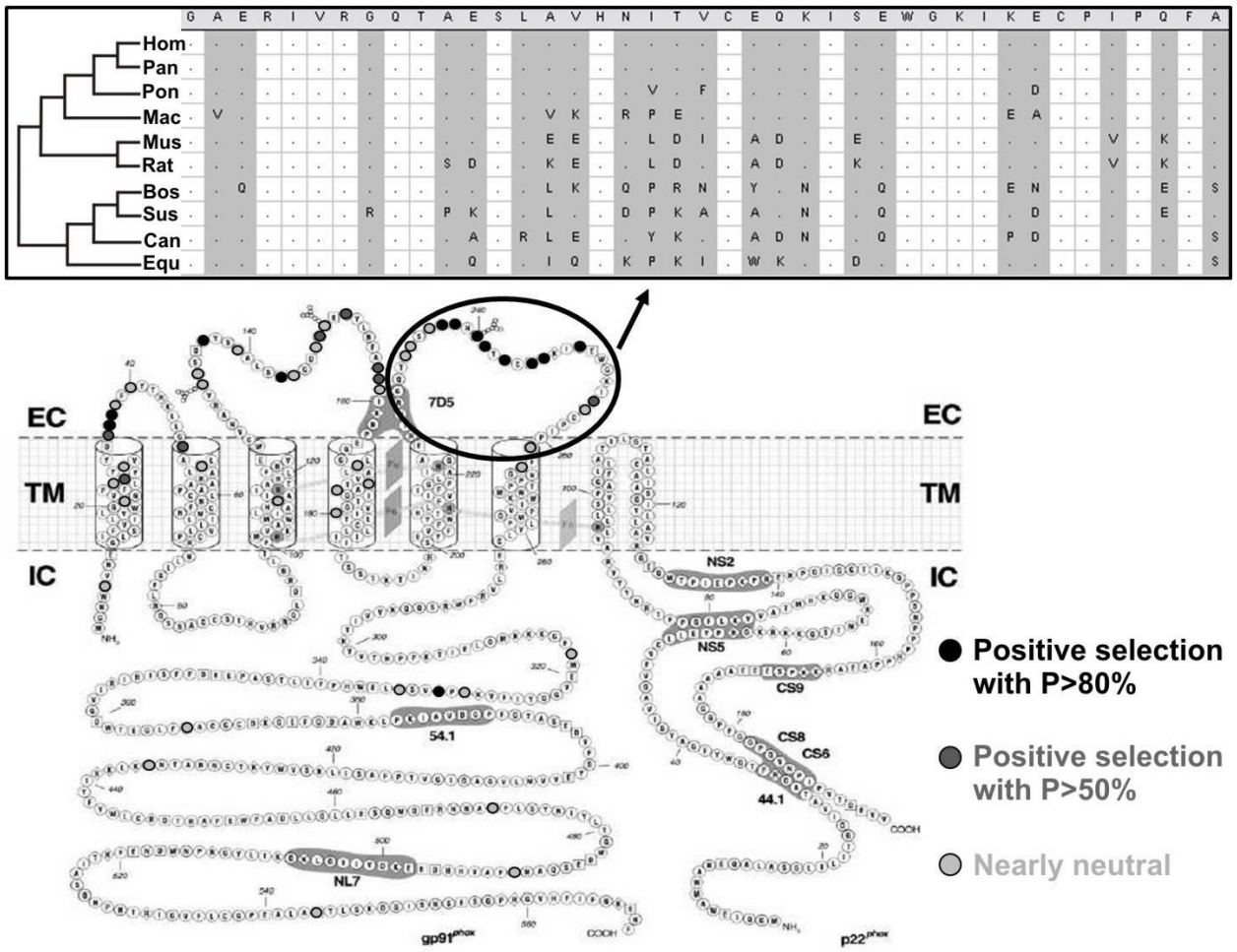
(a) Representation of the inactivated (left) and activated (right) forms of the phagocyte NADPH oxidase components. The activated form is responsible for the respiratory burst (adapted from Heyworth et al. 1999, Nature Encyclopedia of Life Sciences). The coding genes are: gp91 (*CYBB*, Xp21.1), p22 (*CYBA*, 16q24), p67 (*NCF2*, 1q25), p40 (*NCF4*, 22q13.1) e p47 (*NCF1*, 7q11.23). (b) Protein domains of the NADPH oxidase components and their interactions during activation of the NADPH oxidase (from Sumimoto et al. [2005]). TM: Trans-membrane, PRR: Proline-rich region, AIR: Autoinhibitory region, TPR: Tetratricopeptide repeat, AD: Activation domain, SH3: Src homology 3, PX: Phox homology, PB1: Phox and Bem1.

Figure 2



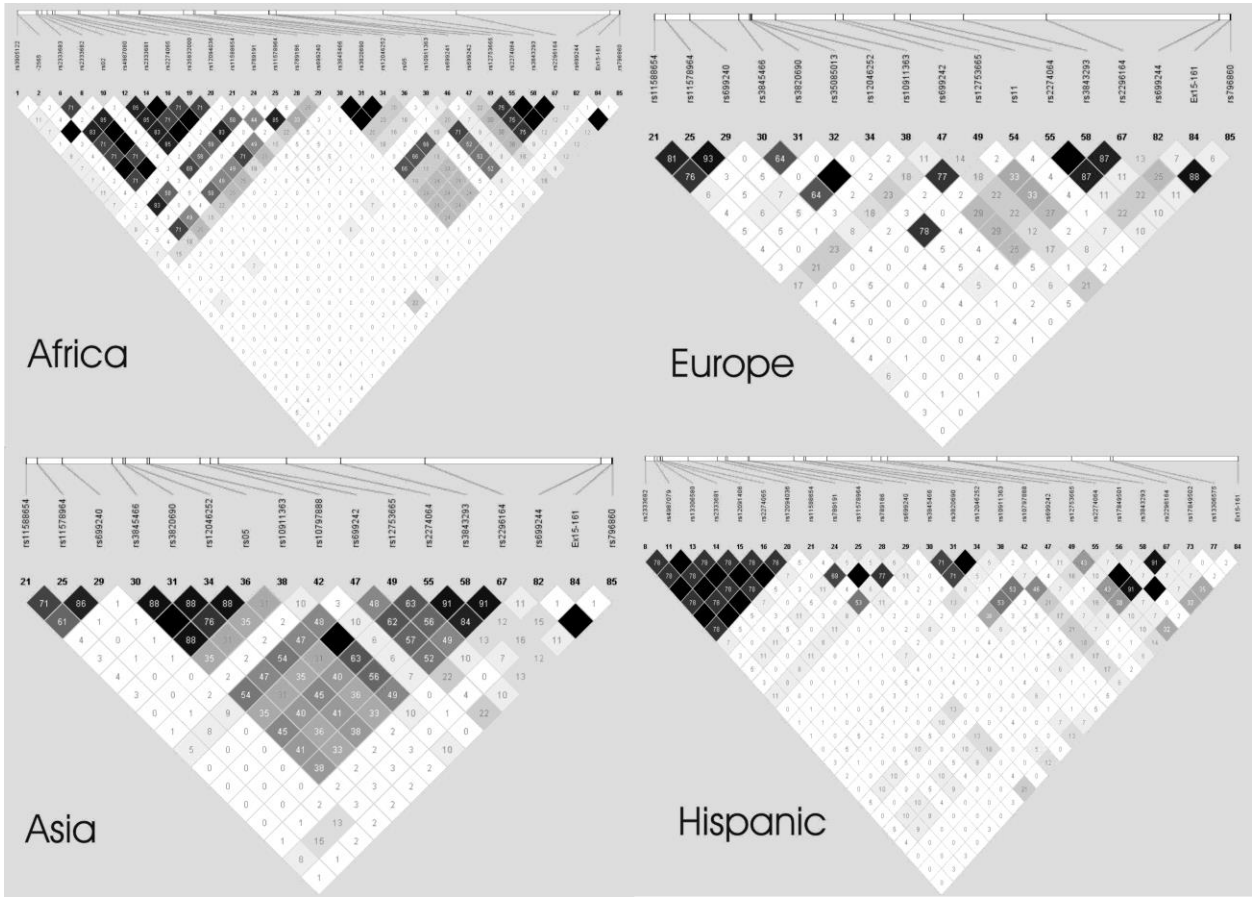
Inferred types of natural selection for codons of the NADPH genes at the evolutionary time scale of mammals. For each gene, three classes of sites (black, dark gray and light gray) are considered, each evolving under different ω values (evidenced for each gene in the figure). These classes correspond to the model M3 of Yang (2007a) with three classes of sites. Given our data, this model is more likely than alternative models of evolution that assume simpler scenarios such as a unique ω for the entire gene (see Supplementary Material for details about methods, results under alternative models and discussion of comparisons among models). The three classes correspond to different types and levels of natural selection, from strong purifying selection (in the lightest gray) to positive selection (ω > 1). For each codon, the probabilities of belonging to each of the three classes of ω correspond to the height of the corresponding color in the vertical bar. For instance, codon 37 of *CYBB* with probability 0.000 belong to the class of ω = 0.026 (light gray, which means purifying selection), with probability 0.195 belong to the class of ω = 1.15 (dark gray, which means near neutrality) and with probability 0.805 belong to the class of ω = 5.11 (black, which means positive selection). In this case, there is a reasonable evidence of positive selection on this codon.

Figure 3



Mapping of natural selection mapping across *CYBB* (encoding for gp91) along mammalian evolution, as inferred using the PAML method by Yang (2007a). The evidenced topologies for gp91 and p22 are reproduced from Taylor et al. (2004). Gray and black aminoacids have evolved under positive selection with probabilities >50% and >80%, respectively. Most of this aminoacids are concentrated on the extracellular part of the protein. The upper part of the figure shows the gp91 alignment for the region evidenced by the black ellipse, where there is a high level of aminoacid variation between species.

Figure 4



Pairwise linkage disequilibrium (LD) for *NCF2* common SNPs (minor allele frequency ≥ 0.05) in the four studied populations, measured by the r^2 statistics. Black squares represent $r^2 = 1$ (i.e. complete linkage disequilibrium). Decreasing r^2 values from 1 to 0 are represented by lighter gray tones. Lower diversity (less common polymorphisms) and higher LD is evident for the Asian population.

3. ESTUDOS DE EPIDEMIOLOGIA GENÉTICA E VARREDURA GENÔMICA (*GENOME-WIDE ASSOCIATION STUDIES*)

Estudos de associação por varredura genômica, conhecidos como *Genome-wide Association Studies* (GWAS), surgiram como uma importante metodologia para a descoberta de regiões no genoma, que apresentam variantes genéticas que conferem risco a diferentes doenças como, por exemplo, diferentes tipos de câncer e diabetes tipo2. O sucesso dos GWAs nos últimos três anos pode ser atribuído ao surgimento de novas tecnologias de genotipagem em paralelo, que permitem a tipagem de centenas de milhares de SNPs. Esta metodologia tem possibilitado a análise de grandes regiões genômicas em grandes conjuntos de milhares de casos e controles, sem a necessidade de definir uma hipótese ***a priori*** (Chung *et al*., 2010). Uma vez que novas associações genéticas são identificadas, pesquisadores adquirem um conhecimento mais completo da patogênese de uma doença, podendo eventualmente usar essas informações para desenvolver estratégias melhores para detectar e tratar um determinado fenótipo ou desfecho patológico.

Os estudos de varredura genômica são particularmente potentes porque permitem o teste simultâneo de hipóteses de associação com regiões do genoma representadas por cada um dos marcadores genotipados, bem como testes de associação entre regiões do genoma definidas por combinações de variantes (haplótipos). De fato, quando associada a tamanhos amostrais adequados e a uma adequada cobertura da variabilidade genômica por parte do conjunto de marcadores genotipados, a varredura genômica é uma das estratégias mais poderosas para descobrir o envolvimento da variabilidade genômica na patogênese de doenças complexas (Donnelly, 2008).

Dessa forma, os GWAS têm sido importantes para o entendimento da base genética de doenças complexas. A identificação de loci envolvidos na predisposição ao diabetes do tipo 1 (Todd *et al*., 2007), ao diabetes to tipo 2 (Sladek *et al*., 2007), à doença de Crohon (Duerr *et al*., 2006), ao câncer de próstata (Gudmundsson *et al*., 2007) e ao câncer de mama (Easton *et al*., 2007) são exemplos que têm permitido identificar o envolvimento de moléculas insuspeitas na patogênese dessas doenças. Até o presente momento, existem 591 publicações de estudos de GWA que referenciam 2879 SNPs (Manolio *et al*., 2009).

Outra característica importante dos estudos de associação é a grande quantidade de dados que eles geram e o potencial de integração desses dados com informações já existentes, tanto geradas por outros projetos da área como informações biológicas complementares, como vias metabólicas (Elbers *et al*., 2009). Por esta razão, os GWAS colocam desafios estatísticos e computacionais formidáveis para a análise de dados.

3.1 Publicações

3.1.1 Artigo IV

Genome-wide association studies in cancer—current and future directions

Estudos de associação genômica se tornaram uma importante ferramenta na descoberta de regiões que contem variações genéticas que conferem risco para diferentes tipos de câncer. O sucesso deste tipo de estudo nos últimos três anos foi principalmente devido à convergência de novas tecnologias que são capazes de genotipar centenas de milhares de SNPs junto com a anotação eficiente dessas variações genéticas.

Com este trabalho tive a oportunidade de discutir as principais iniciativas que utilizavam estudos de varredura genômica (GWAs), sua aplicações e perspectivas na elucidação da complexa arquitetura da susceptibilidade a doenças complexas, com ênfase em câncer.

Carcinogenesis vol.31 no.1 pp.111–120, 2010
doi:10.1093/carcin/bgp273
Advance Access publication November 11, 2009

Genome-wide association studies in cancer—current and future directions

Charles C.Chung[1], Wagner C.S.Magalhaes[1,2], Jesus Gonzalez-Bosquet[1] and Stephen J.Chanock[1,*]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, 20892-4608, USA and [2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CEP 31270-910, Belo Horizonte, MG, Brazil

*To whom correspondence should be addressed. Tel: +1 301 435 7559; Fax: +1 301 402 3134;
Email: chanocks@mail.nih.gov

Genome-wide association studies (GWAS) have emerged as an important tool for discovering regions of the genome that harbor genetic variants that confer risk for different types of cancers. The success of GWAS in the last 3 years is due to the convergence of new technologies that can genotype hundreds of thousands of single-nucleotide polymorphism markers together with comprehensive annotation of genetic variation. This approach has provided the opportunity to scan across the genome in a sufficiently large set of cases and controls without a set of prior hypotheses in search of susceptibility alleles with low effect sizes. Generally, the susceptibility alleles discovered thus far are common, namely, with a frequency in one or more population of >10% and each allele confers a small contribution to the overall risk for the disease. For nearly all regions conclusively identified by GWAS, the per allele effect sizes estimated are <1.3. Consequently, the findings of GWAS underscore the complex nature of cancer and have focused attention on a subset of the genetic variants that comprise the genomic architecture of each type of cancer, which already can differ substantially by the number of regions associated with specific types of cancer. For instance, in prostate cancer, there could be >30 distinct regions harboring common susceptibility alleles identified by GWAS, whereas in lung cancer, a disease strongly driven by exposure to tobacco products, so far, only three regions have been conclusively established. To date, >85 regions have been conclusively associated in over a dozen different cancers, yet no more than five regions have been associated with more than one distinct cancer type. GWAS are an important discovery tool that require extensive follow-up to map each region, investigate the biological mechanism underpinning the association and eventually test the optimal markers for assessing risk for a disease or its outcome, such as in pharmacogenomics, the study of the effect of genetic variation on pharmacological interventions. The success of GWAS has opened new horizons for exploration and highlighted the complex genomic architecture of disease susceptibility.

Introduction

The history of human genetics has focused on mapping regions of the genome that can explain part or all of a disease or human trait. With the generation of a draft of the human genome in 2001, geneticists quickly set out to comprehensively annotate the genome and apply the evolving knowledge of the pattern of genetic variation to investigate both monogenic, Mendelian disorders and complex diseases, the latter of which by nature are polygenic (1–4). Until recently, the scope and breath of human variation was certainly underappreciated until the advent of early maps of common variants, such as the single-nucleotide polymorphism (SNP), the most common variant in the genome (1,5–7). It is notable that a comprehensive set of genetic variation has shifted the analysis paradigm to finding genetic contributions to complex disease, whereas the capacity to capture environmental exposures and lifestyle decisions is far more rudimentary, even though these factors are essential for understanding complex diseases and traits.

For many years, human genetics has successfully mapped uncommon mutations with large effect sizes in studies conducted in families or special populations, such as the *BRCA1/BRCA2* mutations in Ashkenazi women with breast cancer and ovarian cancer (8). The search for highly penetrant mutations in familial aggregation has been based on genetic linkage analysis, an approach that has used microsatellite markers across the genome to scan for markers that segregate within a family (9,10). Based on the identification of linkage peaks using rigorous statistical approaches, follow-up of regions was pursued based on strong signals. Because of the wide spacing of markers across the genome, signals often pointed to regions over multiple megabases that in turn required sequencing large regions of the genome in search of the causative mutations, a daunting task in scope and until recently hampered by technical limitations. Nonetheless, successes in families loaded with melanoma, breast cancer and sets of cancers (Li-Fraumeni Syndrome) (8,11–14) are notable and provided an important substantiation of the approach of using markers indirectly. In retrospect, the use of markers to conclusively identify regions for detailed analysis has been an important lesson for mapping germ line genetic variants associated with risk for cancer, but the approach yielded only mutations with very strong effects.

Over the past 20 years, a parallel approach has been pursued to discover common genetic variants that confer susceptibility to different types of cancers. Initially, association studies were conducted using a handful of annotated genetic variants for which a strong hypothesis could be formulated. In a genetic association study, the analysis consists of a comparison of the distribution of a marker allele between cases and controls, in search of a statistical difference that can be reflected in an estimated effect size—usually quite small compared with mapped linkage signals due to highly penetrant mutations. Naively, at first, investigators searched for alleles with high estimated effect sizes (e.g. per allele odds ratios > 2.0), but with time, it has become apparent that common alleles confer small risk overall in sufficiently large case–control studies of unrelated subjects, the primary study design for association analyses (15).

Nominally, investigators focused on SNPs that altered the coding sequence and resulted in a non-synonymous change, namely a shift in the amino acid sequence of the protein. The approach was predicated on a more simplistic model: changes in the amino acid content would lead to a pronounced (e.g. measurable) change in function and thus influence the disease or trait of interest. Due to the inadequately sized studies, issues of study design and the overestimation of effect size, nearly all published candidate gene association studies, probably represent false positives. In this regard, the candidate gene approach has yielded very few notable findings, namely those that are conclusive and do not represent false positives. To date, perhaps a handful have been adequately replicated and confirmed in follow-up studies. For example, *GSTM1* null and *NAT2* slow acetylator genotypes have been associated with increased overall risk of bladder cancer and could account for up to 31% of the disease because of their high prevalence (16). Similarly, candidate genes have shown robust findings for a promoter SNP in *TNF* in non-Hodgkin's lymphoma and a coding variant in *CASP8* in breast cancer (17,18). But overall, very few candidate studies have yielded convincing results worthy of the enormous investment of time to pursue the biological basis of the association.

Abbreviations: CNV, copy number variation; GWAS, genome-wide association studies; LD, linkage disequilibrium; MAF, minor allele frequency; PSA, prostate serum antigen; SNP, single-nucleotide polymorphism.

In the early part of the new millennium, candidate gene studies expanded in scope, looking at sets of genetic markers across a gene of interest. This transition adopted the use of sets of markers defined on the basis of genetic correlation, known as linkage disequilibrium (LD) discussed below. Often, markers are located in introns or intergenic regions, raising the possibility that genetic variants could alter expression or regulation of a gene, thus not only widening the spectrum of variants to be examined but also increasing the scope of underlying mechanisms. As this approach began to find variants associated with cancer risk, the focus was on markers for risk. For examples, Garcia-Closas *et al.* (19) identified a promising marker near the *VCAM1* gene in association with bladder cancer as part of an exploration of genes in several pathways related to cancer biology. Again, the approach was hypothesis driven, in that specific genes were chosen for the best markers but the scope was enlarging and increasing the number and types of variants explored (20).

In 1996, Risch and Merikangas argued that for complex diseases, such as most cancers, large scale linkage studies will be both difficult and not as well powered to detect susceptibility alleles with low estimated effect sizes, of the type that are probably to contribute in a polygenic model (15,21,22). Instead, they suggested that large-scale association testing could be more efficient and more effective (15,21) in the discovery phase. Moreover, the practicality of collecting large sets of family pedigrees was identified as a daunting, and perhaps overwhelming challenge. Indeed, the age of genome-wide association studies (GWAS) has established the association study as an integral tool for discovering the contribution of common genetic susceptibility alleles to different types of cancer.

The value of conducting statistically sound studies that are well powered has become a central tenet of the GWAS era because of the enormous risk for false-positive discovery. The threshold for discovery has been established at a high level, known as genome-wide significance, which serves two dual purposes (23,24). First, it necessitates careful consideration of the power to detect the effect sizes expected to be observed in the study. Second, the high bar of genome-wide significance protects against the probability of a false-positive finding (25,26). The latter is critical because GWAS are discovery tools that point investigators toward long arduous follow-up studies for unraveling the underlying biology and the pursuit of markers for risk assessment (27).

Background

The scope of genetic variation

Based on the international annotation projects and the sequencing of nearly a dozen full human genomes, the spectrum of human genetic variation is enormous with respect to the types of genetic variation and the magnitude of variants in any given genome (28–34). Although two genomes are estimated to differ by $<0.5\%$, there are at least several million differences, only a small subset of which contributes to disease risk while the majority is probably vestigial. The most common type of variation is a single-nucleotide base substitution, known as the SNP. Next generation sequence analysis has begun to identify the large set of small insertions or deletions in sequence (30,35,36). Progressively, larger structural alterations and copy number variants are fewer in absolute number but impact more bases across the genome (Figure 1).

Most common variants namely those with a minor allele frequency (MAF) $>5\%$ are common to all populations, although the distribution of allele frequencies can vary greatly across the globe (37). Ascertainment estimates for lower frequency variants depend on both the number of subjects as well as the population genetic history of those examined. With next generation sequencing applied to high-profile regions in large numbers, greater complexity in different human populations is emerging, particularly with variants of lower frequency (36,38,39). Interestingly, the scope of structural variants is much greater than previously recognized, though the majority of large-scale polymorphisms appear to be less common, namely <1–5% in unrelated populations, unlike SNPs and insertions and deletions, of which there are millions with frequencies $>5\%$. Accordingly, the GWAS approach in unrelated subjects has been most successfully applied to SNPs and it has been far less successful applied to structural variants, also known as copy number variations (CNVs).

The most common sequence variation in the germ line genome is SNP, which, by definition, is observed in at least 1% of a population. By definition, the MAF is a relative term and applies to the allele with the lower frequency at a locus in a reference population. In many instances, there can be major differences in MAFs between populations with distinct histories. For the common SNPs (MAF $>5\%$), $<10\%$ of SNPs are specific to a given population (28,37). This observation suggests the common ancestry of common SNPs. The literature suggests that there are at least 10 million SNPs with

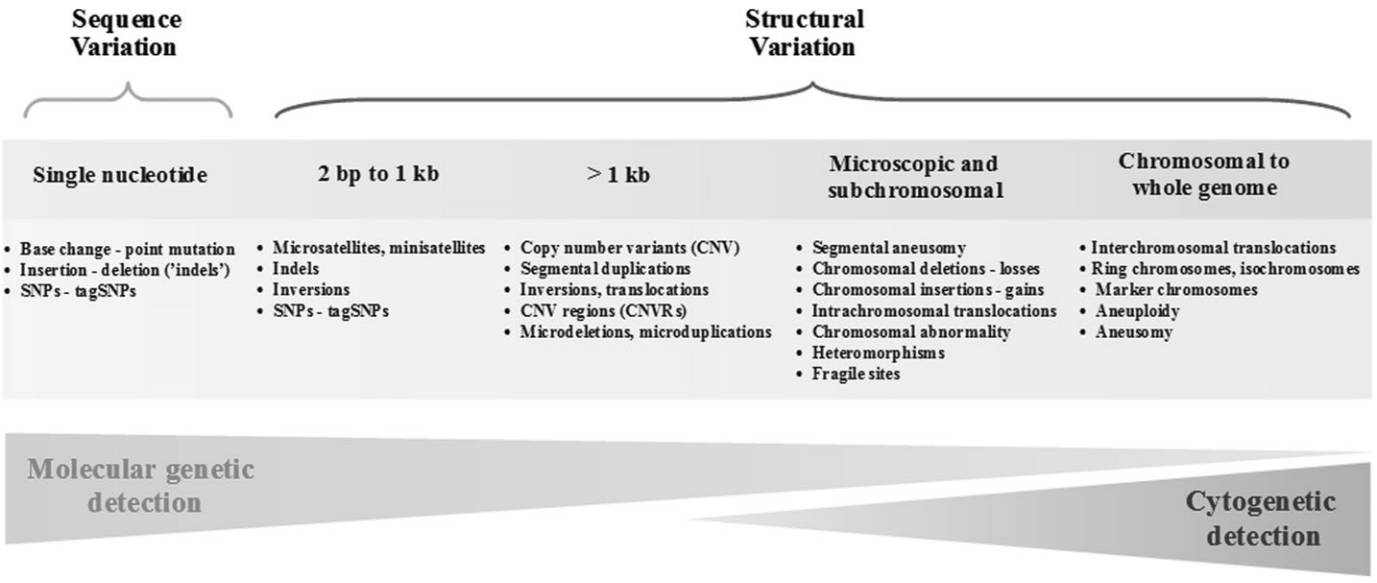


Fig. 1. Types of genetic variations in the human genome. Common types of genetic variations can be categorized into two major groups—those that involve single base changes (e.g. SNPs) and those that alter more than one base (e.g. microsatellites or structural variants).

a MAF >1% (40–42) and 5 million SNPs with a MAF >10% (3,4,40) but recent large-scale sequencing efforts, such as the 1000 Genome project, indicate that these estimates are low (www.1000genomes.org/) (43). In fact, there could be double or triple the earlier estimates. Lastly, there is a small subset of SNPs that are tri-allelic; at a given base on the reference genome, there can be three different bases, though these are rare, they can be formidable technical challenges for quality control metrics.

It is estimated that between 50 000 and 250 000 common SNPs could be biologically active, as non-synonymous coding variants or regulators of gene expression or splicing (7,15). For candidate gene studies, there was a premium assigned to SNPs in coding regions, usually based on *in silico* predictions. These coding SNPs, known as cSNPs, can be divided into non-synonymous variety (which alters the predicted amino acid codon) and synonymous SNPs (which do not alter the codon sequence). The latter are far more common and less probably alter function. Though intense interest has been directed at non-synonymous SNPs, few have been conclusively associated with human diseases and even fewer have corroborative biological data to provide plausibility for the association (7,15). There has been considerable effort to predict the effect of a non-synonymous cSNP and putative conformational protein changes, but the biological significance is based on laboratory evidence only. Recently, it has emerged that there are subset of SNPs that alter regulation or expression of a gene. These regulatory SNPs are difficult to identify using informatic tools and thus have to be defined on the basis of laboratory data (44).

More than 5 million human SNPs of the international public repository for SNPs, known as dbSNP (www.ncbi.nih.gov/SNP/), have been validated to date with genotyping assays by the SNP Consortium and the International HapMap Project (1,28). Until recently, sequence validation was applied to a small subset but this is about to shift with the completion of the 1000 Genome Project, so that the majority of entries will be sequence based (45,46). Historically, many variants in dbSNP are monoallelic, due to either genotyping error or, more probably, sequencing errors (47,48). It is notable that the reported SNPs have been biased toward high-frequency variants in populations of European ancestry. The catalog of uncommon variation, namely SNPs with MAF under 1%, is incomplete but the 1000 Genome Project is expected to generate a catalog of variants between 0.5 and 5% frequency, which will complement the International HapMap of common variants above 5–10%. Already, the latest build of dbSNP has >20 million variants, mainly less common ones. In addition, dbSNP contains downloads from many disease-specific mutation databases, which will make the curation and utility of less common variants even more daunting for analytical approaches toward prioritization of variants for study. Still, the contribution of uncommon variants represents an untapped portion of the genomic architecture and will necessitate new approaches toward mining these variants for cancer susceptibility. Highly penetrant disease mutations are cataloged in a public database, the Online Mendelian Inheritance in Man or OMIM (www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM/).

The spectrum of genetic variation in the genome can range from single base substitutions to small insertions/deletions to structural variations that can be cytologically observed. The short tandem repeat, also known as the microsatellite, represents a class of polymorphisms used in linkage analysis that are defined by repeats of two or more nucleotides but display notable differences in the frequencies of the repeat units. Typically, they are located in non-coding regions. However, most large-scale structural variation is submicroscopic and ranges in size from a few base pairs to thousands of base pairs (49,50). Collectively, the submicroscopic variants are known as CNVs, a focus of intense interest in large-scale association studies. Estimates of segmental duplications in the genome have been suggested to approach 10% of the genome, but most are not common enough to be effectively analyzed using current GWAS (51–53). Current surveys suggest that CNVs are less common than previously reported (54,55) and in fact, perhaps, three-quarters of common CNVs are in LD with common SNPs (55).

Correlation of common genetic variants

It has been observed that the majority of SNPs are not inherited independently but segments on a chromosome, inherited from generation to generation (41,56,57). A central concept in germ line genetics is the inheritance of correlated markers on the same chromosome, known as LD. It is defined as the non-random association between allelic markers on a chromosome and is classically measured using one of two estimators, D' or r^2 (58). Individual SNPs that are strongly correlated with each other are said to be in LD, but with time and geographic distribution, LD can erode by recombination events (e.g. exchange of genetic material) during meiosis (59).

Haplotypes are defined as sets of SNPs or polymorphisms (e.g. insertions, deletions or large copy events) in strong LD, in which one or more can serve as surrogates for the other markers on the haplotype. A haplotype can be determined in most cases with family trios but in GWAS or large association studies, family structure is usually not available. Still, the offspring haplotype phase can be determined if the parental genotypes are known or established by biochemical methods and then applied to study to best estimate the common haplotypes (58). However, the phasing of haplotypes is more challenging in unrelated subjects but accurate estimates based by well-developed statistical methods that can account for the ambiguity of unobserved haplotypes can provide haplotypes with assigned probabilities (58). Some have argued that haplotypes are preferable for candidate gene studies but for GWAS, the approach is laborious and less nimble in analyzing the thousands of markers genotyped. The methods are not as robust for conducting analysis across thousands of variants.

The appreciation of applying LD to the millions of SNPs observed in human populations that has given rise to the fundamental principle of GWAS, testing across the genome with well-chosen markers that serve as surrogates for untested markers (60–62). The 'indirect approach' represents the first step in identifying regions with strong association with cancer or a human trait and relegates the investigation of the optimal variants to study for understanding the biological basis of the association signal (59). The commonly used approach to select optimal SNPs is the 'greedy algorithm', which estimates highly correlated SNPs, on the basis of MAFs and creates heuristic bins of 'tagged' SNPs. It is the set of tags that function as proxies for the highly correlated untested variants (60).

Practical issues in GWAS

GWAS have emerged as a powerful tool to identify susceptibility loci with low effect sizes in unrelated subjects with specific cancers and related outcomes. Though epidemiologic design is important, in the discovery phase, there has been a relaxation of epidemiologic rigor in order to discover novel regions, mainly because of the need to gather a sufficiently large enough data set to detect low effect sizes. Often, groups have used convenient or publicly available controls for the discovery analysis in GWAS (23), of which the Wellcome Trust Case Control Consortium has been a notable example. These steps could come at a cost, such as a slightly higher rate of false positives, or in related manner, the apparent contradiction of regions or loci that do not robustly replicate in separate scans, suggesting subtle, but real differences related to selection and exposure criteria. Consequently, the estimates are slightly unstable and maybe refined as better studies if analyzed with high quality epidemiologic and environmental exposure data. In order to meet the requirements of a sufficiently large enough data set to observe significant differences between cases and controls, many scans, particularly for rarer cancers, have had to amalgamate data sets.

Replication of results is critical in a separate comparable set of studies (63). The value of replication is to guard against the blizzard of false positives observed with common alleles with low effect sizes. By scaling the studies, GWAS can effectively shed the majority of false positives. The industry standard that has emerged has targeted genome-wide statistical significance for a GWAS with a P value less than between 5×10^{-7} and 1×10^{-8} using either a trend or genotype test, adjusted for minimal cofactors/covariates (23,64–66).

Because GWAS are conducted in unrelated subjects, there has been intense interest in the background population substructure of cases and controls. The capacity to examine thousands of markers with minimal or no LD can be used to effectively discriminate differences in population substructure (67–69). Population stratification is present when there is a measurable difference in the distribution of alleles between subgroups that have different population histories, which can certainly alter association analyses, providing false-positive findings, such as in early case–control studies, in which the cases and controls were drawn from individuals of different populations. Stratification between cases and controls based on differences in exposures can also be problematic, but less so in GWAS. The ability to detect stratification with sets of markers depends on the allele frequencies in each subgroup (70). Subjects with admixture coefficients >15–20% can be removed from association analyses (71) based on attempt to separate subjects into groups and determining the distribution of shared alleles. Further, detection of population stratification is conducted on the GWAS data set to adjust simultaneously for a fixed number of top-ranked principal components resulting from a principal component analysis (67). The search for underlying subgroups in stratified samples can be investigated with genetic markers not linked to the phenotype, using a principal component analysis that yields eigenvectors, used to adjust for possible inflation of test statistics due to stratification (67,72,73).

One of the fundamental reasons for the success of GWAS has been the foresight to collect biospecimens in case–control and cohort studies over the past decades, each of which affords advantages for studying exposures or avoiding survivorship bias. Since the high throughput genotype platforms that analyze thousands of commercially determined SNPs and now CNVs demand high performance DNA, most investigators have used native DNA—either from blood or buccal cells. The latter works quite well when optimally collected and extracted (74). Neither whole genome amplified DNA can be effectively used in GWAS or can materials from tumor tissue (or its adjacent region) due to problems with allelic imbalance. High-quality genotypes are generated using widely accepted quality control metrics for SNP completion, sample completion, heterozygosity scores, testing for fitness for proportion of Hardy–Weinberg equilibrium (70) and assay verification with a second technology (75).

Scanning the genome with SNPs can be performed with commercially available fixed products that provide hundreds of thousands of SNPs, chosen either on the basis of the tag strategy, spacing across the genome or inclusion of obligate SNPs either known or predicted to be functionally important. Great importance has been attached to the extent of 'coverage' afforded by the fixed content chips, which for each commercial product has translated into higher cost for greater coverage (24). The bias of the chips has been to select SNPs that most efficiently tag common SNPs in individuals of European background based on the successive builds of the International HapMap Project (Figure 2). Specifically, the level of coverage is generally measured by determining the percentage of 'bins' tagged by SNPs (with MAF > 5 or 10%) for each of the three HapMap II populations, individuals of European background (known as CEU), Yoruban of West Africa (YRI) and East Asians (CHN and JPN) (24,59,60). Over 500 regions of the genome have now been conclusively associated (e.g. report signals with P value $<5 \times 10^{-7}$) in >100 human diseases or traits (76–78).

The analysis of dense genotyping data can be carried out with publicly available tools in either Genotype Library and Utilities (GLU) or PLINK (79), each of which permits archiving, manipulation and basic analyses of data sets, including assessment of population

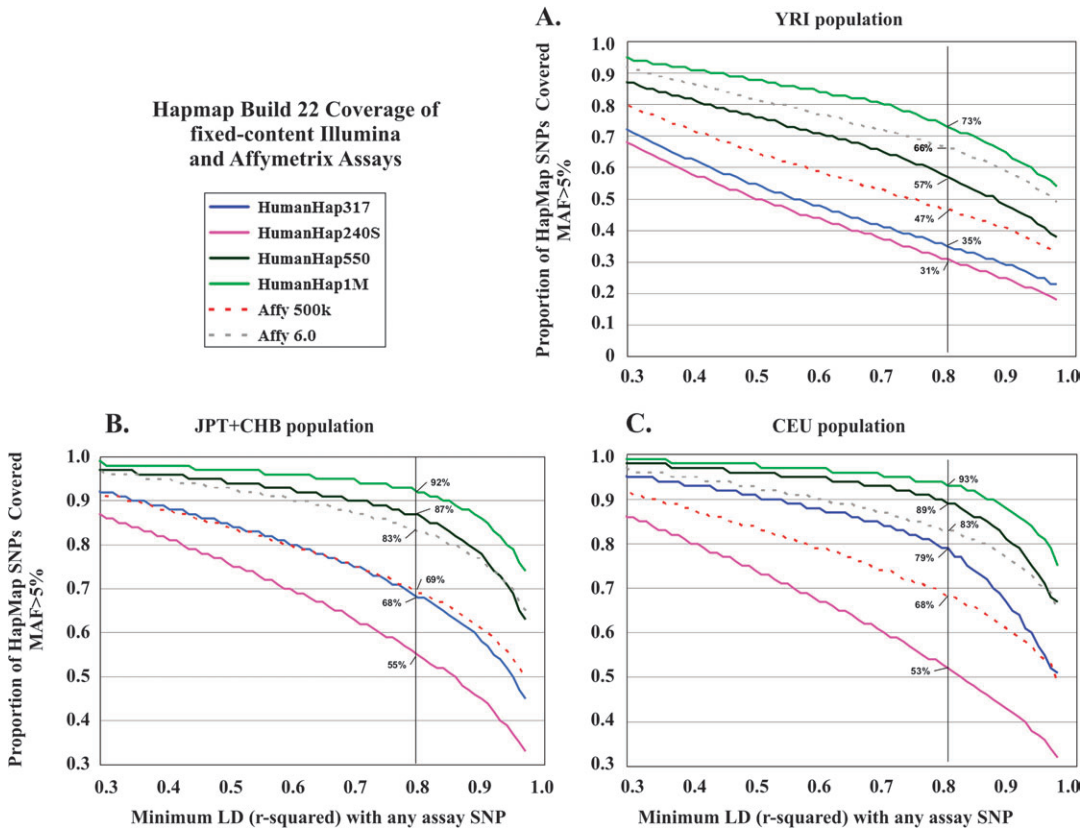


Fig. 2. Coverage of various genotyping platforms on HapMap II SNPs. The coverage of commercially available genotyping platforms in HapMap populations are plotted based on estimates of linkage disequilibrium using r^2, the correlation coefficient. A vertical bar depicts the cut off of an $r^2 = 0.8$, which is commonly used as a threshold to effectively tag monitored SNPs. The three HapMap populations of Phase II are labeled and the percentage estimated at the threshold is provided. (**A**): Coverage plot in Yoruban population (Ibadan, Nigeria), (**B**): coverage plot in Japanese (Tokyo, Japan) and Han Chinese (Bejing, China) and (**C**): coverage plot of US residents with northern and western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH).

substructure and association testing for SNPs. CNVs are more challenging because the primary image files have to be analyzed and quality control metrics applied to predict CNVs with varying degrees of probability. It is this latter issue, together with the evolving annotation of CNVs, which has hampered the widespread application of this type of analysis to yield association results comparable to those from common SNPs. Consequently, only a handful of common CNVs have been conclusively associated with complex diseases. In cancer GWAS, only one conclusive finding has been reported, the association of a region on chromosome 1 with the rare pediatric cancer neuroblastoma (80).

The first look at GWAS findings in cancer

Theme and variations

The age of GWAS and cancer have quickly ushered in a new era of discovery of regions that harbor germ line genetic variants (common and uncommon) associated with susceptibility to specific cancers. Currently, >75 regions of the genome (some harboring multiple independent signals) have been conclusively associated with susceptibility to specific cancers. Notably, in a handful of few circumstances, more than one type of cancer maps to the same set of genetic variants but overall, it appears that the contribution of common germ line variation has a strong component of tissue specificity. It is also notable that no single locus identified by the current crop of etiologically driven GWAS has also been shown to influence outcome, as measured by progression, disease stage, metastases or survivorship. This latter observation suggests that the germ line factors responsible for development of a cancer could differ from those genetic factors that sustain carcinogenesis or lead to progression. It is interesting to note that for the 29 independent loci identified in prostate cancer GWAS, so far, not a single locus exclusively associates with the more aggressive form of the disease (65,66,81–84). In the Cancer Genetic Markers of Susceptibility Initiative of a GWAS in prostate cancer, the analysis plan specifically addressed the early and advanced forms of prostate cancer, yet did not identify a locus specific to disease state (65,66,84). Consequently, it will be necessary to conduct distinct GWAS in studies designed to address these important outcomes, but it will most probably require new collections and collaborative networks to achieve the required numbers to discover the low to moderate effect alleles influencing cancer outcomes.

It was unanticipated that GWAS studies in certain cancers would yield many novel regions (e.g. prostate cancer with perhaps 29, breast cancer with 13 and colon with 10) (64,66,75,81–93), whereas other cancers strongly associated with environmental exposures have yielded so few regions: three for lung cancer in primarily smokers and three in bladder cancer despite analysis of sufficiently large data sets. Thus, it is plausible that the effect of tobacco use is substantially stronger than any single region with low estimated effect sizes (below 1.3 in GWAS). The lung cancer findings are also notable in that the strongest signal on chromosome 15q25 maps to a region that has also been identified in GWAS of smoking phenotypes (94–97). Prior to GWAS, it was also considered on the list of candidate genes because it contains nicotine receptors (e.g. *CHNRA3* and *CHRNA5*) (98,99). Further studies are urgently needed in non-smoking cases and controls to discriminate between signals that could be driven by tobacco exposure versus primary carcinogenesis (94). Fine-mapping studies in different populations may accelerate the pinpointing of the set of variants in this region requiring further study to understand the biology underlying the association study.

There are few notable exceptions to the observation that the per allele estimated effect is <1.5 for alleles discovered in cancer GWAS (100). In fact, most are <1.3, and it is anticipated that more will be discovered in the vicinity of 1.1–1.2 as consortial activities permit meta-analyses with larger sets of scanned subjects (Figure 3). Still, it was notable that two recent testicular cancer scans each identified two regions with effect sizes considerably greater than what had been observed previously in cancer GWAS. The loci mapped to regions on chromosomes 5 and 12 that harbored candidate genes previously implicated in testicular development, the ligand for the receptor tyrosine kinase (*KITLG)* and sprouty 4 (*SPRY4*). Moreover, the studies were notable for the high effect sizes detected for chromosome 5, namely >2.5, as well as the biological plausibility of the candidate genes (101,102). This was not surprising in light of the marked increase risk for family members (103,104). Another cancer with a familial aggregation, thyroid cancer, also yielded alleles with relatively high estimated effect sizes, and interestingly, they were detected in a small primary scan (105).

In select GWAS, the findings have pointed to genes previously investigated in that cancer. Pancreatic cancer is a highly lethal disease with a 5-year relative survival of <5% (106), with known risk factors of family history of pancreatic cancer, type 2 diabetes mellitus and cigarette smoking. Interestingly, the first reported GWAS in pancreatic cancer identified a variant in an intron of the ABO blood group antigen, which confirmed a finding suggested 50 years ago (107,108).

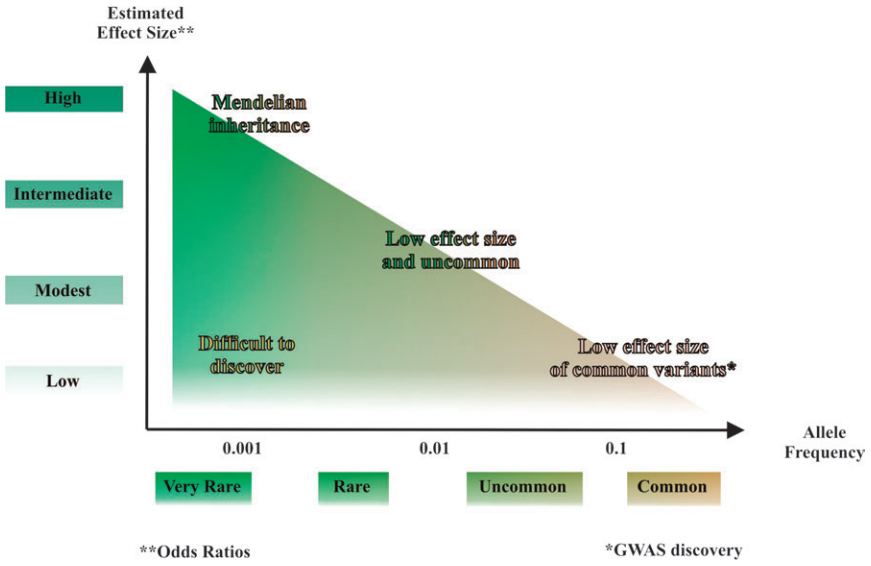


Fig. 3. The relationship between the estimated effect size and the allele frequency of disease susceptibility locus. The majority of disease susceptibility loci identified by GWAS in different cancers have low effect size (per allele estimated effect size of 1.1–1.3).

This is a striking example of how a GWAS hit points to a finding previously described in the epidemiology literature and has been confirmed with a recent study, in which comparable effect sizes have been observed by known blood type (109).

In prostate cancer, the signal on chromosome 10q13 points to a variant in the promoter of the *MSMB* gene, which encodes a protein, PSP94, under intense investigation as a biomarker for prostate cancer (65,89). The T allele of rs10993994, 57 bp centromeric to the first exon of the *MSMB* gene, showed significant association with prostate cancer in two independent studies (65,89), and it is known to have influence in the *MSMB* gene expression (prostate secretory protein 94, PSP94) in tumor (110,111). Now that the region has been extensively resequenced, further investigation of additional variants in strong LD with rs10993994 is warranted and it is possible that a neighboring gene, *NCOA4*, could also be a candidate gene for analysis because it is an androgen receptor coactivator.

A GWAS of neuroblastoma, a rare pediatric cancer, has implicated three different chromosomal regions, one of which is a copy number variation at chromosome 1q21.1 (80,112,113). The first region is at 6p22 and it is plausible that the risk alleles have dosage effect on the severity of disease by subgrouping patients into patients of metastatic stage 4, patients with somatic *MYCN* amplification and patients with relapse. The second region is at 2q35 within the *BARD1* gene (112).

Despite the enormous effort focused on choosing candidate genes or pathways, based on current models, so far, the results of cancer GWAS have pointed to primarily new or unknown regions and genes. However, there are a few notable exceptions, such as two GWAS of pediatric lymphoblastic leukemia, which have uncovered three sets of markers pointing to genes involved in B-cell development (114,115), but the clustering of related genes has not been observed. Moreover, for a disease such as breast cancer, which has been epidemiologically linked to hormones, surprisingly, none of the major signals map to regions harboring estrogen/progesterone genes in women of European background. However, in a scan of Asian women, a GWAS convincingly discovered markers near the estrogen receptor alpha (known as *ESR1*) (93).

Discovering more complexity

GWAS have uncovered a series of possible interesting and unexpected relationships between different diseases. For example, three of the regions identified in prostate cancer GWAS also map to type two diabetes susceptibility regions. For some time, there has been a controversial literature reporting an inverse relationship between type two diabetes and prostate cancer; it is further speculated that the protection against prostate cancer is more apparent several years after the diagnosis of diabetes. For two of regions, the markers appear to be inversely related, namely the apparent risk allele for prostate cancer is protective for diabetes for *HNF1B* on chromosome 17q24 and for *THADA* on chromosome 2p21. The signal on chromosome 7p15 localizes to intron 2 of *JAZF1*, a very large gene, whereas the diabetes signal, as well SNPs for height, body stature and systemic lupus erythematosus are localized to a distinct region >200 kb away in intron 1 with no residual LD, suggesting different variants.

Differences in study design can lead to important observations related to both the genetic and environmental contributions to cancer etiology. In one notable instance, two distinct GWAS efforts in prostate cancer have yielded different results for a region of chromosome, 19q13.33, that harbors the gene responsible for the prostate serum antigen (PSA), used by many, but not all for screening for prostate cancer (116,117). In one study, that used clinically advanced cases with controls that had low PSA levels, a strong signal for a SNP in *KLK3* was observed, replicating with a substantially lower degree of statistical significance in the follow-up studies, whereas in Cancer Genetic Markers of Susceptibility Initiative, comprised of mainly cohort studies, there was little effect for prostate cancer risk (39,89,118,119). In fact, the Cancer Genetic Markers of Susceptibility Initiative analysis reported that the SNP in the region of *KLK3* was associated with PSA levels, raising the possibility that the locus could be related to PSA levels instead of prostate carcinogenesis, though it is possible it could be a both but further studies are needed. Indeed, now that the *KLK3* region has been resequenced, it will be possible to investigate this issue with the optimal markers (36).

Most studies have relied on combining data from different designs and often combining histologic or molecular subtypes of a classically defined cancer. The result has been to identify regions that appear to be associated with biological processes common to the development of a tissue-specific type of cancer. For example, the follow-up analysis of the initial set of signals identified in breast cancer GWAS suggests that there could be a differential effect for some regions based on estrogen receptor status for some regions (120). The preponderance of estrogen receptor-positive cases in the discovery studies certainly could have contributed to this observation, but additional reports have identified regions with stronger effects in estrogen receptor-positive subjects (92). In other GWAS, subtype GWAS have yielded convincing findings for a histologic subtype, such as the chromosome 5p15.33 locus in lung cancer (in predominately smokers), which is significantly associated in the adenocarcinoma subtype but not in squamous cell carcinoma (121,122). Similarly, in non-Hodgkin's lymphoma, distinct regions have been identified in the chronic lymphocytic leukemia (114) and follicular subtypes (123). On the other hand, for the associations with high effect sizes in testicular cancer, there was no appreciable difference by subtype analysis for seminoma and nonseminoma cancers, suggesting the common contribution of the two regions to testicular carcinogenesis (101,102,124).

Based on follow-up fine mapping of the regions, often using HapMap chosen SNPs or those defined by comprehensive resequence analysis (36,38,39), intense effort has focused on the investigation of the genomic architecture of each GWAS region. It is plausible that more than one common variant, each with small effect sizes, could contribute to cancer susceptibility and in fact, this has been demonstrated in three regions identified in prostate cancer susceptibility. For 8q24, there are at least four distinct prostate cancer susceptibility loci in men of European background (66,82,84,85,90,125). In men of other backgrounds (e.g. African, East Asian or Latino/admixed), it is possible that even more population-specific loci could be important and perhaps partially explain some of the disease disparity among different ethnic groups (85,90). For the *HNF1B* locus on chromosome 17q24, further mapping identified a second independent signal (126). Similarly, the gene desert of 11q13 harbors at least two independent signals and perhaps more (127).

Cancer GWAS Nexus regions

8q24, a cancer susceptibility region for many unrelated cancers

A region of ~600 kb, centromeric to the well studied, *MYC* oncogene, is a region that has been repeatedly discovered to harbor distinct independent markers associated with cancer risk (Figure 4). *MYC* encodes for nuclear phosphoprotein that involves in growth regulation, cell differentiation and apoptosis, and its amplification/overexpression is a frequent event in bladder tumors (128,129). The findings have unexpectedly found that prostate, breast, colorectal, bladder and perhaps ovarian cancers are associated with common genetic variants in this region (66,75,82,88,90,130–134). The region is also notable because it is frequently amplified in epithelial cancers and does not harbor candidate genes, but instead several pseudogenes, whose function and presence are not well established. In this regard, the findings of 8q24 attest to the complexity of the region and the likelihood that regulatory elements of both *MYC* and other regions could underlie the cancer susceptibility.

The 8q24 region was first implicated as a prostate cancer risk locus by a genome-wide linkage scan in Icelandic men, followed by identification of an allele of the microsatellite marker, DG8S737, and A allele of rs1447295 from replication association studies in three case–control samples of European ancestry from Iceland, Sweden and USA (125). The region was also discovered by an admixture mapping

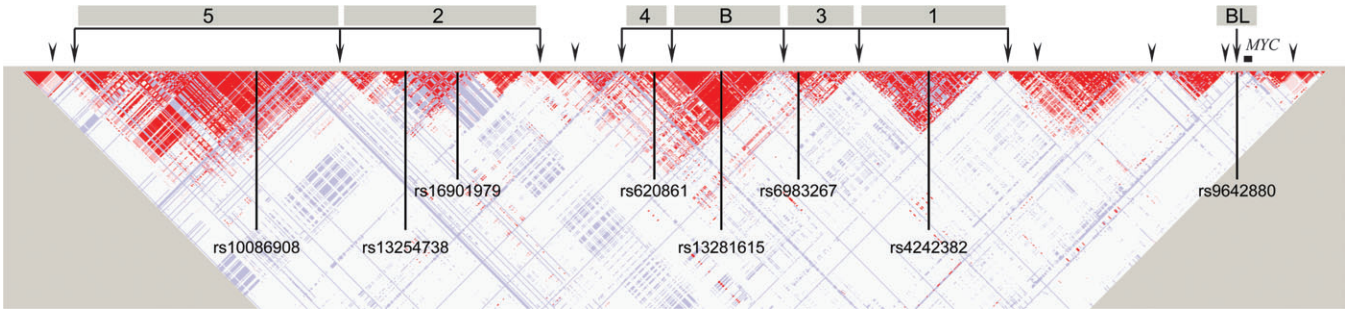


Fig. 4. Linkage disequilibrium pattern and cancer susceptibility loci indentified in 8q24 region. The 8q24 region harbors multiple cancer susceptibility loci identified by GWAS. The linkage disequilibrium heat map was drawn using HapMap I + II release 22 CEU data from 127 948 to 128 950 kb genomic region (reference build 36.3). The arrowheads indicate probable recombination hotspots according to the HapMap I + II. Five distinct regions have been associated with prostate cancer risk (regions 1–5). Region 3 is also conclusively associated with colorectal cancer and precancerous colorectal adenomas. Region B harbors a breast cancer susceptibility locus rs13281615, and BL indicate a bladder cancer susceptibility locus rs9642880, which is telomeric to the region 1, and ~30 kb centromeric to the *MYC* oncogene.

in African-Americans (135). The SNP, rs1447295, was reconfirmed by a large nested case–control study using 6637 cases and 7361 matched controls (91). Independent of the rs1447295, which marked as 'region 1', two independent loci, rs16901979 and rs6983267, marked as region 2 and region 3, respectively, centromeric to the region 1 were identified by three independent studies (66,82,90). Notably, the rs16901979 showed clear association in African-Americans with higher risk allele frequency than Europeans. In two recent studies, another independent prostate cancer susceptibility locus rs620861 was identified, located in between region 2 and region 3 and overlapping with a region previously identified in a breast cancer GWAS (81,84,136).

For colorectal cancer, four different studies reported the same variant, rs6983267 (in region 3 of prostate cancer), as the strongest signal by GWAS (88,90,132,137). Recently, published work has begun to generate insights in the functional nature of the rs6983267 variant, which has only two other variants in strong LD compared with rsw1447295 with 49 variants in strong LD (36,138,139). The two studies suggest that in colorectal cancer, rs6983267 shows long-range interaction with MYC as well as possible enhancement of the Wnt-signaling pathway. Interestingly, the prostate specific effect is more complex and as of now, not well explained except for the presence of multiple regions across the 600 kb of 8q24.

Kiemeney *et al.* (130) reported that the T allele of rs9642880 located ~30 kb upstream of *MYC* oncogene showed significant association with bladder cancer (odds ratio = 1.22, $P = 9.34 \times 10^{-12}$). Wu *et al.* (140) reported that rs2294008 located in exon 1 of *PSCA* on the other side of *MYC* is significantly associated with bladder cancer risk. Since the SNP, rs2294008, is located in the exon 1 of *PSCA* and yields a missense variant that alters the start codon, Wu *et al.* further performed an *in vitro* reporter assay using the four most frequent haplotypes of the *PSCA* 5′ upstream region including rs2294008 and showed significantly lower promoter activity of the T allele-containing haplotypes.

5p15.33

Common variants in the *TERT-CLPTM1L* locus on 5p15.33 have been identified by GWAS to harbor susceptibility alleles for cancer of the brain and lung (96,97,122,141,142). For lung cancer, it appears that the signal is strongly associated with the adenocarcinoma subtype and not squamous or other subtypes (122). In the region, there is an attractive candidate gene, *TERT*, the reverse transcriptase component of the telomerase a gene that is critical for telomere replication and stabilization by controlling telomere length. TERT promotes epithelial proliferation and telomere maintenance has been implicated in the progression from *KRAS*-activated adenoma to adenocarcinoma in a murine model (143,144). There is additional evidence for associations with cancer of the bladder, prostate, uterine cervix and skin including basal cell carcinoma and melanoma based on candidate studies in follow-up of GWAS hits (145).

This region is particularly interesting because of the scope and spectrum of allele frequencies associated with diseases. Mutations in the *TERT* gene have been described in acute myelogenous leukemia and in the inherited bone marrow failure family pedigrees with dyskeratosis congenita, a cancer predisposition syndromes (146,147). Mutations in the *TERT* gene have also been described in patients with idiopathic pulmonary fibrosis (148,149) and in families with hematologic disorders and serious liver fibrosis (150). Mutations in *TERT* have also been shown to result in shorter telomeres and explain a subset of those with familial idiopathic pulmonary fibrosis (151).

Conclusions

The age of genome-wide association studies in cancer have ushered in a new era of discovery of regions of the genome harboring common genetic susceptibility alleles that require extensive effort to map the signal to define the optimal variants for investigating the biological basis of the association. For nearly all signals identified, the markers have not immediately uncovered variants that can easily explain the signal and in most cases, appear to be variants not in coding regions that instead of shifting the amino acid sequence, probably alter the regulation of one or more complex genetic processes. In this regard, GWAS are the first step toward identifying novel regions and pathways associated with both primary carcinogenesis and probably gene–environment interactions.

To make sense of the known GWAS signals and to find more signals, some that could explain major disparities in incidence and outcomes by ethnic backgrounds, it will be critical to conduct GWAS in populations with distinct population genetic histories (and different underlying LD structures) as well as to map known hits in other populations. The age of GWAS has not only uncovered new regions but perhaps provided insights in a subset of the regions that require refined analyses, such as the effect of tobaccos usage and lung cancer risk to unravel the complex nature of these types of cancer.

The recent genomic revolution has produced a comprehensive map of genetic variation that has enabled research to scan the genome in search of statistically sound signals worthy of follow-up. However, the ability to survey environmental and lifestyle exposures is not nearly as advanced, thus hampering the opportunity to explore the dynamic relationship between genomic variants and the environment. Lastly, the age of GWAS is actually the beginning of a new age, one characterized by many new regions of the genome worthy of pursuit as candidate genes to explore the common as well as uncommon variants that contribute to the risk of different cancers.

Acknowledgements

Conflict of Interest Statement: None declared.

References

1. The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426,** 789–796.
2. Collins,F.S. *et al*. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
3. Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Venter,J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
5. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437,** 1299–1320.
6. Sachidanandam,R. *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
7. Chanock,S. (2001) Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis. Markers*, **17**, 89–98.
8. Miki,Y. *et al*. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**, 66–71.
9. NIH/CEPH Collaborative Mapping Group. (1992) A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. *Science*, **258**, 67–86.
10. Elston,R.C. *et al*. (2001) Overview of model-free methods for linkage analysis. *Adv. Genet.*, **42**, 135–150.
11. Malkin,D. *et al*. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, **250**, 1233–1238.
12. Hussussian,C.J. *et al*. (1994) Germline p16 mutations in familial melanoma. *Nat. Genet.*, **8**, 15–21.
13. Kamb,A. *et al*. (1994) Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat. Genet.*, **8**, 23–26.
14. Wooster,R. *et al*. (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**, 789–792.
15. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
16. Garcia-Closas,M. *et al*. (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*, **366**, 649–659.
17. Rothman,N. *et al*. (2006) Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. *Lancet Oncol.*, **7**, 27–38.
18. Cox,A. *et al*. (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.*, **39**, 352–358.
19. Garcia-Closas,M. *et al*. (2007) Large-scale evaluation of candidate genes identifies associations between VEGF polymorphisms and bladder cancer risk. *PLoS Genet.*, **3**, e29.
20. Dunning,A.M. *et al*. (2009) Association of ESR1 gene tagging SNPs with breast cancer risk. *Hum. Mol. Genet.*, **18**, 1131–1139.
21. Risch,N. (2001) The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol. Biomarkers Prev.*, **10**, 733–741.
22. Risch,N. *et al*. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
23. The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447,** 661–678.
24. Barrett,J.C. *et al*. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.
25. O'Berg,M.T. (1980) Epidemiologic study of workers exposed to acrylonitrile. *J. Occup. Med.*, **22**, 245–252.
26. Wolff,M.S. *et al*. (1993) Blood levels of organochlorine residues and risk of breast cancer. *J. Natl. Cancer Inst.*, **85**, 648–652.
27. Erichsen,H.C. *et al*. (2004) SNPs in cancer research and treatment. *Br. J. Cancer*, **90**, 747–751.
28. Frazer,K.A. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
29. Kidd,J.M. *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
30. Levy,S. *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
31. Ng,S.B. *et al*. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
32. Wang,J. *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
33. Wheeler,D.A. *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
34. Kim,J.I. *et al*. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.
35. Harismendy,O. *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32.
36. Yeager,M. *et al*. (2008) Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum. Genet.*, **124**, 161–170.
37. Hinds,D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
38. Yeager,M. *et al*. (2009) Comprehensive resequence analysis of a 97 kb region of chromosome 10q11.2 containing the MSMB gene associated with prostate cancer. *Hum. Genet.*, **126**, 743–750.
39. Parikh,H. *et al*. (2009) A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33. *Hum. Genet.,* in press.
40. Kruglyak,L. *et al*. (2001) Variation is the spice of life. *Nat. Genet.*, **27**, 234–236.
41. Reich,D.E. *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
42. Reich,D.E. *et al*. (2003) Quality and completeness of SNP databases. *Nat. Genet.*, **33**, 457–458.
43. Hayden,E.C. (2008) International genome project launched. *Nature*, **451**, 378–379.
44. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nat. Genet.*, **33**, 439–440.
45. Packer,B.R. *et al*. (2006) SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.*, **34**, D617–D621.
46. Stephens,M. *et al*. (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.*, **38**, 375–381.
47. Marth,G. *et al*. (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl Acad. Sci. USA*, **100**, 376–381.
48. Marth,G.T. *et al*. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.
49. McCarroll,S.A. *et al*. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.
50. Scherer,S.W. *et al*. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
51. Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
52. Bailey,J.A. *et al*. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
53. Bailey,J.A. *et al*. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
54. Buckley,P.G. *et al*. (2005) Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet.*, **21**, 315–317.
55. McCarroll,S.A. *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
56. Bonnen,P.E. *et al*. (2002) Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res.*, **12**, 1846–1853.
57. Sabeti,P.C. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
58. Slatkin,M. (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, **9**, 477–485.
59. Orr,N. *et al*. (2008) Common genetic variation and human disease. *Adv. Genet.*, **62**, 1–32.
60. Carlson,C.S. *et al*. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
61. Cardon,L.R. *et al*. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.
62. Johnson,G.C. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
63. Chanock,S.J. *et al*. (2007) Replicating genotype-phenotype associations. *Nature*, **447**, 655–660.
64. Thomas,G. *et al*. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.*, **41**, 579–584.

65. Thomas,G. *et al*. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, **40**, 310–315.
66. Yeager,M. *et al*. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645–649.
67. Yu,K. *et al*. (2008) Population substructure and control selection in genome-wide association studies. *PLoS One*, **3**, e2551.
68. Patterson,N. *et al*. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
69. Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
70. Ryckman,K. *et al*. (2008) Calculation and use of the Hardy-Weinberg model in association studies. *Curr. Protoc. Hum. Genet.*, **57**, 1.18.1–1.18.11.
71. Falush,D. *et al*. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
72. Devlin,B. *et al*. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
73. Pritchard,J.K. *et al*. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, **65**, 220–228.
74. Feigelson,H.S. *et al*. (2007) Successful genome-wide scan in paired blood and buccal samples. *Cancer Epidemiol. Biomarkers Prev.*, **16**, 1023–1025.
75. Easton,D.F. *et al*. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.
76. Hindorff,L.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
77. Manolio,T.A. *et al*. (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.*, **118**, 1590–1605.
78. Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
79. Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
80. Diskin,S.J. *et al*. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, **459**, 987–991.
81. Gudmundsson,J. *et al*. (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1122–1126.
82. Gudmundsson,J. *et al*. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.*, **39**, 631–637.
83. Gudmundsson,J. *et al*. (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.*, **40**, 281–283.
84. Yeager,M. *et al*. (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **41**, 1055–1057.
85. Eeles,R.A. *et al*. (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.*, **41**, 1116–1121.
86. Houlston,R.S. *et al*. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet*, **40**, 1426–1435.
87. Hunter,D.J. *et al*. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.
88. Zanke,B.W. *et al*. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **39**, 989–994.
89. Eeles,R.A. *et al*. (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.*, **40**, 316–321.
90. Haiman,C.A. *et al*. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.*, **39**, 638–644.
91. Schumacher,F.R. *et al*. (2007) A common 8q24 variant in prostate and breast cancer from a large nested case-control study. *Cancer Res.*, **67**, 2951–2956.
92. Stacey,S.N. *et al*. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **39**, 865–869.
93. Zheng,W. *et al*. (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.*, **41**, 324–328.
94. Chanock,S.J. *et al*. (2008) Genomics: when the smoke clears. *Nature*, **452**, 537–538.
95. Hung,R.J. *et al*. (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, **452**, 633–637.
96. McKay,J.D. *et al*. (2008) Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.*, **40**, 1404–1406.
97. Wang,Y. *et al*. (2008) Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.*, **40**, 1407–1409.
98. Bierut,L.J. *et al*. (2007) Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.*, **16**, 24–35.
99. Caporaso,N. *et al*. (2009) Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One*, **4**, e4653.
100. Easton,D.F. *et al*. (2008) Genome-wide association studies in cancer. *Hum. Mol. Genet.*, **17**, R109–R115.
101. Kanetsky,P.A. *et al*. (2009) Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat. Genet.*, **41**, 811–815.
102. Rapley,E.A. *et al*. (2009) A genome-wide association study of testicular germ cell tumor. *Nat. Genet.*, **41**, 807–810.
103. Skinner,D.G. (1983) *Urological Cancer.* Grune & Stratton, New York.
104. Swerdlow,A.J. *et al*. (1997) Risks of breast and testicular cancers in young adult twins in England and Wales: evidence on prenatal and genetic aetiology. *Lancet*, **350**, 1723–1728.
105. Gudmundsson,J. *et al*. (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat. Genet.*, **41**, 460–464.
106. Jemal,A. *et al*. (2008) Cancer statistics, 2008. *CA Cancer J. Clin.*, **58**, 71–96.
107. Amundadottir,L. *et al*. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.*, **41**, 986–990.
108. Bodmer,W. *et al*. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.
109. Wolpin,B.M. *et al*. (2009) ABO blood group and the risk of pancreatic cancer. *J. Natl Cancer Inst.*, **101**, 424–431.
110. Chang,B.L. *et al*. (2009) Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Hum. Mol. Genet.*, **18**, 1368–1375.
111. Liu,P. *et al*. (2008) Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J. Natl Cancer Inst.*, **100**, 1326–1330.
112. Capasso,M. *et al*. (2009) Common variations in BARD1 influence susceptibility to high-risk neuroblastoma. *Nat. Genet.*, **41**, 718–723.
113. Maris,J.M. *et al*. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.*, **358**, 2585–2593.
114. Papaemmanuil,E. *et al*. (2009) Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat. Genet.*, **41**, 1006–1010.
115. Trevino,L.R. *et al*. (2009) Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat. Genet.*, **41**, 1001–1005.
116. Andriole,G.L. *et al*. (2009) Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, **360**, 1310–1319.
117. Schroder,F.H. *et al*. (2009) Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, **360**, 1320–1328.
118. Ahn,J. *et al*. (2008) Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. *Nat. Genet.*, **40**, 1032–1034; author reply 1035–1036.
119. Eeles,R. *et al*. (2008) Reply to "Variation in KLK genes, prostate-specific antigen and risk of prostate cancer". *Nat. Genet.*, **40**, 1035–1036.
120. Garcia-Closas,M. *et al*. (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.*, **4**, e1000054.
121. Broderick,P. *et al*. (2009) Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res.*, **69**, 6633–6641.
122. Landi,M.T. *et al*. (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.*, **85**, 679–691.
123. Skibola,C.F. *et al*. (2009) Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat. Genet.*, **41**, 873–875.
124. Chanock,S. (2009) High marks for GWAS. *Nat. Genet.*, **41**, 765–766.
125. Amundadottir,L.T. *et al*. (2006) A common variant associated with prostate cancer in European and African populations. *Nat. Genet.*, **38**, 652–658.
126. Sun,J. *et al*. (2008) Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. *Nat. Genet.*, **40**, 1153–1155.
127. Zheng,S.L. *et al*. (2009) Two independent prostate cancer risk-associated loci at 11q13. *Cancer Epidemiol. Biomarkers Prev.*, **18**, 1815–1820.
128. DePinho,R.A. *et al*. (1991) myc family oncogenes in the development of normal and neoplastic cells. *Adv. Cancer Res.*, **57**, 1–46.

129. Mhawech-Fauceglia,P. *et al*. (2006) Genetic alterations in urothelial bladder carcinoma: an updated review. *Cancer*, **106**, 1205–1216.
130. Kiemeney,L.A. *et al*. (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.*, **40**, 1307–1312.
131. Ghoussaini,M. *et al*. (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.*, **100**, 962–966.
132. Tomlinson,I. *et al*. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.*, **39**, 984–988.
133. Tomlinson,I.P. *et al*. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.*, **40**, 623–630.
134. Tenesa,A. *et al*. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.*, **40**, 631–637.
135. Freedman,M.L. *et al*. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA*, **103**, 14068–14073.
136. Al Olama,A.A. *et al*. (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1058–1060.
137. Gruber,S.B. *et al*. (2007) Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol. Ther.*, **6**, 1143–1147.
138. Tuupanen,S. *et al*. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.*, **41**, 885–890.
139. Pomerantz,M.M. *et al*. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
140. Wu,X. *et al*. (2009) Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat. Genet.*, **41**, 991–995.
141. Wrensch,M. *et al*. (2009) Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat. Genet.*, **41**, 905–908.
142. Shete,S. *et al*. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.*, **41**, 899–904.
143. Choi,J. *et al*. (2008) TERT promotes epithelial proliferation through transcriptional control of a Myc- and Wnt-related developmental program. *PLoS Genet.*, **4**, e10.
144. Sweet-Cordero,A. *et al*. (2006) Comparison of gene expression and DNA copy number changes in a murine model of lung cancer. *Genes Chromosomes Cancer*, **45**, 338–348.
145. Rafnar,T. *et al*. (2009) Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat. Genet.*, **41**, 221–227.
146. Calado,R.T. *et al*. (2009) Constitutional hypomorphic telomerase mutations in patients with acute myeloid leukemia. *Proc. Natl Acad. Sci. USA*, **106**, 1187–1192.
147. Yamaguchi,H. *et al*. (2005) Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia. *N. Engl. J. Med.*, **352**, 1413–1424.
148. Tsakiri,K.D. *et al*. (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc. Natl Acad. Sci. USA*, **104**, 7552–7557.
149. Mushiroda,T. *et al*. (2008) A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. *J. Med. Genet.*, **45**, 654–656.
150. Calado,R.T. *et al*. (2009) A spectrum of severe familial liver disorders associate with telomerase mutations. *PLoS ONE*, **4**, e7926.
151. Armanios,M.Y. *et al*. (2007) Telomerase mutations in families with idiopathic pulmonary fibrosis. *N. Engl. J. Med.*, **356**, 1317–1326.

Received October 30, 2009; revised October 30, 2009;
accepted October 30, 2009

4. CONSIDERAÇÕES FINAIS

Neste trabalho abordamos a questão da grande quantidade de informação biológica disponível devido, principalmente, ao avanço do desenvolvimento de novas técnicas de genotipagem em paralelo e sequenciamento em larga escala (NGS). Desenvolvemos uma plataforma bioiinformática para tratar a questão de organização estruturada dessa informação e manipulação de forma eficiente desses dados para sua utilização em análises comumente empregadas em genética de populações. Ao longo dessa dissertação utilizei as ferramentas desenvolvidas para auxiliar no entendimento dos processos evolutivos, que moldam a variabilidade genética, com ênfase na ação da seleção natural, utilizando genes de intersse bioimédico. Finalmente, mostramos como a bioinformática pode auxiliar na organização e análise de dados biológicos através do desenvolvimento de metodologias mais sofisticas e robustas de análise e interpretação auxiliando a organização e exploração mais eficiente dos dados gerados.

5. REFERÊENCIAS

ALTSHULER, D. *et al.* A haplotype map of the human genome. *Nature* [S.I.], v. 437, n. 7063, p. 1299-1320, Oct 27 2005.

_____. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* [S.I.], v. 407, n. 6803, p. 513-6, Sep 28 2000.

ANDRES, A. M. *et al.* Targets of Balancing Selection in the Human Genome. *Molecular Biology and Evolution* [S.I.], v. 26, n. 12, p. 2755-2764, Dec 2009.

BALARESQUE, P. L. *et al.* Challenges in human genetic diversity: demographic history and adaptation. *Human Molecular Genetics* [S.I.], v. 16, p. R134-R139, Oct 15 2007.

BAMSHAD, M. J. *et al.* Human population genetic structure and inference of group membership. *Am J Hum Genet* [S.I.], v. 72, n. 3, p. 578-89, Mar 2003.

BIRD, C. P. *et al.* Fast-evolving noncoding sequences in the human genome. *Genome Biol* [S.I.], v. 8, n. 6, p. R118, 2007.

BRINKMANN, B. *et al.* Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* [S.I.], v. 62, n. 6, p. 1408-1415, Jun 1998.

CAMPBELL, M. C.; TISHKOFF, S. A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* [S.I.], v. 9, p. 403-33, 2008.

CARLSON, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* [S.I.], v. 74, n. 1, p. 106-120, Jan 2004.

CHANOCK, S. J. *et al.* The Respiratory Burst Oxidase. *Journal of Biological Chemistry* [S.I.], v. 269, n. 40, p. 24519-24522, Oct 7 1994.

CHARLESWORTH, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* [S.I.], v. 10, n. 3, p. 195-205, Mar 2009.

CROW, J. F. Anecdotal, historical and critical commentaries on genetics twenty-five years ago in genetics: motoo kimura and molecular evolution. *Genetics* [S.I.], v. 116, n. 2, p. 183-4, Jun 1987.

EWENS, W. J. The sampling theory of selectively neutral alleles. *Theor Popul Biol* [S.I.], v. 3, n. 1, p. 87-112, Mar 1972.

EWENS, W. J.; FELDMAN, M. W. Analysis of neutrality in protein polymorphism. *Science* [S.I.], v. 183, n. 123, p. 446-8, Feb 1 1974.

EXCOFFIER, L. Human demographic history: refining the recent African origin model. *Current Opinion in Genetics & Development* [S.I.], v. 12, n. 6, p. 675-682, Dec 2002.

EXCOFFIER, L.; RAY, N. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution* [S.I.], v. 23, n. 7, p. 347-351, Jul 2008.

FAY, J. C.; WU, C. I. Hitchhiking under positive Darwinian selection. *Genetics* [S.I.], v. 155, n. 3, p. 1405-1413, Jul 2000.

_____. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet* [S.I.], v. 4, p. 213-35, 2003.

FEARNHEAD, P. SequenceLDhot: detecting recombination hotspots. *Bioinformatics* [S.I.], v. 22, n. 24, p. 3061-3066, Dec 15 2006.

FRAZER, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* [S.I.], v. 449, n. 7164, p. 851-U3, Oct 18 2007.

FRYXELL, K. J.; MOON, W. J. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution* [S.I.], v. 22, n. 3, p. 650-658, Mar 2005.

FU, Y. X.; LI, W. H. Statistical Tests of Neutrality of Mutations. *Genetics* [S.I.], v. 133, n. 3, p. 693-709, Mar 1993.

GOLDSTEIN, D. B.; CHIKHI, L. Human migrations and population structure: What we know and why it matters. *Annual Review of Genomics and Human Genetics* [S.I.], v. 3, p. 129-152, 2002.

GRANTHAM, R. Amino-Acid Difference Formula to Help Explain Protein Evolution. *Science* [S.I.], v. 185, n. 4154, p. 862-864, 1974.

HARRIS, E. E.; MEYER, D. The molecular signature of selection underlying human adaptations. *Yearbook of Physical Anthropology, Vol. 49 2006* [S.I.], v. 49, p. 89-130, 2006.

HEIN, J. *et al. Gene Genealogies, Variation and Evolution - A PRIMER IN COALESCENT THEORY*. First Edition. ed. Oxford: OXFORD - University Press, 2005.

HELLMANN, I. *et al.* A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* [S.I.], v. 72, n. 6, p. 1527-35, Jun 2003.

_____. Why do human diversity levels vary at a megabase scale? *Genome Research* [S.I.], v. 15, n. 9, p. 1222-1231, Sep 2005.

HEYWORTH, P. G. *et al.* Chronic granulomatous disease. *Current Opinion in Immunology* [S.I.], v. 15, n. 5, p. 578-584, Oct 2003.

HINDS, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* [S.I.], v. 307, n. 5712, p. 1072-9, Feb 18 2005.

HUDSON, R. R. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* [S.I.], v. 23, n. 2, p. 183-201, Apr 1983.

_____. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* [S.I.], v. 18, n. 2, p. 337-338, Feb 2002.

HUDSON, R. R. *et al.* A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* [S.I.], v. 116, n. 1, p. 153-159, May 1987.

HURST, L. D. FUNDAMENTAL CONCEPTS IN GENETICS Genetics and the understanding of selection. *Nature Reviews Genetics* [S.I.], v. 10, n. 2, p. 83-93, Feb 2009.

KIMMEL, M. *et al.* Signatures of population expansion in microsatellite repeat data. *Genetics* [S.I.], v. 148, n. 4, p. 1921-1930, Apr 1998.

KIMURA, M. Genetic Variability Maintained in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Isoalleles. *Genetical Research* [S.I.], v. 11, n. 3, p. 247-&, 1968.

_____. Neutral Theory as a Supplement to Darwinism. *Trends in Biochemical Sciences* [S.I.], v. 1, n. 11, p. N248-N249, 1976.

_____. Neutral Theory of Molecular Evolution and Polymorphism. *Scientia* [S.I.], v. 112, n. 9-12, p. 687-721, 1977a.

_____. Preponderance of Synonymous Changes as Evidence for Neutral Theory of Molecular Evolution. *Nature* [S.I.], v. 267, n. 5608, p. 275-276, 1977b.

KINGMAN, J. F. Origins of the coalescent. 1974-1982. *Genetics* [S.I.], v. 156, n. 4, p. 1461-3, Dec 2000.

KREITMAN, M.; DI RIENZO, A. Balancing claims for balancing selection. *Trends Genet* [S.I.], v. 20, n. 7, p. 300-4, Jul 2004.

MCDONALD, J. H.; KREITMAN, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* [S.I.], v. 351, n. 6328, p. 652-4, Jun 20 1991.

MEAD, S. *et al.* Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science* [S.I.], v. 300, n. 5619, p. 640-643, Apr 25 2003.

MYERS, S. *et al.* A fine-scale map of recombination rates and hotspots across the human genome. *Science* [S.I.], v. 310, n. 5746, p. 321-324, Oct 14 2005.

MYLES, S. *et al.* Identification and analysis of genomic regions with large between-population differentiation in humans. *Ann Hum Genet* [S.I.], v. 72, n. Pt 1, p. 99-110, Jan 2008.

NACHMAN, M. W.; CROWELL, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* [S.I.], v. 156, n. 1, p. 297-304, Sep 2000.

NIELSEN, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* [S.I.], v. 154, n. 2, p. 931-42, Feb 2000.

_____. Molecular signatures of natural selection. *Annual Review of Genetics* [S.I.], v. 39, p. 197-218, 2005.

NIELSEN, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* [S.I.], v. 3, n. 6, p. e170, Jun 2005.

_____. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* [S.I.], v. 168, n. 4, p. 2373-82, Dec 2004.

_____. Darwinian and demographic forces affecting human protein coding genes. *Genome Research* [S.I.], v. 19, n. 5, p. 838-849, May 2009.

NIELSEN, R.; SIGNOROVITCH, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor Popul Biol* [S.I.], v. 63, n. 3, p. 245-55, May 2003.

NOVEMBRE, J.; DI RIENZO, A. Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet* [S.I.], v. 10, n. 11, p. 745-55, Nov 2009.

PICKRELL, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* [S.I.], v. 19, n. 5, p. 826-37, May 2009.

PICOULT-NEWBERG, L. *et al.* Mining SNPs from EST databases. *Genome Res* [S.I.], v. 9, n. 2, p. 167-74, Feb 1999.

PLUZHNIKOV, A. *et al.* Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* [S.I.], v. 161, n. 3, p. 1209-18, Jul 2002.

ROACH, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* [S.I.], v. 328, n. 5978, p. 636-9, Apr 30 2010.

ROSENBERG, N. A. *et al.* Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* [S.I.], v. 73, n. 6, p. 1402-1422, Dec 2003.

SABETI, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* [S.I.], v. 419, n. 6909, p. 832-837, Oct 24 2002.

_____. Positive natural selection in the human lineage. *Science* [S.I.], v. 312, n. 5780, p. 1614-1620, Jun 16 2006.

_____. The case for selection at CCR5-Delta 32. *Plos Biology* [S.I.], v. 3, n. 11, p. 1963-1969, Nov 2005.

SACHIDANANDAM, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* [S.I.], v. 409, n. 6822, p. 928-933, Feb 15 2001.

SAVAGE, S. A. *et al.* Genetic variation, nucleotide diversity, and linkage disequilibrium in seven telomere stability genes suggest that these genes may be under constraint. *Human Mutation* [S.I.], v. 26, n. 4, p. 343-350, Oct 2005.

SLATKIN, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* [S.I.], v. 9, n. 6, p. 477-85, Jun 2008.

SOLDEVILA, M. *et al.* Assessing the signatures of selection in PRNP from polymorphism data: results support Kreitman and Di Rienzo's opinion. *Trends Genet* [S.I.], v. 21, n. 7, p. 389-91, Jul 2005.

SPENCER, C. C. A. *et al.* The influence of recombination on human genetic diversity. *Plos Genetics* [S.I.], v. 2, n. 9, p. 1375-1385, Sep 2006.

STUMPF, M. P.; MCVEAN, G. A. Estimating recombination rates from population-genetic data. *Nat Rev Genet* [S.I.], v. 4, n. 12, p. 959-68, Dec 2003.

TAJIMA, F. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* [S.I.], v. 123, n. 3, p. 585-595, Nov 1989.

VOIGHT, B. F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America* [S.I.], v. 102, n. 51, p. 18508-18513, Dec 20 2005.

_____. A map of recent positive selection in the human genome. *PLoS Biol* [S.I.], v. 4, n. 3, p. e72, Mar 2006.

WAKELEY, J. *et al.* The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. *Am J Hum Genet* [S.I.], v. 69, n. 6, p. 1332-47, Dec 2001.

WANG, Y.; RANNALA, B. Population genomic inference of recombination rates and hotspots. *Proc Natl Acad Sci U S A* [S.I.], v. 106, n. 15, p. 6215-9, Apr 14 2009.

WATTERSON, G. A. An analysis of multi-allelic data. *Genetics* [S.I.], v. 88, n. 1, p. 171-9, Jan 1978a.

_____. The homozygosity test of neutrality. *Genetics* [S.I.], v. 88, n. 2, p. 405-17, Feb 1978b.

WRIGHT, S. I.; CHARLESWORTH, B. The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics* [S.I.], v. 168, n. 2, p. 1071-1076, Oct 2004.

YANG, Z. G. Adaptive Molecular Evolution. In: BALDING DJ, B. M. A. C. C. (Ed.). *Handbook of Statistical Genetics* Susex, UK: John Wiley & Sons, Ltd, 2007. p. 377-406.

YANG, Z. H. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* [S.I.], v. 24, n. 8, p. 1586-1591, Aug 2007.