Giordano Bruno Soares Souza

NOVAS ABORDAGENS PARA INTEGRAÇÃO DE BANCOS DE DADOS E DESENVOLVIMENTO DE FERRAMENTAS BIOINFORMÁTICAS PARA ESTUDOS DE GENÉTICA DE POPULAÇÕES

Tese apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do Grau de Doutor.

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos

Co-Orientadora: Dra. Maíra Ribeiro Rodrigues

Belo Horizonte

2014

043 Souza, Giordano Bruno Soares.
Novas abordagens para integração de bancos de dados e desenvolvimento de ferramentas bioinformáticas para estudos de genética de populações [manuscrito] / Giordano Bruno Soares Souza. – 2014.
213 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Eduardo Martín Tarazona Santos. Coorientadora: Dra. Maíra Ribeiro Rodrigues.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Biologia Computacional. 2. Grupo com Ancestrais Nativos do Continente Americano. 3. Polimorfismo de Nucleotídeo Único. 4. Características da População. I. Santos, Eduardo Martín Tarazona. II. Rodrigues, Maíra Ribeiro. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Rosilene Moreira Coelho de Sá – CRB 6 - 2726

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

43/2014
entrada
1º/2010
CPF:
065.989.496-37

ATA DA DEFESA DE TESE

Giordano Bruno Soares Souza

Às nove horas do dia **02 de abril de 2014**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: **"NOVAS ABORDAGENS PARA INTEGRAÇÃO DE BANCOS DE DADOS E DESENVOLVIMENTO DE FERRAMENTAS BIOINFORMÁTICAS PARA ESTUDOS DE GENÉTICA DE POPULAÇÕES"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Eduardo Martin Tarazona Santos**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	Indicação
Dr. Eduardo Martin Tarazona Santos	UFMG	aprovado
Dra. Maíra Ribeiro Rodrigues	UFMG	aprovado
Dr. Pedro Olmo Stancioli Vaz de Melo	UFMG	aprovado
Dr. Renato Martins Assunção	UFMG	aprovado
Dr. Emmanuel Dias Neto	Hospital A.C Camargo	Aprovado
Dr. Nelson Jurandi Rosa Fagundes	UFRGS	APROVADO

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

Belo Horizonte, 02 de abril de 2014.

Dr. Eduardo Martin Tarazona Santos - Orientador _____

Dra. Maíra Ribeiro Rodrigues - Co-Orientadora _____

Dr. Pedro Olmo Stancioli Vaz de Melo _____

Dr. Renato Martins Assunção _____

Dr. Emmanuel Dias Neto _____

Dr. Nelson Jurandi Rosa Fagundes _____

"Temos todos duas vidas:
A verdadeira, que é a que sonhámos na infância,
E que continuamos sonhando, adultos, num substrato de névoa;
A falsa, que é a que vivemos em convivência com os outros,
Que é a prática, a útil,
Aquela em que acabam por nos meter num caixão.

Na outra não há caixões, nem mortes.
Há só ilustrações de infância:
Grandes livros coloridos, para ver mas não ler;
Grandes páginas de cores para recordar mais tarde.
Na outra somos nós,
Na outra vivemos."
Fernando Pessoa

Este trabalho é dedicado à pessoa que me permitiu por tanto tempo desfrutar da outra vida, agradeço profundamente pelas cores, livros, cadernos, pelas coisas que me ensinou a amar e a respeitar, pelo incentivo e liberdade. Ao meu pai, Élcio de Souza (*in memorian*).

AGRADECIMENTOS

Agradecer, por vezes, é uma tarefa um tanto ingrata, seja por não ser possível reconhecer todos os atos que me ajudaram a percorrer este caminho, seja pelos lapsos de memória. Por vezes, concluo que todo agradecimento, em primeiro lugar, deve ser universal, devido a feliz coincidência de ações que nos permitem alcançar o resultado almejado. Desta forma, o meu agradecimento é sincero – e não uma precaução a prováveis esquecimentos – a todos que contribuíram neste processo, direta ou indiretamente. Não posso me furtar, no entanto, de reconhecer e agradecer a participação especial de algumas pessoas que participaram intensamente deste processo.

Aos meus pais, Marisa e Élcio (*in memorian*) pelo suporte, incentivo, amor e paciência, este último, um agradecimento especial ao meu pai, que durante toda nossa convivência jamais se furtou a responder às minhas perguntas e a debater os mais diferentes assuntos. Este reconhecimento se faz tão necessário, pois a enorme disposição para responder as minhas questões (e diariamente eram inúmeras), fez com que a minha curiosidade se mantivesse viva até hoje. E, hoje, neste caminho, vejo o quão essencial é este instinto. Manifesto minha gratidão, também, à minha irmã pelo companheirismo e pelos agradáveis momentos de bom-humor, aos meus avós, pelo carinho, confidência, boas estórias e brincadeiras. Agradeço, ainda, aos meus tios e primos pelo conforto nos momentos mais difíceis. Faz-se necessário agradecer aos amigos da família que se esforçaram para diminuir o sofrimento do meu pai durante a enfermidade, cada visita trazia consigo alento para um ambiente tão inóspito e monótono. E também àqueles funcionários do hospital que honraram o compromisso ético de garantir a saúde e o bem-estar do meu pai.

À Camila, minha companheira em toda esta trajetória, pelo amor, apoio, conforto, disponibilidade e sensibilidade; com quem compartilho os melhores momentos e que me ajuda a resistir aos menos favoráveis. Aos meus amigos pelos momentos de descontração, alento, confiança e camaradagem.

Ao meu orientador, o Prof. Dr. Eduardo Tarazona, pelas diversas oportunidades a mim concedidas, pela visão científica apurada, pelo conhecimento compartilhado, pela dedicação e empenho no âmbito da pesquisa.

À minha co-orientadora, Dra. Maíra Rodrigues, por todas as dicas, atalhos e conhecimento no âmbito computacional, à oportunidade concedida para aprender sobre sistemas multiagentes e, porque não, orientação a objeto. E ainda pela paciência e ternura no relacionamento com os demais integrantes do laboratório.

Aos meus colegas de laboratório, atuais ou passados, pela colaboração, companheirismo e ambiente agradável para se trabalhar. Dentre os membros atuais, minha estima por Wagner Magalhães, Fernanda Kehdy, Camila Zolini, Moara Machado, Marília Scliar, Natália Araújo, Mateus Gouveia, Fernanda Soares, Gilderlânio Araújo, Tiago Peixoto, Roxana Zamudio, Laélia Pinto e Renan Moreira. Dentre os ex-membros, agradeço especialmente àqueles que trabalharam diretamente nos mesmos projetos: Juliana Chevitarese, Latife Pereira, Andrea Marrero, Guilherme Kingma e Maria Clara Fernandes, minha co-orientadora de iniciação científica. E aos demais precedentes: Luciana Werneck, Hanaísa Sant'Anna, Thaís Muniz, Donnys Silva, Márcia Iannini, Lívia Lemos, Allan Sene, Bruno Araújo, Fernanda Lyon e Rodrigo Redondo.

Agradeço institucionalmente ao NCI-NIH, na figura do Dr. Stephen Shanock, pelos dados disponibilizados para análises e ao Hemominas, na figura de Marina Lobato e Maria Clara Fernandes, pela oportunidade de trabalhar no projeto de Ancestralidade Genômica de indivíduos com Anemia Falciforme. Agradeço também o corpo docente e discente da bioinformática, e a CAPES pelo apoio financeiro.

"Contudo, a credulidade, a aversão à dúvida, a temeridade no responder, o vangloriar-se com o saber, a timidez no contradizer, o agir por interesse, a preguiça nas investigações pessoais, o fetichismo verbal, o deter-se em conhecimentos parciais: isto e coisas semelhantes impediram um casamento feliz do entendimento humano com a natureza das coisas e o acasalaram, em vez disso, a conceitos vãos e experimentos erráticos; o fruto e a posteridade de tão gloriosa união pode-se facilmente imaginar." Adorno e Horkheimer

RESUMO

A variabilidade genética está associada à diferenciação fenotípica encontrada entre as populações humanas. Ainda que a maior parte da diversidade humana se dê dentro das populações humanas e a interação com o ambiente seja fundamental na determinação do fenótipo, a identificação de variantes genéticas muito diferenciadas entre populações humanas é essencial em estudos de evolução e biomédicos. Entretanto, as ferramentas analíticas e as fontes de dados são bastante heterogêneas, e faz-se necessário desenvolver ferramentas bioinformáticas acessíveis que permitam aos pesquisadores lidar com a grande quantidade e diversidade de dados e análises. O presente trabalho descreve ferramentas bioinformáticas desenvolvidas no Laboratório de Diversidade Genética Humana (LDGH) da UFMG com a participação do candidato a Doutor: estas incluem pipelines de análises e a plataforma bioinformática DivergenomeTools, que permite a conversão de arquivos através de pipelines flexíveis. Estas ferramentas foram aplicadas em três artigos sobre a diversidade genética de populações nativas e miscigenadas da América Latina, e na identificação de 69891 variantes genéticas (SNPs) altamente diferenciadas em populações Nativo-Americanas em relações às populações da Europa, África Ocidental e Leste Asiático. Estas variantes foram anotadas utilizando a ferramenta de integração de bancos de dados MASSA (Multi-Agent System for Snp Annotation), desenvolvida pelo candidato. Baseada na tecnologia de sistemas multiagente, MASSA permite a execução de tarefas de forma paralela e cooperativa, contornando o problema da distribuição, tamanho e heterogeneidade dos dados biológicos. A versão atual de MASSA integra informações dos bancos de dados dbSNP (repositório de SNPs), UCSC (genômica), GO (ontologias), HGNC (repositório de nomenclaturas gênicas), OMIM (fenótipos), PGKB (farmacogenética), Reactome (vias metabólicas) e PolyPhen/SIFT/Provean (impacto funcional das substituições nucleotídicas). Através da análise de enriquecimento realizada por MASSA, identificaram-se termos sobrerrepresentados na anotação dos genes diferenciados em Nativo-Americanos. Estas análises confirmaram conhecimentos prévios sobre a estrutura genética destas populações, tais como: alta diversidade nos polimorfismos associados à diabetes tipo 2 e indícios de seleção positiva em genes implicados na resposta imune e na atividade do sistema nervoso. Além disso, permitiram identificar novas especificidades, como, por exemplo: grupos de genes diferenciados envolvidos na produção de receptores de membrana ionotrópicos associados à eficácia do tratamento e severidade de fenótipos cognitivos; e nucleoporinas associadas à susceptibilidade viral e transporte de carboidratos. Em perspectiva, estamos aprimorando as

análises de enriquecimento, aplicando enfoques estatísticos e computacionais que considerem as especificidades dos dados de diversidade genômica e a estrutura do conhecimento biológico. Estas análises permitem aprimorar nossos conhecimentos sobre a biologia dos Nativos Americanos, um grupo étnico negligenciado nas iniciativas para o estudo da diversidade genômica humana e um dos focos das pesquisas do LDGH.

Palavras-chave: Nativo-americanos, Ferramentas bioinformática, Anotação, Enriquecimento, SNPs, Sistema Multiagente, Estruturação Populacional

ABSTRACT

Genetic diversity is associated with the phenotypic differentiation among human populations. Even if most human diversity occurs within populations and the interaction with the environment is crucial to determine the phenotype, the study of variants that differentiated between human populations is essential in evolutionary studies and biomedical research. However, the analytical tools and data are heterogeneous in biology, reinforcing the need of new approaches to pipeline construction and tools that allow researchers to deal with the large amount and diversity of data and analyzes. In this work, we describe bioinformatics tools developed in the Laboratory of Human Genetic Diversity of the UFMG with the involvement of the PhD candidate: these tools include analyses pipelines and the bioinformatics platform DivergenomeTools that allows file conversions through a flexible pipeline. These tools have been applied in three articles on genetic diversity of native and admixture populations of Latin America, and for the identification of 69891 polymorphisms (SNPs) highly differentiated in Native American populations in respect to the West Africa, Europe and East Asia populations. These genetic variants were annotated using the database integration tool MASSA (Multi-Agent System for SNP Annotation), developed by the PhD candidate. Based in the multi-agent technology, MASSA allows parallel and cooperative execution of tasks, bypassing the problem of the distribution, size and heterogeneity of the biological data. MASSA current version integrates information from these databases: dbSNP (SNP repository), UCSC (genomics), GO (ontologies), HGNC (gene names repository), OMIM (phenotypes), PGKB (pharmacogenetics), Reactome (metabolic pathways) and PolyPhen/SIFT/Provean (functional impact of nucleotide substitutions). Through the enrichment analysis performed by MASSA, we identified enriched terms in the annotation of differentiated genes in Native American. These analyzes confirmed previous knowledge on the genetic structure of natives, such as: high diversity in polymorphisms associated with type 2 diabetes and evidence of positive selection in genes involved in immune response and nervous system activity. We also identified new insights such as genes involved in the production of ionotropic membrane receptors related to the efficacy of treatment and severity of cognitive phenotypes; and nucleoporines associated with viral infection and carbohydrate transport. In perspective, we are improving the enrichment analysis, by implementing approaches that take into account the specificities of genomic diversity data and biological knowledge structure. These analyses allow us to enhance our knowledge of the biology of Native Americans, an ethnic group

neglected in the initiatives for study of human genomic diversity, which is one of the focuses of the research of the LDGH.

Keywords: Native-Americans, Bioinformatic tools, Annotation, Enrichment Analysis, Multiagent System, SNPs, Population Structure

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1000Genomes (1kG)	*The 1000 Genomes Project*
ACP (PCA)	Análise de Componentes Principais
AMOVA	Análise de Variância Molecular
ANOVA	Análise de Variância
API	*Application Program Interface*
ASW	Afro americanos do Sudeste dos Estados Unidos da América
BD (DB)	Banco de Dados
CEPH	*Centre d'Etude du Polymorphisme Humain*
CEU	Eurodescentes da região de Utah
CHB	Chineses Han de Beijing
CNV (CNP)	*Copy Number Variation*
CpG	(*Cytosine-phosphate-Guanine*) Ilhas de Citosinas seguidas por Guaninas
DL (LD)	Desequilíbrio de Ligação
DLM	Desequilíbrio de Ligação gerado por Miscigenação
DNA	Ácido desoxirribonucleico
EAS	Leste Asiático
EHW (HWE)	Equilíbrio de Hardy-Weinberg
EM	*Expected Maximization*
EUR	Europa
FIPA	*Foundation for Intelligent Physical Agents*
FLT	*Flat File database*
GC	*Genomic Control*
GLU	*Genotypes and Library Utilities*
GO	*Gene Ontology*
GSEA	*Gene Set Enrichment Analysis*
GWAS	*Genome Wide Association Study*
HapMap	*The International HapMap Project (Haplotype Map)*
HGDP	*Human Genome Diversity Project*
HGNC	*HUGO Gene Nomenclature Committee*
HMM	*Hidden Markov Model*
HTML	*HyperText Markup Language*
HUGO	*Human Genome Organisation*

IBS	*Identity by State*
InDel	Inserções/Deleções
I/O	Input/Output
JADE	*Java Agent DEvelopment Framework*
JDBC	*The Java Database Connectivity*
JSC	*Java Statistical Classes*
JSON	*JavaScript Object Notation*
JPT	Japoneses de Tóquio
LDGH	Laboratório de Diversidade Genética Humana
LWK	Luhya de Webuye, Quênia
MASSA	*Multi-Agent System for Snp Annotation*
MEA	*Modular Enrichment Analysis*
MIA (AIM)	Marcador Informativo de Ancestralidade
MDS	Escalonamento Multidimensional
MKK	Maasai de Kinyawa, Quênia
NAT	Nativo-Americanos
NCI	*National Cancer Institute*
NGS	*Next-Generation Sequencing*
NHGRI	*National Human Genome Research Institute*
NIH	*National Institute of Heath*
OMIM	*Online Mendelian Inheritance in Man*
OPA	*Oligonucleotide Pool Assay (Illumina GoldenGate)*
OPA-N	Conjunto de dados HGDP e NAT genotipados com OPA *SNPCancer*
OPA-II	Conjunto de dados HGDP genotipados com OPA *Innate Immunity*
OPA-NHL	Conjunto de dados HGDP genotipados com OPA *Non-Hodgkin Linfoma*
OWL	*Web Ontology Language*
PCR	*Polymerase Chain-Reaction*
PCR	*Principal Component Regression*
PharmGKB (PGKB)	*The Pharmacogenetics Knowledge Base*
RAO	*Recent African Origin*
RNA	Ácido Ribonucleico
RNAi	Ácido Ribonucleico de Interferência
RNAm	Ácido Ribonucleico Mensageiro
RNAmi (RNAmicro)	Micro Ácido Ribonucleico

RNAnc	Ácido Ribonucleico Não Codificante
SA	*Structured Association*
SDAT	*Sample Datafile* (Matriz de genótipos – Indivíduos x Loci)
SEA	*Singular Enrichment Analysis*
SGBD	Sistema de Gerenciamento de Bancos de Dados
SIFT	*Sorting Intolerant From Tolerant*
SQL	*Structured Query Language*
SMA (MAS)	Sistema Multiagente
SNP	*Single Nucleotide Polymorphism*
STR	*Short Tandem Repeat*
TSI	Toscanos da Itália
TSV	*Tab-Separated Values*
UCSC	*University of California, Santa Cruz*
UTR	*Untranslated Region*
WAFR	África Ocidental
WS	*WebService*
XML	*Extensible Markup Language*
YRI	Iorubas de Ibadan, Nigéria

LISTA DE SÍMBOLOS

Φ	Estimativa baseada em distâncias genéticas e frequências alélicas, calculada a partir da análise de variância e análoga às estatísticas-F.
K	Número de dimensões ou clusters
α	Nível de significância
F_{ST}	Estimativa de divergência genética entre populações
F_{IS}	Estimativa de diversidade genética intrapopulacional
F_{CT}	Estimativa de divergência genética entre grupos populacionais
F_{SC}	Estimativa de divergência genética entre populações dentro do grupo populacional
Θ	Estimativa baseada em frequências alélicas, calculada a partir da análise de variância e análoga às estatísticas-F.
σ^2	Variância
H_E	Heterozigosidade Esperada
f	Frequência
r^2	Índice de correlação entre alelos
pB	Pares de Bases
kB	kiloBases (1000 pares de bases)
pR	Taxa de paralelização
δ	Distância entre as frequências alélicas

SUMÁRIO

INTRODUÇÃO

Um dos principais desafios da genética de populações humanas reside na interpretação da diversidade genética em termos demográficos e, concomitantemente, no entendimento das bases genéticas da adaptação fenotípica (BALARESQUE; BALLEREAU; JOBLING, 2007). Assim, a identificação de variantes genéticas com grandes diferenças na distribuição de frequências alélicas entre grupos populacionais é um dos principais propósitos da pesquisa genética (LEWONTIN, 1972).

Diferentes tipos de polimorfismos têm sido utilizados para descrever a diversidade genética das populações humanas, dentre eles podemos citar: grupos sanguíneos (LEWONTIN, 1972); microssatélites, ou STRs (*Short Tandem Repeats*) (ROSENBERG et al., 2002); pequenas inserções e deleções, InDels (BASTOS-RODRIGUES; PIMENTA; PENA, 2006); deleções ou duplicações genômicas, CNVs (*Copy Number Variation*) (JAKOBSSON et al., 2008); e polimorfismos de uma única base nucleotídica, SNPs (*Single Nucleotide Polymorphism*) (LI et al., 2008). Os SNPs são a maior e mais comum fonte de variação genética, sendo comumente utilizados não só na caracterização da história demográfica das populações, mas também no esclarecimento das bases genéticas das manifestações fenotípicas (BROMBERG; CAPRIOTTI, 2012).

A variabilidade genética dentro ou entre populações pode ser estimada de diferentes maneiras, entretanto, o ponto nevrálgico destas análises é o cálculo das diferenças alélicas entre os indivíduos. Desta forma, o paradigma da genética de populações assenta-se sobre as estimativas de diversidade alélica e o estudo do comportamento destas diferenças ao longo do tempo.

Uma das medidas de variabilidade genética em uma população é a Heterozigosidade Esperada, que é igual à probabilidade de não identidade entre dois alelos escolhidos aleatoriamente. Essa medida apresenta duas componentes: a variância intralocus relacionadas às frequências alélicas e ao tamanho amostral; e a variância interlocus associada a eventos evolutivos tais como mutação, seleção natural e deriva genética (NEI; ROYCHOUDHURY, 1974). O viés de averiguação, ou seja, o desvio sistemático das estatísticas genético-populacionais em relação ao esperado teórico é causado, por exemplo, pela maior

probabilidade de se amostrar polimorfismos comuns durante o processo de identificação de SNPs em uma amostragem restrita de indivíduos sendo um fator de imprecisão relevante nas estimativas de heterozigosidade. Além disso, outra implicação do viés é que as frequências dessas variantes podem estar superestimadas ou subestimadas em relação a outras populações (CLARK et al., 2005; LACHANCE; TISHKOFF, 2013; NIELSEN, 2004; ROGERS; JORDE, 1996).

O Princípio de Hardy-Weinberg prediz que uma população comportando-se de maneira mendeliana, mantém as frequências alélicas inalteradas com o decorrer do tempo. Desta forma, dada as frequências de dois alelos p e q, a proporção das combinações destes alelos na população é estimada por $p^2 : 2pq : q^2$, onde cada combinação de alelos é conhecida como genótipo, e as proporções, como frequências genotípicas. Entretanto, vários fatores são capazes de afastar as frequências alélicas do postulado, tais como deriva genética, efeito Wahlund, casamentos preferenciais, seleção natural e endocruzamentos, neste último, o desvio se dá nas frequências genotípicas (LI; GRAUBARD, 2009). Particularmente, os quatro últimos fatores podem ser quantificados pelas estatísticas F (ROBERTSON; HILL, 1984). Além disso, tais desvios podem ocorrer por dificuldades em genotipar heterozigotos, mutações em sítios de ligação de *primers* de PCR e devido a polimorfismos comuns do tipo deleção. Dessa maneira, o E-HW é usualmente utilizado como medida de controle de qualidade (BALDING, 2006).

A deriva genética é o resultado das flutuações estocásticas sofridas pelas frequências alélicas ao longo das gerações. Tal efeito decorre da amostragem aleatória dos alelos na formação das novas gerações durante o processo de reprodução e é fortemente influenciado pelo tamanho efetivo populacional (MASEL, 2011).

O efeito Wahlund ocorre quando há subpopulações crípticas com frequências alélicas distintas em uma dada população. Desta forma, há um déficit no número de heterozigotos (ou excesso de homozigotos) observados na população em relação ao que seria esperado de acordo com o Equilíbrio de Hardy-Weinberg.

A panmixia, ou o acasalamento aleatório entre os indivíduos de uma população, é um dos pressupostos do Equilíbrio de Hardy-Weinberg para que as frequências alélicas se mantenham constantes ao longo das gerações. Quando a escolha dos parceiros reprodutivos

não é aleatória, diz-se que os acasalamentos são preferenciais, um caso especial é o endocruzamento, ou seja, o acasalamento entre indivíduos geneticamente mais próximos que a média populacional.

O desequilíbrio de ligação consiste na correlação não aleatória entre os alelos de diferentes variantes, desta forma, a partir da informação sobre o estado alélico de uma variante é possível estimar o estado alélico de outros sítios próximos. Em retrospectiva, isto ocorre devido à ancestralidade compartilhada entre as sequências, ou seja, em prospectiva, cada novo evento de mutação gera uma nova linhagem (ramificação na genealogia) desta sequência, e a nova mutação estará ligada aos demais alelos desta sequência até que um novo evento evolutivo desfaça a unidade destes elementos. Esta combinação de alelos numa sequência recebe o nome de haplótipo. Os fatores evolutivos capazes de desfazer a associação entre variantes genéticas numa sequência são a mutação, citada acima, e a recombinação, evento no qual as sequências, durante o processo de *crossing-over*, permutam trechos, segmentos, de DNA formando novas configurações entre os alelos. Outros eventos evolutivos, tais como seleção natural, deriva genética, fluxo gênico e demografia são cruciais na manutenção ou extinção das linhagens de sequências. O estudo do desequilíbrio de ligação e dos haplótipos é de particular interesse nos estudos de associação, uma vez que as variantes causais de um determinado fenótipo não precisam ser diretamente testadas, pois a detecção da associação pode ocorrer em quaisquer polimorfismos correlacionados às variantes causais. Além disso, outro ganho proporcionado pelo desequilíbrio de ligação é permitir a inferência do genótipo de um loco quando este não é diretamente testado. Este processo denominado imputação permite o aumento do poder estatístico de GWAS (Genome-Wide Association Studies), auxilia na fase de *fine-mapping*, ou seja, na determinação da variante causal de um fenótipo e facilita os estudos de meta-análise (BROWNING; BROWNING, 2011; BROWNING, 2008; CLARK, 2004; ROSENBERG; NORDBORG, 2002; THE INTERNATIONAL HAPMAP CONSORTIUM, 2003, 2005, 2007).

Outros dois conceitos importantes associados à genealogia das sequências são: a identidade por estado (IBS – *Identity By State*) e a identidade por descendência (IBD – *Identity By Descent*). IBS refere-se àqueles alelos que contém o mesmo estado, dessa forma, a probabilidade de que alelos sejam idênticos, numa determinada amostragem, está associada à frequência dos mesmos na população. Quando alelos compartilham não apenas o estado, mas também a genealogia, ou seja, foram gerados pelo mesmo evento de mutação, são

denominados IBD. Os cálculos provenientes destes conceitos são utilizados nas estimativas de similaridade e grau de parentesco entre os indivíduos (BROWNING; BROWNING, 2012; ROSENBERG; NORDBORG, 2002).

A seleção natural pode ser subdividida em três classes: positiva, purificadora e balanceadora. A seleção positiva, ou seleção Darwiniana, está relacionada ao incremento da frequência de um alelo que aumente o *fitness* do indivíduo (HURST, 2009). Dessa forma, fenótipos que apresentam grande diferenciação entre populações, possivelmente, estão relacionados a polimorfismos apresentando grandes diferenças nas frequências alélicas (MYLES et al., 2008). Regiões sob seleção positiva têm alto desequilíbrio de ligação, isso devido à elevação das frequências no alelo selecionado ser mais rápida do que a recombinação no local onde ele está situado (SABETI et al., 2002). A seleção purificadora, ou negativa, elimina mutações deletérias. De acordo com a premissa que a seleção natural gera indivíduos com valor adaptativo satisfatório, provavelmente, esse tipo de seleção é a mais comum, pois a introdução de mudanças poderia desfazer mecanismos essenciais à manutenção do *fitness*. A seleção balanceadora atua no sentido de favorecer a diversidade através da codominância, seleção dependente da frequência ou coevolução parasita-hospedeiro cíclica. Alelos não são fixados e não podem ser classificados como deletérios ou vantajosos nesse modo de seleção (HURST, 2009).

A Teoria Neutra prediz que a seleção positiva é um evento raro e que a deriva genética e a seleção purificadora predominam no nível molecular, desta forma, grande parte da variabilidade entre os indivíduos se dá pela ação da deriva genética (AKASHI; OSADA; OHTA, 2012; HUGHES, 2008). Entretanto, devido à limitação em explicar taxas médias de evolução diferenciais entre taxa e tipos de mutação (por exemplo, sinônimas e não sinônimas), a Teoria Neutra foi sendo substituída pela Teoria Quase-Neutralidade. Sob essa teoria, grande parte da variabilidade entre populações ocorre devido à deriva genética (HURST, 2009). Entretanto, um dos principais obstáculos referentes às inferências evolutivas repousa justamente na identificação de quais variantes evoluem por deriva genética e quais, devido às pressões seletivas (BALARESQUE; BALLEREAU; JOBLING, 2007).

Dentre as metodologias que permitem estimar a diversidade inter e intrapopulacional, o conjunto de estatísticas-F e cálculos análogos, desenvolvidas a partir do trabalho de Sewall Wright e Gustave Malécot, constituem uma das maneiras mais difundidas e usuais de se

analisar a estrutura genética populacional (BHATIA et al., 2013). A partir dos conceitos desenvolvidos por ambos, diferentes análises foram criadas para estimar a divergência entre demes, nas últimas décadas, as estimativas mais utilizadas se baseiam na Análise de Variância (ANOVA). Neste cálculo, desenvolvido por Weir e Cockerham, é introduzida a distinção entre a variabilidade genética – devida às diferenças nas frequências alélicas das populações – e a amostral – variância devida ao processo de amostragem, além de se explicitar a relação entre a estruturação populacional e os componentes da variância intra e interpopulacional (HOLSINGER; WEIR, 2009; WEIR; HILL, 2002).

A Análise de Variância Molecular se baseia no cálculo da diferenciação entre haplótipos, podendo, contudo, ser estendida a diferentes tipos de dados moleculares (EXCOFFIER; SMOUSE; QUATTRO, 1992). A ideia central de AMOVA é análoga à análise de variância proposta por Weir e Cockerham, onde as variâncias das frequências alélicas são calculadas e as médias entre dois ou mais grupos são testadas quanto à homogeneidade (HOLSINGER; WEIR, 2009). A AMOVA permite a partição da variância genética entre vários loci em dois ou mais componentes, ou seja, níveis hierárquicos: intrapopulacional e interpopulacionais. Nestes modelos, estatísticas análogas às Estatísticas F são designadas Estatísticas Φ (EXCOFFIER; SMOUSE; QUATTRO, 1992).

As estimativas de ancestralidade possuem duas perspectivas quanto à estruturação: global e local. As estimativas globais buscam esclarecer a ancestralidade, ou as proporções desta, em relação a um indivíduo, gerando porcentagens correspondentes a cada uma das ancestralidades presentes em um indivíduo. Além disso, as estimativas globais permitem assinalar indivíduos às populações ou identificar a estruturação presente em um grupo de indivíduos, estimando, inclusive, o número de clusters, ou populações, presentes naquela amostragem. As estimativas locais referem-se às análises que pretendem elucidar a ancestralidade de segmentos cromossômicos em um indivíduo. Deste modo, as estimativas de ancestralidade são capazes de auxiliar a resolução de questões relativas à miscigenação, estruturação populacional, estudos de associação e seleção natural (LIU et al., 2013b).

Diferentes estatísticas e modelos têm sido propostos para inferir a ancestralidade global e local; além disso, os softwares também se distinguem quanto à confiabilidade dos resultados, tempo de execução e à inferência do número de clusters. Dentre os métodos com foco na ancestralidade global tem-se: STRUCTURE (FALUSH; STEPHENS; PRITCHARD,

2003; PRITCHARD; STEPHENS; DONNELLY, 2000), o mais utilizado, onde a distribuição a posteriori para o cálculo das probabilidades de ancestralidade dos genótipos é construída a partir de simulações de Monte Carlo via Cadeias de Markov e a significância do modelo é estimada para cada valor de K (número de clusters); ADMIXTURE (ALEXANDER; NOVEMBRE; LANGE, 2009) onde o valor de K que melhor se adapta aos dados é obtido a partir de validação cruzada e estima as probabilidades de ancestralidade dos genótipos observados a partir de cálculos da máxima verossimilhança; e *frappe* (TANG et al., 2005), também baseado em um modelo de máxima verossimilhança, entretanto sem a possibilidade de estimar o K ótimo. Dentre os softwares supracitados, a escalabilidade de STRUCTURE é proibitiva e a acurácia de ADMIXTURE é superior à de *frappe* (LIU et al., 2013b).

As técnicas de álgebra linear, uma importante área matemática envolvida na redução da dimensionalidade e classificação, têm sido utilizadas na discriminação da estrutura populacional em estudos de associação e genética de populações desde os trabalhos seminais de Cavalli-Sforza e colegas (FRANÇOIS et al., 2010). Recentemente, estas técnicas têm retomado a popularidade como ferramentas para sumarizar estudos genômicos de larga escala, provendo covariáveis que podem ser utilizadas como correção da estrutura populacional em estudos de associação genômicos e por revelar os principais fatores que explicam a estruturação da variabilidade genética em grandes amostras (FRANÇOIS et al., 2010). A Análise de Componentes Principais (ACP – *Principal Component Analysis* – PCA) é uma das técnicas mais populares no campo da genômica, sendo comumente utilizada na identificação da estruturação da variabilidade genética entre diferentes localidades geográficas e na discriminação do "*background*" étnico em estudos de associação (MCVEAN, 2009). Além da ACP, outras técnicas têm sido utilizadas na descrição da variabilidade genética humana, tais como a Decomposição em Valores Singulares (*Singular Value Decomposition* – SVD) (LIU; ZHAO, 2006) e o Escalonamento Multidimensional (*Multidimensional Scaling* – MDS) (LIU et al., 2013a). Em comum, essas técnicas realizam uma transformação linear que decompõe uma série de variáveis em novas variáveis, não correlacionadas, capazes de fornecer informações relativas à estrutura e classificação dos dados.

O processo de reconstrução da história evolutiva humana seja a partir da descrição dos efeitos demográficos ocorridos no homem anatomicamente moderno, ou, do estudo de populações humanas contemporâneas permite a construção de um modelo nulo de evolução a partir de variantes seletivamente neutras e facilita a identificação dos polimorfismos genéticos

que contribuem para a adaptação humana ou o desenvolvimento de doenças (GARRIGAN; HAMMER, 2006). O modelo proposto de expansão humana a partir da África (RAO – *Recent African Origin*) propõe que as populações humanas sofreram múltiplos eventos fundadores durante a colonização de novas áreas. Dessa forma, a heterozigosidade tende a diminuir de acordo com a distância a partir da África (NOVEMBRE; DI RIENZO, 2009). Contudo, efeitos de gargalo, ou *bottlenecks*, seguidos por expansão espacial podem levar ao aumento da frequência de um alelo enquanto novas populações colonizam áreas próximas, fenômeno conhecido por *allele surfing* (HOFER et al., 2009). Esse fenômeno ocorre devido à ação da deriva genética que ocorre durante a expansão populacional e dificulta as inferências sobre a ação da seleção positiva nas populações (NOVEMBRE; DI RIENZO, 2009). Toda esta conjuntura implica no fato de que as populações africanas são as mais diversas e que todas as demais populações apresentam apenas amostragens desta variabilidade presente na África. As populações das Américas e Oceania são as menos diversas devido à maior ação da deriva genética, uma vez que estas foram as últimas áreas a serem colonizadas (FRIEDLAENDER et al., 2008; ROSENBERG et al., 2002; SCLIAR et al., 2012; WANG et al., 2007).

Há pouco mais de uma década, as coletas de dados genéticos constituíam um esforço fragmentado, entretanto a era dos grandes projetos de diversidade genômica humana inicia-se com o Projeto de Diversidade Genética Humana (HGDP) (CAVALLI-SFORZA, 2005). Este projeto gerenciado pela fundação Jean Dausset – CEPH (*Centre d'Etude du Polymorphisme Humain*) disponibiliza linhagens celulares de 1056 indivíduos provenientes de 52 populações espalhadas por 5 continentes e tem como objetivo não apenas a coleta, armazenamento e distribuição, mas também proporcionar recursos para o estudo da variabilidade humana e compartilhar os resultados das genotipagens realizadas (CANN, 1998). Outras iniciativas similares foram criadas posteriormente, como o projeto HapMap (ALTSHULER et al., 2010; THE INTERNATIONAL HAPMAP CONSORTIUM, 2005, 2007), cujo objetivo é identificar e catalogar variantes genéticas, além de estabelecer um mapa confiável das variantes em desequilíbrio de ligação (mapa de haplótipos) em 11 populações na fase III do projeto (4 nas fases I e II). Outro projeto, o 1000 Genomes pretende criar um compreensivo catálogo de diferentes tipos de variantes genéticas obtidas a partir do sequenciamento e genotipagem de 1092 indivíduos provenientes de 14 populações (ABECASIS et al., 2012).

Entretanto, as populações miscigenadas e autóctones da América Latina encontram-se subrepresentadas tanto nos painéis de indivíduos quanto nos estudos genômicos. Tomando

como exemplo os estudos genômicos de associação (GWAS), aproximadamente 96% dos indivíduos participantes são eurodescendentes (BUSTAMANTE; BURCHARD; DE LA VEGA, 2011). Tais fatos demonstram a necessidade de se estudar os padrões de diversidade e susceptibilidade à doenças nas populações miscigenadas e aborígenes americanas.

Quanto à estrutura genética, as populações da América do Sul apresentam alta correlação entre a distância geográfica e a heterozigosidade esperada, sendo que os valores de diversidade genética decrescem a partir da África. Tais dados corroboraram a hipótese de origem humana a partir da África e que a diáspora e expansão da população humana se deram através de sucessivos efeitos fundadores (RAMACHANDRAN et al., 2005).

Duas rotas podem ter sido usadas por populações pleistocênicas para alcançar o continente americano. Uma delas seguiria a partir do Crescente Fértil (Oriente Médio) pela costa do Sudeste Asiático e Mar do Japão (Oceano Pacífico) e a outra rota se daria pela Ásia Central. O trabalho de (ZHANG et al., 2007) demonstra que a colonização do Leste Asiático provavelmente se deu a partir do sul do continente e os níveis de miscigenação entre populações da Ásia Central e populações da região Nordeste da Ásia (direção centro-norte) são maiores que as observadas na direção centro-sul.

O processo histórico de povoamento da América, iniciado no Pleistoceno ainda é controverso em relação ao número de levas migratórias, à rota seguida pelas primeiras populações, à idade dos primeiros assentamentos no continente e ao *pool* gênico dos colonizadores (CORELLA et al., 2007). Atualmente, a hipótese mais aceita é de que apenas uma leva migratória tenha contribuído efetivamente para a formação do pool gênico das populações nativo-americanas e a rota de colonização tenha ocorrido a partir da costa (WANG et al., 2007). Entretanto, há consenso em relação a certos eventos evolutivos que atuaram sobre as populações pioneiras na exploração do continente, tal como a deriva (efeito de gargalo) que imprimiu sua assinatura no genoma das populações nativo-americanas, reduzindo a diversidade genética dentro dos diversos grupos populacionais.

Este efeito fundador experimentado pelas primeiras populações nativo-americanas pode ter impacto direto na saúde das populações nativo-americanas pioneiras. Muitas doenças comuns estão aumentando sua prevalência em populações nativo-americanas, levando a crer que a perda de alelos raros devido à *bottlenecks* não altera significativamente a genética de

doenças complexas (MULLIGAN et al., 2004). Atualmente, existem poucos estudos com foco na dinâmica genético-populacional de polimorfismos deletérios. O excesso de variantes deletérias e o declínio do *fitness* nos eixos das expansões populacionais, fenômeno denominado *expansion load*, tem sido observado em diversas espécies, inclusive nas populações humanas não-africanas. Os mecanismos envolvidos neste fenômeno ainda não são compreendidos e algumas hipóteses, tais como, excesso de variantes raras durante o crescimento populacional ou um efeito gargalo severo a partir da África, não se adequam aos dados observados (PEISCHL et al., 2013).

A discriminação de genes e variantes causais para doenças complexas representa um importante passo em direção à elucidação dos mecanismos genéticos envolvidos na patogênese de doenças complexas e, em alguns casos, na melhora no tratamento, diagnóstico e prevenção de doenças (*International HapMap Consortium*, 2005). E a identificação de SNPs com alta variabilidade entre grupos continentais é crucial em estudos de epidemiologia genética devido a dois fatores especialmente importantes na população brasileira e de maneira geral em populações miscigenadas: estruturação populacional e desequilíbrio de ligação gerado por miscigenação (BALDING, 2006; SMITH; BRIEN, 2005).

A estruturação populacional está relacionada ao risco de associações espúrias em estudos caso-controle devido a diferenças nas frequências alélicas entre as subpopulações constituintes do pool gênico de uma população. Três fatores podem levar à sobrerrepresentação de um grupo entre os casos. O primeiro refere-se ao risco inerente de que um alelo possa ser falsamente associado ao fenótipo caso esse último seja mais frequente numa das populações. Tal fato ocorreria apenas por essa população estar sobrerrepresentada nos casos e o alelo, erroneamente associado, em maior frequência nessa população em relação às outras constituintes do *pool* gênico. Portanto, para muitos alelos candidatos, a associação se dará apenas pelas diferenças demográficas e não pela causalidade da doença. Marcadores que apresentam grandes diferenças entre grupos populacionais estão mais sujeitos a esse tipo de erro, ou seja, à falsa associação. Outro fator passível de introduzir associações espúrias em estudos de associação é a penetrância diferencial devida a variáveis ambientais. Nesse caso, alguns subgrupos podem ter maior penetrância do genótipo causal devido a pressões ambientais diversas, tais como, hábitos alimentares. O terceiro fator está relacionada ao viés de averiguação, ou seja, quando há diferenças na amostragem dos indivíduos constituintes do

estudo devido a fatores não genéticos tais como, acesso à saúde pública, local de moradia, erros de amostragem (BALDING, 2006).

O objetivo do mapeamento de associação em estudos caso-controle é determinar quais variantes genéticas estão associadas a determinados fenótipos. A ideia consiste em que alelos causais, ou outros próximos a eles, caso exista desequilíbrio de ligação, devem ter frequências alélicas diferentes nos grupos de casos e controles (PRITCHARD; DONNELLY, 2001). Comumente, o desenho de estudos caso-controle prevê que os casos e controles sejam selecionados a partir de amostras populacionais não enviesadas pelo uso de indivíduos relacionados. Além disso, os indivíduos dos dois grupos devem ser pareados de acordo com a idade, sexo, etnicidade, dentre outras possíveis variáveis confundidoras (THOMAS; WITTE, 2002). Entretanto, a estruturação populacional e a miscigenação podem invalidar resultados obtidos a partir de populações com distribuição heterogênea de frequências alélicas (PRITCHARD; DONNELLY, 2001).

Outro fator relevante para estudos de associação em populações como a brasileira, entretanto, é um artifício positivo para os estudos de mapeamento em populações miscigenadas. Nessas populações de miscigenação recente, o desequilíbrio de ligação gerado por miscigenação (DLM) pode ser utilizado para o mapeamento genético de doenças complexas (SMITH; BRIEN, 2005). Também nessa abordagem, a diferença de incidência da doença entre as populações parentais é de grande importância. Isso porque as regiões de ancestralidade genômica correspondentes à população em que a doença é mais freqüente serão escaneadas com o intuito de localizar marcadores causais para doença (SELDIN, 2007). Essa estratégia é realizada através da localização dessas regiões a partir de Marcadores Informativos de Ancestralidade (MIAs) e, assim que os fragmentos cromossômicos são localizados, eles são mapeadas em busca de mutações relacionadas ao fenótipo de interesse (SMITH; BRIEN, 2005). Dessa forma, quando há excesso de casos compartilhando um mesmo alelo que é mais comum na população em que a prevalência da doença é maior, este pode ser um sinal de que aquele alelo contribui para o risco de desenvolvimento da doença (BALDING, 2006).

Com o advento das tecnologias de genotipagem e sequenciamento de alto desempenho a baixos custos (NGS – *Next Generation Sequencing*), o número de estudos genômicos de associação e diversidade genética tem crescido exponencialmente, e a capacidade de gerar

dados supera, atualmente, a capacidade de interpretá-los (GOLDSTEIN et al., 2013). Novos gargalos analíticos e computacionais têm surgido devido à esta enorme quantidade de dados gerada. Este novo paradigma, também conhecido como *Big Data*, exige novas metodologias para manipular, processar e compartilhar dados. Ainda que outros campos, como a física e a astronomia, compartilhem gargalos semelhantes, o desafio das ciências biomédicas é ainda maior devido à heterogeneidade e distribuição dos dados (MARX, 2013). Tais dificuldades são ainda mais evidentes para os SNPs, que constituem a maior fonte de variabilidade genética e são os marcadores mais utilizados na genômica nos últimos anos. A enorme quantidade destes torna os processos de visualização, armazenamento, análises e anotação, custosos, sendo crítico desenvolver metodologias capazes de lidar com tamanha quantidade de dados em tempo hábil e livres de erros (BROMBERG; CAPRIOTTI, 2013).

Um problema adicional na interpretação dos dados genômicos surge da falta de integração entre as diferentes análises e ferramentas, sendo que muitas delas sequer são projetadas para interagirem e complementarem outras (BROMBERG, 2013). A formalização de análises em fluxogramas e a integração de ferramentas em *pipelines* diminuem o tempo de execução das análises e a manipulação dos dados, incrementando a eficiência e a reprodutibilidade dos estudos. Em geral, laboratórios de médio e pequeno porte, sem apoio bioinformático, dependem de ferramentas desenvolvidas por outros pesquisadores, entretanto, mesmo a utilização de plataformas como o Galaxy, pode ser difícil para pessoas com pequeno *background* computacional (D'ANTONIO et al., 2013; MARX, 2013). Desta forma, um esforço adicional dos desenvolvedores é necessário para criar plataformas intuitivas, facilmente extensíveis e robustas para lidar com dados de pequena à grande escala.

O *Big Data* na biologia influencia ainda a mudança do paradigma analítico. De um campo essencialmente descritivo, para uma disciplina orientada à integração e comparação de grandes quantidades de dados. Deste modo, a formalização de atributos biológicos, por exemplo, através de ontologias, tem se tornado cada vez mais importante na integração e descrição do conhecimento biológico (JENSEN; BORK, 2010). A integração do conhecimento biológico ocorre através da mineração de dados em larga escala, ou seja, é o processo de utilizar as informações de vários conjuntos de dados para responder uma questão biológica. No contexto da associação entre variantes genéticas e fenótipos, a integração de bancos de dados exige: metodologias eficientes para interrogar as bases, ou seja, escaláveis;

integrar informações de vários dados experimentais; e retornar dados dos repositórios apropriados (GOLDSTEIN et al., 2013; THORISSON; MUILU; BROOKES, 2009).

Formalmente, este processo de adicionar informações relevantes aos dados é conhecido como anotação. Ou seja, metadados são criados buscando aumentar a quantidade e qualidade das informações associadas aos dados originais. O processo de anotação é essencial na caracterização funcional dos elementos genômicos, envolvendo-se desde a identificação de genes até a caracterização fenotípica destes mesmos elementos. As complexidades associadas ao processo de anotação envolvem, além da quantidade e heterogeneidade, o acesso, a qualidade e complexidade dos dados, e a representação do conhecimento. Esta última refere-se ao processo de sumarizar a complexidade de análises e dados em hipóteses e conclusões inteligíveis (THORISSON; MUILU; BROOKES, 2009).

A análise de enriquecimento é uma maneira conveniente de sumarizar e extrair conhecimento de um conjunto de dados assentada sobre o princípio de que os genes relacionados a alguma função biológica são mais propensos a serem selecionados em estudos de triagem funcionais ou evolutivos, por exemplo. Ou seja, uma lista de genes pré-selecionados é comparada a uma lista controle de referência, e a partir do processo de anotação destes genes, espera-se que, havendo interações funcionais na primeira lista, os termos associados sejam mais frequentes nesta amostra do que na população – lista referência. Em resumo, a análise de enriquecimento indica quão improvável é a combinação de genes numa lista caso esses fossem selecionados ao acaso. As ferramentas disponíveis para a análise de enriquecimento podem ser divididas em três categorias mais comuns: Análises de Enriquecimento Singulares (SEA); Análises de Enriquecimento baseadas em conjuntos de genes (GSEA); e Análises de Enriquecimento Modulares (MEA). SEA é o método mais tradicional, onde cada termo associado a um conjunto de genes é testado iterativamente, assim, a contagem de genes associados a um termo é efetuada na amostra e na lista referência e a significância estatística é calculada através de distribuições binomiais ou hipergeométricas e testes como Qui-Quadrado e Teste Exato de Fisher. GSEA é bastante comum na interpretação dos resultados de expressão gênica, as principais diferenças em relação à SEA são: a utilização dos resultados experimentais no cálculo de significância; e a ausência de valores de corte para compor a lista a ser analisada. Em geral, utilizam-se os testes de Kolmogorov-Smirnov, Teste z, Teste t e análises de permutação. MEA diferencia-se de SEA por incorporar as relações termo-a-termo utilizando algoritmos de clusterização e descoberta

de redes, desta forma, são consideradas não apenas as anotações dos genes, mas as relações entre os termos, permitindo a descoberta de associações crípticas entre funções biológicas. (HUANG; SHERMAN; LEMPICKI, 2009; HUNG et al., 2012). Novas metodologias, como o enriquecimento de conjuntos de vias metabólicas (PSEA – *Pathway Set Enrichment Analysis*) (CARBONETTO; STEPHENS, 2013; DE FILIPPO et al., 2012; GLAAB et al., 2012; SUN et al., 2013) e de conjuntos de SNPs (SSEA – *SNP Set Enrichment Analysis*) (HUNG et al., 2012; LARSON; SCHAID, 2014) têm surgido nos últimos anos, aumentando o escopo das análises e reforçando a importância desse tipo de análise.

A partir da necessidade de se entender o impacto funcional da variabilidade humana e devido aos desafios impostos pela grande quantidade de dados e informações, a epidemiologia genômica e a genética de populações são campos promissores nos quais a transdisciplinaridade se faz necessária para clarificar processos complexos e dinâmicos. Atualmente, dispomos de muitas bases de dados nestas áreas, contendo variados tipos de informação, porém, fragmentados. Faz-se necessário o desenvolvimento de ferramentas bioinformáticas capazes de integrar os diferentes tipos de informação biológica e, isto permitirá uma interpretação mais abrangente dos mecanismos envolvidos na prevalência e desenvolvimento de doenças complexas. A integração de bases de dados e análises genéticas pode, não apenas responder e validar questões complexas no contexto biológico, como também inspirar novos paradigmas computacionais. Além disso, aplicações com grande volume de dados relacionados, e grande potencial de distribuição e colaboração apresentam um desafio para o desenvolvimento de ferramentas computacionais de integração eficientes.

O presente trabalho intenta desenvolver ferramentas bioinformáticas que, através da integração de análises estatísticas e informações de diferentes bases de dados biológicos, permitam incrementar a capacidade analítica de estudos de genética de populações, em especial, nas latino-americanas no contexto deste estudo, permitindo a descrição da estruturação populacional, a identificação e a caracterização fenotípica de variantes genéticas que apresentem evidências de seleção natural e que tenham interesse biomédico nessas populações. Os passos necessários para atingir este objetivo são: i) formalizar metodologias de análise de dados em genética de populações através do desenvolvimento de ferramentas e *pipelines* robustos, eficientes e amigáveis de modo a permitir a reprodutibilidade de dados e análises; ii) descrever a estrutura genética de populações autóctones e miscigenadas da América Latina através de diferentes estimativas com intuito de esclarecer os padrões de

diversidade genética das mesmas; iii) identificar polimorfismos com frequências alélicas divergentes nas populações nativo-americanos em relação às africanas, europeias e asiáticas através de estimativas de divergência populacional visando a identificação de variantes candidatas à seleção natural, responsáveis pela susceptibilidade à fenótipos ou ambos; iv) desenvolver uma plataforma que permita a integração das anotações funcionais e análises de enriquecimento através da utilização de uma tecnologia computacional eficiente intentando a geração de novos níveis de conhecimento.

CAPÍTULO 1 - APLICAÇÕES BIOINFORMÁTICAS NO ESTUDO DA VARIABILIDADE GENÔMICA: DESENVOLVIMENTO DE FERRAMENTAS E FLUXOGRAMAS PARA ESTUDOS DE GENÉTICA DE POPULAÇÕES

Com o objetivo de padronizar e facilitar as análises genético-populacionais o nosso grupo desenvolveu plataformas e fluxogramas para automatizar e integrar diferentes ferramentas e softwares. Foram selecionados softwares livres, robustos e aptos a serem integrados aos fluxogramas de análise. Neste capítulo apresentamos os principais componentes dos *pipelines* de análises de genética de populações desenvolvidos no LDGH, com particular atenção na necessidade de que estes pipelines sejam flexíveis e facilitem o seu aprimoramento e a colaboração entre diferentes pesquisadores.

Para a execução dos fluxogramas criados, duas funcionalidades tiveram de ser implementadas: a criação de scripts de conversão de dados, de forma a compatibilizar a entrada e saída de alguns softwares componentes dos fluxogramas; e a criação de scripts para armazenamento de dados em um banco de dados específico. Esse banco de dados, chamado Divergenomedb, é uma base de dados relacional capaz de armazenar dados genotípicos e fenotípicos de diferentes projetos e grupos de pesquisa. Ele faz parte de uma plataforma chamada Divergenome, a qual foi desenvolvida como projeto de doutorado de (MAGALHÃES, 2011), e permite o armazenamento e a manipulação de dados de diferentes projetos de genética de populações e epidemiologia genética. A plataforma Divergenome possui ainda um segundo componente, o DivergenomeTools (MACHADO et al., 2011; RODRIGUES et al., 2012), uma ferramenta que oferece funções de conversão de formatos de dados diversos, compatíveis com diferentes softwares de genética de populações. Parte dos scripts de conversão de dados que compõe a ferramenta DivergenomeTools, bem como os scripts para inserção de dados no Divergenomedb, foram criados no contexto desta tese, e são especificados nas seções 1.1 e 1.2. Além disso, todos os dados apresentados nesta tese, à exceção dos dados do EPIGEN que pertencem às respectivas coortes, estão armazenados no Divergenomedb.

Nas seções 1.1 e 1.2 é descrita a plataforma Divergenome, ponto inicial para as diversas análises realizadas no LDGH. Na seção 1.3 é apresentado o fluxograma de análises concebido pelo doutorando com o objetivo de auxiliar investigações da genética de populações para conjuntos de dados de pequena a média escala. E diante do desafio imposto pelo projeto EPIGEN-Brasil, novos métodos de estimar a ancestralidade foram integrados ao fluxograma de análises privilegiando a utilização de softwares capazes de lidar em tempo hábil com grandes quantidades de dados. Estes fluxogramas são descritos na seção 1.4. No Anexo A é descrita a simbologia dos fluxogramas utilizados neste trabalho.

1.1 DIVERGENOME: A BIOINFORMATICS PLATFORM TO ASSIST POPULATION GENETICS AND GENETIC EPIDEMIOLOGY STUDIES

1.1.1 RESUMO TRADUZIDO

Iniciativas genômicas de larga-escala como os projetos HapMap e 1000Genomes fiam-se em um robusto suporte bioinformático para auxiliar a produção e análises dos dados. Entretanto, poucas plataformas bioinformáticas orientadas a grupos de pesquisa menores existem para armazenar, manipular, compartilhar e integrar dados de diferentes fontes, assim como para auxiliá-los a realizar as análises de forma eficaz. Nós desenvolvemos tal plataforma bioinformática, DIVERGENOME, para assistir estudos genético-epidemiológicos e de genética de populações executados por grupos de pesquisa de pequeno e médio porte. A plataforma é composta por dois componentes integrados, o banco de dados relacional (DIVERGENOMEdb), e um conjunto de ferramentas para converter formatos de arquivos requeridos por softwares populares nas áreas de genética de populações e epidemiologia genética (DIVERGENOMEtools). No DIVERGENOMEdb, as informações sobre genótipos, polimorfismos, protocolos laboratoriais, fenótipos, indivíduos e populações estão organizados em projetos. E estes projetos podem ser interrogados de acordo com as permissões de acesso. Neste artigo, validamos a plataforma DIVERGENOME através de um estudo de caso relacionado às análises de diversidade populacional do gene *SLC24A* em populações humanas. O uso de DIVERGENOME por indivíduos sem conhecimento bionformático é facilitado pela interface Web intuitiva e carregamento automático de dados, permitindo consultas complexas serem interrogadas facilmente e a conversão direta de formatos de

arquivos (não disponíveis em plataformas similares). DIVERGENOME baseia-se em código aberto, acesso livre e pode ser acessada *online* (pggenetica.icb.ufmg.br/divergenome) ou sediada localmente.

1.1.2 ATIVIDADES REALIZADAS

Neste projeto, eu contribuí com as seguintes atividades: a) testar exaustivamente o sistema, buscando identificar erros e inconsistências, para isto, 1) todos os dados apresentados nesta tese foram incorporados ao banco Divergenomedb e, posteriormente, estes dados foram interrogados e comparados ao conjunto de dados original, 2) os arquivos de entrada, para a maior parte das análises genético-populacionais apresentadas nesta tese, foram criados a partir do conjunto de scripts presentes em DivergenomeTools buscando identificar e corrigir falhas; b) fornecer novos códigos para conversão de formatos ao módulo DivergenomeTools, sendo estas ferramentas identificadas na figura 4; c) fornecer scripts para permitir a inserção de dados no Divergenomedb a partir de arquivos TSV (*tab-separated values*).

DIVERGENOME: A Bioinformatics Platform to Assist Population Genetics and Genetic Epidemiology Studies

Wagner C. S. Magalhães,[1†*] Maíra R. Rodrigues,[1†] Donnys Silva,[1] Giordano Soares-Souza,[1] Márcia L. Iannini,[1] Gustavo C. Cerqueira,[2] Alessandra C. Faria-Campos,[3] and Eduardo Tarazona-Santos[1*]

[1]*Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Brazil*
[2]*Broad Institute, Cambridge, Maryland*
[3]*Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Brazil*

Large-scale genomics initiatives such as the HapMap project and the 1000-genomes rely on powerful bioinformatics support to assist data production and analysis. Contrastingly, few bioinformatics platforms oriented to smaller research groups exist to store, handle, share, and integrate data from different sources, as well as to assist these scientists to perform their analyses efficiently. We developed such a bioinformatics platform, DIVERGENOME, to assist population genetics and genetic epidemiology studies performed by small- to medium-sized research groups. The platform is composed of two integrated components, a relational database (DIVERGENOMEdb), and a set of tools to convert data formats as required by popular software in population genetics and genetic epidemiology (DIVERGENOMEtools). In DIVERGENOMEdb, information on genotypes, polymorphism, laboratory protocols, individuals, populations, and phenotypes is organized in projects. These can be queried according to permissions. Here, we validated DIVERGENOME through a use case regarding the analysis of *SLC2A4* genetic diversity in human populations. DIVERGENOME, with its intuitive Web interface and automatic data loading capability, facilitates its use by individuals without bioinformatics background, allowing complex queries to be easily interrogated and straightforward data format conversions (not available in similar platforms). DIVERGENOME is open source, freely available, and can be accessed online (pggenetica.icb.ufmg.br/divergenome) or hosted locally. *Genet. Epidemiol.* 36:360–367, 2012.

Key words: databases; genetic epidemiology; population genetics; tools; format conversion

†These authors contributed equally to this work.
Supporting Information is available in the online issue at wileyonlinelibrary.com.
*Correspondence to: 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. E-mail: edutars@icb.ufmg.br or Wagner C. Santos Magalhães, 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. E-mail: wcsmagalhaes@gmail.com
Received 15 December 2011; Revised 8 February 2012; Accepted 8 February 2012
Published online 16 April 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21629

INTRODUCTION

The production of biological data by high-throughput technologies has revolutionized biology [Bell et al., 2009]. In genetics, classical and emerging scientific questions are being addressed using single nucleotide polymorphism (SNP) and copy number variation (CNV) genotyping and Next Generation Sequencing (NGS) platforms [Alkan et al., 2011; Altshuler et al., 2010; Harismendy et al., 2009; Mardis and Wilson, 2009]. Today, the body of investigators in biology is composed of a few big research groups that produce high-throughput data, and thousands of small- and medium-sized groups that, in addition to producing smaller amounts of data themselves, also use and integrate the data generated by the large research groups to resolve relevant scientific questions [Fagundes et al., 2007; Tarazona-Santos et al., 2010]. While large-scale genomics initiatives such as the HapMap [Frazer et al., 2007], CGEMs (http://cgems.cancer.gov/), and the 1000-genomes [Durbin et al., 2010] rely on powerful computational and bioinformatics support to assist in the production and analysis of data, there are very few bioinformatics platforms oriented to small and medium groups to store, handle, share, and integrate data from different sources, as well as to assist these scientists in performing their analyses efficiently. As a consequence, these tasks are frequently executed suboptimally, with data files handled manually, which is an error-prone operation seldom coupled with adequate quality control procedures.

Here, we present DIVERGENOME, an open-source, freely available bioinformatics platform, to assist population genetics and genetic epidemiology studies performed by small and medium research groups and to facilitate the integration of the data produced by these groups with those derived from large human genome initiatives. We validated this platform in several population

genetics projects and herein we present a use case. The platform is composed of two components: DIVERGENOMEdb and DIVERGENOMEtools. DIVERGENOMEdb is a relational database that allows the user to safely store individual genotypes from different types of polymorphisms as well as different types of phenotypes. Genotypes can be linked to genomic annotations regarding the studied loci and to descriptions of the laboratory protocols used to generate them. Individuals can be organized in populations. DIVERGENOMEtools is a set of data format conversion tools that handles a sort of input files, facilitating the use of various popular population genetics and genetic epidemiology software. 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. These different tools are potentially combined in a *dynamic conversion pipeline* that increases the number of data format conversions available to the user, following the novel conceptual framework described by Rodrigues et al. (in press). This approach makes DIVERGENOMEtools an easily extendable system that can keep up with the constant developments in the fields of population genetics and genetic epidemiology, and encourage collaborative developments. Moreover, DIVERGENOME has an intuitive and user-friendly interface that includes options such as automatic loading and recovery of data, allowing its use by small- to medium-scale research groups without a robust bioinformatics background. Because DIVERGENOMEdb and DIVERGENOMEtools are integrated, data extracted from the database may be easily converted to formats handled by different analysis tools. Although DIVERGENOME is focused on small to medium research groups, this does not preclude the use of the platform, its individual components, its database design, and the dynamic and extensible pipeline framework by research groups with a robust bioinformatics support.

MATERIAL AND METHODS

DIVERGENOMEDB DESIGN AND BUILDING

DIVERGENOMEdb is hosted using the database management system MySQL version 5.1.45 (http://www.mysql.org/). 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 also be downloaded and hosted locally.

There are no software requirements for users of the online version of the platform. Users who want to run DIVERGENOME locally need to install the following freely available software: the Apache HTTP Web server (http://www.apache.org/), the PHP scripting language (http://www.php.net/), the MySQL database management system (http://www.mysql.org/), and the Perl programming language (http://www.perl.org/). These software programs are already installed in most Web servers from both Windows and Unix-based systems.

DIVERGENOMETOOLS

This is a set of data format conversion tools that may be dynamically assembled to form a pipeline. Conversion pipelines are composed dynamically by compiling a set of conversion tools on demand, depending on the functionality required by the user. DIVERGENOME tools was implemented using the novel general framework developed by our group (Rodrigues et al., submitted), which uses a graph-based approach to implement extensible and low-maintenance pipelines that require different combinations of steps in each execution, which is common in population genetics and genetic epidemiology analyses. In this graph-based approach, the connectivity of pipeline components is represented with a directed graph in which components are the graph edges (conversion tools), their inputs and outputs are the graph nodes, and the paths through the graph are pipelines. The implementation of DIVERGENOMEtools also relies on special data structures (i.e., a tool registry that contains information about the tools and metadata with descriptions of file formats) and a dynamic pipeline algorithm. The tool registry, the graph representation, and the dynamic pipeline algorithm for automatic pipeline composition allow DIVERGENOMEtools to be easily extended without changes to the core pipeline code. The pipeline framework was written in Java and each of the independent conversion modules was written using the Perl programming language.

WEB INTERFACE

DIVERGENOME may be accessed for data storage or recovery through a Web-based interface offering users a simple interaction and friendly navigation. The queries for data recovery are built dynamically by selecting a combination of tables, each one holding complementary information to be queried. This interface implements scripts that perform requests to the MySQL server and the Apache Web server (http://www.apache.org). The whole system is available for download (http://www.pggenetica.icb.ufmg.br/divergenome), including the interface and accompanying scripts. After recovering the data from DIVERGENOMEdb, these may be downloaded and converted to the different formats available in DIVERGENOMEtools. To guarantee portability and accessibility, the system was tested in the Windows (Vista and 7) and Linux CentOS 4.6 operating systems and in the Internet Explorer 9, Chrome, Mozilla Firefox 8, and Opera version (for Windows) web browsers.

RESULTS

DATABASE

To store and link information on genotypes, polymorphisms, individuals, populations, and individual phenotypes and even more important, to organize all these data in the format of Projects, we developed DIVERGENOMEdb, composed of 22 tables that are divided in three parts (Figure 1). The first part (Figure 1A) stores data from individuals, their populations, type of variables collected from them as well as their values, and information about the biological samples available for the individuals. The second part (Figure 1C) stores information on polymorphisms, their annotations (e.g., dbSNP rs code, gene names, reference sequences, and other links), and their possible values (i.e., genotypes). The third part (Figure 1B) defines Projects that are a group of individuals (derived from the first part) screened for a set of polymorphisms or a genomic region

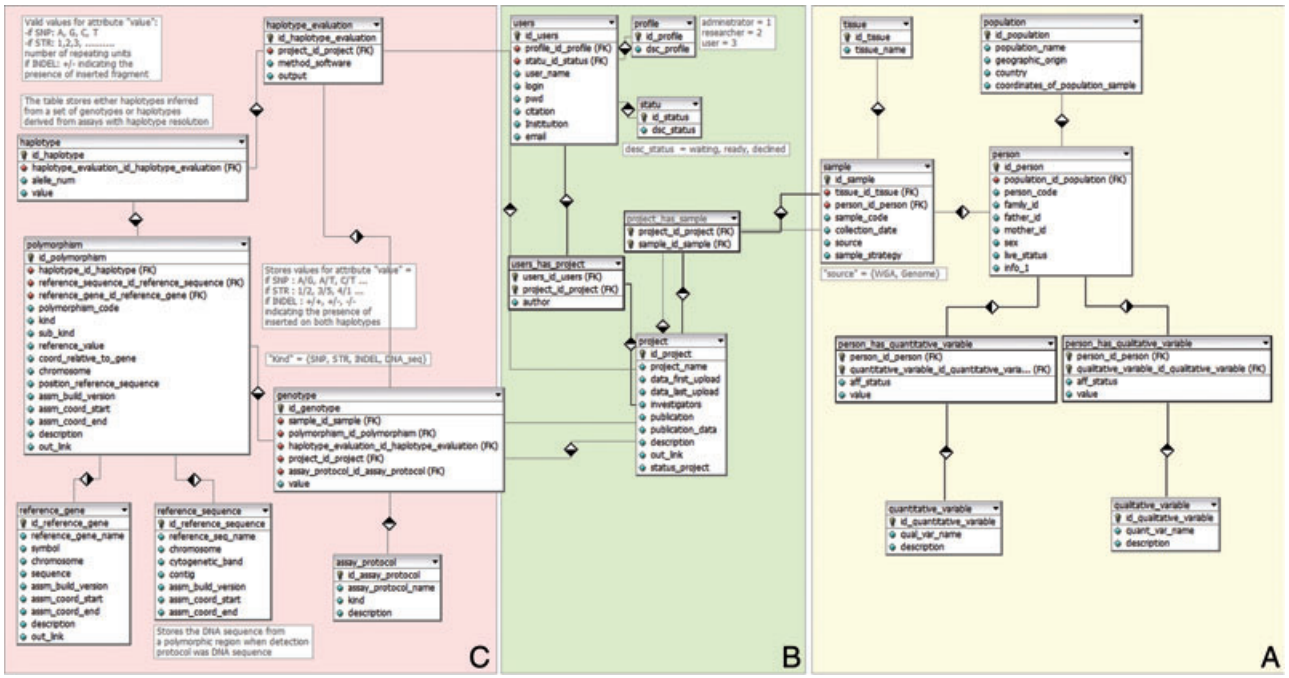


Fig. 1. Entity relationship diagram (ERD) of DIVERGENOMEdb. (A) Data from individuals, their populations, type of variables collected from them and their values, as well as information about the biological samples available for the individuals. (C) Information of polymorphisms, as well as their annotations (e.g., availability of dbSNP rs, gene name, reference sequence, and other links). (B) Projects that are a group of individuals (derived from Part A) screened for a set of polymorphisms or a genomic region. Projects manage all information described for the database's parts A and C.

and its individual genotypes. DIVERGENOME Projects manage and integrate all information of polymorphisms (Figure 1A) and individuals (Figure 1C) of the database. Projects can be defined by their coordinator (e.g., Principal Investigators) as public, when the managed data are intended to be visualized by unregistered users (e.g., for published data), or as private, when data should be accessed only by users who have been granted permission by the project coordinator.

Genetic variation information stored on DIVERGENOMEdb can be retrieved and used to run several types of population genetics and genetic epidemiology software with the assistance of DIVERGENOMEtools.

DATABASE REGISTRATION AND DATA ENTRY

In DIVERGENOMEdb, data entry and modification are possible only for registered users (Figure 2). There are three levels of registered users, as outlined in the following hierarchical order: (i) Administrators have full access to all database functionalities and contents; (ii) Project Coordinators have data entry and modification rights as well as the registration and creation of accounts for project members within Projects; (iii) Project members with access rights can download and search public data as well as data from their respective projects. Unregistered users can search and download public data only.

Data entry is carried out only by Administrators and Project Coordinators, as stated before, using the Web interface (described in the following sections). At the moment, it is possible to upload files in CSV (comma-separated value) and tab-delimited formats. Users can upload their own data as well as complementary data from different public data sources.

Although DIVERGENOME may be accessed via Internet, data visibility depends on the status of the Project that contains the data. Therefore, Projects with a public status are recommended for public or published databases only. Databases containing sensitive individual information including genotypes or clinical data should be stored in private Projects, with restricted visibility. Another option for storing sensitive data, while they are not public, is to keep them in a local version of DIVERGENOME with Intranet access only.

DATA ACCESS

Once datasets are uploaded into DIVERGENOMEdb, users can access those at a single point and handle them using DIVERGENOMEtools. Data from Projects with a public status can be accessed by all users. By contrast, data from private Projects can only be accessed by authorized users (i.e., Project members with password-protected access).

Additional information can be accessed at the platform's Website: http://pggenetica.icb.ufmg.com.br/divergenome/

DIVERGENOMETOOLS

DIVERGENOMEtools is currently able to produce input files for 10 different software programs commonly used in population genetics and genetic epidemiology: PHASE [Stephens et al., 2001], FastPHASE [Scheet and Stephens,

Fig. 2. Screenshots from DIVERGENOME's web interface displaying its main functionalities. From the initial page, users can query public or private datasets as well as use the DIVERGENOMEdb conversion tools. In clockwise direction, the figure shows: the registration screenshot required to access private data, project administrator page screenshot where it is possible to manage user grants, the page used to assign members to a specific project. DIVERGENOMEdb query page showing a query for "individuals" and "populations," and at the center DIVERGENOMEtools screenshot displaying the formats handled by our tool.

2006], DNAsp [Rozas et al., 2003], Haploview [Barrett et al., 2005], STRUCTURE [Pritchard et al., 2000], Haplopainter [Thiele and Nurnberg, 2005], SWEEP [Sabeti et al., 2002], GLU (http://code.google.com/p/glu-genetics/), PLINK [Purcell et al., 2007], and R packages, such as Hierfstat [de Meeus and Goudet, 2007] and Adegenet [Jombart and Ahmed, 2011]. It also accepts popular file formats, such as SDAT, Nexus, and Prettybase. It works by combining 15 different format conversion tools into pipelines that perform specific conversions. In total, 26 different conversion pipelines are available to the user. Each conversion tool has its own internal control that validates the input file before it is converted to the desired file format. If the format is not correct an error message is returned. Noteworthy, the modular and dynamic design of DIVERGENOMEtools (Rodrigues et al., in press) facilitates future extensions of the conversion pipelines to include additional functionalities. The dynamic composition of conversion pipelines is internal to the system, and the user has only to select on the Web interface his original input file format and desired output file format. We have improved the application shown in Rodrigues et al. (in press) by refining the user interface. When the user selects an input file format, output formats that are not available for conversion are disabled from the interface. Also, to facilitate data format selection, we have incorporated metadata about the supported file formats, so that the user can see descriptions of the file formats before selecting them.

WEB INTERFACE

One of the unique features of DIVERGENOMEdb is its query interface. It was designed keeping in mind the complexity of data from genetic epidemiology and population genetics studies that need to be stored and queried. To allow recovering datasets with flexibility, the interface of DIVERGENOME offers a variety of searching options and administration functions: (i) users can combine different types of information (e.g., SNPs, populations, genes, clinical variables) in a single search to integrate different sets of information from distinct database tables (Figure 2). This was achieved by implementing a dynamic query mechanism that combines database tables on demand, depending on the information required by the user; (ii) users can define the level of information detail that will appear in the search results by clicking and checking the information that will compose the search result requested. This is useful when users are retrieving data that will be used in analyses requiring a specific set of information. For example, to recover data in pedigree format, which requires information about individuals, it is possible to select specific contents on the Web interface, such as individual ID, family ID, father ID, mother ID, sex, and live status. It is also possible to add information stored in the table polymorphism, for instance the polymorphism id (e.g., rs dbSNPs code) and its genotype values (Figure 2); (iii) project leaders can manage their projects, including adding new sets of data, and can give permission to new users to access this data (Figure 2).

VALIDATION

DIVERGENOMEdb was designed to provide flexibility for all users, which includes efficient storage and recovery; particularly, it allows integration of distinct datasets produced by genetic diversity projects.

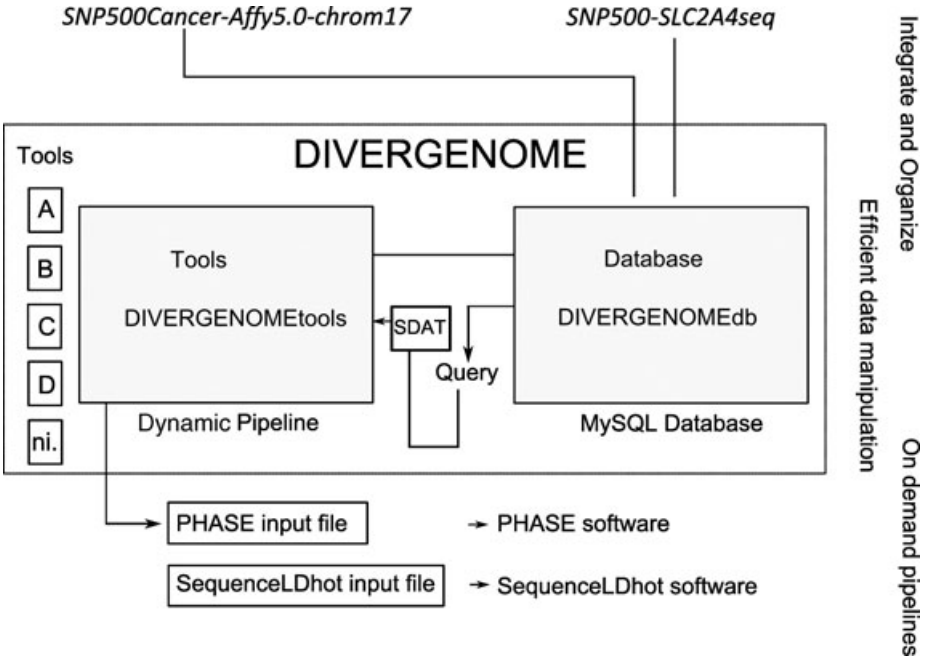


Fig. 3. A case study of DIVERGENOME: diversity of human *SLC2A4* (solute carrier family 2, [facilitated glucose transporter], member 4). Two different types of data for the same set of globally diverse individuals were used: the first concerns SNPs across chromosome 17 obtained from the Affymetrix 5.0 array (SNP500Cancer-affy5.0-chrom17) and the second, sequence data for the gene SLC2A4 (SNP500-SLC2A4seq). These datasets are merged to study long-range linkage disequilibrium and to test the hypothesis involving the existence of a recombination hotspot within the gene. To perform this analysis, we first queried the databases, and then using DIVERGENOMEtools converted the data retrieved in SDAT format to the format required to run the PHASE software. After that, output files from PHASE were used to run SequenceLDhot.

As a use case of DIVERGENOME, we illustrate an analysis performed for the study of the diversity of the gene *SLC2A4* (solute carrier family 2, [facilitated glucose transporter]), member 4 [Tarazona-Santos et al., 2010]. We had two different types of data for the same set of globally diverse individuals that compose the SNP500Cancer Panel (24 African ancestry, 23 admixed Latin American, 31 Europeans, and 24 Asians-Oceanians [Packer et al., 2006]): resequencing data for all these individuals for SLC2A4 (Dataset 1 stored as Project *SNP500-SLC2A4seq*), and genotypes for more than 380,000 genome-wide SNPs for the same individuals, obtained with the Affymetrix 5.0 array [Hughes et al., 2008] (Dataset 2 stored as Project *SNP500Cancer-Affy5.0-chrom17* for SNPs from chromosome 17). We wanted to test the hypothesis that a hotspot of recombination exists within the gene. We performed this test analyzing long-range linkage disequilibrium (LD) across a genomic region of ∼1 Mb that included *SLC2A4* and using the approximate marginal likelihood method implemented in the software SequenceLDhot [Fearnhead, 2006]. To perform the long-range LD analysis, we needed to integrate *SLC2A4* resequencing data with a subset of Dataset 2 that contained only ∼50 SNPs flanking the *SLC2A4* region (0.5 Mb upstream and 0.5 Mb downstream of the gene). We performed the following tasks in DIVERGENOME:

Task 1: We uploaded information about polymorphisms, genotypes, gene, reference sequence, individuals, and populations in DIVERGENOMEdb, organized as Projects *SNP500-SLC2A4seq* and *SNP500Cancer-Affy5.0-chrom17*.

Task 2: We queried the Project *SNP500Cancer-Affy5.0-chrom17* for SNPs inside and flanking *SLC2A4* on both sides (0.5 Mb upstream and 0.5 Mb downstream of the gene), filtering by position.

Task 3: We added this subset of queried SNPs flanking *SLC2A4* to the Project *SNP500-SLC2A4seq*, creating a new expanded dataset.

Task 4: This new Project *SNP500-SLC2A4seq* dataset was queried for the list of common SNPs (MAF $>$ 0.05 in at least one of the SNP500Cancer populations) inside and flanking the *SLC2A4* region on both sides (+ 0.5 Mb upstream and 0.5 Mb downstream of the gene). The output was saved in SDAT format (individuals in the rows and SNPs in the columns). In our case, we knew the common SNPs, but it would be also possible to extract all the SNPs, to use DIVERGENOMEtools to prepare the input for a population genetics software that calculates allele frequencies, and then to perform the query again for the list of common SNPs, or to manually exclude from the SDAT file the noncommon SNPs.

Task 5: The SDAT file was then converted to the PHASE input file format using DIVERGENOMEtools to run the PHASE software that performs phase inferences. The result files of PHASE were then used to run the software SequenceLDhot.

The whole process is visualized in Figure 3. These analyses allowed us to reject the hypothesis of the presence of a recombination hotspot in this genomic region. Projects *SNP500-SLC2A4seq* and *SNP500Cancer-Affy5.0-chrom17* are currently public in DIVERGENOME and a detailed tutorial

TABLE I. Databases for human genetic variation studies: purposes and functionalities

Databases	Aims	Types of data	Requirements	Data storage and accessibility	Tools integrated to the system	Extensibility	Web interface
DIVERGENOME	Assist population genetics and genetic epidemiology studies performed by small-medium research groups, by providing storage, query, and format conversion functionalities.	SNPs, Indels, STRs and CNPs	No requirements for the web version for the standalone version only: • Database management system MySQL • Apache HTTP Web server • PHP scripting language • Perl programming language	WEB and standalone	DIVERGENOMEtools that is able to produce input files for 10 different software programs commonly used in population genetics and genetic epidemiology. It works by combining 15 different format conversion tools into pipelines that perform specific conversions.	The modular and dynamic design of DIVERGENOME-tools facilitates future extensions of the conversion pipelines to include additional functionalities. New conversion tools can be added by experienced users.	DIVERGENOME´s interface offers a variety of searching and filtering options and administration functions.
dbGAP [Mailman et al., 2007]	Official NCBI tool that archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.	SNPs	No requirements for the web version. No standalone version	WEB or controlled FTP access	No integrated tools	The system cannot be extended by the user.	Organizes data in the form of flat files. Limited number of searching and filtering options. Focused on bucked downloads.
SNPATOR [Mailman et al., 2007]	Originally designed to help the CeGen genotyping facility users to handle, retrieve, transform, and analyze genetic data generated by the genotyping facilities of the institution. Has been conceived as a tool to management and analyze genomic SNP data.	SNPs	No requirements for the web version. No standalone version	Only WEB	Tools that allow retrieving the browsed data in seven distinct formats	The system cannot be extended by the user, but changes can requested to the system administrator at the CeGen.	Limited number of searching and filtering options
LOVD [Fokkema et al., 2011]	Web-based software for the collection, display, and curation of DNA variants in locus-specific databases (LSDBs)	SNPs, STRs, CNPs, and indels	No requirements for the web version. For the standalone version: • Database management system MySQL • Apache HTTP Web server • PHP scripting language	WEB and standalone	Data visualization tools, Reference sequence parser	Developers can incorporate additional tools into the source code and experienced users can also extend the database structure.	Offers a variety of searching and filtering options, but the main search criteria is "gene," since it hosts manly gene-specific databases. Does not include as search options the customized features.
GENETIC database [Becker et al., 2004]	Collect, standardize, and archive genetic association study data and provide easily accessible to the scientific community	SNPs and indels	No requirements for the web version. No standalone version	Only WEB	No integrated tools	The system cannot be extended by the user.	Organizes data in the form of flat files . Limited number of searching and filtering options.

to reproduce these analyses is available as supplementary material.

DISCUSSION

We developed an open-source, freely available bioinformatics platform, DIVERGENOME, to assist population genetics and genetic epidemiology studies. We validated the platform through several population genetics projects [Tarazona-Santos et al., 2010; Scliar et al., 2012; Pereira et al., in press], and showed as a use case the test of a population genetics hypothesis of the genetic diversity of *SLC2A4*, integrating two sources of resequencing and SNP data.

Table I shows a comparison between DIVERGENOME and other bioinformatics platforms developed for related purposes: dbGAP [Mailman et al., 2007], Genetic Database [Becker et al., 2004], LOVD [Fokkema et al., 2011], and SNPator [Navarro et al., 2008]. Relative strengths of DIVERGENOME are: (i) its completely open access nature; (ii) the integration between the database and conversion tools functionalities; (iii) the easy extensibility of DIVERGENOMEtools, based on the framework by Rodrigues et al. (in press); (iv) high flexibility for queries provided by the DIVERGENOMEdb interface; (v) the possibility of obtaining individual genotyping data in different formats; and (vi) flexibility to define different levels of access to the data. Conversely, a limitation of DIVERGENOME compared with LOVD is that our database does not allow the final user to insert or customize new database schema tables, while the dynamic structure of LOVD database allows managers to add custom columns. However, this feature limits the LOVD database, making it difficult to automatically update the graphical interface or searching criteria with the added tables or columns.

The intuitive Web interface and automatic data-loading capability of DIVERGENOME facilitate its use by individuals who do not have a strong bioinformatics background. Moreover, turnover of Postdocs and students in research groups frequently leads to data loss [Bourne, 2010], which could be prevented by using DIVERGENOME to store and organize all the data generated in a genetics laboratory. DIVERGENOMEdb could also be used to make datasets used in published papers publicly available. In this case, the scientific community may use DIVERGENOMEtools to obtain published data in different formats, facilitating replication of results and reanalysis of data by peers.

The combination of openness, extensibility, a user-friendly interface, tools for database integration, and format conversion functionalities makes DIVERGENOME a powerful tool oriented to population genetics and genetic epidemiology research groups. In addition to completed and ongoing medium-scale studies that were performed with the support of DIVERGENOME, we will use DIVERGENOME to perform a set of population genomics and genome-wide association studies including more than 6,000 Brazilians (the EPIGEN-Brazil project). Our group is permanently incorporating new options in DIVERGENOMEdb and DIVERGENOMEtools, which include better possibilities for data integration and combined analyses with data from the 1000-genomes project and other large genomic initiatives.

ACKNOWLEDGMENTS

We are grateful to Douglas Santos and Eduardo Galvão for their informatics and technical assistance, and to Dr. Sérgio D Pena, Dr. Alexandre Pereira, Dr. Emanuel Dias Neto, Dr. Mirella Moro, and members of the Laboratory of Translational Genomics and the Core Genotyping Facility from the National Cancer Institute for their suggestions and criticisms. Members of the Laboratory of Human Genetic Diversity collaborated in testing DIVERGENOME and provided suggestions. We are also grateful to Dr. Peter E.M. Taschner (LOVD), Dr. Mike Feolo (dbGAP), and Dr. Carlos Morcillo (SNPator) for clarifying aspects of their bioinformatics platforms. Fogarty International Center and National Cancer Institute (5R01TW007894) funded this study. The study and its participants also received funding and fellowships from the following Brazilian agencies: Brazilian National Research Council (CNPq), Ministry of Education (CAPES), Ministry of Health (PNPD-Saúde Program), the Minas Gerais State Research Agency (FAPEMIG) and the EPIGEN-Brazil Project (Ministry of Health – FINEP).

REFERENCES

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363–376.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Gonzaga-Jauregui C, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265.

Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. Nat Genet 36(5):431–432.

Bell G, Hey T, Szalay A. 2009. Computer science. Beyond the data deluge. Science 323(5919):1297–1298.

Bourne PE. 2010. What do I want from the publisher of the future? PLoS Comput Biol 6(5):1–3.

de Meeus T, Goudet J. 2007. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. Infect Genet Evol 7(6):731–735.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA 104(45):17614–17619.

Fearnhead P. 2006. SequenceLDhot: detecting recombination hotspots. Bioinformatics 22(24):3061–3066.

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32(5):557–563.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu FL, Yang HM, Zeng CQ, Gao Y, Hu HR, Hu WT, Li CH, Lin W, Liu SO, Pan H, Tang XL, Wang J, Wang W, Yu J, Zhang B, Zhang QR, Zhao HB, Izhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao ZJ, Huang W, Chu X, He YG, Jin L, Liu YF, Shen YY, Sun WW, Wang HF, Wang Y, Wang Y, Xiong XY, Xu L, Waye MMY, Tsui SKW, Wong JTF, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai DM, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tarn PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Mailer J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan WH, Li Y, Munro HM, Qin ZHS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Consortium IH. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861.

Harismendy O, Ng PC, Strausberg RL, Wang XY, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S and others. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10(3):1–13.

Hughes AL, Welch R, Puri V, Matthews C, Haque K, Chanock SJ, Yeager M. 2008. Genome-wide SNP typing reveals signatures of population history. Genomics 92(1):1–8.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 21:3070–3071.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L and others. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39(10):1181–1186.

Mardis ER, Wilson RK. 2009. Cancer genome sequencing: a review. Hum Mol Genet 18:R163–R168.

Navarro A, Morcillo-Suarez C, Alegre J, Sangros R, Gazave E, de Cid R, Milne R, Amigo J, Ferrer-Admetlla A, Moreno-Estrada A, Gardner M, Casals F, Perez-Lezaun A, Comas D, Bosch E, Calafell F, Bertranpetit J. 2008. SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. Bioinformatics 24(14):1643–1644.

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

Pereira L, Zamudio R, Soares-Souza G, Herrera P, Cabrera L, Hooper CC, Cok J, Combe J, Vargas G, Prado WA, Schneider S, Kehdy F, Rodrigues MR, Chanock SJ, Berg DE, Gilman RH, Tarazona-Santos E. In press. Socioeconomic and Nutritional Factors Account For The Association of Gastric Cancer with Amerindian Ancestry in A Latin American Admixed Population.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575.

Rodrigues MR, Magalhães WCS, Machado M, Tarazona-Santos EA. In press. Graph-based Approach for Designing Extensible Pipelines.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19(18):2496–2497.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832–837.

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(4):629–644.

Scliar MO, Soares-Souza GB, Chevitarese J, Lemos L, Magalhães WC, Fagundes NJ, Bonatto SL, Yeager M, Chanock SJ, Tarazona-Santos E. 2012. The population genetics of quechuas, the largest native South American group: Autosomal sequences, SNPs, and microsatellites evidence high level of diversity. Am J Phys Anthropol. 147(3):443-451.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978–989.

Tarazona-Santos E, Fabbri C, Yeager M, Magalhaes WC, Burdett L, Crenshaw A, Pettener D, Chanock SJ. 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):1–10.

Thiele H, Nurnberg P. 2005. HaploPainter: a tool for drawing pedigrees with complex haplotypes. Bioinformatics 21(8):1730–1732.

1.2 DIVERGENOME TOOLS

O conjunto de ferramentas presente no módulo DivergenomeTools tem sido continuamente aperfeiçoado e estendido. A primeira versão (MACHADO et al., 2011) lidava, essencialmente, com formatos relacionados a estudos de resequenciamento, permitindo, por exemplo, a transformação do *output* da suíte de aplicativos Phred-Phrap-Consed em um arquivo de entrada para o software de genética de populações DNAsp, mas esta ferramenta tinha o inconveniente de ser um fluxograma único, fixo e inflexível. Posteriormente, a Dra. Maíra Rodrigues desenvolveu um novo paradigma conceitual baseado em grafos ao DivergenomeTools, transformando o outrora estático conjunto de ferramentas em um *pipeline* dinâmico capaz de definir as relações existentes entre as diferentes ferramentas de conversão de arquivos e, então, realizar a transformação entre dois formatos (RODRIGUES et al., 2012).

O artigo ***A graph-based approach for designing extensible pipelines*** publicado na BMC Bioinformatics é descrito pela revista com altamente acessado, apresentando até o dia 13 de fevereiro 2278 acessos, dos quais 580 no último ano (Figura 1). O índice Altmetric indica que esse é o 159º artigo mais acessado da revista BMC Bioinformatics e o nono na comparação com os artigos publicados à mesma época, obtendo pontuação 9. O código do projeto encontra-se armazenado no sítio http://code.google.com/p/dynamic-pipeline, e a partir da ferramenta Google Analytics é possível caracterizar os acessos ao projeto e medir o impacto do DivergenomeTools na comunidade científica. A figura 2 apresenta a discriminação geográfica dos acessos, indicando a ampla distribuição dos 626 acessos. Na figura 3 estão descritas as estatísticas de acesso em termos numéricos, e os países que mais acessaram o projeto nos últimos meses: Índia (140), Estados Unidos (116) e Brasil (44).

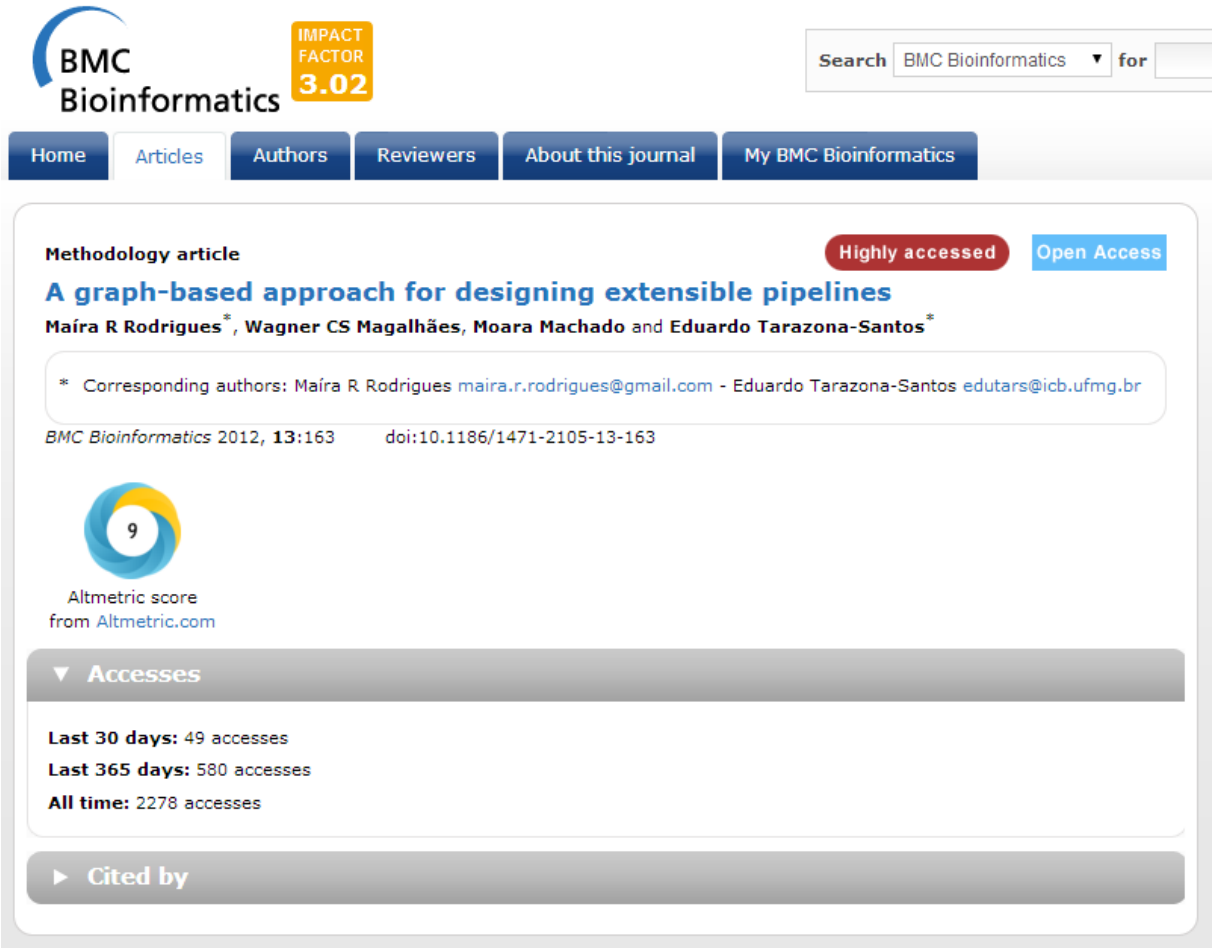


Figura 1: Estatísticas de acesso do artigo *A graph-based approach for designing extensible pipelines* obtidas em fevereiro de 2014 (RODRIGUES et al., 2012)**.**

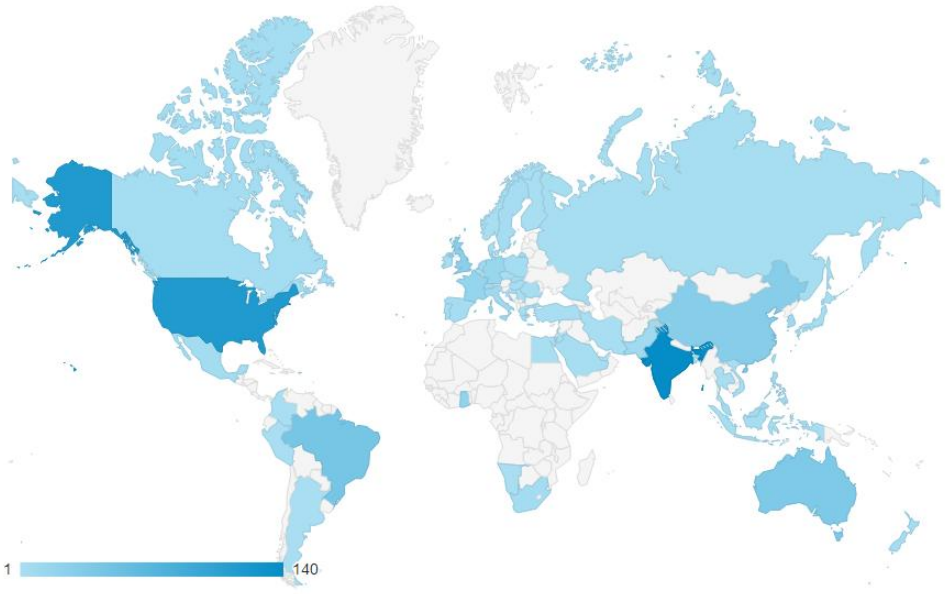


Figura 2: Discriminação geográfica do acesso ao *Divergenome tools* até fevereiro de 2014.

País/território	Aquisição Visitas
	626 Porcentagem do total: 100,00% (626)
1. India	140
2. United States	116
3. Brazil	44
4. Australia	37
5. China	30
6. (not set)	28
7. United Kingdom	26
8. Ghana	21
9. France	18
10. Germany	15

Figura 3: Distribuição mundial do acesso ao projeto *Divergenome tools* até fevereiro de 2014.

.

Atualmente, com a minha participação, estamos trabalhando na expansão do módulo DivergenomeTools, aprimorando as ferramentas de conversão para lidar com a grande quantidade de dados gerada por arranjos de genotipagem e dados de NGSs. Contribuí com novos scripts de conversão de arquivos visando ampliar a gama de formatos disponíveis e tornar a utilização do fluxograma de análises genéticas descrito a seguir mais intuitivo e amigável ao usuário.

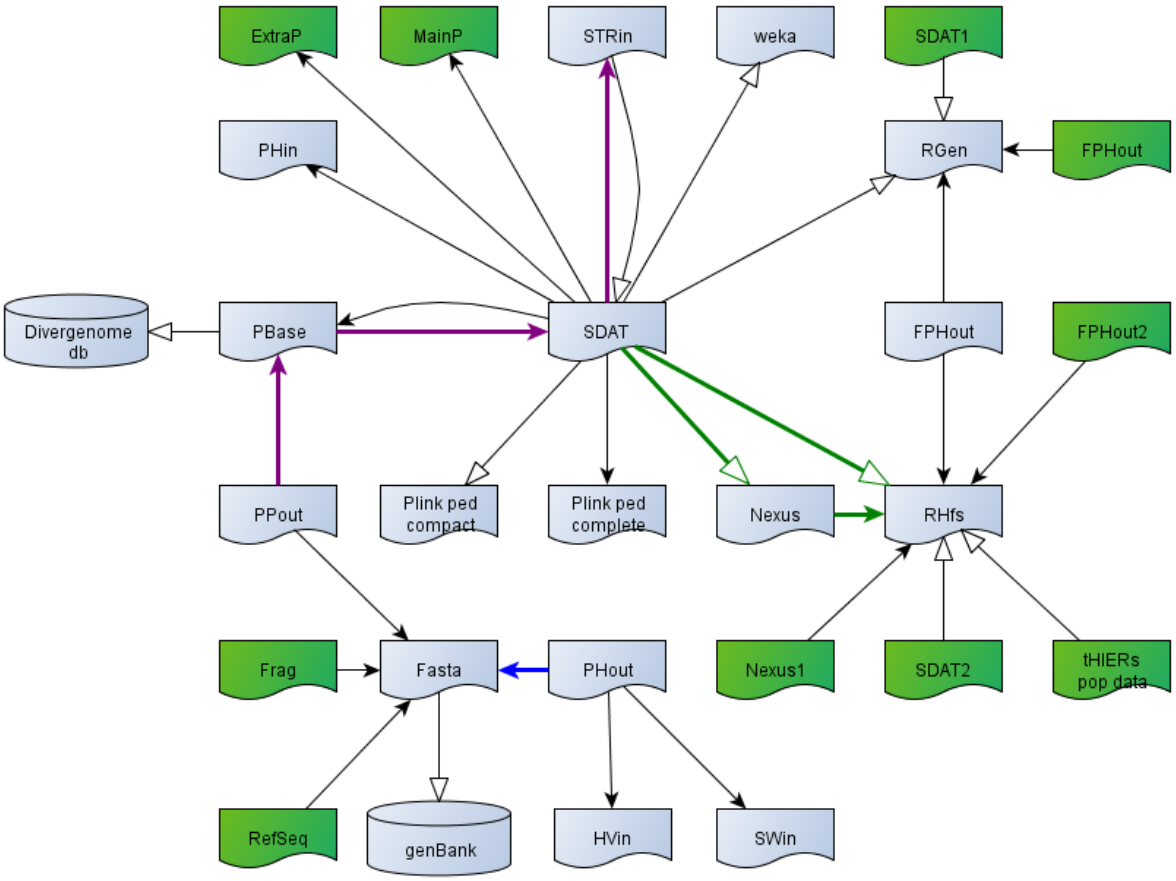


Figura 4: Grafo do *pipeline* dinâmico *DivergenomeTools*. As conversões indicadas por setas abertas foram desenvolvidas parcialmente ou integralmente por Soares-Souza. Os arquivos em verde correspondem àqueles que disponibilizam informações adicionais necessárias para a conversão do formato, por exemplo, para a criação de RHfs – input do pacote Hierfstat (ambiente R) – é necessário um arquivo SDAT ou Nexus para cada população ou uma lista de identificação de indivíduos e populações (*tHIERs pop data*). Três exemplos de conversão são indicados na figura: a seta azul indica a conversão do output do software PHASE (PHout) em input do software DnaSP (Fasta); as setas roxas indicam a conversão do output do software PolyPhred (PPout) em input do software STRUCTURE (STRin); as setas verdes indicam os possíveis caminhos para a conversão de um arqvuivo SDAT no input do pacote Hierfstat (RHfs).

1.3 FLUXOGRAMA DE ANÁLISES GENÉTICO-POPULACIONAIS

O fluxograma apresentado nesta seção foi utilizado nos artigos: ***The Population Genetics of Quechuas, the Largest Native South American Group: Autosomal Sequences, SNPs, and Microsatellites Evidence High Level of Diversity*** (SCLIAR et al., 2012), ***Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population*** (PEREIRA et al., 2012) e

Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers (DA SILVA et al., 2011). Este fluxograma também suporta os resultados apresentados no capítulo 3 (Estrutura genética e evolução dos polimorfismos em populações nativo-americanas). O *pipeline* de análises proposto integra programas que realizam análises genético-populacionais visando a caracterização da diversidade e estruturação a partir de quatro enfoques distintos: a) populacional; b) polimorfismos; c) genes e d) haplótipos.

Os estudos populacionais envolvem a descrição da diversidade genética de populações autóctones ou miscigenadas em um contexto regional e global. Nestes estudos são inferidos os níveis de diversidade intra e interpopulacionais através de diferentes estimadores, além da aferição dos percentuais de miscigenação populacionais e individuais.

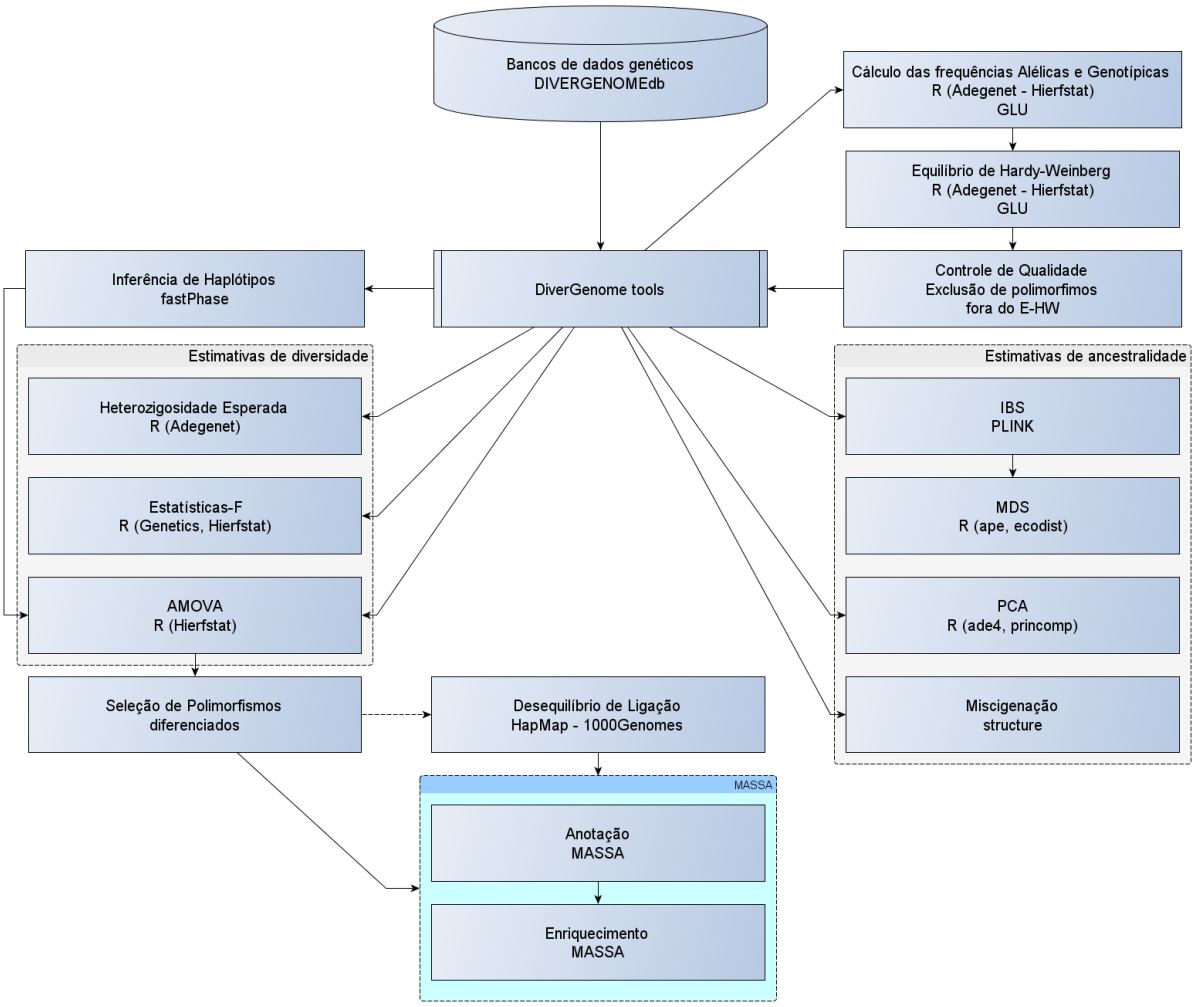


Figura 5: Fluxograma de análises genético-populacionais. O fluxograma acima indica a ordem das análises genético-populacionais realizadas. O processo pré-definido *DivergenomeTools* foi descrito anteriormente na Figura 4. As análises do agrupamento Estimativa de Ancestralidade possuem alternativas para dados de larga escala e são descritas e

exemplificadas no contexto do projeto EPIGEN-Brasil na seção 1.4. Clarificação da simbologia utilizada no fluxograma é apresentada no Anexo A.

Nas análises focadas na evolução dos polimorfismos, busca-se selecionar SNPs que tenham frequências muito diferentes entre as populações. A estruturação presente em alguns polimorfismos é essencial tanto na epidemiologia genética quanto na evolução. Isto porque variantes que apresentem diferenças nas frequências alélicas entre populações podem levar à associações espúrias em estudos de associação ou estarem relacionadas a eventos de seleção natural.

As investigações baseadas em genes têm por objetivo aumentar a robustez dos resultados obtidos, numa tentativa de minimizar os efeitos da deriva genética sobre a interpretação da estruturação genética.

Por fim, os fluxogramas com foco nas análises por haplótipos também tem por objetivo consolidar os resultados da estruturação observada, mas com uma vantagem adicional em relação às análises baseadas em genes: são capazes de fornecer informações com diferentes níveis de resolução, permitindo elucidar os padrões de estruturação intra e intergênicos.

O fluxograma completo é apresentado na Figura 5 demonstrando o encadeamento de ações: da extração dos dados da base Divergenome db à anotação das variantes estruturadas, entre estes dois extremos estão evidenciados os métodos e ferramentas utilizados nas estimativas de variabilidade genética. Nas subseções subsequentes são descritas as metodologias, softwares e parâmetros utilizados em cada uma das análises apresentadas.

1.3.1 CÁLCULO DAS FREQUÊNCIAS ALÉLICAS E GENOTÍPICAS E EQUILÍBRIO DE HARDY-WEINBERG

O cálculo das frequências alélicas e genotípicas e o teste de significância para o teste de Hardy-Weinberg são calculados através do módulo *qc.summary* do software GLU versão 1.0b3, desenvolvido no Cancer Genomics Research Laboratory, sob a liderança do Dr. Kevin Jacobs (https://code.google.com/p/glu-genetics/). O E-HW é obtido através do parâmetro --

hwp, e o nível de significância é estimado através da correção de Bonferroni, desta forma o valor de p é fixado através da divisão do nível descritivo para uma amostra (α = 0,05) pelo número total de loci, ou seja, o valor de corte é α = 0,05/nLoci, onde nLoci é a contagem de polimorfismos presente no conjunto de dados analisado.

O nível de significância estipulado nesse trabalho é conservativo, uma vez que os testes não são completamente independentes, isto devido à correlação não aleatória entre os polimorfismos, ou seja, o desequilíbrio de ligação entre os loci, especialmente dentro dos genes e nas regiões que apresentam assinatura de seleção natural positiva recente. A correção de Bonferroni não é calculada a partir do número de genes, pois o DL pode não ser completo dentro dos genes, particularmente, os longos e há regiões intergências com alto DL. Os loci presentes nos cromossomos sexuais e no genoma mitocondrial foram excluídos das análises devido à dinâmica de segregação diferencial destes cromossomos, esta diferença leva a alterações nas frequências genotípicas esperadas impedindo a correta assunção do E-HW.

1.3.2 ESTATÍSTICAS-F

Os valores para as estatísticas-F são obtidos através da função *fstat* do pacote Adegenet. Esta função é um encapsulamento da função *varcomp.glob* do pacote Hierfstat. Os valores globais de F_{ST}, F_{IS} e F_{IT} são estimadas de acordo com o algoritmo de (YANG, 1998) através da seguinte equação:

$$F_{ji} = \frac{\sigma^2 i(j)}{\sigma^2_{\sum i}}$$

Sendo, $\sigma^2 i$ a variância do componente i, $\sigma^2_{\sum i} = \sum_{k=(j+1)}^{i} \sigma_k^2$ a soma dos componentes da variância desde o nível mais baixo até o nível i e $\sigma^2 i(j) = \sum_{k=(j+i)}^{i} = \sigma_k^2$.

No caso mais simples, os índices $_I$, $_S$ e $_T$ se referem a indivíduos, subpopulações e a população total. Nas análises hierárquicas são acrescidos novos níveis populacionais, onde, por exemplo, o índice $_C$ indica a partição de variância dos grupos populacionais em F_{CT} e F_{SC}.

Para os cálculos de F_{ST} par-a-par entre duas populações (A e B) utiliza-se a função *pairwise.fst*, implementada a partir do cálculo de (NEI, 1973) no qual as heterozigosidades (Hs) das populações A e B são balanceadas pelos tamanhos dos grupos (n) e Ht indica a heterozigosidade total na amostra como demostrado na equação:

$$F_{ST}(A,B) = \frac{(H_t - (n_A H_s(A) + n_B H_S(B))/n_A + n_B))}{H_t}$$

1.3.3 ANÁLISE VARIÂNCIA MOLECULAR – AMOVA

A Análise de Variância Molecular (AMOVA) é calculada através dos dados genotípicos a partir do pacote Hierfstat da plataforma R. A significância dos valores é estimada a partir de 1000 permutações. Além das vantagens relativas à utilização da plataforma R, o pacote Hierfstat foi escolhido por utilizar o algoritmo generalista de (YANG, 1998) que permite a utilização de mais níveis hierárquicos que os algoritmos implementados nos programas GDA (LEWIS; ZAYKIN, 2002) e Arlequin (EXCOFFIER; LAVAL; SCHNEIDER, 2005).

É importante ressaltar que a nomenclatura AMOVA é utilizada para reforçar a noção de que a análise de variância é obtida através de níveis hierárquicos e a partir de dados moleculares. Originalmente, o termo AMOVA refere-se ao cálculo proposto por (EXCOFFIER; SMOUSE; QUATTRO, 1992) e que leva em consideração, não apenas as frequências alélicas e genotípicas, mas também, pode-se incorporar a informação de dados moleculares como, por exemplo, a distância evolutiva entre os haplótipos. A permuta entre termos é usual nas medidas de diferenciação populacional, por exemplo, as siglas θ (WEIR; COCKERHAM, 1984), Φ_{ST} (EXCOFFIER; SMOUSE; QUATTRO, 1992) e G_{ST} (NEI, 1973) referem-se a cálculos diferentes, mas são usualmente intercambiadas com o F_{ST}, favorecendo não só o entendimento do resultado apresentado, mas também reforçando o modelo teórico ao qual os estimadores fazem referência.

É possível estimar os componentes de variância através de duas funções: *varcomp* e *varcomp.glob*. A última é capaz de retornar as estimativas para cada loco, entretanto, é apenas um encapsulamento da função *varcomp*. Desta forma, optou-se por utilizar a primeira devido

à paralelização implementada no script HierFastat que não seria possível para a função *varcomp.glob*.

O modelo tri-hierárquico é utilizado para descrever a variância entre três componentes: ente grupos populacionais (F_{CT}), entre subpopulações dentro dos grupos populacionais (F_{SC}) e entre todas as subpopulações (F_{ST}).

1.3.4 HETEROZIGOSIDADE ESPERADA

A heterozigosidade esperada dos loci foi obtida através da função *Hs* do pacote Adegenet (JOMBART; AHMED, 2011) versão 1.2-8. A diversidade é assim calculada:

$$\frac{1}{K}\sum_{k=1}^{K}\left(1-\sum_{i=1}^{m(k)}f_i^2\right)$$

Dado m(k) o número de alelos k, K a contagem total de loci e *fi* a frequência total do alelo *i* em uma dada população.

1.3.5 CÁLCULO DE IDENTIDADE POR ESTADO (IBS)

IBS refere-se àqueles alelos que contém o mesmo estado, dessa forma, a probabilidade de que alelos sejam idênticos, numa determinada amostragem, está associada à frequência dos mesmos na população. Para o cálculo de distância entre indivíduos é gerada uma matriz de similaridade baseada nos valores de identidade por estado (1-IBS) através do parâmetro --mds-plot K do software *Plink* 1.07 (PURCELL et al., 2007). O parâmetro (K) determina o número de dimensões a serem extraídas da análise de escalonamento multidimensional (MDS).

1.3.6 ESCALONAMENTO MULTIDIMENSIONAL (MDS)

As matrizes de distância genética interindividuais são sumarizadas através do escalonamento multidimensional. O pacote *ecodist* (GOSLEE; URBAN, 2007) realiza o cálculo de distância euclidiana entre os indivíduos através da função *dist*. A função *nmds* permite encontrar uma configuração que melhor represente a dissimilaridade entre os indivíduos utilizando um dado número de dimensões. Além disso, a função *nmds* retorna valores de estresse que medem a qualidade do ajustamento (*goodness of fit*) à representação espacial.

1.3.7 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

A análise de componentes principais pode ser realizada a partir de diferentes fontes dados: genótipos, distâncias genéticas ou variáveis ambientais. O script para a análise de CP a partir de dados genotípicos foi desenvolvido por (CHEVITARESE, 2009), ex-integrante do nosso grupo de pesquisa. Este script foi adaptado com o objetivo de integrá-lo ao fluxograma de análises. São utilizados os pacotes do ambiente R Adegenet (JOMBART; AHMED, 2011), responsável pela transformação dos dados genéticos em dados numéricos, e *Ade4* (CHESSEL; DUFOUR; THIOULOUSE, 2004; DRAY; DUFOUR; CHESSEL, 2007; DRAY; DUFOUR, 2007), responsável pela redução da dimensionalidade. O PCA pode ser realizado para dados individuais (objeto *genind* gerado pelo Adegenet) através do comando *dudi.pca* ou populacionais (objeto *genpop*), função *dudi.coa*. A instrução *s.label* do *Ade4* permite a plotagem dos componentes principais, entretanto, optamos por utilizar a função *plot* padrão do R por esta proporcionar maior controle sobre o design da figura plotada.

Os Componentes Principais a partir de distâncias genéticas e variáveis ambientais, por exemplo, dados socioeconômicos, são obtidos através da função *princomp* do conjunto base do ambiente R.

1.3.8 ESTIMATIVAS DE MISCIGENAÇÃO

As estimativas de miscigenação são calculadas pelo software STRUCTURE (FALUSH; STEPHENS; PRITCHARD, 2003; PRITCHARD et al., 2000) utilizando uma abordagem Bayesiana que permite identificar e caracterizar a estruturação populacional. A

aferição da miscigenação pode ser realizada nos níveis individuais e populacionais, não apenas indicando à qual população um indivíduo pertence, mas também as taxas de ancestralidade genômicas. Os parâmetros de corrida de STRUCTURE variam para os diferentes estudos, e as especificações de corrida podem ser conferidas nos artigos onde o software foi utilizado.

1.3.9 DESEQUILÍBRIO DE LIGAÇÃO

As informações e os dados de desequilíbrio são obtidos através da ferramenta SNAP (*SNP Annotation and Proxy Search*) (JOHNSON et al., 2008). Para obter os SNPs proxy selecionamos aqueles com valores de correlação (r^2) superiores à 0,8, distância máxima de 500Kb e painel populacional CHB+JPT (Chineses Han e Japoneses) sequenciado pelo projeto 1000Genomes. A escolha do painel populacional CHB+JPT visa obter blocos de DL mais adequados e próximos ao que seria observado nas populações Nativo-Americanas. Entretanto, para as populações miscigenadas latino-americanas, pode se utilizar o painel CEU (Euro-descendentes de Utah – EUA). É possível ainda utilizar o painel populacional MEX (Mexicanos) genotipado pelo projeto HapMap fase 3.

1.4 FLUXOGRAMA DE ESTIMATIVAS DE ANCESTRALIDADE PARA DADOS DE GRANDE PORTE (EPIGEN)

1.4.1 EPIGEN-BRASIL

O consórcio de epidemiologia genética brasileiro, EPIGEN-Brasil, é composto por cinco grupos de pesquisa: Fundação Oswaldo Cruz (Belo Horizonte – MG); Universidade Federal de Pelotas; Universidade Federal de Minas Gerais; Universidade de São Paulo – INCOR; e Universidade Federal da Bahia. O projeto dispõe dos dados de 3 coortes: 1) 1309 infantes da coorte de crianças de Salvador (Projeto SCAALA); 2) 1442 anciões da coorte de idosos de Bambuí; e 3) 3736 indivíduos da coorte de nascidos vivos de Pelotas. O EPIGEN tem por objetivo: a) examinar a estrutura genômica e a ancestralidade das populações participantes; b) investigar os efeitos conjuntos da arquitetura genética e do ambiente na ocorrência de doenças complexas, com ênfase na ancestralidade e desigualdades sociais. Para

tal, 6487 e 265 indivíduos foram genotipados, respectivamente, pelos *arrays Illumina HumanOmni2.5* e *Illumina HumanOmni5*. Além disso, 10 indivíduos de cada coorte, perfazendo um total de 30, tiveram o genoma completamente sequenciado.

1.4.2 AMOSTRAGEM

Para as análises de estruturação genética foram incorporadas novas populações ao conjunto de dados composto pelas coortes participantes do projeto EPIGEN. Foram acrescentadas 6 populações do banco HapMap, sendo 3 populações africanas (LWK, MKK e YRI), 1 afrodescendente (ASW), 1 europeia (TSI) e 1 eurodescendente (CEU); e 4 populações nativo-americanos do CEPH-HGDP (Pima, Maia, Karitiana e Surui).

1.4.3 EQUIPES DE TRABALHO DO PROJETO EPIGEN-BRASIL

Em função da operabilidade e capacidade analítica do consórcio EPIGEN foram criadas cinco equipes de trabalho: a) Análises Básicas; b) Imputação e inferência haplotípica; c) Integração, disponibilização e enriquecimento de dados; d) Fluxogramas básicos de análises de associação; e) Estrutura populacional, ancestralidade e *admixture mapping*. Estas equipes, coordenadas por pós-doutorandos e supervisionadas pelos pesquisadores-chefes de cada um dos grupos participantes, atuam num dado escopo analítico (a-e), sendo as competências de cada uma delas discriminadas a seguir:

a) Análises Básicas: limpeza dos dados, controle de qualidade, determinação da estrutura familiar oculta, investigação das anormalidades cromossômicas e disponibilização dos dados congelados, isto é, prontos para uso.
b) Imputação e inferência haplotípica: definição da metodologia de imputação dos dados, construção de haplótipos e disponibilização das fases cromossômicas.
c) Integração, disponibilização e enriquecimento de dados: gerenciamento de dados, repositório de bancos, relatórios de análise, coordenação do desenvolvimento de pipelines analíticos, definição de prioridades sobre alocação de recursos e estrutura computacional, anotação de variantes genéticas, integração de informações moleculares e fenotípicas e enriquecimento de classes biológicas.

d) Fluxogramas básicos de análises de associação: fornecer modelos básicos para análises de associação, estabelecer metodologias de controle de qualidade para estudos de associação e ajustes da estruturação populacional.
e) Estrutura populacional, ancestralidade e *admixture mapping*: fornecer modelos básicos para análises de estrutura populacional, descrição inicial da estrutura genética observada nas coortes, definir metodologias para mapeamento por miscigenação e descrição dos componentes de ancestralidade.

Devido à grande quantidade de dados disponibilizada pelo EPIGEN, um novo *pipeline* (Figura 6) foi construído com o intuito de realizar as análises de forma mais eficaz, confiável e rápida. Em especial para as análises de estrutura populacional, alguns dos softwares indicados no fluxograma da seção 1.3 foram substituídos por programas mais robustos e eficientes. Nas subseções subsequentes são descritas as novas metodologias e ferramentas utilizadas na determinação da estrutura genética das populações participantes do consórcio.

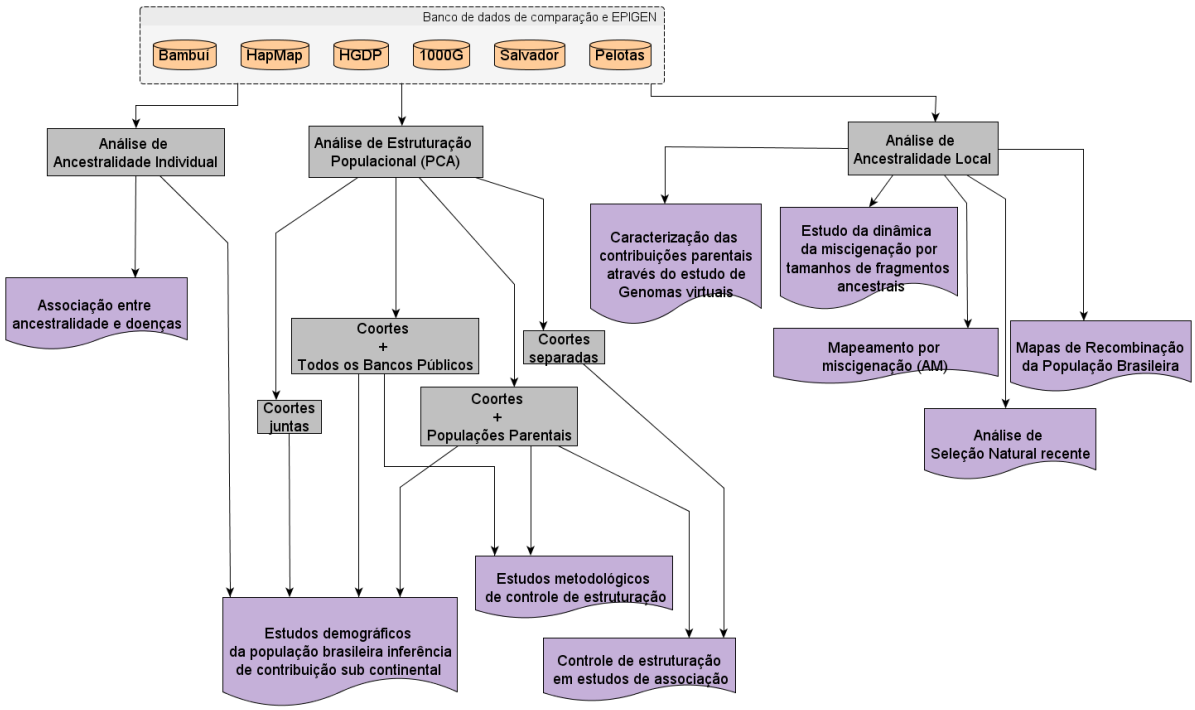


Figura 6: Fluxograma de análises para estimativa de ancestralidade em dados de larga-escala.

1.4.4 ANÁLISE DE COMPONENTES PRINCIPAIS

A decomposição da matriz de genótipos e a retenção das componentes principais é realizada através do software EIGENSTRAT v4.2. Os gráficos são criados a partir da função *plot* do ambiente R.

1.4.5 ANÁLISE DE ANCESTRALIDADE CROMOSSÔMICA E INDIVIDUAL

As taxas de miscigenação individual foram calculadas a partir do programa Admixture (ALEXANDER; NOVEMBRE; LANGE, 2009), neste, a máxima verossimilhança destas taxas é estimada através do algoritmo EM (*Expectation Maximization*). Por não examinar as taxas de desequilíbrio de ligação, os cálculos desta ferramenta são mais rápidos do que aqueles efetuados no STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000). As estimativas de miscigenação populacional foram calculadas a partir das médias de miscigenação individuais para cada população.

A partir do software Multimix (CHURCHHOUSE; MARCHINI, 2013) que utiliza um método baseado em HMM (*Hidden Markov Model*) será realizada a inferência da ancestralidade local ao longo de cada cromossomo. O algoritmo elegante desta ferramenta permite a inclusão de duas ou mais populações parentais, de modelos de desequilíbrio de ligação e pode realizar estimativas em dados faseados e não faseados.

Os gráficos exibindo as proporções individuais e populacionais de mistura são concebidos a partir da função *plot* do ambiente R.

CAPÍTULO 2 - DIVERSIDADE E ESTRUTURA GENÉTICA DE POPULAÇÕES AUTÓCTONES E MISCIGENADAS DO CONTINENTE AMERICANO

Neste capítulo são apresentados os artigos referentes à descrição da diversidade e estruturação das populações americanas, sendo um artigo metodológico e três empíricos, contendo aplicações das ferramentas e fluxogramas explicados no capítulo anterior, sobre as seguintes populações: miscigenadas - Minas Gerais e Peru; autóctone – Quechua (Peru). O artigo **The Population Genetics of Quechuas, the Largest Native South American Group: Autosomal Sequences, SNPs, and Microsatellites Evidence High Level of Diversity** descreve a variabilidade e estrutura genética da população nativo-americana Quechua. Neste artigo o presente autor compartilha a primeira autoria com a doutora Marília Scliar. O artigo **Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: an application to populations of the State of Minas Gerais (Brazil)** apresenta os detalhes metodológicos da criação de um painel de marcadores de ancestralidade de baixo custo para as populações latino-americanas. Este painel será utilizado para descrever os níveis de miscigenação de portadores de Anemia Falciforme no artigo **Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers**. Em **Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population** foi descrita a estrutura genética da população de Lima (Peru) e investigada a relação entre ancestralidade ameríndia e o câncer gástrico nestes indivíduos.

2.1 THE POPULATION GENETICS OF QUECHUAS, THE LARGEST NATIVE SOUTH AMERICAN GROUP: AUTOSOMAL SEQUENCES, SNPS, AND MICROSATELLITES EVIDENCE HIGH LEVEL OF DIVERSITY

2.1.1 RESUMO TRADUZIDO

Elucidar o padrão de diversidade genética para populações não europeias é necessário para conceder aos indivíduos destes grupos os benefícios da pesquisa genética. Na era das grandes iniciativas genômicas, as populações Nativo-americanas têm sido negligenciadas, em

particular, a Quechua, o maior grupo ameríndio sul-americano estabelecido ao longo da cordilheira dos Andes. A diversidade genética da população Quechua foi caracterizada em um contexto global, utilizando sequências autossômicas não-codificantes (nove loci não ligados totalizando 16 kb), 351 SNPs não ligados e 678 microssatélites, além disso, testou-se as predições do modelo de evolução proposto por (TARAZONA-SANTOS et al., 2001). Os níveis de ancestralidade europeia são inferiores a 5% e os de ancestralidade africana quase indetectáveis. As maiores distâncias genéticas se dão entre africanos e melanésios ou quéchuas, o que é concordante com a origem humana na África e com o fato da América do Sul ser o último continente a ser povoado. A diversidade na população Quechua é comparável a das populações eurasianas e o espectro de frequência alélica obtido a partir dos dados de ressequenciamento não reflete a redução na proporção de alelos raros. Dessa forma, a população Quechua é um vasto reservatório de variantes genéticas comuns e raras dos nativos-americanos da América do Sul. Estes resultados corroboram e complementam o modelo evolucionário dos nativos sul-americanos proposto a partir de dados do cromossomo Y. Esse modelo prediz uma alta diversidade genômica devida aos altos níveis de fluxo gênico entre as populações andinas e o tamanho efetivo populacional de longo prazo.

2.1.2 ATIVIDADES REALIZADAS

Neste artigo, realizei as seguintes atividades: (1) Cálculo das frequências alélicas e genotípicas, heterozigosidade observada e esperada, teste para o equilíbrio de Hardy-Weinberg e cálculo de desequilíbrio de ligação para o conjunto de dados de SNPs; (2) Cálculos de F_{ST} par-a-par para as populações do conjunto de dados SNPs; (3) Cálculo de IBS para os indivíduos do conjunto de dados SNPs; (4) PCA e MDS para os três conjuntos de dados analisados (Sequências, AIMs e SNPs); (5) Teste de correlação entre as matrizes de distância genética; (6) Análises de miscigenação nos quéchuas utilizando AIMs e SNPs; (7) Escrita do artigo e elaboração das figuras.

The Population Genetics of Quechuas, the Largest Native South American Group: Autosomal Sequences, SNPs, and Microsatellites Evidence High Level of Diversity

Marilia O. Scliar,[1†] Giordano B. Soares-Souza,[1†] Juliana Chevitarese,[1] Livia Lemos,[1] Wagner C.S. Magalhães,[1] Nelson J. Fagundes,[2] Sandro L. Bonatto,[3] Meredith Yeager,[4,5] Stephen J. Chanock,[6] and Eduardo Tarazona-Santos[1*]

[1]*Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486, Belo Horizonte, Minas Gerais, CEP 31270-910, Brazil*
[2]*Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, Porto Alegre, Rio Grande do Sul, CEP 91501-970, Brazil*
[3]*Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga 6681, Caixa Postal 1429, Porto Alegre, Rio Grande do Sul, CEP 90619-900, Brazil*
[4]*Intramural Research Support Program, SAIC Frederick, NCI-FCRDC, Frederick, MD 21702*
[5]*Core Genotype Facility, NCI, NIH, Gaithersburg, MD*
[6]*Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), Gaithersburg, MD 20877*

KEY WORDS native American; Andes; microsatellites; autosomal noncoding sequence; HGDP–CEPH

ABSTRACT Elucidating the pattern of genetic diversity for non-European populations is necessary to make the benefits of human genetics research available to individuals from these groups. In the era of large human genomic initiatives, Native American populations have been neglected, in particular, the Quechua, the largest South Amerindian group settled along the Andes. We characterized the genetic diversity of a Quechua population in a global setting, using autosomal noncoding sequences (nine unlinked loci for a total of 16 kb), 351 unlinked SNPs and 678 microsatellites and tested predictions of the model of the evolution of Native Americans proposed by (Tarazona-Santos et al.: Am J Hum Genet 68 (2001) 1485–1496). European admixture is <5% and African ancestry is barely detectable in the studied population. The largest genetic distances were between African versus Quechua or Melanesian populations, which is concordant with the African origin of modern humans and the fact that South America was the last part of the world to be peopled. The diversity in the Quechua population is comparable with that of Eurasian populations, and the allele frequency spectrum based on resequencing data does not reflect a reduction in the proportion of rare alleles. Thus, the Quechua population is a large reservoir of common and rare genetic variants of South Amerindians. These results are consistent with and complement our evolutionary model of South Amerindians (Tarazona-Santos et al.: Am J Hum Genet 68 (2001) 1485–1496), proposed based on Y-chromosome data, which predicts high genomic diversity due to the high level of gene flow between Andean populations and their long-term effective population size. Am J Phys Anthropol 147:443–451, 2012.

Elucidating the genetic diversity in human populations is crucial to understand their evolutionary history, the genetic architecture of rare and complex diseases and ultimately, to materialize the promise of genomic medicine. For this purpose, unprecedented amounts of data are being generated in some human groups. Regrettably, this perspective is much closer for some individuals and populations than for other people belonging to neglected populations in human genetics research. For instance, in the last genomic revolution of human genetics (i.e., that of the Genome-Wide Association Studies), 96% of the thousands of participants in these studies are individuals of European origin (Bustamante et al., 2011).

Additional Supporting Information may be found in the online version of this article.

Grant sponsors: Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil), Fundação de Amparo a Pesquisa de Minas Gerais (Brazil), Brazilian Ministry of Education (Agency for the Development of Graduate Education-CAPES) and National Cancer Institute.

†These authors contributed equally to this work.

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

Received 11 October 2011; accepted 9 December 2011

DOI 10.1002/ajpa.22013
Published online 27 January 2012 in Wiley Online Library (wileyonlinelibrary.com).

During the seventies, some Native Americans were the focus of intensive investigation using protein markers (reviewed in Salzano and Callegari-Jacques, 1988; Cavalli-Sforza et al., 1994; Luiselli et al., 2000; Long and Bortolini, 2011). While several scholars studied tribes from Eastern South America (Brazil, Venezuela and Guyana) to test theoretical population genetics models as well as to study human microevolution and its interaction with culture and diseases (Ward and Neel, 1970; Salzano et al., 1972; Spielman et al., 1977; Neel, 1978; Smouse and Long 1992); other Amerindians were relatively neglected, in particular from Western South America. In the era of DNA and large human genomic initiatives, Native American populations have been neglected (Bustamante et al., 2011; Wall et al., 2011), not being included in the HapMap Project (Frazer et al., 2007). In the Human Genome Diversity Panel (HGDP-CEPH, Cann et al., 2002), a widely used tool for population genetics studies, the included South Amerindian populations are outliers that poorly represent the genetic diversity of South Amerindians. Although several single-locus studies have been performed on mtDNA and Y-chromosome (Tarazona-Santos et al., 2001; Fuselli et al., 2003; Fagundes et al., 2008; Yang et al., 2010), only one study producing large-scale (in number of makers and samples) microsatellite data has been published on Amerindians (Wang et al., 2007). Recently, Wall et al. (2011) interestingly suggested that Latin American admixed populations can provide information about the pattern of diversity of Native Americans, through the identification of genomic regions derived from these populations.

South Amerindians have been considered to be a set of small groups or tribes in which genetic drift determined a high level of differentiation. In fact, the highest continental F_{ST} value is observed among autochthonous populations of South America (Cavalli-Sforza et al., 1994). During the last 10 years, we and others (Tarazona-Santos et al., 2001; Fuselli et al., 2003; Corella et al., 2007; Wang et al., 2007; Lewis and Long, 2008; Yang et al., 2010; Hunley and Healy, 2011) have shown that the pattern of genetic diversity in South America is more complex than previously thought. We proposed a model for the evolution of South Amerindian populations (Tarazona-Santos et al., 2001), which suggested contrasting regional patterns of genetic drift and gene flow: Western populations of South America, in particular those from the Quechua linguistic group in the Andes, show a combination of higher long-term effective sizes and gene flow than eastern populations, settled in Brazil. This has resulted in homogenization of the gene pool and the maintenance of genetic diversity in the Quechuas, in contrast to the divergence of populations and the recurrent loss of variants within the latter. As a consequence, the Quechua could act as a reservoir of the genetic variation of South Amerindians. Also, they are the most widespread linguistic group of the Americas, with more than 12 million speakers in Colombia, Ecuador, Peru, Bolivia, Chile and Argentina (Salzano and Bortolini, 2002).

The available resequencing public datasets are currently biased toward gene-centric regions of the human genome (Garrigan and Hammer, 2006; Wall et al., 2011). The initial pilot phase of the 1000 Genome Project included low pass (less than 4X coverage) and deep-coverage exomes (Consortium T1000 GP, 2010). Data from resequenced autosomal noncoding sequences (from now on called ANS) mapped far from known genes are particularly useful to make robust inferences about the evolutionary history of human populations, because of their following characteristics: (i) they are unlikely targets of natural selection (Wall et al., 2008); (ii) if several unlinked regions are resequenced, they provide independent information that is less subject to erratic results that may derive from the study of a single locus (Cox et al., 2009); (iii) data from targeted resequencing may be free of ascertainment bias, which is a limitation that affects SNPs genotyping studies regardless of their scale (Chikhi, 2008; Albrechtsen et al., 2010); (iv) SNP data from non-repetitive regions are the most common source of human genetic variation (Frazer et al., 2009). Finally, data addressing ANS are important to compare patterns of diversity with that of coding regions, particularly for the category of rare variants, which could account for a proportion of the "missing hereditability" of complex diseases.

In this article, we combined novel and published autosomal data to test a prediction of the evolutionary model by Tarazona-Santos et al. (2001, based on Y-chromosome) and Fuselli et al. (2003, based on mtDNA): that autosomal data from the Quechua population should also evidence high genetic diversity, independently of European/African admixture. We characterized the genetic diversity of the Quechua in a global setting using ANS (this study, Wall et al., 2008), SNPs (this study) and microsatellites (Wang et al., 2007), presenting the first resequencing autosomal noncoding data spread across the genome (nine loci comprising 16,465 bp per individual) for a Native American non-structured population, namely, a sample collected from a Quechua ($n = 11$) population from the Peruvian Central Andes. Respect to the study of the diversity of Native Americans, our sampling scheme, which reasonably approaches the sampling of a non-structured population, complements the sampling of Fagundes et al. (2007), who studied resequencing data for one individual from each of 10 different Amerindian populations. We also present new data on 351 unlinked gene-centric SNPs for the same populations.

MATERIALS AND METHODS

In this article, we studied Native American individuals from the farmer community of Tayacaja, Province of Huancavelica, in the Peruvian Central Andes, reported in Luiselli et al. (2000) and Tarazona-Santos et al. (2001). Samples were collected with informed consent and under approval of Ethics Committees of participating institutions.

Resequencing dataset

We bi-directionally resequenced the 10 autosomal noncoding loci from Frisse et al. (2001) from nine different autosomes. Each region includes a segment of ~1–2 kb at each end of a ~10-kb segment, a design Frisse et al. (2001) called locus-pair. We resequenced a total of 20,007 bp for each individual, including the 10 locus-pairs, except for region 6, which was resequenced only in the second half segment. We added to our sample ($n = 10$, from Tarazona-Santos et al., 2001) one Quechua individual from Arequipa, also from the Central Andean region of Peru, resequenced by Fagundes et al. (2007). The coordinates of the resequenced and analyzed regions, using as reference the build 37.2 of the Human Genome, and

the primers used for PCR and resequencing, are available in Supporting Information Table S1. Sequencing reactions were performed using the Big Dye terminator v.3.1. and run on automated capillary sequencers ABI 3130. Sequences were analyzed following the pipeline described in Machado et al. (2011). Only 1.3% of genotype calls in polymorphic nucleotide positions contained missing data. We compared our sequences with the nine overlapping regions resequenced by Wall et al. (2008) for 16 French Basque, 16 Han, 9 Melanesian, 14 Biaka Pygmy, 14 Mandenka and 9 San, all from the HGDP-CEPH panel, resulting in a dataset with seven populations aligned for 16,465 bp, the region for which we performed the population genetics analyses. The Quechua dataset has been submitted to GenBank under accession numbers JN813918 to JN814117.

SNP dataset

We report genotypes for 351 gene-centric unlinked SNPs (one for each of 351 genes) in 22 Quechua individuals and six samples from the same HGDP-CEPH populations studied by resequencing. These SNPs are part of a larger unpublished dataset generated using the Illumina GoldenGate Platform (San Diego) utilizing the Cancer SNP Panel (Illumina, San Diego). Supporting Information Table S2 presents detailed information about the 351 SNPs. These genotyped SNPs were included in this study after they passed through quality control (QC) procedures performed in the Genotype Library and Utilities software (GLU, v.1.0b1) that involved a larger database of 1198 individuals from 60 populations. The file with the genotypes for all individuals is available in Supporting Information Dataset S1.

Microsatellite dataset

We also analyzed the same populations from the dataset by Wang et al. (2007) for 678 autosomal microsatellites for the same six HGDP-CEPH populations and for 20 Quechua individuals from the Central Andean region of Peru (but a different population than our Tayacaja samples).

Whenever possible, the same individuals from the HGDP-CEPH panel were selected for the three datasets (Supporting Information Table S3 details the identification of the studied individuals). All three used datasets are also available for download in the DIVERGENOME bioinformatics platform (Magalhães et al. submitted; http://pggenetica.icb.ufmg.br/divergenome/pagina/index.php).

Admixture analyses for the Quechua population

We estimated admixture in 35 individuals from the Quechua population of Tayacaja using the Bayesian approach implemented in Structure software (Pritchard et al., 2000; Falush et al., 2003), using two datasets: (a) the SNPs dataset described above for 69 individuals, from the following populations: Mandenka (n = 23), French Basque (n = 24), and Quechuas (n = 22); and (b) the set of 48 Ancestry Informative Markers (AIMs, Santos et al., 2010) for 79 individuals from the following populations: African Americans (n = 24), Euro-descendants from USA (n = 31) and Quechuas (n = 24). We performed three runs assuming three clusters (K = 3), using the correlated allele frequencies and admixture models, with 100,000 iterations after a burn-in of length 10,000; and lambda set to 1.0 and α parameters to be estimated for each of the three clusters. We estimated the Spearman correlation between the results of the different runs with Hmisc package for R (Harrell, 2010). We estimated population admixture as the average of the individual admixtures, for which the 90% credibility intervals were estimated. If admixture for an individual was estimated using both datasets, we reported both admixture estimates, otherwise, only the result from one of the datasets was reported.

Statistical analyses

The intrapopulational genetic diversity was summarized by the nucleotide diversity (Tajima, 1983) and the Watterson θ estimator (Watterson, 1975) for the ANS data, and by the mean expected heterozygosity under Hardy-Weinberg equilibrium for microsatellites and SNPs. We used the Arlequin 3.5 software (Excoffier and Lischer, 2010) to do these estimates for the ANS and microsatellite data and the Genetics package for R environment (Warnes et al., 2003) for the SNP data. We also used the Arlequin 3.5 software to calculate the pairwise differences (Nei and Li, 1979) among individuals for the ANS data, and the F_{ST} (Weir and Cockerham, 1984; Michalakis and Excoffier, 1996) among populations for the ANS (with each SNP considered as a locus) and microsatellites. The polysat package for R was used to calculate a modified Lynch distance between individuals (Lynch, 1990; Clark and Jasieniuk, 2011) for the microsatellites. For the SNP data, the PLINK software (Purcell et al., 2007) was used to estimate the identity-by-state (IBS) similarity index between individuals, and the F_{ST} among populations was calculated with the adegenet package for R (Jombart, 2008). The genetic distance matrices between individuals and populations were summarized by a non-metric multidimensional scaling plot (MDS), with the Ecodist package in R (Goslee and Urban, 2007), which also estimated the stress, as a measure of goodness of fit of the configuration, and the total variance explained by the MDS configuration (r^2). The correlations between F_{ST} matrices of genetic distances were estimated by Mantel's test implemented in ape R package (Paradis et al., 2004). To test for deviations from a neutral model with constant population size in the ANS dataset, we estimated the observed and expected frequency spectra and the Tajima's D (Tajima, 1989) statistics using DNAsp v.5 software (Librado and Rozas, 2009). P values for Tajima's D were determined by 1000 coalescent simulations under constant population size, conditioning on the observed nucleotide diversity. All R commands used for the analyses are available in Text S1 to allow reproducibility of the results.

RESULTS AND DISCUSSION

Admixture analysis

Tayacaja, where the samples were collected, is a typical rural area in the Peruvian Central Andes where farming communities were settled before the arrival of the Spaniards in the sixteenth century. On the basis of genetic data, we estimated for the first time that European admixture in this population is less than 5% (Fig. 1) and African ancestry is almost null (~2%), which is consistent with the history of this region and our biodemographic studies that reported a low level of interethnic marriages in the last few centuries (Pettener et al., 1998; Luiselli et al., 2000). The distribution of this

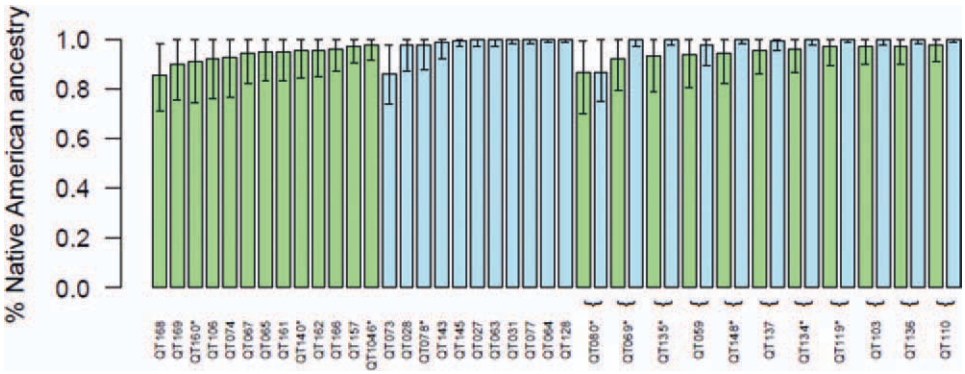


Fig. 1. Native American ancestry for Quechua individuals from Tayacaja and their 90% credibility intervals estimated using the software Structure and assuming that three populations contributed to their genetic structure: African, European, and Native American. These results are based on two panels: 351 unlinked SNPs for individuals of the SNPs dataset of this article (n = 22, blue bars) or on 48 Ancestry Informative Markers by Santos et al. (2010; n = 24, green bars). For 11 individuals, ancestry was estimated using both sets of markers, and in this case both estimates are reported. Individuals marked with an asterisk are those resequenced for ~20 kb of autosomal noncoding regions in this study. We ran the program Structure three times for each dataset. The point estimates as well as the lower and upper limits of their 90% credibility intervals were highly correlated [for pairwise comparisons between runs, for the 351 SNPs: the coefficient correlation r was always >0.94 and $P < 0.0001$; for AIMs from Santos et al. (2010), r was always > 0.88 and $P < 0.0001$ for Native American and European admixture; and r always >0.73 and $P < 0.0001$ for African admixture].

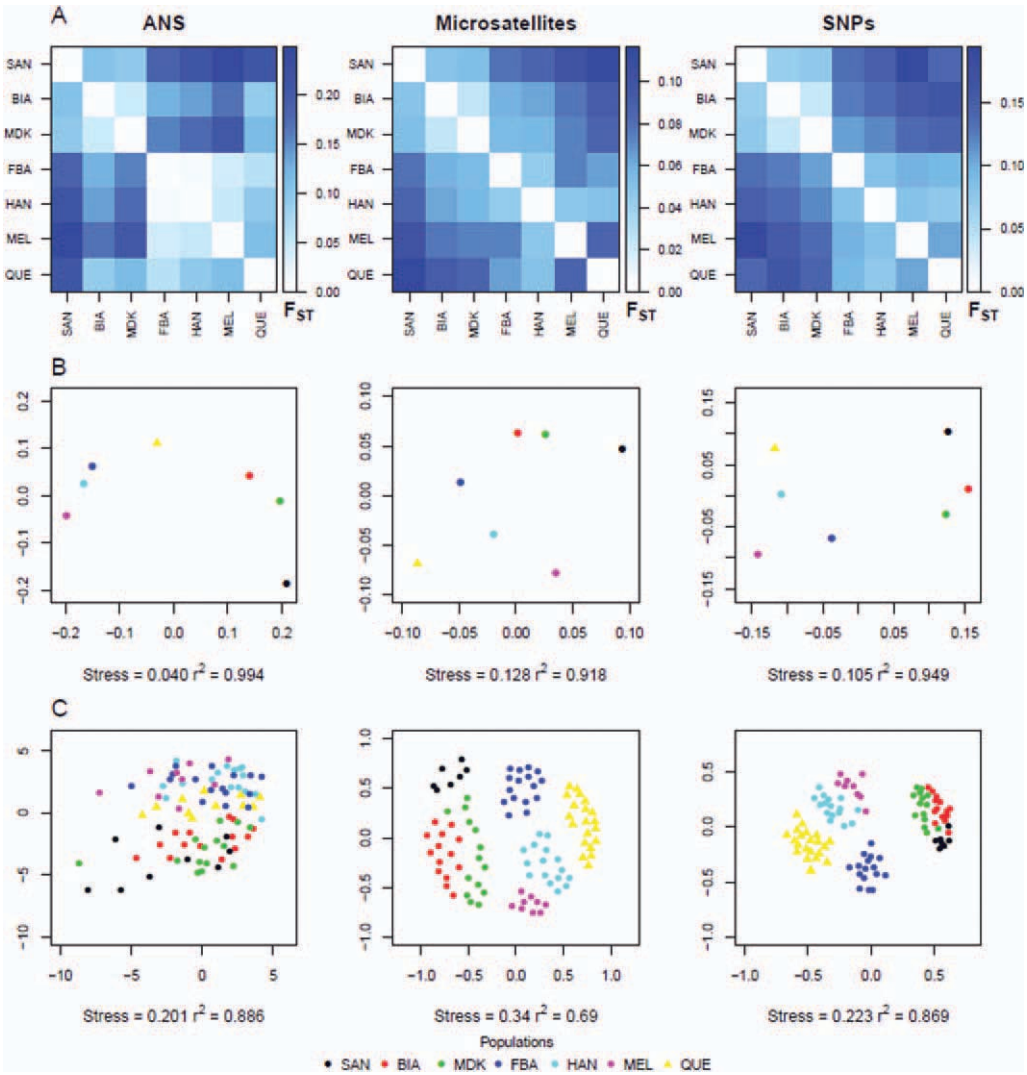


Fig. 3.

TABLE 1. Averages of summary statistics for autosomal noncoding sequences, SNPs, and microsatellites datasets

	ANS							SNPs		Microsatellites	
Population	S[a]	Sh[b]	P[c]	TD[d]	π in % (SE)	θ_W in % (SE)	Rank[e]	He[f]	Rank[e]	He[f]	Rank[e]
Quechua	47	–	4	0.169	0.087 (0.030)	0.078 (0.023)	4	0.284	3	0.671	6
Melanesian	44	38	6	**0.809**	0.092 (0.037)	0.077 (0.020)	5	0.279	4	0.670	7
Han	46	39	5	0.072	0.073 (0.029)	0.069 (0.012)	6	0.308	2	0.712	5
Basque	41	37	3	**0.671**	0.081 (0.022)	0.062 (0.014)	7	0.321	1	0.720	4
Biaka	71	33	21	-0.396	0.103 (0.043)	0.111 (0.030)	2	0.245	6	0.759	1
Mandenka	61	37	12	-0.266	0.102 (0.047)	0.095 (0.036)	3	0.259	5	0.753	2
San	64	31	21	0.088	0.118 (0.042)	0.113 (0.037)	1	0.211	7	0.740	3

Supporting Information Table S4 presents the results for each ANS locus.SE, standard errors.
[a] Number of segregating sites.
[b] Number of shared SNPs between Quechuas and each of the other populations.
[c] Number of private SNPs.
[d] Tajima's D statistic, significant values ($P < 0.05$) are in bold.
[e] Rank in ascending order of diversity. For ANS the rank is based on the θ_W values
[f] Mean expected heterozygosity over loci.

admixture in a sample of 35 Tayacaja individuals shows that (Fig. 1): European admixture is less than 5% in most individuals (83%), between 5% and 10% in 12% of the individuals and higher than 10% in only 5% of the studied individuals. This level of population admixture is very low compared with other Native American populations of farmers where admixture has been estimated using genetic data (Sans, 2000; Salzano and Bortolini, 2002). Still, robust estimates of individual admixture based on enough markers, such as those presented herein, are scanty in Amerindian populations (despite of Wang et al., 2007). In particular, the European admixture in our Quechua sample from the Andes seem to be lower than in most US and Canada Native American populations, and similar to Mexican native populations with comparable dimensions (Collins-Schramm et al., 2004; Wang et al., 2007; Bryc et al., 2010). The Quechua individuals in the Wang et al. (2007) survey using microsatellites also showed a higher level of European admixture than our sample (15%, Hunley and Healy, 2011). From our total sample of Tayacaja Quechua individuals, 10 samples (reported in Fig. 1 with asterisks) were selected for the resequencing experiment blindly, with respect to admixture results.

Intrapopulation diversity and allele frequency spectra

For comparisons, we analyzed three sets of genetic data in the Quechuas and in six worldwide samples from the HGDP-CEPH panel: French Basque, Chinese Han, Melanesians, and African Mandenka, San and Biaka. The three datasets include: (i) 16,465 bp of resequencing data from ANS distributed in nine independent loci (Frisse et al., 2001); (ii) 351 gene-centric unlinked SNPs; (iii) 678 microsatellite data from Wang et al. (2007). Detailed results for the new data generated in the datasets (i) and (ii) are available as Supporting Information.

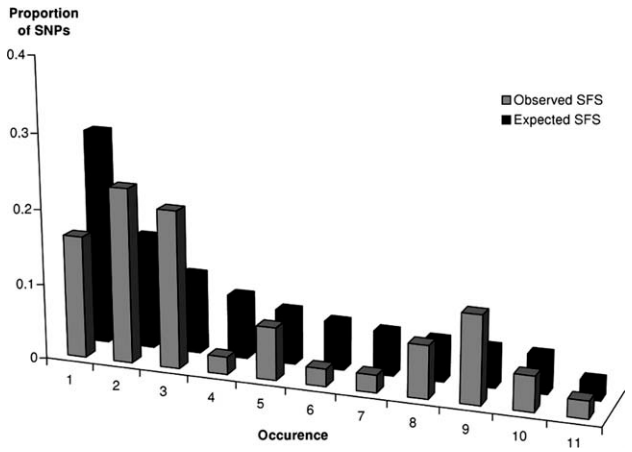


Fig. 2. Observed (gray) and expected (black) site frequency spectra (SFS) calculated from the autosomal noncoding sequences database for the Quechua population. Expected SFS is estimated under neutrality, constant-size, and panmixia.

Table 1 shows that, as expected, the ANS and microsatellites databases show a reduced genetic diversity in non-African relative to African populations, but the relative diversity of the Quechua is different among these two datasets. Quechua are less diverse than Eurasian populations for the microsatellites, but more diverse than Eurasian populations for the ANS. This result suggests that when resequencing data are analyzed, the diversity of some Native American populations could be higher than previously thought. Nonetheless, our results should be interpreted with caution, because the variance of our estimates is high due to the low number of independent loci that we analyzed. Also, while the Quechua is one of the most diverse Native American populations (Fuselli et al., 2003), the Basque is one of the least diverse European populations (Veer-

Fig. 3. Inter- and intra-population diversity plots. **A**: Graphical representation of the genetic distance matrix of seven populations based on pairwise F_{ST}. **B**: Multidimensional scaling plots (MDS) of seven populations based on pairwise F_{ST}. **C**: MDS plots of individuals from seven populations based on pairwise differences for autosomal noncoding sequences, Lynch distance for microsatellites and identity-by-state for SNPs. Populations: SAN, San/South Africa; BIA, Biaka/West Africa; MAN, Mandenka/West Africa; FBA, French Basque/Central Europe; HAN, Han/East Asia; MEL, Melanesian/Oceania; QUE, Quechua/West of South America. Datasets: ANS, Autosomal Noncoding Sequences.

amah et al., 2011). On the other hand, the SNPs dataset shows a pattern of diversity that is the opposite of that expected under the evolutionary history of human populations: the largest diversity among non-Africans and the lowest among Africans, which reflects a strong ascertained bias in the selection of these SNPs (Packer et al., 2006).The unbiased allele frequency spectrum for the studied populations obtained from ANS was summarized by the Tajima's D test and is presented in detail in Figure 2 for the Quechua. It fits the neutral expectation and does not contain evidence of past population expansions (i.e., excess of rare alleles) or bottlenecks (i.e., deficit of rare alleles). Although these results contrast with those obtained from mtDNA in other Quechua samples, which suggest a demographic expansion for these populations (Fuselli et al., 2003; Yang et al., 2010), these differences seem due to the fact that the patterns of diversity of these markers respond differently to the same demographic event, in part due to their different effective population sizes (Fay and Wu, 1999). The positive and significant values of the Tajima's D statistic for Basques and Melanesians, whose samples have similar sizes, suggest that the studied regions have enough power to detect important demographic events. Currently, after more than one thousand genome-wide association studies of complex diseases focused on common variants, there is an increasing interest in rare variants, which may explain the still "missing hereditability" accounting for complex diseases in human populations. Both the facts that the diversity revealed by ANS in the Quechua population is comparable with that observed in Eurasian populations, and that its allele frequency spectrum does not evidence a reduction in the proportion of rare alleles, suggest that the Quechua population is a reservoir of common and rare genetic variants of South Amerindians, and that it may be an interesting target for genetic association studies focused on rare variants. These results, obtained in the same Tayacaja population, are consistent with our evolutionary model of South Amerindian populations (Tarazona-Santos et al., 2001; Fuselli et al., 2003), which had the limitation of relying on Y-chromosome and mtDNA only. The model predicts that, because of the high level of gene flow between Andean populations and their long-term effective population size, they should bear high genomic diversity.

Within- and between-populations differentiation

Consistent with all previous studies of human genetic variation, we found that most genetic variation is shared among all populations. Indeed, the observed worldwide F_{ST} were 0.14 for ANS, 0.19 for SNPs and 0.06 for microsatellites; and the three matrix of pairwise genetic distances were highly correlated (Mantel's test significant for the three pairwise comparisons, pairwise Spearman correlations always > 0.62, P always < 0.02 obtained by a permutation test, Fig. 3A, and see Excoffier and Hamilton, 2003, for an explanation of the low values observed for microsatellites). The largest pairwise F_{ST} values were consistently observed between African and Quechua or Melanesian populations. These results agree with the African origin of modern humans and the fact that South America was the last part of the world to be peopled (Ramachandran et al., 2005). When we represented the matrices of pairwise F_{ST} by a non-metric Multidimensional Scaling (MDS, Fig. 3B), the three datasets consistently show that, in accordance with other studies (Li et al., 2008; Xing et al., 2010), the most evident differentiation is between Africans and non-Africans. For microsatellites and SNPs, the Quechua population showed increased differentiation in respect to Han, Basques, Melanesians and Africans, which is consistent with the Pleistocene Asian origin of Native American groups (Gonzalez-Jose et al., 2008). Although the pairwise F_{ST} values calculated from ANS data and its MDS representation do not reveal the expected higher similarity of the Quechua population to Han than to Basques (Fig. 3A–B), this may be due to the high similarity between Basques and Han samples ($F_{ST} = 0.014$).

Finally, we also used MDS (Fig. 3C) of the matrix of inter-individual pairwise distances as a representation of the power of our data to allocate individuals to the population where they proceeded from. We observed that 351 unlinked SNPs and 678 microsatellites allow us reasonably to discriminate individuals from different populations, including the Quechua; however, the ANS data (~16 kb in nine unlinked loci) do not allow it, probably because our ANS data include only 47 SNPs in Quechuas and 40–70 SNPs in the different worldwide populations. These results suggest that resequencing more regions is necessary to allocate individuals to their original populations. These results are in agreement with Witherspoon et al. (2007), who showed that although the proportion of human genetic variation due to differences between populations is low, and individuals from different populations can be genetically more similar than individuals from the same population, enough genetic data can allow an accurate classification of individuals into populations.

Admixture and Native American populations

An interesting issue is that Quechua populations are not restricted to rural areas of the Andean region. Large cities from Ecuador, Peru, Bolivia, and Northern Argentina host populations whose cultural identification as *mestizo* ignores their large Amerindian genetic background. These urban populations are frequently formed by immigrants arrived from rural areas. We estimated that the genetic contribution of Amerindians to a *mestizo* sample from the shantytown of Las Pampas in Lima is around 80% (Fuselli et al., 2007). These results suggest large cities of Western South America host millions of individuals with predominant Amerindian (likely Quechua) genetic background. Contemporary international South-to-North migrations are also spreading the genetic background of Quechuas and other Amerindian groups worldwide, and it is expected that the almost one million United States immigrants coming from Andean countries (U.S. Census Bureau, 2005–2009) have high levels of Quechua genetic background. It would not be surprising if these populations, classified as "Hispano/Latino" in the United States, had more Amerindian ancestry than US individuals classified as Native American.

We observed a low level of European admixture in our Quechua population. However, recent admixture is an issue to be considered with caution in inferences about the evolutionary history of Native Americans, as demonstrated by a study of European- and African-Americans, where recent admixture was explicitly included in the model (Nielsen et al., 2009), as well as in the recent

analysis by Hunley and Healy (2011) for Native Americans. The effect of European admixture is to reduce the F_{ST} between Native Americans and Europeans. Supporting Information Table S5 illustrates the expected changes in allele frequencies after 20 generations (~500 years) of continuous gene flow from Europeans to a hypothetical Native American population, assuming current F_{ST} values similar to that observed in this study between Quechuas and Europeans (0.06–0.10). These calculations are only illustrative, and are based on a simple island model of population structure [Eq. (9.1c) from Hedrick, 2005]. Because both Quechua and European populations are relatively large, we ignored changes in allele frequencies due to genetic drift during the last 20 generations. For the level of admixture observed in this study (<5%), changes in allele frequencies due to admixture are <0.02–0.03. For hypothetical admixture levels of 15% and 25%, changes in allele frequencies may be 0.04–0.05 and 0.07–0.10, respectively. Further studies and simulations are pending to understand how these changes affect evolutionary inferences about Native Americans that ignore admixture.

Considering admixture in evolutionary inferences about Native Americans is important in the context of the recent article by Wall et al. (2011), who suggested that admixed Latin Americans can provide information about the genetic structure of Amerindians through the identification of genomic regions derived from autochthonous populations. These authors studied a sample of 22 admixed Mexican-Americans from the NIEHS SNP initiative (Livingston et al., 2004), combining genomewide SNP data (Affymetrix 6.0 array) and resequenced data for 244 genes. They estimated local ancestry across each chromosome for each individual and identified genomic regions for which both homologous chromosomes had Amerindian ancestry, building a database of resequencing data for 163 gene-centric regions with Amerindian ancestry for six Mexican-American individuals. They then used these data to make evolutionary inferences about the time and mode of the first peopling of the Americas from Asia. However, this approach ignores the uncertainty of their ancestry estimates and therefore, their putative Native American dataset likely contains some European admixture (that we cannot assess because they do not report the cutoff they used to consider a diploid region as Native American). For instance, if they consider as Native American regions with >0.7 probability of a diploid genomic region to have this ancestry, it implies that in the dataset the probability that this region has at least one chromosome with European ancestry is 0.30. If the cutoff used by Wall et al. was low, the dataset contains a high level of admixture, and in subsequent evolutionary inferences, because European sequences are similar to Asian ones, this may explain the low values estimated for the time of separation (T) among the ancestral populations of current Native Americans and Asians. If admixture is properly modeled in evolutionary inferences, the approach by Wall et al. (2011) is a promising strategy to recover the genomic diversity of Native American populations that does not exist anymore, through the study of *mestizos* currently settled in the same geographic region. This approach has been named "homopatric targeting" by Gonçalves et al. (2010), and is suitable in Latin America because Wang et al. (2008) have showed that admixture predominantly involves Natives from the area where the *mestizos* are located.

CONCLUSIONS

In conclusion, our results, based on three types of autosomal markers, consistently suggest that the Quechua population, which is the largest Native American linguistic group from South America, although very differentiated in comparison to other worldwide populations, bears a considerable level of genetic diversity, as predicted by our evolutionary model for the evolution of South Amerindians (Tarazona-Santos et al., 2001; Fuselli et al., 2003). The combined study of Native American groups such as the Quechua and of admixed populations may generate large genetic datasets that may allow robust evolutionary inferences about the evolution of Native Americans, as well as understanding of the genetic architecture of complex diseases that are more common in these populations, such as type II diabetes (Winkler et al., 2010), rheumatoid arthritis (Daha et al., 2011), and lupus erythematosus (Borchers et al., 2010). We estimated a very low level of European admixture (<5%) and almost a null African admixture in our Quechua sample, a pattern that seems to be present even in some areas of large urban populations from Andean countries. Therefore, the Quechua population, being a large Native American population and a reservoir of the genetic diversity of Native Americans, is an excellent target both for evolutionary and genetic epidemiology studies.

ACKNOWLEDGMENTS

The authors are grateful to Sarah Lustigman and Sidney Santos for the genotyping of the Quechua samples with the AIMs; to Prof. Fabricio Santos's and Prof. Santuza Teixeira's laboratories for the access to DNA sequencers, to Prof. Michael Hammer for access to the worldwide dataset of ANS of Wall et al. (2008), and to the two reviewers who helped to improve the early version of the manuscript.

LITERATURE CITED

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27:2534–2547.

Borchers AT, Naguwa SM, Shoenfeld Y, Gershwin ME. 2010. The geoepidemiology of systemic lupus erythematosus. Autoimmun Rev 9:A277–A287.

Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H. 2010. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc Natl Acad Sci USA 107(Suppl 2):8954–8961.

Bustamante CD, Burchard EG, De la Vega FM. 2011. Genomics for the world. Nature 475:163–165.

Cann HM, de Toma C, Cazes L, Legrand M-F, 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.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton, NJ: Princeton University Press.

Chikhi L. 2008. Genetic markers: how accurate can genetic data be? Heredity 101:471–472.

Clark LV, Jasieniuk M. 2011. POLYSAT: an R package for polyploid microsatellite analysis. Mol Ecol Resour 11:562–566.

Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell L, Hanson RL, Knowler WC, Silva G, Belmont JW, Seldin MF. 2004. Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet 114:263–271.

Consortium T1000 GP. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Corella A, Bert F, Pérez-Pérez A, Gené M, Turbón D. 2007. Mitochondrial DNA diversity of the Amerindian populations living in the Andean Piedmont of Bolivia: Chimane, Moseten, Aymara and Quechua. Ann Hum Biol 34:34–55.

Cox MP, Morales DA, Woerner AE, Sozanski J, Wall JD, Hammer MF. 2009. Autosomal resequence data reveal Late Stone Age signals of population expansion in sub-Saharan African foraging and farming populations. PLoS One 4:e6366.

Daha NA, Willemze A, Robinson DB, Oen KG, Smolik I, Hart D, Ghidey W, Houwing-Duistermaat JJ, Siminovitch K, Huizinga TW, El-Gabalawy HS, Toes RE. 2011. Genetic interaction in the susceptibility of rheumatoid arthritis. Ann Rheum Dis 70:A84–A84.

Excoffier L, Hamilton G. 2003. Comment on "Genetic structure of human populations." Science 300:1877.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567.

Fagundes NJR, Kanitz R, Eckert R, Valls ACS, Bogo MR, Salzano FM, Smith DG, Silva WA, Zago MA, Ribeiro-dos-Santos AK, Santos SEB, Petzl-Erler ML, Bonatto SL. 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet 82:583–592.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA 104:17614–17619.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Fay JC, Wu CI. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol 16:1003–1005.

Frazer K, Ballinger DG, Cox DR, Hinds D, Stuve LL, Gibbs R, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler D, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Frazer K, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251.

Frisse L, Hudson RR, Bartoszewicz, Wall JD, Donfack J, Di Rienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69:831–843.

Fuselli S, Gilman RH, Chanock SJ, Bonatto SL, De Stefano G, Evans CA, Labuda D, Luiselli D, Salzano FM, Soto G, Vallejo G, Sajantila A, Pettener D, Tarazona-Santos E. 2007. Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. Pharmacogenomics J 7:144–152.

Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Mol Biol Evol 20:1682–1691.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nature reviews. Genetics 7:669–680.

González-José R, Bortolini MC, Santos FR, Bonatto SL. 2008. The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. Am J Phys Anthropol 137:175–187.

Gonçalves VF, Parra FC, Gonçalves-Dornelas H, Rodrigues-Carvalho C, Silva HP, Pena SD. 2010. Recovering mitochondrial DNA lineages of extinct Amerindian nations in extant homopatric Brazilian populations. Investig Genet 1:13.

Goslee SC, Urban DL. 2007. The ecodist package for dissimilarity-based analysis of ecological data. J Stat Soft 22.

Harrell FE. 2010.http://biostat.mc.vanderbilt.edu/trac/Hmisc.

Hedrick PW. 2005. Genetics of populations. Boston: Jones & Bartlett Publishers.

Hunley K, Healy M. 2011. The impact of founder effects, gene flow, and European admixture on Native American genetic diversity. Am J Phys Anthropol 1–9.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Lewis CM, Long JC. 2008. Native South American genetic structure and prehistory inferred from hierarchical modeling of mtDNA. Mol Biol Evol 25:478–486.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. 2004. Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831.

Long JC, Bortolini MC. 2011. New developments in the origins and evolution of Native American populations. Am J Phys Anthropol 1–4.

Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, Pettener D. 2000. Genetic structure of Quechua-speakers of the Central Andes and geographic patterns of gene frequencies in South Amerindian populations. Am J Phys Anthropol 113:5–17.

Lynch M. 1990. The similarity index and DNA fingerprinting. Mol Biol Evol 7:478–484.

Machado M, Magalhães WC, Sene A, Araújo B, Faria-Campos AC, Chanock SJ, Scott L, Oliveira G, Tarazona-Santos E, Rodrigues MR. 2011. Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies. Invest Genet 2:3.

Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064.

Neel JV. 1978. The population structure of an Amerindian tribe, the Yanomama. Annu Rev Genet 12:365–413.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG. 2009. Darwinian and demographic forces affecting human protein coding genes. Genome Res 19:838–849.

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 fre-

quency analysis for genetic variation in candidate genes. Nucleic Acids Res 34:D617–D621.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.
Pettener D, Pastor S, Tarazona-Santos E. 1998. Surnames and genetic structure of a high-altitude Quechua community from the Ichu River Valley, Peruvian Central Andes, 1825–1914. Hum Biol 70:865–887.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.
Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102:15942–15947.
Salzano FM, Bortolini MC. 2002. The evolution and genetics of Latin American populations. Cambridge: Cambridge University Press.
Salzano FM, Callegari-Jacques SM. 1988. South American Indians: a case study in evolution. Oxford: Clarendon Press.
Salzano FM, Neel JV, Weitkamp LR, Woodall JP. 1972. Serum proteins, hemoglobins, and erythrocyte enzymes of Brazilian Cayapo Indians. Hum Biol 44:443–458.
Sans M. 2000. Admixture studies in Latin America: from the 20th to the 21st century. Hum Biol 72:155–177.
Santos NPC, Ribeiro-Rodrigues EM, Ribeiro-Dos-Santos AKC, Pereira R, Gusmão L, Amorim A, Guerreiro JF, Zago MA, Matte C, Hutz MH, Santos SEB. 2010. Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. Hum Mutat 31:184–190.
Smouse PE, Long JC. 1992, Matrix correlation analysis in anthropology and genetics. Am J Phys Anthropol 35:187–213.
Spielman RS, Neel JY, Li FHF. 1977. Inbreeding estimation from population data: models, procedures, and implications. Genetics 85:355–371.
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.
Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, Rickards O, Tyler-Smith C, Pena SD, Santos FR. 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am J Hum Genet 68:1485–1496.
U.S. Census Bureau 2005–2009 http://factfinder.census.gov/servlet/DTTable?_bm=y&-geo_id=01000US&-ds_name=ACS_2009_1YR_G00_&-_lang=en&-mt_name=ACS_2009_1YR_G2000_B03001&-format=&-CONTEXT=dt.
Veeramah KR, Tönjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, Stumvoll M. 2011. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. Eur J Hum Genet 19:995–1001.
Wall JD, Cox MP, Mendez FL. 2008. A novel DNA sequence database for analyzing human demographic history. Genome Res 18:1354–1361.
Wall JD, Jiang R, Gignoux C, Chen GK, Eng C, Huntsman S, Marjoram P. 2011. Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. Mol Biol Evol 28:2231–2237.
Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K. 2007. Genetic variation and population structure in Native Americans. PLoS Genet 3:e185.
Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, Camrena B, Nicolini H, Klitz W, Barrantes R, Molina JA, Freimer NB, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Dipierri JE, Alfaro EL, Bailliet G, Bianchi NO, Llop E, Rothhammer F, Excoffier L, Ruiz-Linares A. 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet 4:e1000037.
Ward RH, Neel JV. 1970. Gene frequencies and microdifferentiation among the Makiritare Indians. IV. A comparison of a genetic network with ethnohistory and migration matrices; a new Index of Genetic Isolation. Am J Hum Genet 22:538–561.
Warnes GR, Gorjanc G, Leisch F MM. 2003. The genetics package: population genetics. R News 3:9–13.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276.
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Soc Stud Evol 38:1358–1370.
Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, Jorde LB. 2007. Genetic similarities within and between human populations. Genetics 176:351–359.
Winkler C, Nelson GW, Smith MW. 2010. Admixture mapping comes of age. Annu Rev Genomics Hum Genet 11:65–89.
Xing J, Watkins WS, Shlien A, Walker E, Huff CD, Witherspoon DJ, Zhang Y, Simonson TS, Weiss RB, Schiffman JD, Malkin D, Woodward SR, Jorde LB. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. Genomics 96:199–210.
Yang NN, Mazières S, Bravi C, Ray N, Wang S, Burley M-W, Bedoya G, Rojas W, Parra MV, Molina J, Gallo C, Poletti G, Hill K, Hurtado AM, Petzl-Erler ML, Tsuneto LT, Klitz W, Barrantes R, Llop E, Rothhammer F, Labuda D, Salzano FM, Bortolini M-C, Excoffier L, Dugoujon JM, Ruiz-Linares A. 2010. Contrasting patterns of nuclear and mtDNA diversity in Native American populations. Ann Hum Genet 1–14.

2.2 DEVELOPMENT OF TWO MULTIPLEX MINI-SEQUENCING PANELS OF ANCESTRY INFORMATIVE SNPS FOR STUDIES IN LATIN AMERICANS: AN APPLICATION TO POPULATIONS OF THE STATE OF MINAS GERAIS (BRAZIL)

2.2.1 RESUMO TRADUZIDO

A miscigenação ocorre quando indivíduos de populações parentais isoladas por centenas de gerações formam uma nova população híbrida. Atualmente, o interesse em medir a ancestralidade biogeográfica tem se ampliado: da antropologia às ciências forenses, à genômica personalizada mercantil e às questões legais das minorias, sendo imprescindível nos estudos de associação realizados em populações miscigenadas. Marcadores com frequências altamente diferenciadas entre populações humanas são informativos de ancestralidade e são denominados Marcadores Informativos de Ancestralidade (MIAs). Para as populações latino-americanas tri-híbridas é necessária a informação sobre a ancestralidade africana, europeia e nativo-americana. Foram desenvolvidos dois painéis multiplex de AIMs com o intuito de estimar a miscigenação das populações latino-americanas. Estes AIMs são genotipados por duas reações de minisequenciamento sendo apropriados a pesquisadores de laboratórios de pequeno e médio porte. O desempenho deste painel foi testado comparando os resultados aos obtidos a partir de 108 AIMs nos mesmos indivíduos para os quais o DNA está disponível a outros pesquisadores. Enfatiza-se que a comparação sempre deve ser realizada ao longo do processo de desenvolvimento de um novo painel de ancestralidade. A nível populacional, os 14 MIAs São úteis para estimar a ancestralidade europeia, ainda que sobrestimem a miscigenação africana e subestime a nativo-americana. Combinado com mais MIAs, o painel desenvolvido pode ser utilizado para inferir os níveis de ancestralidade. Os painéis foram utilizados para estimar os padrões de miscigenação em duas populações urbanas do estado de Minas Gerais no sudeste do Brasil: Montes Claros e Manhuaçu. A partir destes dados obtivemos um perfil atual da estrutura genética no contexto da história demográfica destas cidades.

2.2.2 ATIVIDADES REALIZADAS

As atividades realizadas neste trabalho são compartilhadas com as do artigo **Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers**, por isso, optou-se por relatá-las uma única vez, na próxima seção.

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: an application to populations of the State of Minas Gerais (Brazil)

M.C.F. Silva[1,2*], L.W. Zuccherato[2*], G.B. Soares-Souza[1,2], Z.M. Vieira[1], L. Cabrera[3], P. Herrera[3], J. Balqui[3,4], C. Romero[3,4], H. Jahuira[3,4], R.H. Gilman[3,5], M.L. Martins[1] and E. Tarazona-Santos[2]

[1]Fundação Hemominas, Belo Horizonte, MG, Brasil
[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[3]Asociación Benéfica PRISMA, Urbanización Maranga, Lima, Peru
[4]Laboratorio de Investigación en Enfermedades Infecciosas, Universidad Peruana Cayetano Heredia, San Martín de Porres, Lima, Peru
[5]Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

*These authors contributed equally to this study.
Corresponding author: E. Tarazona-Santos
E-mail: edutars@icb.ufmg.br

Genet. Mol. Res. 9 (4): 2069-2085 (2010)
Received May 21, 2010
Accepted July 23, 2010
Published October 19, 2010
DOI 10.4238/vol9-4gmr911

ABSTRACT. Admixture occurs when individuals from parental populations that have been isolated for hundreds of generations form a new hybrid population. Currently, interest in measuring biogeographic ancestry has spread from anthropology to forensic sciences, direct-to-consumers personal genomics, and civil rights issues of minorities, and it is critical for genetic epidemiology studies of admixed populations. Markers with highly differentiated frequencies among human populations are informative of ancestry and are called ancestry informative markers (AIMs). For tri-hybrid Latin American populations, ancestry information is required for Africans,

Europeans and Native Americans. We developed two multiplex panels of AIMs (for 14 SNPs) to be genotyped by two mini-sequencing reactions, suitable for investigators of medium-small laboratories to estimate admixture of Latin American populations. We tested the performance of these AIMs by comparing results obtained with our 14 AIMs with those obtained using 108 AIMs genotyped in the same individuals, for which DNA samples is available for other investigators. We emphasize that this type of comparison should be made when new admixture/population structure panels are developed. At the population level, our 14 AIMs were useful to estimate European admixture, though they overestimated African admixture and underestimated Native American admixture. Combined with more AIMs, our panel could be used to infer individual admixture. We used our panel to infer the pattern of admixture in two urban populations (Montes Claros and Manhuaçu) of the State of Minas Gerais (southeastern Brazil), obtaining a snapshot of their genetic structure in the context of their demographic history.

Key words: Admixture; Latin American; Mini-sequencing

INTRODUCTION

Admixture occurs when individuals from populations that have been isolated for hundreds of generations (i.e., parental populations) form a new hybrid population. Latin-Americans (Hispanics/Latino in the United States), African-Americans and Caribbeans in the Americas (Sans, 2000; Salzano and Bortolini, 2002; Benn-Torres et al., 2008; Herrera-Paz et al., 2010), Central Asian (Comas et al., 1998) and South African colored (Patterson et al., 2010) are examples of admixed human populations. Chromosomes of admixed individuals may be conceived as mosaics of chunks with different ancestry, which sizes reduce along time by recombination among chromosomes with different ancestry (Falush et al., 2003). The recent availability of millions of markers across the human genome for different populations has made it possible to infer admixture (also called biogeographic ancestry) not only for populations, as has been traditional, but also for individuals and for specific genomic regions along a chromosome (Via et al., 2009).

The interest in biogeographic ancestry estimations is not limited to anthropology anymore; it has spread to forensic sciences, direct-to-consumers personal genomics and civil rights issues of minorities (Lee et al., 2009). Estimating biogeographic ancestry is critical for genetic epidemiology of admixed populations (Tarazona-Santos et al., 2007). In fact, in a well-designed case-control association study, cases and controls should be sampled from the same population and thereby be ethnically homogeneous. Otherwise, spurious statistical association for any allele more common in a parental population may result if the disease is more prevalent in this population and therefore, individuals with a predominant ancestry of this population are oversampled among cases. Hence, the first step in a case-control association study in an admixed population should be to measure admixture of the participants and ascertain if cases and controls are ethnically different (i.e., if population stratification exists). In this case, it is possible to test the statistical association among genetic variants and biomedical traits controlling for population stratification at the population level using genomic control methods, or at the individual level using regression analysis or structured association methods (Tarazona-Santos et al., 2007).

In this context of broad interest in admixture studies, it is important to recognize that measuring admixture is methodologically complex (Chakraborty, 1986; Choisy et al., 2004). For instance, difficulties in estimating admixture increase from population to individual and chromosomal levels (Pritchard et al., 2000; Falush et al., 2003). Another complicating factor is the number of parental populations that have contributed to the gene pool of the admixed population/individual. Markers with frequencies that are highly differentiated among populations are particularly informative of ancestry and are called ancestry informative markers (AIMs). If enough AIMs are genotyped, they allow estimates of admixture at the levels of population, individual and chromosome regions. The use of AIMs reduces the number of markers that need to be genotyped to infer admixture at population and individual levels, compared to the genotyping of randomly selected markers (Rosenberg et al., 2003; Parra et al., 2003; Pfaff et al., 2004). For tri-hybrid Latin American populations, ancestry information is required for African, European and Native American populations. However, there is no unique and optimal set of markers (or AIMs) for all Latin American populations, because informativeness depends on the combination of allele frequencies in the parental populations and on admixture proportions (Pfaff et al., 2004). In general, for Latin American tri-hybrid populations, the best AIMs have a very different frequency in one of the three parental populations and similar frequencies in the other two.

The number of markers that are necessary to estimate population admixture or individual ancestry depends on the informativeness of the markers and the required accuracy. Currently, Affymetrix and Illumina commercial arrays allow to genotype up to ~10^6 markers scattered in the human genome, at a cost of few hundreds of US dollars per individual (Chung et al., 2010). Even if most of these single-nucleotide polymorphisms (SNPs) are not AIMs, with this resolution it is possible to estimate individual and chromosomal region ancestries with high accuracy. Although the cost of genome-wide genotyping is declining, this possibility is still limited to a few research groups, in particular for admixture studies at the population level, which require large sample sizes. For small-medium laboratories, it would be advantageous to use small-medium size panels of AIMs for studies at population and individual levels, at a cost of few dozens of dollars per individual. Kosoy et al. (2009) have shown that panels of 24 AIMs are useful to ascertain the origin of subjects from particular continents and to correct for population stratification at the population level. Some low-medium cost multiplex panels of AIMs have been published (Lins et al., 2010; Santos et al., 2010); however, these may vary in their informativeness. Therefore, the best option for an investigator performing an admixture study is to assess which combination of AIMs is most informative for the target population.

We developed two multiplex panels of AIMs that include 14 SNPs to estimate admixture in Latin American populations. We tested the performance of these 14 AIMs by comparing admixture estimates obtained with this set of markers, with estimates obtained using 108 AIMs (that we assume to be more accurate). We used our panels to infer the pattern of admixture in two populations of the State of Minas Gerais (southeastern Brazil), obtaining a snapshot of their genetic structure in the context of their demographic history.

MATERIAL AND METHODS

Selection of AIMs for the two panels

To design two panels of AIMs to be genotyped by multiplex mini-sequencing reac-

tions, we pre-selected a large set of candidate AIMs by two procedures: 1) 250 unlinked AIMs were selected based on their informativeness (index I_a of Rosenberg et al., 2003) from the admixture mapping panel of Tian et al. (2006) to assess African/European admixture. 2) 150 SNPs informative of Native American admixture were selected from the SNP500 Cancer resource (Packer et al., 2006, http://variantgps.nci.nih.gov/cgfseq/pages/snp500.do), based on differences in allele frequencies between European, African and Pima-Maya Native American populations. These SNPs were pre-selected by avoiding physical proximity in the human genome. We assessed compatibility for multiplex polymerase chain reaction (PCR) amplification using the Muplex resource (Rachlin et al., 2005), which is a convenient web-enabled system that, starting from a set of targeted sequences, automatically designs sub-sets of primers that will likely co-amplify in multiplex PCR assays under a number of conditions imposed by the investigator. After applying these criteria, we selected the following two SNP panels to be tested experimentally: AFR (Africans) (rs2697520, rs8035530, rs1372115, rs2789823, rs241679, rs7512316, rs9626698, rs1443985, rs6046024, rs735480) and AMR (Native Americans) (rs8058694, rs691968, rs2234636, rs3760657, rs2619681, rs2569190, rs800292, rs2518967, rs2088102, rs700518). We also evaluated the specificity of primers using the electronic PCR tool (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/reverse.cgi). Among these SNPs, the following were excluded for further genotyping because of their high rate of missing data or because of a lack of reproducible results: rs2789823, rs7512316, rs6046024, rs2569190, rs2518967, and rs700518.

Genotyping

Genotyping by mini-sequencing consists of three steps (Carvalho and Pena, 2005): 1) amplification of regions flanking the SNPs by multiplex PCR; 2) multiplex mini-sequencing; 3) analysis of mini-sequencing products by capillary electrophoresis.

Amplification of regions flanking the SNPs by multiplex PCR

Primers designed using Muplex performed well when experimentally tested and are available as Supplementary Material (Table S1). PCR was performed in a volume of 25 μL with 100 ng genomic DNA, 0.2 μM of each primer and 1X of a commercial master-mix (Qiagen Multiplex PCR Master Mix or 1.5 U Platinum Taq DNA Polymerase from Invitrogen plus 1X STR buffer from Promega). Amplification consisted of 95°C for 5 min, followed by 30 cycles of 30 s at 94°C, 90 s at 57°C, 90 s at 72°C, and a final extension for 10 min at 72°C. After the amplification, we performed enzymatic purification of the PCR product (i.e., removal of remaining PCR primers and dNTPs before the mini-sequencing reaction, using respectively exonuclease I and shrimp alkaline phosphatase, as detailed in the Supplementary Material).

Multiplex mini-sequencing

For each locus of a multiplex panel, a mini-sequencing primer with the 3'-end adjacent to the target SNP was designed to anneal with the PCR product (Figure 1B). A mini-sequencing reaction extends this primer, producing different products for each allele, which

are visualized by capillary electrophoresis. Different loci are differentiated by the sizes of the mini-sequencing primers, which include PUC 18 plasmid sequence tails with a specific size for each locus. The mini-sequencing primers are available as Supplementary Material (Table S1). We selected the SNPs for these assays to allow genotyping both by: a) monochromatic mini-sequencing (Carvalho and Pena, 2005) and b) polychromatic mini-sequencing. Monochromatic mini-sequencing uses a homemade mix with one of the nucleotides in the form of fluorescent ddNTP (in our case ddATP-FAM) and the other three as dNTPs (Figure 1C). In this way, it is only possible to genotype SNPs with an allele complementary to the ddNTP (i.e., T in our case), in which case the primer is extended by a single ddNTP. The mini-sequencing product for the other allele will be extended until the next position containing the same nucleotide (i.e., T in our case) in the sequence, and therefore alleles are differentiated by size (Figure 1E). Polychromatic mini-sequencing does not use dNTPs; it only uses four ddNTPs with different fluorescence (and therefore, primers will always be extended by a single ddNTP; Figure 1D). Thus, alleles will be distinguished by colors (Figure 1F). The commercial kits SNaPshot (Applied Biosystems) and SNuPe (GE Healthcare) are available for polychromatic mini-sequencing. To use the same set of mini-sequencing primers in monochromatic or polychromatic protocols, we avoided combinations of A/G and C/T polymorphisms in the same multiplex reaction, since these variants cannot be co-genotyped by the monochromatic protocol, an option for investigators who prefer to use homemade reagents.

We performed the multiplex monochromatic mini-sequencing in a 13.5-µL volume with 2 µL purified PCR product and 0.37 µM of each primer, 0.46 µM ddATP labeled with fluorescein (Perkin Elmer Life Sciences), 0.46 µM of unlabeled dCTP, dTTP and dGTP (GE Healthcare), 1X Thermo Sequenase reaction buffer, and 1 U Thermo Sequenase DNA Polymerase (GE Healthcare). The thermal cycling consisted of 2 min at 80°C for denaturation, followed by 30 cycles of 30 s at 95°C, 30 s at 55°C and 20 s at 72°C. The polychromatic protocol (SnaPshot Multiplex System, Applied Biosystems) consisted of a 5-µL volume reaction containing 1 µL purified PCR product, 1 µL SNaPshot™ Kit Reaction Mix and 2 µL primer mix (at a concentration of 1 µM of each primer). The thermal cycling consisted of 2 min at 96°C for denaturation, followed by 25 cycles of 10 s at 95°C, 5 s at 55°C and 30 s at 60°C. After the mini-sequencing reaction, we performed an enzymatic purification of the reaction (see Supplementary Material for details).

A*nalysis of the mini-sequencing products by capillary electrophoresis*

For the monochromatic protocol, a mixture of 2.0 µL mini-sequencing products diluted twice, plus 7.75 µL Tween 20 at 0.1% and 0.25 µL of the size standard ET-ROX 550 (GE Heathcare) was applied in a Megabace DNA sequencer (GE Heathcare). The run parameters were: injection voltage of 3 Kv, injection time of 80 s, run voltage of 10 Kv and run time of 75 min. The analyses were done with the Fragment Profiler software (GE Heathcare). For the polychromatic option, a mixture of 1.0 µL of the SNaPshot product, 8.9 µL Hi-Di formamide and 0.1 µL Liz120 Size Standard (Applied Biosystems) was applied in an ABI 3130 DNA sequencer (Applied Biosystems). The run parameters were: injection voltage of 1.2 Kv, injection time of 18 s, run voltage of 15 Kv and run time of 800 s (capillary size of 36 cm). In this case the analyses were done with the package GeneScan Analysis 3.7 or Genotyper 3.7 software (Applied Biosystems).

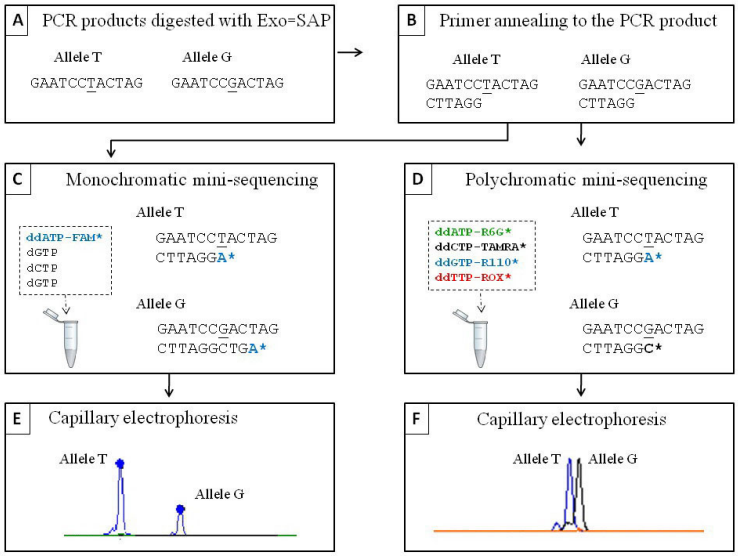


Figure 1. Representation of genotyping by monochromatic (A,B,C,E) or polychromatic (A,B,D,F) mini-sequencing. See text for details. PCR = polymerase chain reaction; Exo-SAP = a reaction containing *E. coli* exonuclease I + shrimp alkaline phosphatase.

Samples of parental and admixed populations

We used the following three sets of individuals as putative parental populations of the Latin American admixed samples: 1) 31 European ancestry and 2) 24 African ancestry from the SNP500Cancer panel (http://variantgps.nci.nih.gov/cgfseq/pages/snp500.do; Packer et al., 2006). We also used 3) 85 Peruvian Native Americans settled between the eastern slope of the Andes and the Amazon tropical forest (in the region called High Forest or "Selva Alta"). Some of these individuals are from the region of Cusco and belong to the communities of Shimaa (N = 30) and Monte Carmelo (N = 15) from the Matsiguenga linguistic group, and some of them (N = 40) reside in Ashaninka villages along the Tambo River (Region of Junin).

Admixed samples included three sets of individuals: 1) 23 Latin American catalogued as Hispanic in the SNP500Cancer initiative, 2) 24 Brazilian individuals from the city of Montes Claros, at north of the State of Minas Gerais, and 3) 30 Brazilian individuals from the city of Manhuaçu, eastern Minas Gerais. Brazilian samples were from healthy and unrelated blood donors attending centers of the Minas Gerais Blood Bank in their respective cities. The inclusion of European, African and Latin American individuals from the SNP500Cancer initiative is convenient because they have been genotyped for a large set of polymorphisms in the context of the SNP500Cancer initiative, and this information can be used to assess the informativeness of the panels of AIMs that we developed. Institutional Review Boards from the participant institutions approved this study.

Statistical and population genetics analyses

We estimated population admixture using: 1) The gene identity method developed by Chakraborty (1985), as implemented in the ADMIX95 software (developed by Bernardo Bertoni and available at http://www.genetica.fmed.edu.uy/software.htm). This method takes into account sampling error and the effect of genetic drift in the parental and admixed populations (Chakraborty, 1986) and 2) The

coalescent-based method by Dupanloup and Bertorelle (2001), which, in addition to sampling and drift errors in parental and admixed populations, considers the degree of divergence at the time of admixture.

Individual admixture was estimated using the Bayesian clustering algorithms developed by Pritchard and implemented in the STRUCTURE v2.3.2 program (Pritchard et al., 2000; Hubisz et al., 2009). We assumed that three parental populations (K = 3 clusters) contributed to the genome of the admixed individuals. STRUCTURE estimates individual admixture conditioning in Hardy-Weinberg and linkage equilibrium in each of the K = 3 clusters, which represent the parental populations. We ran the program using a burn-in period of 100,000, and 100,000 repetitions of MCMC after burning. We used prior population information for individuals from the parental populations to assist clustering (USEPOPINFO = 1) and assumed the admixture model for individuals from the admixed populations, inferring the alpha parameter for each population. We also used the parameters GENSBACK = 2 and MIGRPRIOR = 0.05. Moreover, we assumed that allele frequencies were correlated and that the different populations have different levels of differentiation (F_{ST}, with prior mean = 0.01 and standard deviation = 0.05). Based on individual admixture estimations obtained with STRUCTURE, population admixture was averaged over individuals.

RESULTS

The allele frequencies for the 14 AIMs of this study in European, African and Native American populations (for which public data are available), as well as for the parental and admixed samples, are given in Table 1. In general, there were large differences between continental groups and small differences within these groups, which confirm that the selected markers are informative for ancestry.

Table 1. Frequencies of the genotyped single-nucleotide polymorphisms (SNPs) for the different populations.

SNP	Population																
	AFR1	AFR	YRI	NiloS	Edo	AFA	EUR1	EUR	CEU	DEU	AMR	MCA	ASH	SHI	MCU	MOC	HISP
AFR panel																	
rs1443985	0.15	0.09	0.07	0.11	0.13	0.21	0.85	0.91	0.90	0.94	0.84	0.85	0.76	0.92	0.71	0.57	0.71
rs9626698	0.50	0.79	0.79	0.76	0.81	0.57	0.00	0.04	0.03	0.04	0.00	0.00	0.00	0.00	0.12	0.17	0.12
rs2416791	0.76	0.05	0.03	0.10	0.00	0.18	0.88	0.92	0.92	0.92	0.53	0.64	0.44	0.50	0.71	0.57	-
rs2697520	0.17	0.07	0.05	0.10	0.06	0.28	0.80	0.88	0.89	0.85	0.31	0.31	0.26	0.34	0.36	0.50	0.50
rs8035530	0.70	0.82	0.79	0.86	0.79	0.67	0.08	0.02	0.02	0.02	0.48	0.44	0.53	0.48	0.18	0.25	0.19
rs735480	0.09	0.03	0.00	0.06	0.02	0.21	0.90	0.95	0.95	0.96	0.88	0.82	0.99	0.84	0.82	0.58	0.86
rs1372115	1.00	0.99	1.00	0.97	1.00	0.79	0.37	0.07	0.17	0.23	0.41	0.07	0.62	0.54	0.43	0.56	0.92
AMR panel																	
rs2234636	0.20	0.00	-	-	-	-	0.28	0.28	-	-	0.78	0.77	0.73	0.83	0.35	0.21	0.37
rs3760657	0.09	0.02	-	-	-	-	0.10	0.07	-	-	0.26	0.12	0.40	0.26	0.03	0.06	0.07
rs8058694	0.59	0.23	-	-	-	-	0.88	0.61	-	-	0.70	0.62	0.75	0.72	0.37	0.54	0.39
rs800292	0.63	0.63	-	-	-	-	0.25	0.18	-	-	0.93	1.00	0.87	0.91	0.35	0.52	0.35
rs2619681	0.17	0.06	-	-	-	-	0.13	0.17	-	-	0.86	0.88	0.90	0.79	0.15	0.15	0.26
rs691968	0.41	0.42	-	-	-	-	0.04	0.00	-	-	0.01	0.00	0.00	0.02	0.10	0.09	0.17
rs2088102	0.71	0.65	-	-	-	-	0.54	0.50	-	-	0.33	0.27	0.40	0.33	0.98	0.89	0.34

AFR1 = Africans (SNP500Cancer panel, genotyped in this study); AFR = Africans (Tian et al., 2006); YRI = Yoruban (West Africans from the HapMap project); NiloS = Kanuri (Nilo-Saharan speakers from Nigeria); Edo = Bini (Niger-Congo group of Bantu speakers); AFA = African-Americans (Coriell Institute for Medical Research); EUR1 = European (SNP500Cancer panel, genotyped in this study); EUR = Europeans (Tian et al., 2006); CEU = CEPH European; DEU = European-Americans from New York City; AMR = Average for all Amerindians (genotyped in this study); MCA = Monte Carmelo Amerindians; ASH = Ashaninka Amerindians; SHI = Shimaa Amerindians; HIS = Hispanic (SNP500Cancer panel, genotyped in this study); MCU and MOC = Brazilian samples from Manhuaçu and Montes Claros, respectively. The data presented for Africans (AFR, YRI, NiloS, Edo, AFA) and Europeans (EUR, CEU, DEU) were obtained from Tian et al., 2006 (AFR panel) or SNP500Cancer database (AMR panel).

Population admixture

First, we tested our set of 13 AIMs by estimating population admixture for the Latin American sample of the SNP500Cancer project, hereafter called Hispanics to follow the SNP500Cancer nomenclature. Hughes et al. (2008), using ~350 K SNPs, have shown the predominant European ancestry of this set of individuals. Although our set of AIMs includes 14 SNPs, in the analysis including Hispanics we conservatively considered only 13 AIMs, due to the high missing rate of rs241679, specifically in the Hispanic sample. We compared our estimates with those obtained using 108 AIMs, selected for admixture estimation among thousands of SNPs genotyped in the SNP500Cancer project using the criterion that they show $F_{ST} > 0.20$ between European, African and Native American (Pima and Maya populations from CEPH; Cann et al., 2002) and $F_{ST} < 0.10$ between populations within these groups (unpublished data; see Table S2 of the Supplementary Material, for additional details). Individual genotypes and allele frequencies for these 108 SNPs in the parental populations are available in the SNP500Cancer website. We assume that these 108 AIMs provide a more accurate estimate of admixture than our 13 AIMs. Both sets of 108 and 13 AIMs confirm the predominant European ancestry of the Hispanic sample. By using the 13 AIMs in a sample with predominant European ancestry, we obtained accurate estimates of European admixture at the population level (Table 2). However, our set of 13 AIMs seems to overestimate African and underestimate Native American admixture. The methods developed by Chakraborty (1985) and Dupanloup and Bertorelle (2001) to infer population admixture, consistently estimate higher African admixture and lower European admixture than the STRUCTURE method (focused on individual admixture) (Table 1).

Table 2. Population admixture estimation obtained by three methods of analysis for Hispanic and Brazilian samples from Montes Claros and Manhuaçu, Minas Gerais.

	Parental populations					
	African		European		Native American	
	Point estimate	95% CI or SD	Point estimate	95% CI or SD	Point estimate	95% CI or SD
Dupanloup and Bertorelle (2001)						
Hispanic-13 SNPs	0.31	0.05	0.62	0.06	0.07	0.05
Hispanic-108 SNPs	0.15	0.03	0.66	0.03	0.20	0.02
Montes Claros-BR	0.41	0.05	0.54	0.07	0.05	0.06
Manhuaçu-BR	0.27	0.04	0.63	0.06	0.11	0.06
Chakraborty (1985)						
Hispanic-13 SNPs	0.34	0.02	0.58	0.02	0.08	0.02
Hispanic-108 SNPs	0.16	0.00	0.64	0.00	0.20	<0.01
Montes Claros-BR	0.41	0.00	0.54	0.00	0.05	<0.01
Manhuaçu-BR	0.27	0.01	0.63	0.01	0.09	0.01
Pritchard et al. (2000)						
Hispanic-13 SNPs	0.23	-	0.69	-	0.08	-
Hispanic-108 SNPs	0.09	-	0.75	-	0.16	-
Montes Claros-BR	0.39	-	0.52	-	0.09	-
Manhuaçu-BR	0.19	-	0.73	-	0.08	-

CI = confidence interval; SD = standard deviation; SNPs = single-nucleotide polymorphisms.

After testing the performance of our 14 AIMs to estimate population admixture, we interpreted admixture estimation in the Brazilian Minas Gerais samples of Montes Claros and Manhuaçu. Both populations have a predominant European admixture (>50%). On the basis of our test with the Hispanic sample, we consider estimates of African and Native American ancestry for the Minas Gerais samples as maximum and minimum values, respectively. In agreement with its geographical proximity to northern Bahia State, the Montes Claros population showed a higher African contribution than Manhuaçu.

Individual admixture

The 14 AIMs used in this study correctly assigned the individuals from the parental populations to their actual group, even when no information was given to the STRUCTURE algorithm about their population of origin (i.e., when we ran the program using all the parameters specified in the Material and Methods section, but specifying USEPOPINFO = 0; Figure S1 of the Supplementary Material). This implies that individuals with ancestry near 100% for any of the parental populations are correctly identified by our small panel of AIMs. However, this does not imply that admixture will be accurately estimated in more admixed individuals.

To assess the accuracy of individual admixture inferences using the 14 AIMs, we ran the STRUCTURE software in the Hispanic sample by using the set of 108 AIMs (Figure 2), and compared these results with those obtained by subsets of 80, 50 and 20 randomly selected SNPs, and also with our set of 13 AIMs (Figure 3). Using subsets of 50 or 80 AIMs, the correlation of admixture estimations was higher than 85%, while with fewer than 20 AIMs, this correlation was less than 66% (Table 3). Thus, even if our set of 13 AIMs contains information to estimate population admixture, individual admixture inferences for admixed individuals are not accurate with less than 20 AIMs. This is also evident when the Bayesian 90% probability interval (90% PI) of individual admixture generated by the STRUCTURE software is plotted against individual admixture (Figure 3); for admixed individuals, the 90% PI of individual admixture rises dramatically when reducing the number of AIMs below 50.

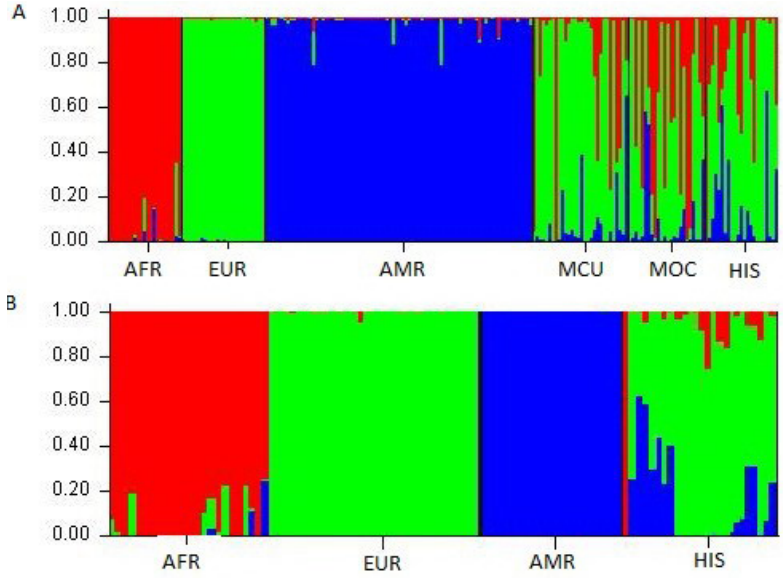


Figure 2. Individual admixture estimated by STRUCTURE with 14 ancestry informative markers (AIMs) for this study (A) and for the Hispanics using 108 AIMs (B). Each individual is represented by a vertical bar. Inferred African, European and Native American admixture is represented by red, green and blue, respectively. AFR = Africans; EUR = Europeans; AMR = Native Americans; MCU = Manhuaçu; MOC = Montes Claros; HIS = Hispanics from SNP500Cancer. Admixture estimates obtained using 108 AIMs considered only Monte Carmelo as Native American parental population.

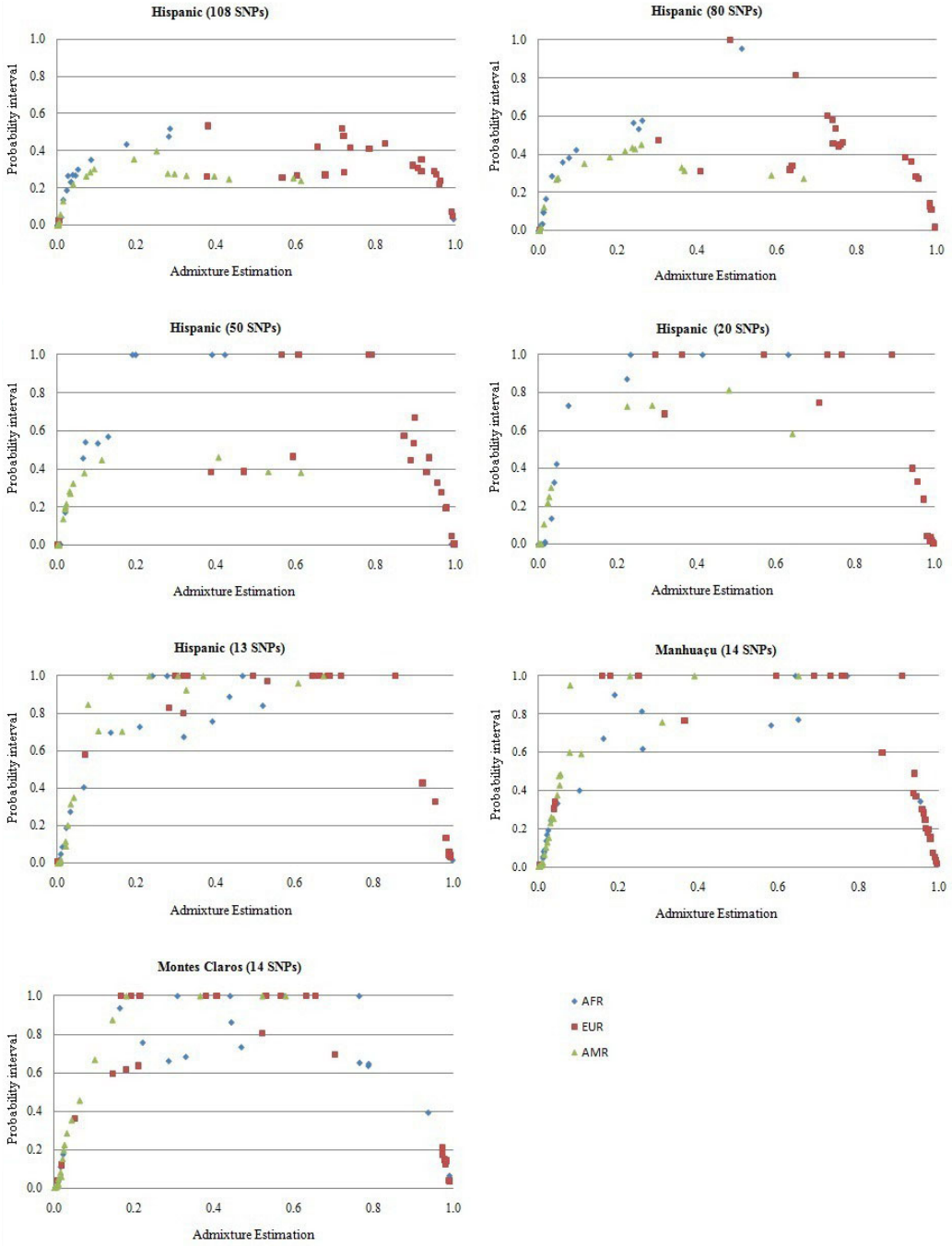


Figure 3. Variation of 90% probability intervals of individual admixture estimates as a function of point estimates of individual admixture, as inferred by the STRUCTURE algorithm with different numbers of single-nucleotide polymorphisms (SNPs) and in different populations. AFR = Africans; EUR = Europeans; AMR = Native Americans.

Table 3. Correlation coefficients between individual admixture point estimations in the Hispanic sample by using 108 ancestry informative markers (AIMs), and admixture estimations using randomly selecting subsets of 80, 50, 20 AIMs, and also by using our set of 13 AIMs.

AIM subsets	Parental population					
	African		European		Amerindian	
	Correlation coefficient	P	Correlation coefficient	P	Correlation coefficient	P
80 SNPs	0.93	>0.001	0.93	>0.001	0.96	<0.001
50 SNPs	0.96	>0.001	0.93	>0.001	0.88	<0.001
20 SNPs	0.58	0.004	0.66	0.001	0.64	0.001
13 SNPs	0.60	0.003	0.49	0.019	0.52	0.011

SNPs = single-nucleotide polymorphisms.

DISCUSSION

We presented two new multiplex panels of AIMs (for a total of 14 SNPs) developed to assist investigators of small-medium size laboratories to estimate admixture in Latin American populations. We gave methodological information in detail to allow other investigators to use these panels, to use individual genotypes of the parental populations in other admixture studies, or to follow the same steps to design additional panels of AIMs more suitable for specific populations to obtain more accurate estimates of admixture. Specifically, we recommend the use of the Muplex resource (Rachlin et al., 2005) to design primers for multiplex amplification and subsequent genotyping. Muplex may also be used to design multiplex panels of insertion-deletion, that have proven to be cost-effective markers for admixture and population structure studies (Bastos-Rodrigues et al., 2006; Santos et al., 2010).

A methodological issue when estimating admixture and population structure using few markers is to test the accuracy of the results. This may be achieved by using a reference set of samples that have been genotyped for the small number of markers (in this case the 14 AIMs) and for a large set of polymorphisms. This strategy has been followed by Bastos-Rodrigues et al. (2006), using the reference HGDP-CEPH panel of DNA samples (Cann et al., 2002) to test the performance of 48 INDELS to study the genetic structure of human populations. We tested the performance of our 14 AIMs by comparing population and individual admixture estimations obtained with this set of markers with those obtained using 108 AIMs (that we assume to be more accurate) in the Hispanic sample of the SNP500Cancer project. This sample, as well as the European and African ancestry used as parental populations, and the Pima and Maya samples used to select the 108 AIMs, is an appropriate reference because they are available as immortalized cells in the Coriell repository (www.coriell.org), which can provide unlimited good-quality DNA to reproduce or extend our results. Immortalized cells are available for reference samples, such as the CEPH-HGDP (Cann et al., 2002), the SNP500Cancer (Packer et al., 2006) and HapMap (the International HapMap Consortium 2007), for which large genome-wide genotype datasets are publicly available. It is important that new sets of markers developed to study admixture or population structure be tested using these resources. By testing the performance of our set of markers, we identified their strengths and limitations. At the population level they are appropriate to estimate European admixture, they overestimate African ancestry and underestimate Native American ancestry. As expected because of the use of few markers, they do not provide adequate estimates of individual admixture, except to identify individuals from the parental populations. However, the ability to identify individuals from the parental populations (European, African or Native American)

should not be generalized to the power to accurately estimate ancestry of admixed individuals.

A pervasive methodological issue in admixture studies is the identification of appropriate parental European, African and Native American populations (Glass and Li, 1953; Chakraborty, 1986). With the use of AIMs, this issue is mitigated by the choice of markers that have frequencies that are very different among the parental groups, but are very homogeneous within them. The selection of markers with these characteristics is facilitated by the availability of datasets of genome-wide surveys for different populations, which include the 52 worldwide populations of the CEPH-HGDP panel and the populations of Phase III of HapMap (International HapMap Consortium, 2010), although Native Americans are still under-represented in these studies. Most of the 14 AIMs that we selected reasonably fit the pattern of differentiation required for AIMs (Table 1). Thus, the effect of our suboptimal choice of parental populations (a limitation shared with most admixture studies) is partially counterbalanced by our use of AIMs.

We estimated admixture in the populations of Montes Claros and Manhuaçu of the State of Minas Gerais (southeastern Brazil), which hosted one of the largest Brazilian populations of African ancestry slaves during the Colonial period. In the geographic area of the region of Manhuaçu (eastern part of the state), slaves were 40% of the population in 1840 (Luna and Klein, 2004), but since the end of the 19th century, this geographic region received a large number of European ancestry immigrants attracted by a flourishing agricultural economy, mainly based on coffee (Botelho et al., 2007). The predominant European ancestry in the Manhuaçu sample complements these historical records, suggesting that recent European immigrants had a substantial impact in the genetic structure of this population. Montes Claros is located in the northern part of Minas Gerais. Though it is a region with one of the smallest populations of African ancestry slaves during the Colonial period (15% of the total population in 1833; Botelho, 1994; Luna and Klein, 2004), our results suggest a substantial African contribution. This may be related to its geographical proximity to Bahia, currently the Brazilian state with the largest proportion of self-identified "Black" individuals (IBGE, 2007). Our results suggest that at least in the State of Minas Gerais (one of the largest in extension and population in Brazil), historical demographic data about African ancestry slaves are not good indicators of the contribution of African ancestry to the current urban local population.

In conclusion, we developed two multiplex panels informative to estimate African (seven SNPs) and Native American (seven SNPs) ancestry, useful to assist investigators of small laboratories in studying the genetic structure of Latin American populations. Our panel of 14 AIMs allows accurate estimation of European population ancestry, but has limited power to estimate individual admixture. Thus, it is a useful tool to be used in combination with other available sets of markers to assess admixture and the genetics structure of Latin American populations (Bastos-Rodrigues et al., 2006; Kosoy et al., 2009; Lai et al., 2009; Lins et al., 2010). Flexibility in measuring admixture is important because depending on the degree of admixture of the targeted populations, the optimal set of markers to infer admixture may vary (Pfaff et al., 2004). Assessing the informativeness, strengths and limitations of new panels of AIMs is necessary to make correct inferences about admixture processes in Latin American populations.

ACKNOWLEDGMENTS

We are grateful to all blood donors and to Telma Regina Guedes Machado (Hemocentro Regional de Montes Claros) and Simone Avelino Rodrigues (Núcleo Regional de Ma-

nhuaçu) from Fundação Hemominas for sample collections, Laboratório de Biodiversidade e Evolução Molecular and Prof. Fabricio Santos, Núcleo de Análise de Genomas e Expressão Gênica (NAGE) and Prof. Santuza Teixeira, Laboratório de Genética Bioquímica and Prof. Gloria Franco for access to the Megabace sequencers, and Rinaldo Pereira and Túlio Lins for collaboration with the Applied Biosystems protocols. Research supported by the National Institutes of Health - Fogarty International Center (1R01TW007894-01 to E. Tarazona-Santos), Brazilian National Research Council (CNPq), Brazilian Ministry of Education (CAPES) and Minas Gerais State Foundation in Aid of Research (FAPEMIG).

REFERENCES

Bastos-Rodrigues L, Pimenta JR and Pena SD (2006). The genetic structure of human populations studied through short insertion-deletion polymorphisms. *Ann. Hum. Genet.* 70: 658-665.

Benn-Torres J, Bonilla C, Robbins CM, Waterman L, et al. (2008). Admixture and population stratification in African Caribbean populations. *Ann. Hum. Genet.* 72: 90-98.

Botelho TR (1994). Famílias e Escravarias: Demografia e Família Escrava no Norte de Minas Gerais no Século XIX. Master's thesis, Universidade de São Paulo, São Paulo.

Botelho TR, Braga MP and de Andrade CV (2007). Immigration and family in Minas Gerais at the end of the 19th century. *Rev. Bras. Hist.* 27: 155-176.

Cann HM, de Toma C, Cazes L, Legrand MF, et al. (2002). A human genome diversity cell line panel. *Science* 296: 261-262.

Carvalho CM and Pena SD (2005). Optimization of a multiplex minisequencing protocol for population studies and medical genetics. *Genet. Mol. Res.* 4: 115-125.

Chakraborty R (1985). Gene Identity in Racial Hybrids and Estimation of Admixture Rates. In: Genetic Differentiation in Human and Other Animal Populations (Ahuja YR and Neel JV, eds.). Indian Anthropological Association, Delhi, 171-180.

Chakraborty R (1986). Gene admixture in human populations: models and predictions. *Am. J. Phys. Anthropol.* 29: 1-43.

Choisy M, Franck P and Cornuet JM (2004). Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol. Ecol.* 13: 955-968.

Chung CC, Magalhaes WC, Gonzalez-Bosquet J and Chanock SJ (2010). Genome-wide association studies in cancer - current and future directions. *Carcinogenesis* 31: 111-120.

Comas D, Calafell F, Mateu E, Perez-Lezaun A, et al. (1998). Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am. J. Hum. Genet.* 63: 1824-1838.

Dupanloup I and Bertorelle G (2001). Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol. Biol. Evol.* 18: 672-675.

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

Glass B and Li CC (1953). The dynamics of racial intermixture; an analysis based on the American Negro. *Am. J. Hum. Genet.* 5: 1-20.

Herrera-Paz EF, Matamoros M and Carracedo A (2010). The Garifuna (Black Carib) people of the Atlantic coasts of Honduras: Population dynamics, structure, and phylogenetic relations inferred from genetic data, migration matrices, and isonymy. *Am. J. Hum. Biol.* 22: 36-44.

Hubisz MJ, Falush D, Stephens M and Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Res.* 9: 1322-1332.

Hughes AL, Welch R, Puri V, Matthews C, et al. (2008). Genome-wide SNP typing reveals signatures of population history. *Genomics* 92: 1-8.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2007). Síntese de Indicadores Sociais: Uma Análise das Condições de Vida da População Brasileira. Available at [http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/indicadoresminimos/sinteseindicsociais2007/indic_sociais2007.pdf]. Accessed April 12, 2010.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

Kosoy R, Nassir R, Tian C, White PA, et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* 30: 69-78.

Lai CQ, Tucker KL, Choudhry S, Parnell LD, et al. (2009). Population admixture associated with disease prevalence in the

Boston Puerto Rican health study. *Hum. Genet.* 125: 199-209.
Lee SS, Bolnick DA, Duster T, Ossorio P, et al. (2009). Genetics. The illusive gold standard in genetic ancestry testing. *Science* 325: 38-39.
Lins TC, Vieira RG, Abreu BS, Grattapaglia D, et al. (2010). Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. *Am. J. Hum. Biol.* 22: 187-192.
Luna FV and Klein HS (2004). Economy and slave society: Minas Gerais and São Paulo in 1830. *Rev. Bras. Est. Pop.* 21: 173-193.
Packer BR, Yeager M, Burdett L, Welch 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.
Parra FC, Amado RC, Lambertucci JR, Rocha J, et al. (2003). Color and genomic ancestry in Brazilians. *Proc. Natl. Acad. Sci. U. S. A.* 100: 177-182.
Patterson N, Petersen DC, van der Ross RE, Sudoyo H, et al. (2010). Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* 19: 411-419.
Pfaff CL, Barnholtz-Sloan J, Wagner JK and Long JC (2004). Information on ancestry from genetic markers. *Genet. Epidemiol.* 26: 305-315.
Pritchard JK, Stephens M and Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
Rachlin J, Ding C, Cantor C and Kasif S (2005). MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Res.* 33: W544-W547.
Rosenberg NA, Li LM, Ward R and Pritchard JK (2003). Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73: 1402-1422.
Salzano FM and Bortolini MC (2002). The Evolution and Genetics of Latin American Populations. Cambridge University Press, Cambridge.
Sans M (2000). Admixture studies in Latin America: from the 20th to the 21st century. *Hum. Biol.* 72: 155-177.
Santos NP, Ribeiro-Rodrigues EM, Ribeiro-Dos-Santos AK, Pereira R, et al. (2010). Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum. Mutat.* 31: 184-190.
Tarazona-Santos E, Raimondi S and Fuselli S (2007). Controlling the Effects of Population Stratification by Admixture in Pharmacogenetics. In: Pharmacogenomics in Admixed populations (Guilherme Suarez-Kurtz, ed.). Landes Bioscience, Austin, 1-16.
Tian C, Hinds DA, Shigeta R, Kittles R, et al. (2006). A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *Am. J. Hum. Genet.* 79: 640-649.
Via M, Ziv E and Burchard EG (2009). Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. *Clin. Genet.* 76: 225-235.

Supplementary Material

Table S1. Primers for the polymerase chain reaction (PCR) and mini-sequencing reactions of the African and Native American panels.

SNP	Forward PCR primers	Reverse PCR primers	Mini-sequencing primers
AFR_rs2697520	TGCCTGTGTGTGTGTGTATGT	CGTATTCTTTAAATGGTGCAACA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcaca gcttgtctTTTGTCCTAAAAGTATCCCCTGTTA (95)
AFR_rs8035530	CCTCTTCAGTACCTCTTCCAGTG	CAACAGAAGGGAGAGCAAGAC	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagc GCAATACAAGAAAAGCACCTTCAT (88)
AFR_rs137211	GCTTGCCTATTGATAGAAGGGT	CCCTCTTTCACGATACTTAGCAT	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcCACCACAGTTTTGA TTTCAGTCC (65)
AFR_rs2416791	CCCAGCTTGTGGTCATTTATTA	GATTCTGGCAACGCTGACT	tcgcgcgtttcggtgatgacggtgaaaCATCTACCATCAGCCCAATTC (48)
AFR_rs9626698	GGCAACGCTCCTAGTGATC	CAACAGCCAAGAAATATGGTG	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcac agcttgtctgtaagcgGAGAGGGGTTTTTGCAATAACATTA (102)
AFR_rs1443985	CGGGTGCTATGTGTACTTTCAT	CAAATTCTGCTTCCTGAAATAGA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagctt gtctgtaagcggCTTGAGAAATGTCTATACAAATGCCCTATTAA (110)
AFR_rs735480	TTAGTTCGCTGCTGCTCTAAT	GATGGAGGCCTCAGAGAGA	tcgcgcgtttcggtgatgacggtgaaaacctcGCTTAAAAGTCATGCGCCTAAG (55)
AMR_rs8058694	GATTTCTTCCTGCAGCCTC	TGGAAACCTACACCACCTCTC	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatGGAAGGATTGCATCAC GATACA (62)
AMR_rs2234636	TTCCATGTCCATCCAATGTA	ATGGCTGTGGAGTTCGAA	tcgcgcgtttcggtgatgacggtgaaaacctctgaGCGACTCCAAAAAGAAGAC AA (56)
AMR_rs3760657	AGGAGCAGACAAACAGTCCTAT	GGATGGCAACAATACCTCACT	tcgcgcgtttcggtgatgacggtgaaaaGTGTCAGATGACACAGCAC (47)
AMR_rs2619681	CATGCTTGCCAACACGAT	AACATATGAACAGATGTCCAACA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagcCAACATCATTATTA ATCAGGAAAATTAAAATCACATT (82)
AMR_rs800292	CCAAACATATCCAGAAGGCAC	AATTCAGGCATAATTGCTACAAT	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcaCCCTGGATATAGATCT CTTGGAAAT (68)
AMR_rs2088102	GGATCTTAGCCTCTTCTGCTACA	CCTTCTTCCAGCACAGGG	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagc ttgtctgtaagcggatgccgggaCCCAGGTGCTCAACAGAA (105)
AMR_rs691968	TTTCACTCCTCCAAGGACC	TGATGGCTACACAAGAGCAAA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagc ttgtctgtGCCATCCTGGTGCCTA (88)

We designed mini-sequencing primers with the following characteristics: minimum length of 40 nucleotides, annealing temperature near 50°C, minimum difference of four nucleotides between predicted mini-sequencing product lengths (the length of the primers was modified by the addition of tails with the PUC 18 plasmid sequence to the 5' extremity of the specific primer sequences, which are represented in lowercase in Table S1), absence of predicted primer dimers or hairpin structures. The primers were commercially synthesized and ranged from 47 to 110 nucleotides.

Enzymatic purification of PCR and mini-sequencing products

The purification of 3 µL PCR products was done by a reaction containing 1 unit of the enzyme *Escherichia coli* exonuclease I (*Exo*I, 10 units/µL), 0.9 units of shrimp alkaline phosphatase (SAP, 1 unit/µL) and 0.2 µL 10X SAP reaction buffer. The Exo-SAP reaction was performed in order to eliminate the excess of PCR primers and dNTPs of the PCR products before the mini-sequencing reaction. The reaction was incubated at 37°C for 90 min, followed by inactivation of the enzymes by heating at 80°C for 20 min.

The purification of the monochromatic mini-sequencing products was performed by a reaction containing 0.3 units SAP, 0.2 µL 10X SAP reaction buffer, 1.6 µL H_2O, and 5 µL of the mini-sequencing product. The products of the polychromatic mini-sequencing were purified in a reaction with 0.5 units SAP and 0.5 µL 10X SAP buffer reaction, added directly to 5 µL of the SNaPshot product. The SAP reaction was incubated at 37°C for 60 min, followed by the inactivation of the enzymes by heating at 75°C for 15 min.

Other technical issues

We imposed on the Muplex (Rachlin et al., 2005) the condition that the minimum difference between PCR product lengths within each panel had to be 10 bp. This allows us, during the set up period of the multiplex PCR, to better evaluate if all primers are properly working, using polyacrylamide gel electrophoresis.

Table S2. Set of 108 ancestry informative markers, selected from the SNP500Cancer project database, using the criterion of F_{ST} >0.20 among Europeans, Africans and Native Americans, and F_{ST} <0.10 among populations within these groups.

SNP500Cancer ID	dbSNP ID	SNP500Cancer ID	dbSNP ID	SNP500Cancer ID	dbSNP ID	SNP500Cancer ID	dbSNP ID
ABCA1_17	rs2230808	EPHX2_04	rs1126452	IL4_03	rs2070874	PIM1_03	rs262933
AKR1C3_36	rs7921327	ERCC1_06	rs3212948	IL6R_04	rs8192284	POLB_08	rs2953983
AMACR_03	rs34689	ERCC5_01	rs1047768	IL6_04	rs1800797	POLD1_13	rs1726787
ANKK1_01	rs1800497	ESR1_17	rs2273206	IL7R_01	rs1494555	RAD52_07	rs6413436
APC_09	rs2229992	FANCA_03	rs1061646	INSR_13	rs919275	RAG1_01	rs2227973
AURKA_16	rs10485805	FASLG_01	rs929087	KRT23_03	rs2269858	RB1CC1_24	rs1129660
BCL2L1_02	rs1484994	FBXW7_44	rs2676329	LCAT_05	rs1109166	RERG_24	rs6488766
BCL6_09	rs3774309	FUT2_05	rs603985	LIPC_37	rs1968689	RGS5_01	rs15049
BIC_34	rs4817027	GATA3_25	rs520236	LRP5_01	rs312016	RNASEL_02	rs486907
BRIP1_09	rs1015771	GDF15_02	rs1059369	MATR3_01	rs11738738	SCARB1_03	rs4765621
CASP3_08	rs1049216	GHR_47	rs7712701	MBL2_46	rs10824793	SEPP1_01	rs7579
CASP8_07	rs2293554	GPX2_21	rs2737844	MSH3_07	rs3797896	SLAMF1_03	rs164283
CASR_11	rs4678045	GPX3_28	rs8177426	MTRR_19	rs8659	SLC23A1_09	rs4257763
CAT_02	rs769214	GSTM3_06	rs1537234	MX1_28	rs455599	SLC4A2_02	rs10245199
CAV1_29	rs6950798	HSD17B2_01	rs1424151	MYBL2_03	rs34771484	SLC6A3_10	rs6347
CDK5_16	rs1549760	HSD3B1_24	rs4659182	MYC_02	rs3891248	SOAT2_01	rs2280699
CDKN2A_03	rs3088440	HSD3B2_14	rs12411115	MYNN_01	rs1317082	SOD1_01	rs2070424
CGA_06	rs932742	IFNAR2_06	rs7279064	NCF2_03	rs2274064	SOD3_05	rs2855262
CYP19A1_01	rs700518	IGF1R_05	rs2137680	NCOA3_04	rs2076546	TCTA_04	rs6784820
CYP1A1_14	rs2606345	IGF2_16	rs3213221	NFKB1_02	rs3774937	TERT_02	rs2075786
CYP1B1_27	rs162556	IGFBP5_10	rs1978346	NFKBIE_01	rs483536	TLR2_06	rs4696480
CYP2E1_31	rs8192766	IGFBP6_19	rs822688	NR1H4_05	rs35724	TP73L_13	rs9840360
CYP3A7_01	rs12360	IL13_01	rs20541	OCA2_23	rs1900758	VCAM1_05	rs3176879
DHDH_02	rs4987162	IL15_02	rs10833	PAK6_13	rs2242119	WDR79_06	rs17886268
DRD2_03	rs1079597	IL1B_03	rs1143627	PCNA_10	rs17352	XRCC4_05	rs2075685
EFNB3_02	rs3744262	IL2_03	rs2069763	PCTP_01	rs2114443	XRCC5_12	rs2440
ENPP1_04	rs1044582	IL4R_07	rs1805016	PHB_02	rs4987082	-	rs1719889

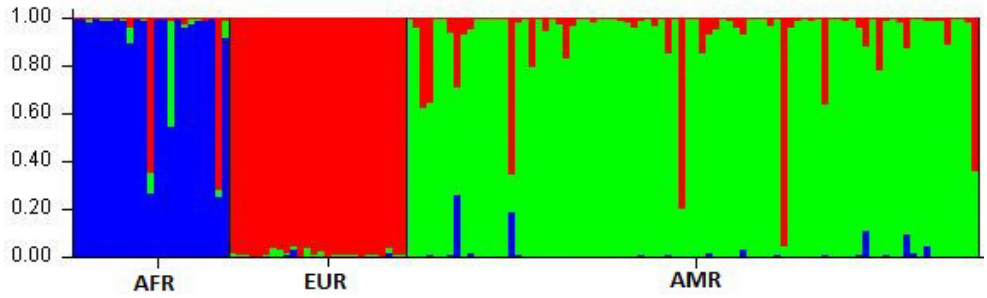


Figure S1. Analysis of parental populations (AFR = Africans; EUR = Europeans; AMR = Native Americans) with the 14 ancestry informative markers selected in our study. Each vertical bar represents an individual subject. Analyses were performed with the admixture model, K = 3 and without any prior population assignment.

2.3 EXTENSIVE ADMIXTURE IN BRAZILIAN SICKLE CELL PATIENTS: IMPLICATIONS FOR THE MAPPING OF GENETIC MODIFIERS

2.3.1 RESUMO

O debate sobre os efeitos da miscigenação nos estudos de associação, em especial, nos de varredura genômicas (*GWAS*) tem sido contumaz, não apenas no alerta sobre os riscos de associação espúria, mas também sobre as possibilidades advindas dos estudos de mapeamento por miscigenação (*Admixture Mapping*). Desta forma, é essencial a descrição dos níveis de miscigenação individual nas populações não autóctones latino-americanas, em especial, na brasileira. Neste estudo foram genotipados 54 marcadores em 200 indivíduos portadores da doença falciforme e 291 doadores de sangue sadios do estado de Minas Gerais. Demonstramos que os níveis de ancestralidade africana em indivíduos portadores de anemia falciforme são bastante heterogêneos, sendo os valores intermediários (15-85%) prevalentes neste grupo (77%). Infere-se, portanto, que classificações raciais ou étnicas não são precisas nesta população para a predição do fenótipo. Assim, ressalta-se a importância de: controlar os níveis de ancestralidade em *GWAS* visando evitar associações espúrias, e utilizar populações latino-americanas em estudos de associação baseados em mapeamento por miscigenação para subfenótipos da doença falciforme.

2.3.2 ATIVIDADES REALIZADAS

As minhas contribuições a este trabalho são: (1) extração de DNA dos pacientes e doadores; (2) quantificação e preparação das alíquotas de DNA das amostras; (3) seleção de *primers* e montagem das reações multiplex dos painéis de SNPs; (4) padronização das reações multiplex; (5) preparação da figura do artigo. Estas ações também se aplicam às atividades realizadas no artigo **Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: an application to populations of the State of Minas Gerais (Brazil)**.

blood

2011 118: 4493-4495
doi:10.1182/blood-2011-06-361915

Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers

Maria Clara F. da Silva, Luciana W. Zuccherato, Flavia C. Lucena, Giordano B. Soares-Souza, Zilma M. Vieira, Sérgio D.J. Pena, Marina L. Martins and Eduardo Tarazona-Santos

Blood (print ISSN 0006-4971, online ISSN 1528-0020), is published weekly by the American Society of Hematology, 2021 L St, NW, Suite 900, Washington DC 20036.

old extending this approach well into the next decade of life. The acute GVHD rate in this group was predictably higher (grade 2 and 3, 7 of 15 = 47%) than their pediatric experience (20%)[2]; however, the GVHD appeared easily treated by single agent prednisone. With the extensive transplant experience that this group and others have reported,[3-7] there is certainly a place for full ablative HSCT in pediatric patients with SCD, and this new experience now suggests that the same is true for young adults eligible for this approach. These accumulating reports continue to confirm a very favorable benefit to risk ratio making this a truly exciting time for patients and physicians contemplating HSCT for SCD.

Myeloablative conditioning and the ensuing risk of GVHD, however, require robust organ function. We now have transplanted 23 patients with severe disease at our center with nonmyeloablative conditioning, their ages ranging from 17 to 65 years. All patients are alive, and engraftment was achieved in 20 (87%). Importantly, 5 of the first 10 patients reported[8] are now off immunosuppression with continued stable mixed chimerism. Equally important, none of the engrafted patients has experienced any GVHD. In addition, 3 patients have produced offspring naturally (1 male and 2 female). This larger experience suggests that with this nonmyeloablative approach, the risk of rejection is similar, the risk of GVHD is lower, long-term immunosuppression is not absolutely required, and stable mixed chimerism is achievable. It is important to note that nearly half (10 of 23) of our patients would be ineligible for myeloablative transplantation because of comorbidities including cirrhosis and poor lung function.

Thus for patients in their second and third decade of life, options include both full and nonmyeloablative transplant conditioning, with the choice depending on organ involvement, potential transplant-related complications, and the desire for future fertility. It is our opinion that patients in this age group should be transplanted as a part of an ongoing clinical trial. HSCT for SCD remains under-utilized and the time to seriously consider this therapeutic option is now.

Matthew Hsieh
National Institute of Diabetes and Digestive and Kidney Diseases, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD

Courtney Fitzhugh
National Institute of Diabetes and Digestive and Kidney Diseases, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD

John F. Tisdale
National Institute of Diabetes and Digestive and Kidney Diseases, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD

Acknowledgements: This work was supported by the intramural research program at the National Institutes of Health.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Dr John F. Tisdale, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, 9000 Rockville Pike, Bdg 10, Rm 9N112, Bethesda, MD 20892-1652; e-mail: johntis@mail.nih.gov.

References

1. Kuentz M, Robin M, Dhedin N, et al. Is there still a place for myeloablative regimen to transplant young adults with sickle cell disease? *Blood.* 2011;118(16):4491-4492.
2. Bernaudin F, Socie G, Kuentz M, et al. Long-term results of related, myeloablative stem cell transplantation to cure sickle cell disease. *Blood.* 2007;110(7):2749-2756.
3. Walters MC, Patience M, Leisenring W, et al. Bone Marrow Transplantation for Sickle Cell Disease. *New Engl J Med.* 1996;335(6):369-376.
4. Walters MC, Patience M, Leisenring W, et al. Stable mixed hematopoietic chimerism after bone marrow transplantation for sickle cell anemia. *Biol Blood Marrow Transplant.* 2001;7(12):665-673.
5. Vermylen C, Cornu G, Ferster A, et al. Haematopoietic stem cell transplantation for sickle cell anaemia: the first 50 patients transplanted in Belgium. *Bone Marrow Transplant.* 1998;22(1):1-6.
6. Brachet C, Azzi N, Demulder A, et al. Hydroxyurea treatment for sickle cell disease: impact on haematopoietic stem cell transplantation's outcome. *Bone Marrow Transplant.* 2004;33(8):799-803.
7. Panepinto JA, Walters MC, Carreras J, et al. Matched-related donor transplantation for sickle cell disease: report from the Center for International Blood and Transplant Research. *Br J Haematol.* 2007;137(5):479-485.
8. Hsieh MM, Kang EM, Fitzhugh CD, et al. Allogeneic hematopoietic stem-cell transplantation for sickle cell disease. *New Engl J Med.* 2009;361(24):2309-2317.

To the editor:

Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers

Genome-wide association studies (GWASs) of sickle cell disease (SCD) patients are a promising tool for identifying genetic modifiers of clinically relevant traits.[1,2] In such a study, Solovieff et al[2] have uncovered a set of SNPs associated with fetal hemoglobin (HbF) concentration, the major modulator of the clinical course of this disease.

After the introduction of the SCD mutations into the Americas by African slaves, African descendants mixed with individuals of Native American and European origin to various extents across the continent. In Latin America, individuals tend to be more admixed than in African-American populations.[3,4] Studying different phenotypes, Solovieff et al[2] did not find an association between fetal hemoglobin concentration and European ancestry, while Creary et al[5] have reported an association between European ancestry and the proportion of erythrocytes containing HbF.

The level of admixture of SCD patients has implications for the design of association studies aimed at identifying new genetic modifiers of SCD clinical manifestations, particularly if these outcomes are associated with ancestry. If both genetic variants and clinical manifestations are associated with ancestry, there are 2 important issues to address. First, an observed association may be spurious if ancestry is not controlled for.[6] Second, the admixture mapping strategy[7] may be used: Thousands of ancestry-informative markers may be genotyped across the genome, and the local ancestry along chromosomes can be inferred. The genetic modifier

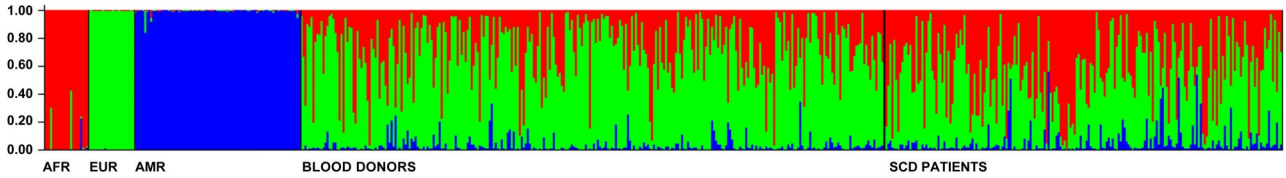


Figure 1. Individual admixture in healthy blood donors and Sickle Cell Disease patients from Minas Gerais. Admixture was estimated using the method by Pritchard et al implemented in the software Structure.[11] Each vertical bar represents an individual and his/her admixture proportions based on the parental populations on the left. Additional methodologic information and results are available as supplemental Methods. Structure was run using the following conditions and parameters: K = 3 (number of parental populations), burn-in period = 100 000, MCMC cycles after burn-in = 100 000, we used a priori information for the individuals from parental populations to assist the clustering (USEPOPINFO = 1), model = ADMIXTURE for the admixed individuals, α parameter was inferred for each population, GENSBACK = 2, MIGRPRIOR = 0.05, allele frequencies was assumed to be correlated.

alleles are expected to be located in genomic regions with an excess of ancestry of the population associated with the more common clinical outcome.

The risk of spurious association and the power of admixture mapping increase both with the level of admixture. However, estimates of ancestry are rare in SCD patients. Excluding GWASs, association studies between SNPs and SCD subphenotypes seldom control for ancestry.

We estimated admixture in 200 patients with SCD and in 291 healthy blood donors; all from the State of Minas Gerais (approximately 20 million inhabitants in South Eastern Brazil). We genotyped 54 SNPs/INDELs validated for admixture studies.[3,8,9] The ancestry of the healthy blood donors, who represent the general population reasonably well, was 33.8% African, 57.7% European and 3.5% Amerindian; whereas SCD patients showed 47.3%, 39.7% and 13.0% of African, European and Amerindian ancestry, respectively (estimated following Dupanloup and Bertorelle,[10] see supplemental Table 1, available on the *Blood* Web site; see the supplemental Materials link at the top of the online article). Considering individual admixture (Figure 1), only 11.05% of SCD patients had $>$ 85% African ancestry. Most of the patients (73.37%) had intermediate levels of admixture (15%-85%), and interestingly, 13.8% had predominant European ancestry ($>$ 85%, and for 13.0%, the lower limit of the 90% credibility interval of European ancestry was $>$ 0.60). Therefore, the prevalence of European ancestry is high, and the individual admixture is very heterogeneous in SCD patients from Brazil. Our results suggest that: (1) despite the association of SCD with African ancestry, the label "ethnic/racial" disease seems inappropriate in this population, (2) in association studies with SCD patients, controlling for ancestry is important to avoid spurious association, and (3) Latin American populations of SCD patients are promising targets for admixture mapping of genetic modifiers of ancestry-associated SCD clinical manifestations.

Acknowledgments: The authors thank the personnel from the Hemominas Foundation for their collaboration in sample collection. The study was approved by Ethical Committees from the Universidad Federal de Minas Gerais and the Hemominas Foundation. This research was funded by Brazilian Federal Agencies CNPq and CAPES, the Minas Gerais State Agency FAPEMIG and by the NCI-Fogarty 1R01TW007894-01 GRIP grant.

Contribution: M.C.F.S., M.L.M. and E.T.S. conceived the study; M.C.F.S., Z.M.V., M.L.M. organized samples collection; M.C.F.S., L.W.Z., F.C. and G.B.S.S. performed experiments or analyzed the data; S.D.J.P. provided resources for the INDELS panel; and M.C.F.S. and E.T.S. wrote the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Eduardo Tarazona-Santos, PhD, Universidade Federal de Minas Gerais Av Antonio Carlos 6627 Pampulha CP 486 UFMG-ICB-Departamento de Biologia Geral Belo Horizonte, MG 31270-901, Brazil; e-mail: edutars@icb.ufmg.br.

Maria Clara F. da Silva
Fundação Hemominas, Minas Gerais, Brazil
Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais, Brazil

Luciana W. Zuccherato
Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais, Brazil

Flavia C. Lucena
Fundação Hemominas, Minas Gerais, Brazil

Giordano B. Soares-Souza
Fundação Hemominas, Minas Gerais, Brazil
Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais, Brazil

Zilma M. Vieira
Fundação Hemominas, Minas Gerais, Brazil

Sérgio D.J. Pena
Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais, Brazil

Marina L. Martins
Fundação Hemominas, Minas Gerais, Brazil

Eduardo Tarazona-Santos
Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais, Brazil

References

1. Sebastiani P, Solovieff N, Hartley SW, et al. Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study. *Am J Hematol.* 2010;85(1):29-35.
2. Solovieff N, Milton JN, Hartley SW, et al. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5′ olfactory receptor gene cluster. *Blood.* 2010;115(9):1815-1822.
3. Pena SD, Di Pietro G, Fuchshuber-Moraes M, et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One.* 2011;6:e17063.
4. Bryc K, Velez C, Karafet T, et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci U S A.* 2010;107(Suppl 2):8954-8961.
5. Creary LE, McKenzie CA, Menzel S, et al. Ethnic differences in F cell levels in Jamaica: a potential tool for identifying new genetic loci controlling fetal haemoglobin. *Br J Haematol.* 2009;144(9):954-960.
6. Tarazona-Santos E, Raimondi S, Fuselli S. Controlling the effects of population

stratification by admixture in pharmacogenetics. In: Suarez-Kurtz G, ed. Pharmacogenomics in Admixed populations. *Landes Bioscience, Austin.* 2007;12-27.

7. Winkler CA, Nelson GW, Smith MW. Admixture mapping comes of age. *Annu Rev Genomics Hum Genet.* 2010;11:65-89.
8. Bastos-Rodrigues L, Pimenta JR, Pena SD. The genetic structure of human populations studied through short insertion-deletion polymorphisms. *Ann Hum Genet.* 2006;70(5):658-665.
9. Silva MC, Zuccherato LW, Soares-Souza GB et al. Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: an application to populations of the State of Minas Gerais (Brazil). *Genet Mol Res.* 2010;9(4):2069-2085.
10. Dupanloup I, Bertorelle G. Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol Biol Evol.* 2001;18(4): 672-675.
11. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000;155(2):945-959.

Response

Genetic admixture in sickle cell disease

Genetic studies require careful thought about ancestry in the study design and analysis to avoid population stratification bias and to maximize the power of finding novel variants.[1] If both the genetic marker and the phenotype vary with respect to ancestry, then a spurious association will occur between the genetic marker and phenotype if one does not adjust properly for ancestry. In this issue of *Blood*, Silva et al examine the level of admixture in a cohort of Brazilian sickle cell patients and find that patients with sickle cell disease have a wide range of African admixture (15%-85%).[2] They correctly acknowledge that one must appropriately adjust for admixture to avoid false positive findings and that this population may be useful for admixture mapping. However, admixture mapping is only useful if the phenotype of interest also varies with admixture and further research is required to establish this relation in this population. In a cohort of African Americans with sickle cell disease we did not find a significant association between admixture (measured by the first principal component from a principal component analysis) and fetal hemoglobin.[3] However, one should note that the African Americans in this study on average did not have high levels of Caucasian admixture.[4] Admixture mapping has been successfully used to find novel genetic variants and regions for other phenotypes that were related to admixture such as white cell counts and prostate cancer.[5,6] Furthermore, examining ancestry in genetic studies of sickle cell disease could lead to novel loci that are either more prevalent or specific to certain ethnic groups. For example, sickle cell patients from the Southwestern Province of Saudi Arabia have fetal hemoglobin (HbF) levels twice as high as African Americans despite having similar *HBB* haplotypes.[7] Furthermore, sickle cell patients from the Eastern Province have even higher levels of HbF (mean [SD] 30.4 ± 6.9).[8] These findings suggest that there are HbF-associated variants that are more prevalent or specific to the Saudi population and that leveraging on ancestry in the genetic analysis can help identify novel variants.

Nadia Solovieff
Boston University School of Public Health, Boston, MA

Martin H. Steinberg
Boston University School of Medicine, Boston Medical Center, Boston, MA

Paola Sebastiani
Boston University School of Public Health, Boston, MA

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Martin H. Steinberg, MD, Boston University School of Medicine, 72 East Concord St, Rm E248, Boston, MA 02118; e-mail: mhsteinb@bu.edu.

References

1. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904-909.
2. Da Silva, MC, Zuccherato L, Lucena F, et al. Extensive admixture in brazilian sickle cell patients: implications for the mapping of genetic modifiers. *Blood.* 2011;118(16):4493-4495.
3. Solovieff N, Milton JN, Hartley SW, et al. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5′ olfactory receptor gene cluster. *Blood.* 2010;115(9):1815-1822.
4. Solovieff N, Hartley SW, Baldwin CT, et al. Ancestry of African Americans with sickle cell disease. *Blood Cells Mol Dis.* 2011;47(1):41-45.
5. Nalls MA, Wilson JG, Patterson NJ, Tandon A, Zmuda JM, et al. Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am J Hum Genet.* 2008; 82(1):81-87.
6. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci U S A.* 2006;103(38):14068-14073.
7. Alsultan A, Solovieff N, Aleem A, et al. Fetal hemoglobin in sickle cell anemia: Saudi patients from the Southwestern province have similar HBB haplotypes but higher HbF levels than African Americans. *Am J Hematol.* 2011;86(7):612-614.
8. Akinsheye I, Alsultan A, Solovieff N, et al. Fetal hemoglobin in sickle cell anemia. *Blood.* 2011;118(1):19-27.

To the editor:

Prothrombin 20210G>A genotype and C-reactive protein level

Thrombin is central not only in procoagulatory processes like fibrinogen or platelet activation but also in other systems that are related to inflammation control.[1] Recently, Flick et al characterized inflammatory responses of transgenic prothrombin mutant ($F2^{WE}$) mice in a collagen-induced arthritis (CIA) model.[2] Mice carrying the $F2^{WE}$ transgene that has dramatically reduced procoagulatory effects on fibrinogen and protease-activated receptor 1 exhibited a significantly attenuated inflammatory joint disease in CIA.

Prothrombin 20210G>A ($F2^{20210G>A}$) is a gain-of-function variant resulting in increased prothrombin expression and elevated

2.4 SOCIOECONOMIC AND NUTRITIONAL FACTORS ACCOUNT FOR THE ASSOCIATION OF GASTRIC CANCER WITH AMERINDIAN ANCESTRY IN A LATIN AMERICAN ADMIXED POPULATION

2.4.1 RESUMO TRADUZIDO

O carcinoma gástrico é um dos mais letais tipos de câncer e apresenta altas taxas de incidência na região Andina da América do Sul. A estrutura genética da população de Lima (Peru) foi avaliada e realizou-se um estudo de associação caso-controle com objetivo de testar a contribuição das ancestralidades ameríndia, europeia e africana no risco de desenvolvimento do câncer gástrico, controlando, para isto, os efeitos de fatores não-genéticos. Um vasto conjunto de informações socioeconômicas, clínicas e nutricionais foi coletado para cada participante do estudo e a ancestralidade foi inferida a partir de um conjunto de 103 marcadores informativos de ancestralidade. Ainda que a população de Lima seja tradicionalmente considerada como mestiça (por ex., miscigenada a partir de africanos, europeus e nativo-americanos), os dados demonstram uma alta fração de ancestralidade nativo-americana (78,4% para os casos e 74,6% para os controles) e taxas muito baixas de ancestralidade africana (<5%). O estudo caso-controle determinou a associação entre altas taxas de ancestralidade ameríndia e o câncer gástrico, ainda que fatores socioeconômicos correlacionados tanto à ancestralidade quanto ao câncer gástrico contribuam para essa associação. Além disso, a alta incidência de câncer gástrico no Peru não parece estar relacionada a alelos de susceptibilidades descritos anteriormente. Em contraposição, os resultados sugerem um papel predominante dos fatores socioeconômicos e das disparidades no acesso à saúde associados à etnicidade. Uma vez que os nativo-americanos formam um grupo negligenciado em estudos genômicos, sugere-se como alvo para estudos epidemiológicos neste grupo étnico as populações das grandes cidades da parte Ocidental da América do Sul, como Lima, que apresentam alta ancestralidade ameríndia.

2.4.2 ATIVIDADES REALIZADAS

Neste artigo participei das seguintes atividades: (1) Controle de qualidade dos dados, cálculo das frequências alélicas e genotípicas, e teste do Equilíbrio de Hardy-Weinberg; (2)

Análise de Componentes Principais a partir dos dados genotípicos para as populações do HapMap, populações nativo-americanas e casos e controles; (3) PCA dos dados socioeconômicos dos casos e controles e PCR para as componentes mais representativas; (4) Estimativas de ancestralidade a partir do software *Structure*; (5) Elaboração das figuras para o artigo.

OPEN ACCESS Freely available online

PLoS one

Socioeconomic and Nutritional Factors Account for the Association of Gastric Cancer with Amerindian Ancestry in a Latin American Admixed Population

Latife Pereira[1], Roxana Zamudio[1], Giordano Soares-Souza[1], Phabiola Herrera[2], Lilia Cabrera[2], Catherine C. Hooper[3], Jaime Cok[4], Juan M. Combe[5], Gloria Vargas[6], William A. Prado[7], Silvana Schneider[8], Fernanda Kehdy[1], Maira R. Rodrigues[1], Stephen J. Chanock[9], Douglas E. Berg[10], Robert H. Gilman[2,3,11], Eduardo Tarazona-Santos[2]*

1 Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, **2** Asociación Benéfica PRISMA, Lima, Peru, **3** Laboratorios de Investigacion y Desarrollo, Facultad de Ciencias, Universidad Peruana Cayetano Heredia, Lima, Peru, **4** Departamento de Patología, Hospital Nacional Cayetano Heredia, Lima, Peru, **5** Departamento de Gastroenterologia, Instituto Nacional de Enfermedades Neoplásicas, Lima, Peru, **6** Servicio de Gastroenterologia, Hospital Nacional Arzobispo Loayza, Lima, Peru, **7** Servicio de Gastroenterologia, Hospital Dos de Mayo, Lima, Peru, **8** Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, **9** Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, Maryland, United States of America, **10** Department of Molecular Microbiology, Washington University Medical School, St Louis, Missouri, United States of America, **11** Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America

Abstract

Gastric cancer is one of the most lethal types of cancer and its incidence varies worldwide, with the Andean region of South America showing high incidence rates. We evaluated the genetic structure of the population from Lima (Peru) and performed a case-control genetic association study to test the contribution of African, European, or Native American ancestry to risk for gastric cancer, controlling for the effect of non-genetic factors. A wide set of socioeconomic, dietary, and clinic information was collected for each participant in the study and ancestry was estimated based on 103 ancestry informative markers. Although the urban population from Lima is usually considered as mestizo (i.e., admixed from Africans, Europeans, and Native Americans), we observed a high fraction of Native American ancestry (78.4% for the cases and 74.6% for the controls) and a very low African ancestry (<5%). We determined that higher Native American individual ancestry is associated with gastric cancer, but socioeconomic factors associated both with gastric cancer and Native American ethnicity account for this association. Therefore, the high incidence of gastric cancer in Peru does not seem to be related to susceptibility alleles common in this population. Instead, our result suggests a predominant role for ethnic-associated socioeconomic factors and disparities in access to health services. Since Native Americans are a neglected group in genomic studies, we suggest that the population from Lima and other large cities from Western South America with high Native American ancestry background may be convenient targets for epidemiological studies focused on this ethnic group.

Citation: Pereira L, Zamudio R, Soares-Souza G, Herrera P, Cabrera L, et al. (2012) Socioeconomic and Nutritional Factors Account for the Association of Gastric Cancer with Amerindian Ancestry in a Latin American Admixed Population. PLoS ONE 7(8): e41200. doi:10.1371/journal.pone.0041200

Editor: Zongli Xu, National Institute of Environmental Health Sciences, United States of America

Received March 1, 2012; **Accepted** June 18, 2012; **Published** August 3, 2012

Funding: Fogarty International Center and National Cancer Institute (5R01TW007894) funded this study. The study and its participants also received funding and fellowships from the following Brazilian agencies: Brazilian National Research Council, Ministry of Education, Ministry of Health (PNPD-Saúde Program), and the Minas Gerais State Reaearch Agency. 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

Introduction

Gastric cancer is one of the most lethal types of cancer, accounting for approximately 800,000 deaths per year, but its incidence varies substantially worldwide [1]. The highest incidence of gastric cancer is observed in East Asia, Eastern Europe, and the Andean region of South America. Indeed, in the Peruvian population, gastric cancer ranks second in incidence among men and third among women (22.6 and 20 cases per 100,000 males and females respectively), being the type of cancer with the highest mortality. Comparatively, the incidence of gastric cancer in Peru is approximately five times higher than in the United States and twice that observed in Brazil [1].

Recavarren-Arce et al. and Correa proposed a progression model for the development of intestinal-type gastric adenocarcinoma, which consist of a transition from superficial gastritis to metaplasia to dysplasia and finally, gastric adenocarcinoma [2,3]. A plethora of socio-economic, environmental, and dietary factors modulate this progression and the individual risk of ultimately developing gastric cancer. Chronic infection of the stomach by the bacterium *Helicobacter pylori* leading to chronic inflammation is a major attributable risk factor [4], although less than 2% of *H. pylori* carriers develop gastric cancer [5]. *Helicobacter pylori* diversity also affects the risk of host gastric cancer, and the presence of the bacterial virulence factor cag+ is one of the most relevant risk factors. While this virulence factor has a frequency of ~60% in European and US populations, it attains more than 90% in the

Peruvian population [6]. Also, while Native American individuals from isolated populations are infected by mostly native strains that resemble Asian strains, due to the Pleistocene Asian origin of Native Americans, individuals living in medium and large urban centers, even if they may have a predominant Native American ancestry, are infected by largely European or hybrid strains brought to the Americas after the 15th century. These strains had largely replaced less virulent or vigorous native strains [7,8].

Poverty also correlates with gastric adenocarcinoma [9] and while elevated consumption of processed or smoked food and salt are risk factors, frequent intake of fresh fruits and vegetables is protective [10,11,12,13]. Human genetic diversity is also relevant [14]. The observed differences in the incidence of gastric cancer worldwide may be due to environmental factors or to the presence of susceptibility genetic variants that are more frequent in populations with high incidence of the disease, but the identification and discrimination of these factors is challenging. Although common susceptibility genetic variants have been identified in European and Chinese populations by genome-wide and candidate-gene association studies in genes such as *PLCE* [15], *IL1B* [14], *IL8* [16,17,18], *IL1RN* [19], and *PTGS2* [20], these variants account for a small portion of the genetic variance associated with sporadic gastric cancer.

Peru, with its high incidence of gastric cancer, has the largest Native American population in South America [21] and large cities such as Lima are populated by people classified as mestizo (i.e., individuals with admixture from Africans, Europeans, and Native Americans). If there is a human genetic basis for the high incidence of gastric cancer in the Andean region, we expect the admixed population from Lima to harbor genetic variants accounting for this high incidence, and if these variants were more common in the Native American genetic background of this population, it would be possible to use the genome-wide strategy of admixture mapping to discover these variants. Admixture mapping studies have recently helped to identify variants associated with prostate cancer [22] in African-American populations, but this approach is yet to be fully applied to Latin American or Latino/Hispanic US populations. In this context, the goals of this case-control genetic association study are: (i) to assess the ethnic composition and its related genetic structure of patients attending large hospitals in Lima; (ii) to test if individual Native American, European and African ancestries are risk factors for gastric cancer, controlling for the effect of non-genetic factors (i.e., socioeconomic, nutritional, and clinical). Socioeconomic, dietary, and clinical information was collected for each participant in the study and ancestry was estimated based on 103 ancestry-informative markers (AIMs). We determined that higher Native American individual ancestry is associated with gastric cancer, but that socioeconomic factors associated both with gastric cancer and ethnicity account for this association. Despite the high incidence of gastric cancer among Peruvians with predominantly Native American ancestry, our results do not point to a clear genetic basis for this discrepancy in incidence. Rather, they suggest a predominant role for ethnic-associated socioeconomic and human ecologic factors and disparities in access to health services.

Results

We recruited individuals attending Gastroenterology Divisions and prescribed for an endoscopy in three large hospitals in Lima (Table S1). Cases were adults referred for endoscopy and whose biopsies were confirmed positive for gastric cancer by histopathological analyses. The control group was composed of individuals whose biopsies were negative for gastric cancer. Individuals with intestinal metaplasia ($n = 46$) were included as controls on the assumption that this type of metaplasia does not increase the risk of developing gastric cancer [23,24,25].

We used a validated set of 103 AIMs [26] to estimate Native American, European and Native American individual ancestry for each of the 241 gastric cancer cases and 300 controls recruited for this study. The Principal Component Analysis (PCA, Figure 1) of the individual genotypes for our Peruvian samples (including 296 Native Americans) and for European, African, and Mexican individuals from HapMap-III shows that the AIMs used discriminate between African, European, and Peruvian Native American parental populations, and that the admixed Mexicans (resident in Los Angeles) and the Peruvian gastric cases and controls attending Lima hospitals are placed between Europeans and Native Americans. Moreover, Peruvian gastric cases and controls are relatively closer to the Peruvian Native American parental populations. Although the urban population from Lima is considered as mestizo (i.e., typically admixed), the groups studied herein showed a very high Native American ancestry (78.4% for the cases and 74.6% for the controls, Box in Figure 1), with a low African ancestry (<5%, Figure S1). Interestingly, there is a positive association between Native American ancestry and gastric cancer (logistic regression, $OR = 3.69$, 95%CI of the OR: 1.34-10.09, $p = 0.011$, $R2 = 0.016$), and consequently a negative association with European ancestry.

As expected, variables that are proxies for poverty (low education level, home quality characteristics such as the use of low quality materials, lack of good appliances, and poor sanitary conditions) were associated with gastric cancer (Table 1). Also, some digestive-related symptoms such as burning ($p<0.0001$), nausea ($p<0.0001$), vomiting ($p<0.0001$), and heaviness ($p<0.0001$) were more frequent in cases than in controls, but these symptoms are likely a consequence of the disease (Table 1). The variable age was not normally distributed ($p<0.01$, Kolmogorov–Smirnov test), the medians (and deviation interquartile) for controls and cases were 61 (Standard deviation (SD): 21) and 65 (SD: 25) years respectively ($p = 0.042$, Mann-Whitney test); therefore, this variable was considered as covariate in further logistic regression analyses.

We synthesized the wide set of non-genetic variables collected in cases and controls using a multivariate factor analysis, to reduce the dimensionality of these 43 non-genetic variables by capturing the correlation among them (Table 1 and Table S2). Interestingly, the first factor (16.38% of the total variance) is dominated by socioeconomic variables but includes a subset of correlated nutritional variables, higher values of the factor corresponding to wealthier conditions. The second factor (5.95% of total variance) includes subsets of socioeconomic and nutritional variables, as well as complaints, such as pain. The third factor (5.26% of the total variance) is dominated by digestive-related symptoms. We used the individual coordinates for each of these three factors to synthetically represent the original set of 43 variables. Both the first "socioeconomic" factor and the second factor are associated with gastric cancer ($OR = 0.68$, $p = 0.0002$ and $OR = 0.7$, $p = 0.0008$, respectively) and also with ethnicity ($p = 0.02$ and $p = 0.00003$ respectively, where higher values of the factor correspond to better socioeconomic conditions). The third "digestive symptoms" factor is associated with gastric cancer ($OR = 2.16$, $p<0.0001$).

The observed association between Native American ancestry and gastric cancer may be due to the effect of confounding socioeconomic or nutritional variables associated with both gastric cancer and ancestry. It is well known that high Native American or African ancestry is associated with poverty in many populations in

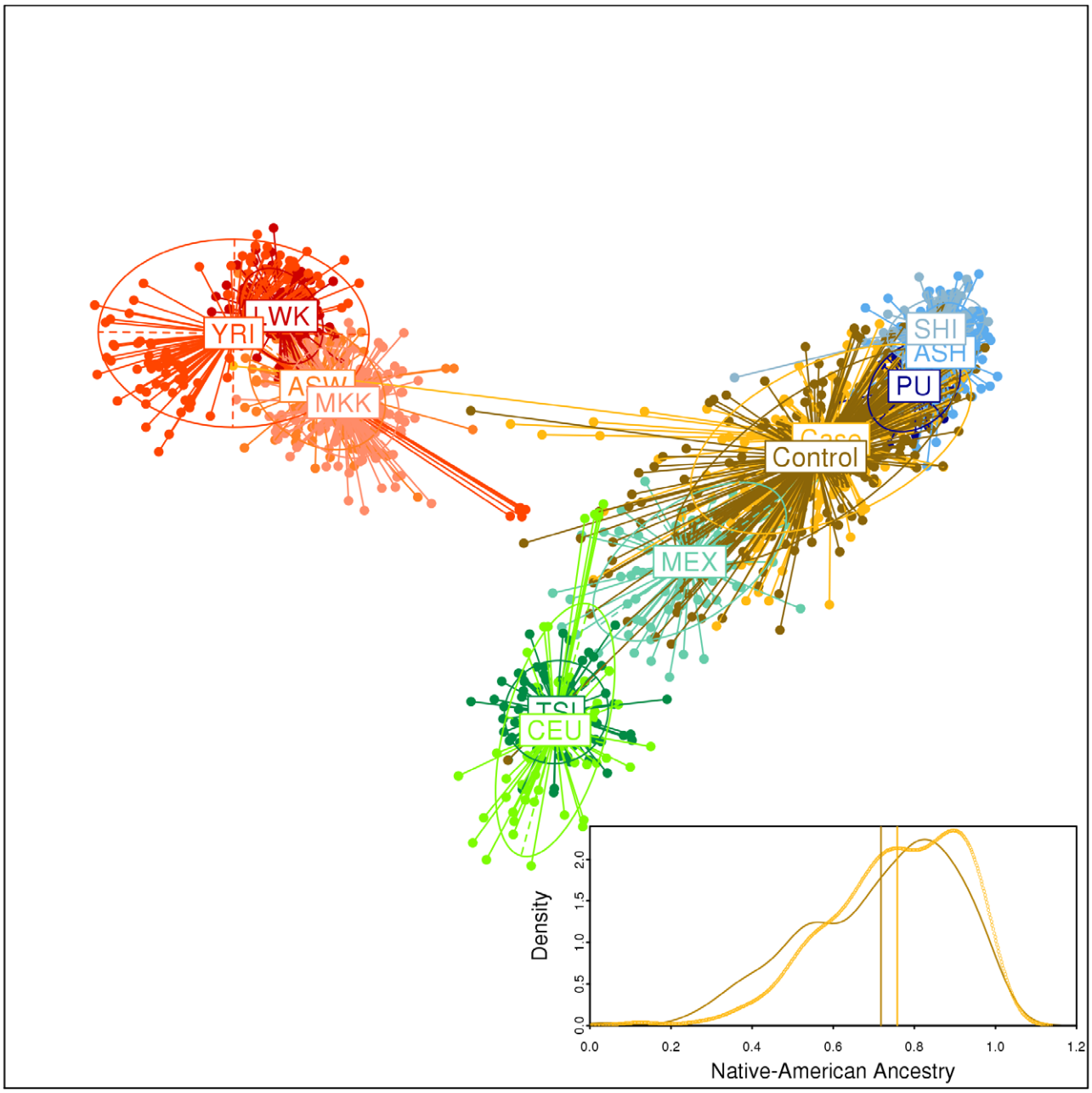


Figure 1. Principal Component Analysis of our Peruvian samples of gastric cancer cases, their controls, and Native Americans in the context of HapMap-III European, African, and Mexican individuals, and distribution of Native American ancestry in cases and controls (box). Each individual was genotyped for 103 ancestry informative markers validated by Yaeger et al. (2008). We represent the first (horizontal) and second (vertical) principal components, which capture the 35.3% and 7.7% of total variance, respectively. HapMap individuals: YRI: Yoruba from Nigeria, LWK: Luhya from Kenya, ASW: African American from Southwest USA, MKK: Maasai from Kenya, CEU: Utah residents with European ancestry, TSI: Toscani from Italy, MEX: Mexican ancestry resident in Los Angeles. Peruvian Native Americans: Shimaa (SHI) and Ashaninkas (ASH) from the Matsiguenga ethnic group, and individuals from Puno in the Andes (PU). The box within the figure shows the density plot and means (vertical lines) of Native American ancestry in the gastric cancer cases (yellow) and controls (brown).
doi:10.1371/journal.pone.0041200.g001

the Americas (see also [27]). Consistently, we observed that Native American ancestry is associated with variables that are indicators of poverty (many of which were also associated with gastric cancer) as well as their synthetic "socioeconomic" first factor of the multivariate analysis (Table 1). When we controlled for all covariates (i.e., the three factors of the multivariate analysis and age), the association between gastric cancer and Native American ancestry does not persist (OR = 1.28, 95%CI of the OR: 0.37–4.47, p = 0.69). Likewise, when we separately controlled for the effect of socioeconomic conditions (the first factor of the multivariate analysis) or age, the association between gastric cancer and Native American ancestry also does not persist (OR = 2.58, 95%CI of the OR: 0.83–8.07 for factor 1 and OR = 1.01, 95%CI of the OR: 0.99–1.02 for age).

Table 1. Socioeconomic, nutritional, and digestive-symptom-related variables and their association with gastric cancer and Native American ancestry.

Variables	P-value of association test with gastric cancer	P-value of association test with Native-American Ancestry
Personal Variables		
Gender	0.0416[b]	0.8621[e]
Ethnicity (self-identification)	0.8692[b]	<0.0001 [e]
Civil status	0.0017[b]	0.0305[f]
Birth in lima	0.3371[b]	0.0007[e]
Socioeconomic Variables		
Education level	0.0013[b]	0.0025[f]
Property of household	0.1241[b]	0.7714[f]
Material of household walls	<0.0001[b]	0.0029[e]
Material of household floor	0.0009[b]	0.0590[e]
Material of household ceiling	<0.0001[b]	0.0023[e]
Type of water supply	0.0007[b]	0.0036[e]
Type of sanitary service	0.0010[b]	<0.0001[e]
Type of garbage collection service	0.0002[b]	0.0056[e]
Fuel used for cooking	0.0149[b]	0.0023[e]
Possession of a refrigerator	<0.0001[b]	0.0020[e]
Possession of a freezer	0.2776[b]	0.5412[e]
Type of energy in the household	0.0057[b]	0.4103[e]
Type of water treatment	0.0069[b]	0.2028[f]
Number of adults in the household	0.0911[c]	0.6643[g]
Number of rooms in the household	0.0665[c]	0.0001[g]
Number of bathroom in the household	0.0009[c]	0.0003[g]
Number of children in the household	0.4894[c]	<0.0001[g]
Number of meals per day	0.9407[c]	0.8532[g]
Number of windows in the household	0.0001[c]	<0.0001[g]
Frequency of eating in a restaurant	0.0690[c]	0.0170[f]
Frequency of eating at the street	0.2345[c]	0.4067[f]
Frequency of eating at home	0.4426[c]	0.9031[f]
Household localization	0.0001[b]	0.0080[f]
Nutritional variables (frequency of consumption of)		
Spicy food	0.8556[c]	0.1903[f]
Steak	0.7363[c]	0.6443[f]
Fish	0.0020[c]	0.5319[f]
Poultry and birds	0.0415[c]	0.5995[f]
Fresh vegetables	0.2226[c]	0.6504[f]
Fresh Fruits	0.0587[c]	0.9235[f]
Tea	0.2864[c]	0.7307[f]
Coffee	0.8658[c]	0.2819[f]
Apple infusion	0.5182[c]	0.0881[f]
Coca leaf infusion	0.3320[c]	0.5237[f]
Symptoms		
Pain	<0.0001[c]	0.7270[f]
Burning	<0.0001[c]	0.4553[f]
Regurgitation	0.0540[c]	0.1220[f]
Nausea	<0.0001[c]	0.0805[f]
Vomit	<0.0001[c]	0.3120[f]
Heaviness	<0.0001[c]	0.2794[f]
Factors from multivariate factor analysis		

Table 1. Cont.

Variables	P-value of association test with gastric cancer	P-value of association test with Native-American Ancestry
Factor 1	0.00002 (OR[a] 0.68, 95%CI: 0.56–0.80)[d]	0.02017[g]
Factor 2	0.00039 (OR 0.70, 95%CI: 0.58–0.85)[d]	0.00003[g]
Factor 3	<0.0001 (OR 1.8895%CI: 1.54–2.29)[d]	0.0638[g]

Association tests reported in the table are: ([a]) OR: Odd ratio, ([b]): χ^2 test, ([c]): G-test, ([d]): logistic regression, ([e]): Mann-Withney, ([f]): Kruskal-Wallis, ([g]): Spearman rank order correlation.

doi:10.1371/journal.pone.0041200.t001

Discussion

We performed a case-control study in the urban admixed population from Lima (Peru) and determined that Native American individual ancestry is associated with gastric cancer. However, this association seems primarily to be due to the association of socioeconomic variables both with gastric cancer and with Native American ancestry. Consistently, although the association of ancestry with gastric cancer is significant ($p = 0.011$), ancestry only explains 1.6% of the variance in disease status. When non-genetic covariates are included, their joint effect with ancestry explains 22.3% of the variance in disease status. Therefore, the high incidence of gastric cancer in Peru [1,28] does not seem to be due to the presence of common susceptibility genetic variants more frequent in Native American populations, but rather to a combination of socioeconomic factors present in this population. However, further studies with larger sample sizes are needed to explore this observation, since the power to detect ancestry genetic effects was limited in the current study. This result is consistent with the relative decrease in gastric cancer incidence in the United States during the last decades, due to the improvement of socioeconomic conditions [29].

Accuracy of ancestry estimations depends on several issues. The first is the number and the nature of markers used to estimate admixture. The 103 AIMs used in this study contain enough information to produce acceptable admixture estimates [26,30]. Galanter et al. have also showed that a panel of more than 88 AIMs contains enough information to estimate individual admixture with accuracy [31]. A second pervasive methodological issue in estimating admixture is the difficulty in using data from the most representative parental African, European, and Native American populations of the admixed group. In this case, we included as proxy for the parental populations European and African individuals from the HapMap project, and a set of Peruvian Native Americans from the Peruvian Andes and neighboring Eastern areas. While this choice may not be optimal, the 103 SNPs used, being AIMs, mitigated this issue because their frequencies are very different among the parental ethnic groups and highly homogeneous within them [26]. Thus, the use of markers with these characteristics renders our results robust to the choice of suboptimal parental populations. Third, different methods to estimate individual admixture may produce slightly different results even starting out from the same dataset. To test the robustness of our admixture results in respect to the admixture estimation methods, we reanalyzed the data using the alternative maximum-likelihood approaches proposed by Tang et al. [32] and implemented in the software Frappe, and the method by Alexander et al. implemented in the software Admixture v. 1.2 [33]. The three methods produced highly correlated results (Figure S2) and the same pattern of association with gastric cancer (data not shown).

Peruvian individuals born in the countryside have more Native American ancestry on average than do residents of large urban centers (Table 1). In a case-control study, cases may frequently include individuals with more Native American ancestry because they are referred from small countryside health centers to large urban hospitals to receive better healthcare. This referral pattern has less effect on controls, and therefore may create a spurious association of Native American ancestry with disease. However, we recorded places of birth for all study participants, and thereby controlled this potential confounding factor: when we included the place of birth (Lima vs. countryside) as a covariate in the logistic regression, the association of Native American ancestry with gastric cancer persisted, although at a lower significance ($p = 0.05$ vs. $p = 0.011$). We conclude that the association between gastric cancer and ancestry is not an artifact of referrals of countryside individuals with high Native American ancestry to our study hospitals. This result emphasizes the importance of gathering birthplace and residence data for genetic association studies with diseases in Latin America, and other regions where most European colonization and admixture occurred in cities, and more autochthonous individuals predominate in rural areas, such as in Melanesia and South Africa.

In this study, we included as controls individuals with intestinal metaplasia ($n = 46$), assuming that this does not increase the risk of developing gastric cancer [23,24,25]. However, this assumption is not universally accepted [34]. When we alternatively assume three ordinal categories of disease risk (i.e., individuals without intestinal metaplasia, with intestinal metaplasia, and with gastric cancer), this progression, assessed by an ordinal logistic regression is also associated with Native American ancestry (OR = 2.83, 95%CI of the OR: 1.10–7.29, $p = 0.031$), but again, this association does not persist when controlled for all covariates (factor 1, 2, and 3 and age) (OR = 1.08, 95%CI of the OR: 0.35–3.34, $p = 0.897$). Thus, our results do not depend on the inclusion of intestinal metaplasia individuals as controls.

An issue in our experimental design is that controls were selected as symptomatic individuals attending a gastroenterology service, undergoing an endoscopy and most of them with a gastric lesion: 6 with histologically normal gastric mucosa, 248 with gastritis, and 46 with metaplasia. It could be argued that the optimal control would be composed only by individuals with normal gastric mucosa. However, only through an endoscopy is it possible to accurately ascertain the absence of gastric lesions, and performing an endoscopy for research purposes only, not motivated by gastric-related symptoms is no longer ethically acceptable. On the other hand, using as controls individuals from the general population who did not undergo endoscopy is not necessarily a better choice since this strategy would have included as controls individuals with undetected gastritis [35] and other similar lesions. Therefore, we believe that our controls are the better operational choice for this study, because they are

individuals attending the same gastroenterology services as the cases, and since they have undergone an endoscopy, we accurately know the status of their gastric mucosa. Also, in Latin America there is a considerable level of population stratification due to socioeconomic level and ancestry, that are correlated in large urban centers in Latin America (in addition to this study, see Avena et al. 2012 [30] for an example in Buenos Aires and Campbell et al. 2012 [27]). By selecting controls among attendants of the same hospital than cases, we mitigate this other potential source of population stratification between cases and controls.

In this study we report a surprisingly high Native American ancestry (>74%) both in controls and cases attending public hospitals from the now cosmopolitan city of Lima, the national capital that was also the capital of the Spaniard Viceroyalty of Peru for five centuries and therefore, the center of Spaniard colonial power. Large cities in Ecuador, Peru, Bolivia, and Northern Argentina host populations whose usual cultural identification as *mestizo* likely ignores their large Amerindian genetic background. These urban populations are frequently peopled by immigrants from rural areas. Our results suggest that large cities of Western South America host millions of individuals of predominantly Amerindian genetic background. Contemporary international South-to-North migrations from South American cities from the Andean region are also spreading the genetic background of Native Americans worldwide, and it is expected that the almost one million United States immigrants coming from Andean countries [36] have high levels of Native American background. It would not be surprising if these populations, classified as "Hispano/Latino" in the United States, had more Amerindian ancestry than US individuals classified as Native American.

In conclusion, we showed that in the urban admixed population from Lima, Native American individual ancestry is associated with gastric cancer, but this is explained by the association of socioeconomic variables with both gastric cancer and Native American ancestry. Despite the high incidence of gastric cancer in the Peruvian population with a very high Native American ancestry, our result shows that this epidemiological observation does not rely on a genetic basis, suggesting a predominant role for socioeconomic factors and disparities in access to health services. We report a surprisingly high Native American ancestry (>74%) in individuals attending hospitals from the now cosmopolitan city of Lima. Since Native Americans are a neglected group in genomic studies, we suggest that the population from Lima and other large cities in Western South America may be convenient targets of epidemiological studies focused on Native American populations. Pursuit of this avenue of research in subsequently larger studies will begin to close the gap [37] in genetic studies and their potential benefits between European individuals and those from other generally less well served populations.

Materials and Methods

Subjects

For this study we initially recruited 576 individuals attending Gastroenterology Divisions of the following three hospitals in Lima, between 2006 and 2009: Arzobispo Loayza, Dos de Mayo (both government-ruled), and the cancer-specialized Instituto Nacional de Enfermedades Neoplásicas. Each participant answered a questionnaire to record age, gender, place of birth and of residence, and other socioeconomic, nutritional, and clinical information (Table 1). Cases were adults referred for endoscopy and whose biopsies were confirmed positive for gastric cancer by histopathological analysis. The control group was also composed of adult individuals referred for endoscopy, but whose biopsies proved negative for gastric cancer. Individuals with premalignant lesions such as dysplasia (n = 3) or with presence of stomach tumors other than gastric cancer (n = 16) were excluded from the study. Individuals with intestinal metaplasia (n = 46) were included as controls (see Discussion). The final number of subjects considered to test in the association study was 241 gastric cancer cases and 300 controls.

We also collected samples from 296 Native American individuals who were used as parental populations to estimate ancestry of the gastric cancer cases and controls. These include 23 farmers from Pichacani (Puno), belonging to the predominant Andean Quechua ethnic group, as well as 87 Shimaa and 186 Ashaninka from the Matsiguenga ethnic group, settled between the Andes and the Amazonian region. For all gastric cancer cases and controls and for the Native Americans, we extracted genomic DNA using the phenol-chlorophorm method described by Sambrook et al. with modifications, or the Gentra Puregene blood kit (Qiagen, USA) [38]. This investigation was approved by IRBs of Asociación Benéfica PRISMA, Universidad Peruana Cayetano Heredia, Johns Hopkins University, Universidade Federal de Minas Gerais, Hospital Arzobispo Loayza, Hospital Dos de Mayo and the Instituto Nacional de Enfermedades Neoplásicas. All participants in the study provided written informed consent.

Ancestry informative markers (AIMs) and genotyping

To estimate ethnicity for each of the study subjects, we genotyped 106 SNPs that are informative for African, European, and Native American ancestry [26]. The genotyping was performed at the Biomedical Genomic Center of the Children's Hospital Oakland Research Institute (University of Minnesota, MN, USA), using the Sequenom iPLEX platform (San Diego, CA, USA). Briefly, it is based on an allele-specific primer extension followed by separation of alternative alleles by mass spectrometry. The genotyping involved four multiplexed assays, three containing 26 SNPs and one containing 28 SNPs. Before genotyping, DNA samples underwent a Quality Control (QC) procedure that consisted of: (1) a non-allelic quantitative-PCR analysis that measures the quantity of PCR-amplifiable DNA and (2) an end-point reading from a Taqman SNP genotyping assay (Applied Biosystems, Palo Alto, CA, USA) that, in addition to providing a second assessment of the ability of PCR to amplify each sample, is a sensitive indicator of sample-to-sample cross-contamination. After we removed the SNPs rs30125 and rs888861, which showed a call rate <95%, the average call rate for the SNPs was 99.7%. Of the 106 markers, 104 robustly generated call rates for at least 95% of samples, but for the SNP rs2592888 there are no genotypes publicly available for the Hapmap populations and the SNP was excluded from further analyses. Thus, we used genotypes for 103 SNPs to estimate admixture (see Table S3 for the complete list, with their allele frequencies in the study populations).

Ancestry estimates

We estimated individual ancestry using the following three parental groups composed of unrelated individuals: (1) West African Yoruba from Nigeria (YRI – 118 individuals from the HapMap II/III project); (2) Utah individuals with European ancestry available at the *Centre d'Etude du Polymorphisme Humain*-CEPH collection (CEU – 60 individuals from the HapMap II Project) and (3) 296 Peruvian Native Americans collected by our group. The genotypes for Africans and Europeans were obtained from the public HapMap database [39], while the Native Americans were genotyped for this study.

We estimated the individual ancestry and its 90% credibility interval using the method implemented in the program Structure

2.3.3 [40,41]. It fits a Bayesian probability model of population structure and admixture using a Markov Chain Monte Carlo (MCMC) procedure, estimating the contribution of K parental populations to the genomes of individuals from the admixed population. We assumed that the HapMap YRI and CEU and our Native American samples were representative of the parental populations and that the gastric cancer cases and controls from Lima were admixed individuals. For this data set, each Structure run had 50,000 burn-in steps followed by 250,000 MCMC steps, and was repeated three times to allow checking for the robustness of the results. This length of the run and the checking procedure exclude the undesirable lack of convergence of the Markov Chains, which happens when the procedure does not properly explore the space of model parameters. All runs were performed assuming three clusters (K = 3), lambda was set to 1.0, and α parameters were estimated for each of the three clusters, GENSBACK = 2, MIGRPRIOR = 0.05 and we did not use a priori information for the individuals from parental populations to assist the clustering (USEPOPINFO = 0).

Statistical analysis

To represent the genetic structure of our samples in the context of parental population diversity, we performed principal component analysis (PCA) of individual genotypes (Figure 1), as implemented in the software Adegenet and Ade4 for R environment [42,43]. We also used Ade4 to apply a clustering method that is based on the PCA two-dimensional representation of individuals and their population centroid, designing bi-dimensional ellipses of dispersion. In addition to gender, age and self-reported ethnicity, a wide set of socioeconomic, nutritional, and clinical information was collected including civil status, place of birth, education level, household conditions, eating habits, frequency of consumption of fruits, vegetables, meat and poultry, infusions, as well as gastric-related symptoms. This information was organized in binary, ordinal, or categorical variables, as detailed in Table S4). We excluded variables that had more than 10% missing data, and a final set of 43 variables were included in subsequent analyses. To reduce the dimensionality of this set of personal, socioeconomic, nutritional, and clinical non-genetic variables we performed a multivariate factor analyses using the Statistical Package for the Social Sciences (SPSS) software (SPSS 19 for Windows, SPSS Inc, Chicago, IL, USA). The factor analysis synthesizes the variance of the original set of variables in a specified minor number of transformed variables (in our case 3), called factors. Each factor captures correlated information on the original dataset but the factors are uncorrelated among them [44].

To test the association between these non-genetic variables or its representation obtained by multivariate factor analyses with gastric cancer (a binary trait) and ancestry (a continuous trait), we used the following statistical tests: logistic regression for continuous vs. binary traits (or Ordinal Logistic Regression for continuous vs. an ordinal dependent variable), G-test for ordinal vs. binary traits, χ^2 test for categorical vs. binary traits, Spearman rank order correlation for continuous vs. continuous traits, and Mann-Withney (2 categories) or Kruskal-Wallis (>2 categories) tests for ordinal vs. continuous traits. These analyses were performed in R environment. Ordinal logistic regression was performed using the 'rms' R package [45].

To test the association of ethnicity with gastric cancer we analyzed 443 individuals (245 cases and 198 controls) for whom we have a histopathological diagnosis, collected personal, socioeconomic, dietary data, and estimated ancestry. We used the logistic regression (observing the R^2 Nagelkerk value – SPSS software) to test the association of ancestry with gastric cancer, which allowed us to control the effect of potential confounding variables (i.e., covariates) that were associated with disease status or ancestry. We included as covariates the original set of non-genetics variables or its factor-analysis synthetic representation.

Supporting Information

Figure S1 Barplot of individual ancestry estimated with the software Structure for Africans (red), Europeans (green), and Native Americans (blue), as well as gastric cancer cases and controls.
(TIF)

Figure S2 Scatterplot and Spearman correlation between individual Native American ancestry estimates by Structure versus Frappe (a) and Admixture (b) methods. Frappe was run with 100,000 maximum iteration of EM, K = 3 and 10,000 optional convergence threshold. Variations of these parameters did not show differences in results. Admixture was run using the default parameters with K = 3.
(TIF)

Table S1 Distribution of cases and controls across hospitals and association with Native American ancestry.
(DOC)

Table S2 Socioeconomic, nutritional, and digestive-symptom-related variables, their LOD scores with the three first factors of the multivariate factor analysis and significance of Spearman correlation between individual values of the variables and coordinates on each factor.
(DOCX)

Table S3 Allele frequencies in the populations included in this study for the 103 Ancestry Informative Markers used in the study.
(DOC)

Table S4 Classification of socioeconomic, nutritional, and digestive-symptom-related variables used in Table 1 and their values.
(DOC)

Acknowledgments

We thank José Claudio Rocha, Ricardo Alves Silva, Dulciene Queiroz, Ana Lúcia Brunialti and Wagner Magalhães for discussions on different parts of the project, and Hanaisa Sant'Anna for their logistic and technical help.

Author Contributions

Conceived and designed the experiments: ETS RHG. Performed the experiments: LP RZ LC CH JMC. Analyzed the data: LP RZ LC CH JC SS GSS FK. Contributed reagents/materials/analysis tools: PH LC CH JMC JC GV WAP SJC DB MRR. Wrote the paper: LP ETS.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, et al. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer.

2. Recavarren-Arce S, Leon-Barua R, Cok J, Berendson R, Gilman RH, et al. (1991) Helicobacter pylori and progressive gastric pathology that predisposes to gastric cancer. Scand J Gastroenterol Suppl 181: 51–57.

3. Correa P (1992) Human gastric carcinogenesis: a multistep and multifactorial process – First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res 52: 6735–6740.
4. Asaka M, Dragosics BA (2004) Helicobacter pylori and gastric malignancies. Helicobacter 9 Suppl 1: 35–41.
5. Atherton JC (2006) The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases. Annu Rev Pathol 1: 63–96.
6. Kersulyte D, Mukhopadhyay AK, Velapatino B, Su W, Pan Z, et al. (2000) Differences in genotypes of Helicobacter pylori from different human populations. J Bacteriol 182: 3210–3218.
7. Kersulyte D, Kalia A, Gilman RH, Mendez M, Herrera P, et al. (2010) Helicobacter pylori from Peruvian amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain. PLoS One 5: e15076.
8. Suzuki M, Kiga K, Kersulyte D, Cok J, Hooper CC, et al. (2011) Attenuated CagA oncoprotein in Helicobacter pylori from Amerindians in Peruvian Amazon. J Biol Chem 286: 29964–29972.
9. Mohebbi M, Wolfe R, Jolley D, Forbes AB, Mahmoodi M, et al. (2011) The spatial distribution of esophageal and gastric cancer in Caspian region of Iran: An ecological analysis of diet and socio-economic influences. Int J Health Geogr 10: 13.
10. Campbell PT, Sloan M, Kreiger N (2008) Dietary patterns and risk of incident gastric adenocarcinoma. Am J Epidemiol 167: 295–304.
11. Yang WG, Chen CB, Wang ZX, Liu YP, Wen XY, et al. (2011) A case-control study on the relationship between salt intake and salty taste and risk of gastric cancer. World J Gastroenterol 17: 2049–2053.
12. Navarro Silvera SA, Mayne ST, Risch HA, Gammon MD, Vaughan T, et al. (2011) Principal component analysis of dietary and lifestyle patterns in relation to risk of subtypes of esophageal and gastric cancer. Ann Epidemiol 21: 543–550.
13. Steevens J, Schouten LJ, Goldbohm RA, van den Brandt PA (2011) Vegetables and fruits consumption and risk of esophageal and gastric cancer subtypes in the Netherlands cohort study. Int J Cancer.
14. Perez-Perez GI, Garza-Gonzalez E, Portal C, Olivares AZ (2005) Role of cytokine polymorphisms in the risk of distal gastric cancer development. Cancer Epidemiol Biomarkers Prev 14: 1869–1873.
15. Abnet CC, Freedman ND, Hu N, Wang Z, Yu K, et al. (2010) A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet 42: 764–767.
16. Savage SA, Abnet CC, Mark SD, Qiao YL, Dong ZW, et al. (2004) Variants of the IL8 and IL8RB genes and risk for gastric cardia adenocarcinoma and esophageal squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev 13: 2251–2257.
17. Taguchi A, Ohmiya N, Shirai K, Mabuchi N, Itoh A, et al. (2005) Interleukin-8 promoter polymorphism increases the risk of atrophic gastritis and gastric cancer in Japan. Cancer Epidemiol Biomarkers Prev 14: 2487–2493.
18. Ohyauchi M, Imatani A, Yonechi M, Asano N, Miura A, et al. (2005) The polymorphism interleukin 8–251 A/T influences the susceptibility of Helicobacter pylori related gastric diseases in the Japanese population. Gut 54: 330–335.
19. Rocha GA, Guerra JB, Rocha AM, Saraiva IE, da Silva DA, et al. (2005) IL1RN polymorphic gene and cagA-positive status independently increase the risk of noncardia gastric carcinoma. Int J Cancer 115: 678–683.
20. Ke-Xiang Z, Yu-Min L, Xun L, Wen-Ce Z, Yong S, et al. (2011) Study on the association of COX-2 genetic polymorphisms with risk of gastric cancer in high incidence Hexi area of Gansu province in China. Mol Biol Rep 38: 649–655.
21. Dean B, Levi JM (2003) At the Risk of Being Heard: Identity, Indigenous Rights, and Postcolonial States. Ann Arbor: University of Michigan Press.
22. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 103: 14068–14073.
23. Filipe MI, Munoz N, Matko I, Kato I, Pompe-Kirn V, et al. (1994) Intestinal metaplasia types and the risk of gastric cancer: a cohort study in Slovenia. Int J Cancer 57: 324–329.
24. Correa P, Piazuelo MB, Wilson KT (2010) Pathology of gastric intestinal metaplasia: clinical implications. Am J Gastroenterol 105: 493–498.
25. Gonzalez CA, Pardo ML, Liso JM, Alonso P, Bonet C, et al. (2010) Gastric cancer occurrence in preneoplastic lesions: a long-term follow-up in a high-risk area in Spain. Int J Cancer 127: 2654–2660.
26. Yaeger R, Avila-Bront A, Abdul K, Nolan PC, Grann VR, et al. (2008) Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa. Cancer Epidemiol Biomarkers Prev 17: 1329–1338.
27. Campbell DD, Parra MV, Duque C, Gallego N, Franco L, et al. (2012) Amerind ancestry, socioeconomic status and the genetics of type 2 diabetes in a Colombian population. PLoS One 7: e33570.
28. Mendoza D, Herrera P, Gilman RH, Lanfranco J, Tapia M, et al. (2008) Variation in the prevalence of gastric cancer in Peru. Int J Cancer 123: 414–420.
29. Jemal A, Center MM, DeSantis C, Ward EM (2010) Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol Biomarkers Prev 19: 1893–1907.
30. Avena S, Via M, Ziv E, Perez-Stable EJ, Gignoux CR, et al. (2012) Heterogeneity in genetic admixture across different regions of Argentina. PLoS One 7: e34695.
31. Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, et al. (2012) Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet 8: e1002554.
32. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28: 289–301.
33. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664.
34. Genta RM, Rugge M (2001) Review article: pre-neoplastic states of the gastric mucosa – a practical approach for the perplexed clinician. Aliment Pharmacol Ther 15 Suppl 1: 43–50.
35. Dooley CP, Cohen H, Fitzgibbons PL, Bauer M, Appleman MD, et al. (1989) Prevalence of Helicobacter pylori infection and histologic gastritis in asymptomatic persons. N Engl J Med 321: 1562–1566.
36. U.S.CensusBureau (2009) 2009 American Community Survey 1-Year Estimates Washington, DC:US Census Bureau.
37. Bustamante CD, Burchard EG, De la Vega FM (2011) Genomics for the world. Nature 475: 163–165.
38. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: A Laboratory Manual. New York: Cold Spring Harbor.
39. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
40. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
41. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9: 1322–1332.
42. Chessel D, Dufour AB, Thiouluose J (2004) The ade4 package – I: one-table methods. R News 4: 5–10.
43. Jombart T, Ahmed I (2011) adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27: 3070–3071.
44. Agresti A (2012) Categorical Data Analysis: John Wiley & Sons. 768 p.
45. Harrell Jr FE (2012) Regression Modeling Strategies.

CAPÍTULO 3 - ESTRUTURA GENÉTICA E EVOLUÇÃO DOS POLIMORFISMOS EM POPULAÇÕES NATIVO-AMERICANAS

A identificação de polimorfismos que apresentem diferenças nas frequências alélicas entre as populações humanas constitui parte essencial na associação entre a diversidade genética e a variabilidade fenotípica. A partir da seleção de polimorfismos divergentes em populações nativo-americanas é possível esclarecer traços da história evolutiva deste grupo populacional, identificando os possíveis impactos dos eventos demográficos e evolutivos na saúde das populações autóctones e, também, das populações miscigenadas da América Latina. Como demonstrado no capítulo 2, as populações andinas e, em especial, a população Quéchua, constitui um dos principais reservatórios da variabilidade genética encontrada nos nativo-americanos da América do Sul. Além disso, em várias populações urbanas da América Espanhola, como Lima no Peru, a ancestralidade dos indivíduos aí residentes é majoritariamente nativo-americana, e mesmo nas populações urbanas do Brasil, como as do estado de Minas Gerais, a ancestralidade ameríndia encontra-se presente.

Desta forma, diferentes configurações demográficas foram utilizadas na descrição da estrutura genética dos loci através dos níveis hierárquicos de variância molecular com o objetivo de identificar polimorfismos que possam levar a resultados falso-positivos em estudos de associação e elucidar aspectos histórico-demográficos e de seleção natural. Para tanto, utilizamos grupos de populações do Leste Asiático, África Ocidental, Europa e Nativo-americanas. Esses agrupamentos populacionais foram construídos de acordo com os principais eventos evolutivos constitutivos das populações americanas, sejam elas autóctones ou miscigenadas.

Em um estudo anterior, a dissertação de mestrado (SOARES-SOUZA, 2010), apresentou resultados sobre a investigação da variabilidade genética em populações humanas. No estudo de 2010 foram utilizados os dados de 1442 SNPs distribuídos por 411 genes, importantes em imunidade, carcinogênese e fármaco-genética, a partir de 1198 indivíduos provenientes de 56 populações distribuídas por 5 continentes, agrupados em 13 grupos populacionais (Anexos B e C). A partir desses dados foram calculadas estimativas de diversidade populacionais e de polimorfismos e os resultados obtidos previamente são

apresentados em resumo nesta seção. A tabela 1 apresenta os valores de diversidade calculados a partir do AMOVA para nove grupos populacionais. Observa-se que a maior parte da variabilidade encontra-se dentro das populações e a menor parte, entre as populações dentro dos grupos.

Tabela 1: Análise de Variabilidade Molecular por Continente

VAR	SAM	CAM	WAFR	EAFR	ORM	EUR	CSA	EAS	OCE
Entre Populações	13,07	7,78	5,27	4,95	1,84	1,17	1,95	1,99	9,49
Dentro das Populações	86,93	92,22	94,73	95,05	98,16	98,83	98,05	98,01	90,51

Os valores descritos nessa tabela correspondem ao cálculo de F_{ST} realizado em cada continente utilizando o modelo de dois componentes de Análise de Variabilidade Molecular (intra e interpopulacional). VAR: Partição da variabilidade. Siglas dos Grupos Populacionais: SAM: América do Sul; CAM: América Central; WAFR: África Ocidental; EAFR: África Oriental; ORM: Oriente Médio; EUR: Europa; CSA: Centro Sul Asiático; EAS: Leste Asiático; OCE: Oceania. Fonte: (SOARES-SOUZA, 2010)

Identificou-se 196 polimorfismos distribuídos por 111 genes com alta divergência interpopulacional na configuração EUR-NAT-WAFR (Europeus – Nativo-Americanos – Africanos do Oeste). Destes marcadores, 35 apresentam alta divergência entre (EUR) – (NAT-WAFR); 53 entre (NAT) – (EUR-WAFR); 75 entre (WAFR) – (EUR-NAT); e para 33, os níveis de variância interpopulacional são semelhantes (Tabela 2).

Tabela 2: Freqüências alélicas para os grupos populacionais África Ocidental, Europa e Nativo-Americanos

Gene	SNP	Gen	EUR	NAT	WAFR	Gene	SNP	Gen	EUR	NAT	WAFR
ABCA1	ABCA1-17	A/T	0,2089	0,1834	0,8839	*IL15*	IL15-02	T/C	0,3671	0,0263	0,1000
AKR1C3	AKR1C3-11	G/A	0,3822	0,8743	0,4732		IL15-06	C/T	0,4808	0,0318	0,3839
	AKR1C3-36	G/A	0,3228	0,8634	0,3393	*IL1B*	IL1B-03	C/T	0,3000	0,8924	0,5926
AMACR	AMACR-03	G/T	0,1424	0,0088	0,5273	*IL2*	IL2-03	T/G	0,3227	0,6994	0,0535
	AMACR-17	G/A	0,4367	0,7485	1	*IL4*	IL4-01	T/C	0,1392	0,6311	0,6574
ANKK1	ANKK1-01	A/G	0,1783	0,6363	0,3909		IL4-03	T/C	0,1338	0,6235	0,4455
APC	APC-09	T/C	0,3354	0,2000	0,8303		IL4-10	A/C	0,1369	0,6324	0,3000
AURKA	AURKA-16	C/T	0,2468	0,7882	0,0893		IL4-11	C/A	0,1378	0,6308	0,2143
BCL2L1	BCL2L1-01	T/G	0,2134	0,0203	0,5463	*IL4R*	IL4R-02	C/A	0,0892	0,0953	0,7091
	BCL2L1-02	C/T	0,2908	0,0303	0,6321		IL4R-07	G/T	0,0621	0,0087	0,3704
BCL6	BCL6-09	T/C	0,3365	0,0714	0,9018	*IL6*	IL6-01	C/G	0,3662	0,0118	0
BIC	BIC-07	T/C	0,2342	0,7308	0,6786		IL6-04	A/G	0,3548	0,0117	0
	BIC-10	T/G	0,2389	0,7289	0,6250	*IL6R*	IL6R-04	C/A	0,3397	0,7083	0,0363

	BIC-15	C/T	0,1709	0,6794	0,4107	*IL7R*	IL7R-01	G/A	0,2866	0,7093	0,0982
	BIC-32	G/A	0,2342	0,7289	0,7364	*INSR*	INSR-13	G/A	0,4363	0,1192	0,8393
	BIC-34	G/A	0,2500	0,7278	0,7411	*KRT23*	KRT23-03	T/C	0,2690	0,7169	0,2130
BRIP1	BRIP1-01	A/G	0,4427	0,8533	0,1339	*LCAT*	LCAT-05	G/A	0,1690	0,1804	0,7600
	BRIP1-09	T/C	0,2057	0,7794	0,3304	*LIPC*	LIPC-04	G/T	0,1139	0,1898	0,7411
CASP3	CASP3-08	C/T	0,2595	0,9556	0,1727		LIPC-37	T/C	0,0601	0	0,3304
CASP8	CASP8-07	G/T	0,0892	0,0116	0,5278	*LRP5*	LRP5-01	T/C	0,2633	0,7107	0,1364
CASR	CASR-11	A/G	0,2821	0,9737	0,2232	*MASP1*	rs696405	A/C	0,3548	0,1316	0,7500
CAT	CAT-02	G/A	0,3636	0,8107	0,4907	*MATR3*	MATR3-01	T/A	0,2911	0,4881	0,9554
CAV1	CAV1-19	T/A	0,3636	0,1893	0,4907	*MBL2*	MBL2-46	C/T	0,4272	0,0439	0,8273
	CAV1-29	C/T	0,3291	0,0117	0,4107	*MSH3*	MSH3-02	A/G	0,0601	0,5292	0,1785
CDK5	CDK5-08	C/A	0,2468	0,8713	0,1909		MSH3-07	G/C	0,1361	0,5407	0,1339
	CDK5-16	A/G	0,2338	0,8713	0,2636	*MTRR*	MTRR-19	T/A	0,3892	0,8982	0,5536
CDKN2A	CDKN2A-03	T/C	0,0886	0,6455	0,1875	*MX1*	MX1-11	G/A	0,4363	0,8970	0,8571
CGA	CGA-06	G/A	0,2816	0,6559	0,1161		MX1-22	C/T	0,3070	0,0146	0,5893
CYP19A1	CYP19A1-01	G/A	0,4841	0,0617	0,1250		MX1-28	G/A	0,500	0,0896	0,1607
	CYP19A1-04	T/G	0,4905	0,9041	0,8661	*MYBL2*	MYBL2-03	A/G	0,0854	0,0058	0,3868
	CYP19A1-06	T/G	0,4905	0,9077	0,8273	*MYC*	MYC-02	A/T	0,1487	0,1928	0,7545
	CYP19A1-08	T/G	0,2722	0,8488	0,2500	*MYNN*	MYNN-01	G/A	0,2994	0,6518	0,0182
	CYP19A1-29	C/T	0,4744	0,0643	0,0818	*NCF2*	NCF2-03	G/A	0,4209	0,8520	0,2856
	CYP19A1-30	G/T	0,0443	0,2235	0,6339		NCF2-04	A/G	0,4236	0,8529	0,2818
	CYP19A1-34	T/C	0,481	0,0636	0,1161	*NCOA3*	NCOA3-04	G/A	0,0886	0,0232	0,4196
	CYP19A1-39	T/C	0,4583	0,0581	0,2054	*NFKB1*	NFKB1-02	C/T	0,3544	0,7156	0,0089
CYP1A1	CYP1A1-14	G/T	0,3000	0,8876	0,9821		NFKB1-33	T/A	0,3590	0,7147	0,1518
CYP1B1	CYP1B1-27	C/T	0,4548	0,6717	0	*NFKBIE*	NFKBIE-02	T/C	0,5000	0,2485	0,900
	CYP1B1-31	C/A	0,1677	0,0268	0,7273	*NR1H4*	NR1H4-05	C/G	0,3766	0,8601	0,6111
CYP2E1	CYP2E1-02	G/C	0,1772	0,0446	0,7232	*OCA2*	OCA2-23	G/A	0,4525	0,9315	0,8750
	CYP2E1-31	G/T	0,0860	0,4357	0,6296	*PAK6*	PAK6-13	A/C	0,3874	0,7981	0,0185
CYP3A7	CYP3A7-01	C/T	0,0949	0,2471	0,6607	*PCNA*	PCNA-10	C/A	0,1677	0,0462	0,7232
DHDH	DHDH-02	G/C	0,2166	0,6627	0,8393	*PCTP*	PCTP-01	G/A	0,1146	0,3314	0,6875
DRD2	DRD2-03	A/G	0,1442	0,6317	0,2232	*PHB*	PHB-02	G/A	0,4522	0,0088	0
EFNB3	EFNB3-02	G/A	0,3981	0,8185	0,9364	*PIM1*	PIM1-03	G/A	0,2949	0,0202	0,4911
ENPP1	ENPP1-04	A/T	0,0570	0,0058	0,4091		rs1233291	C/G	0,2722	0,2246	0,9732
EPHX2	EPHX2-04	C/A	0,2612	0,1441	0,7143		rs1233297	T/C	0,2816	0,2135	0,9022
ERCC1	ERCC1-05	C/T	0,4025	0,8274	0,9464		rs1233299	C/A	0,2707	0,2130	0,8482
	ERCC1-06	G/C	0,3790	0,8485	0,9107		rs1233302	A/C	0,2727	0,2209	0,9259
ERCC5	ERCC5-01	T/C	0,3829	0,8706	0,2767	*PMS1*	rs256550	C/T	0,1346	0,2076	0,7232
ESR1	ESR1-17	T/G	0,0949	0,0152	0,5833		rs256552	G/A	0,1424	0,2078	0,7232
FANCA	FANCA-03	T/C	0,2660	0,8385	0,4727		rs256563	G/A	0,1424	0,2118	0,7182
	FANCA-16	G/C	0,2660	0,8363	0,6545		rs256564	A/G	0,1424	0,2081	0,7232
	FANCA-22	C/T	0,3510	0,8567	0,7321		rs256567	T/C	0,1424	0,2135	0,7143

	FANCA-28	A/G	0,2677	0,8445	0,5189		rs5742938	G/A	0,2707	0,2091	0,8611
	FANCA-35	C/G	0,2742	0,8343	0,5566	*POLB*	POLB-05	C/T	0,1146	0,0149	0,8241
	FANCA-37	A/T	0,3462	0,8410	0,8455		POLB-08	G/A	0,0728	0,0266	0,6818
FASLG	FASLG-01	A/G	0,4172	0,7719	0,0268		POLB-16	G/A	0,0728	0,0203	0,6852
FBXW7	FBXW7-01	A/G	0,2643	0,5473	0,8750	*POLD1*	POLD1-13	C/T	0,0947	0,0507	0,4909
	FBXW7-05	C/T	0,2866	0,0029	0	*RAD51*	rs2412546	G/A	0,4490	0,1036	0,7768
	FBXW7-44	A/G	0,2722	0,4360	0,9732		rs2619679	A/T	0,4430	0,1047	0,7857
FUT2	FUT2-05	C/T	0,4712	0,0349	0,3929		rs2619681	T/C	0,1465	0,7602	0,1827
GATA3	GATA3-25	G/C	0,2532	0,0265	0,5446		rs4924496	T/C	0,3903	0,0298	0,3909
GDF15	GDF15-02	A/T	0,1784	0,6717	0,1696	*RAD52*	RAD52-07	C/T	0,3662	0,2006	0,7857
GHR	GHR-21	T/C	0,1582	0,0087	0,5625	*RAG1*	RAG1-01	G/A	0,1170	0,5977	0,0714
	GHR-47	A/C	0,2373	0,0089	0,6455	*RB1CC1*	RB1CC1-10	C/T	0,1762	0,6598	0,0818
GPX2	GPX2-17	A/G	0,2917	0,3825	0,9273		RB1CC1-24	C/T	0,1815	0,6479	0,0893
	GPX2-21	T/C	0,2885	0,3899	0,9455	*RERG*	RERG-24	G/A	0,2917	0,0059	0,5268
GPX3	GPX3-28	A/G	0,1465	0,0123	0,4519		RERG-37	A/C	0,1266	0,5536	0,0268
GSK3B	rs16830689	G/C	0,2057	0,0233	0,6250		RERG-47	T/C	0,2911	0,0060	0,5636
	rs1719889	T/A	0,2184	0,0234	0,6339	*RGS5*	RGS5-01	C/A	0,0854	0,0058	0,4821
	rs1732170	A/G	0,2848	0,3029	0,9375	*RNASEL*	RNASEL-02	A/G	0,3846	0,0175	0,1339
	rs334535	A/G	0,2057	0,0231	0,6250	*SCARB1*	SCARB1-03	A/G	0,3471	0,8266	0,3482
	rs334559	T/C	0,2025	0,0291	0,6273	*SEPP1*	SEPP1-01	A/G	0,2675	0,6235	0,0804
	rs4072520	T/G	0,2184	0,0231	0,6091	*SLAMF1*	SLAMF1-03	G/A	0,4684	0,0769	0,7946
	rs6770314	T/C	0,2057	0,0203	0,6339	*SLC23A1*	SLC23A1-09	C/T	0,3362	0,4889	0,9898
	rs7617372	C/T	0,2184	0,0291	0,6339	*SLC4A2*	SLC4A2-02	G/A	0,2219	0,8727	0,4107
	rs7620750	T/C	0,2197	0,0233	0,6339	*SLC6A3*	SLC6A3-10	G/A	0,2019	0,0152	0,4727
	rs9851174	T/C	0,2057	0,0234	0,6250	*SOAT2*	SOAT2-01	G/A	0,1847	0,1765	0,7273
	rs9878473	G/A	0,4304	0,3588	1	*SOD1*	SOD1-01	G/A	0,0665	0,4792	0,2091
GSTM3	GSTM3-06	T/G	0,4051	0,8482	0,1545	*SOD3*	SOD3-05	T/C	0,2792	0,5000	0,9643
HSD17B2	HSD17B2-01	G/A	0,0285	0,6441	0,1339	*TCTA*	TCTA-04	G/A	0,4262	0,9765	0,1852
HSD3B1	HSD3B1-18	G/T	0,3608	0,1111	0,7857	*TERT*	TERT-02	T/C	0,3390	0,9639	0,6400
	HSD3B1-22	G/A	0,3726	0,0116	0,0536	*TLR2*	TLR2-04	C/T	0,4588	0,1104	0,5999
	HSD3B1-24	G/T	0,3703	0,1140	0,7857		TLR2-06	T/A	0,4708	0,9649	0,6091
	HSD3B1-25	A/G	0,3408	0,0088	0,0714	*TP73L*	TP73L-13	A/G	0,2184	0,0173	0,5179
	HSD3B1-26	A/G	0,3854	0,0145	0,2411		TP73L-15	G/A	0,2532	0,0145	0,7411
HSD3B2	HSD3B2-14	T/G	0,1571	0,1882	0,7182		TP73L-16	A/G	0,4873	0,8462	0,2411
IFNAR2	IFNAR2-01	A/G	0,3248	0,7292	0,1071		TP73L-26	T/C	0,1677	0,8095	0,2143
	IFNAR2-06	G/T	0,3653	0,7336	0,1518		TP73L-28	T/C	0,4013	0,8364	0,2500
	IFNAR2-10	A/G	0,3280	0,7297	0,1875	*VCAM1*	VCAM1-05	G/A	0,0285	0,0058	0,3125
IGF1R	IGF1R-05	A/G	0,3333	0,1813	0,8036	*WDR79*	WDR79-06	T/A	0,0316	0,0029	0,3727
	IGF1R-18	A/G	0,3344	0,1837	0,8036		WDR79-11	G/C	0,1139	0,1294	0,9272
IGF2	IGF2-16	G/C	0,3726	0,0340	0,6909	*XRCC4*	XRCC4-01	A/G	0,1044	0,5500	0,5546
IGFBP5	IGFBP5-10	C/T	0,3070	0,7122	0,8036		XRCC4-05	T/G	0,4522	0,0434	0,6182

IGFBP6	IGFBP6-19	T/C	0,1154	0,0799	0,6786		XRCC4-10	C/T	0,0981	0,5497	0,5625
IL13	IL13-01	A/G	0,1419	0,7852	0,2641	*XRCC5*	XRCC5-12	A/G	0,3942	0,8639	0,1964
	IL13-06	T/C	0,1474	0,7929	0,8214		XRCC5-19	G/A	0,4487	0,8675	0,3019

Genótipo: As frequências alélicas foram escolhidas de acordo com o alelo de menor frequência na população europeia. **EUR**: Frequência alélica na população europeia. **NAT**: Frequência alélica na população Nativo-Americana. **WAFR**: Frequência alélica para a população da África Ocidental. Os SNPs marcados em azul estão localizados em éxons. As células marcadas em cinza-escuro indicam que aquela população apresenta frequências alélicas mais diferenciadas em relação às demais populações.
Fonte: (Soares-Souza, 2010)

As populações Yakut, Daur, Oroqen e Hezhen, e os grupos Nordeste Asiático – constituído pelas três últimas populações citadas – e o grupo Leste Asiático foram testados quanto à estruturação em relação às populações Nativo-Americanas. Os polimorfismos e genes com alta estruturação são apresentados nas tabelas 3 e 4. O exame de diferenciação entre as populações do Leste Asiático, mais próximas do Estreito de Bering, e as Nativo-americanas demonstrou a existência de 36 polimorfismos pertencentes a 26 genes com alta estruturação. Os Yakut constituem a população mais próxima geográfica e geneticamente das populações Nativo-Americanas, embora essa população tenha se deslocado para atual região de assentamento apenas recentemente (VITEBSKY, 1990), eventos de miscigenação e assimilação de outros povos durante o processo de migração podem ser a origem do padrão observado.

Tabela 3: SNPs entre os maiores valores de F_{CT} para populações do Leste Asiático e Nativo-Americanos

EAS	NEAS	YAK	DAUR	HEZ	ORO
IGF2-16	IGF2-16	CASR-05	GATA3-25	STAT1-01	STAT1-01
TSPO-03	IGF2-02	IGF2-16	IGF2-02	IGF2-02	IGF2-02
IGF2-02	STAT1-01	rs1533593	STAT1-01	CCDC97-03	CD40-01
BRIP1-09	GATA3-25	TLR2-06	CD40-01	CASR-05	CD40-03
CDKN1B-04	BRIP1-09	GATA3-25	GATA3-28	PARP4-19	BRIP1-02
ADH1C-16	HSD17B2-01	MX1-22	CASR-05	ADH1C-16	CASR-05
OPRM1-02	OPRM1-23	CD86-02	CDKN2A-19	CYP19A1-39	CD86-02
CDKN2A-03	TP73L-26	RNASEL-02	CD40-03	ERCC4-15	BRIP1-09
CCND1-03	CCND1-03	RERG-44	BRIP1-09	CD86-02	TP73L-13
CD86-02	PAK6-13	STAT1-01	CDKN2A-20	IL1B-03	GATA3-28

Nomenclatura das populações: EAS (Leste Asiático), NEAS (Nordeste Asiático), YAK (Yakut), DAUR (Daur), HEZ (Hezhen), ORO (Oroqen). SNPs presentes em mais de um conjunto populacional ou subpopulação têm as celas coloridas. Cada cor representa um SNP diferente.
Fonte: (SOARES-SOUZA, 2010)

Tabela 4: Genes entre os maiores valores de F_{CT} para populações do Leste Asiático e Nativo-Americanos

EAS	NEAS	YAK	DAUR	HEZ	ORO
IGF2	IGF2	CASR	GATA3	STAT1	STAT1
TSPO	IGF2	IGF2	IGF2	IGF2	IGF2
IGF2	STAT1	MASP1	STAT1	CCDC97	CD40
BRIP1	GATA3	TLR2	CD40	CASR	CD40
CDKN1B	BRIP1	GATA3	GATA3	PARP4	BRIP1
ADH1C	HSD17B2	MX1	CASR	ADH1C	CASR
OPRM1	OPRM1	CD86	CDKN2A	CYP19A1	CD86
CDKN2A	TP73L	RNASEL	CD40	ERCC4	BRIP1
CCND1	CCND1	RERG	BRIP1	CD86	TP73L
CD86	PAK6	STAT1	CDKN2A	IL1B	GATA3

Nomenclatura das populações: EAS (Leste Asiático), NEAS (Nordeste Asiático), YAK (Yakut), DAUR (Daur), HEZ (Hezhen), ORO (Oroqen). Os nomes dos SNPs com maior divergência foram omitidos em favor do nome dos genes. Genes presentes em mais de um conjunto populacional ou subpopulação têm as celas coloridas. Cada cor representa um gene distinto.
Fonte: (SOARES-SOUZA, 2010)

Nas seções a seguir são relatadas as diferenças metodológicas do presente estudo em relação ao prévio, de 2010. Três conjuntos de dados contendo aproximadamente 650 mil SNPs para as populações do HGDP foram adicionados e, neste estudo, os resultados de divergência são anotados e sumarizados a partir da plataforma MASSA (ver capítulo 4).

3.1 METODOLOGIA

3.1.1 AMOSTRAGEM

A partir da colaboração do LDGH com o laboratório do Dr. Stephen Chanock (*NIH-CGR* – National Institutes of Health-Cancer Genomics Research Lab) nos foram disponibilizados três conjuntos de dados genotipados a partir da plataforma Illumina GoldenGate *oligonucleotide pool assay* (OPA). No presente trabalho estes conjuntos de dados serão daqui em diante assim designados: OPA-Native (OPA-N), OPA-Innate Immunity (OPA-II) e OPA-Non Hodgkin Linfoma (OPA-NHL). Os três conjuntos de dados OPA foram genotipados para as 52 populações do CEPH-HGDP (Centre d'Etude du Polymorphisme Humain – Human Genome Diversity Cell Line Panel) (CANN, 1998; CAVALLI-SFORZA,

2005) e o conjunto de dados OPA-Native incluí 4 populações nativo-americanas do Peru e Equador (Quechua, San Martin, Cayapa e Matsiguenga – dados não publicados, à exceção dos Quechua em (SCLIAR et al., 2012). Integramos às análises os dados disponibilizados por (LI et al., 2008) perfazendo um total de quatro conjuntos de dados descritos na Tabela 5.

Tabela 5: Descrição dos conjuntos de dados utilizados na descrição evolutiva dos polimorfismos em populações nativo-americanas

Descrição	OPA-N	OPA-II	OPA-NHL	Li et al., 2008
Método de genotipagem	Illumina GoldenGate	Illumina GoldenGate	Illumina GoldenGate	Illumina HumanHap650K
Tamanho amostral	1.112	980	1.014	938
Quantidade de SNPs	1.442	1.433	1.457	660.918
Controle de qualidade pós-genotipagem	HWE, Fita	HWE	HWE	HWE

3.1.2 PREPARAÇÃO DAS BASES DE DADOS

Avaliamos a concordância entre os genótipos dos conjuntos OPA-II, OPA-NHL e Li através do módulo de controle de qualidade do software GLU (Genotyping *and Library Utilities*) versão 1.0b3 e não foram encontradas inconsistências significativas entre as bases de dados.

Visando simplificar e aumentar a robustez das análises os conjuntos de dados OPA-II, OPA-NHL e Li foram unificados em único conjunto denominado II-NHL-Li. OPA-N não pode ser incorporado a este conjunto por conter populações ausentes nas demais coleções. Foram excluídos 16657 loci localizados nos cromossomos X, Y e mitocondrial devido às diferenças na dinâmica de segregação e nas taxas evolutivas (HAMMER et al., 2008; WILDER; MOBASHER; HAMMER, 2004).

O conjunto OPA-N foi tratado individualmente e, deste, foram excluídos 195 polimorfismos por inconsistências de fita, duplicação e localização no cromossomo X, maiores detalhes das etapas de controle de qualidade são descritas em (SOARES-SOUZA,

2010). A descrição dos conjuntos populacionais pós-controle de qualidade e união de OPA-II, OPA-NHL e Li pode ser vista na Tabela 6.

Tabela 6: Descrição dos conjuntos populacionais após o controle de qualidade

	OPA-N	II-NHL-Li
SNPs	1247	647151
Genes	392	17979
Indivíduos	1021	966
Indivíduos NAT	132	66
Indivíduos EAS	237	234
Indivíduos EUR	156	156
Indivíduos WAFR	64	64

3.1.3 DEFINIÇÃO DAS POPULAÇÕES E DOS GRUPOS POPULACIONAIS

Os critérios de definição das populações e grupos populacionais podem ser vistos em detalhe em (SOARES-SOUZA, 2010). Em síntese, os indivíduos Bantu do nordeste, sudoeste e sul da África foram agrupados numa única população Bantu (ROSENBERG et al., 2002, 2005), os indivíduos Han do norte e sul da China foram agrupados em uma única população Han (BASTOS-RODRIGUES; PIMENTA; PENA, 2006). As 56 populações foram alocadas em 9 grupos populacionais: América do Sul (AMS), América Central (AMC), Europeus (EUR), Oriente Médio (MDL), África Ocidental (WAFR), África Oriental (EAFR), Centro-Sul da Ásia (CSA), Leste Asiático (EAS) e Oceania (OCE). As populações da América do Sul e Central foram agrupadas para as análises envolvendo polimorfismos em único grupo denominado Nativo-Americanos (NAT). É importante ressaltar que os grupos populacionais foram criados levando em conta não apenas os critérios geográficos, mas também a proximidade genética das populações. Desta forma, a população Mozabite foi alocada no grupo Oriente Médio e a população Bantu no grupo África Ocidental, sendo esta uma alteração em relação às análises realizadas em (SOARES-SOUZA, 2010). Utilizou-se o conjunto de dados H971 (ROSENBERG, 2006) como filtro para a exclusão de indivíduos aparentados ou erroneamente identificados. O número de indivíduos para os grupos populacionais NAT, EAS, EUR, WAFR é discriminado na Tabela 6.

3.1.4 ANÁLISES ESTATÍSTICAS

Os fundamentos das análises estatísticas efetuadas nesta seção foram previamente descritos na seção 1.3. Por isso, os métodos e parâmetros estatísticos serão informados brevemente neste capítulo. A Figura 7 descreve em detalhe os passos utilizados na identificação dos polimorfismos diferenciados em Nativo-americanos e é um recorte do fluxograma apresentado na seção 1.3 (Figura 5). Os procedimentos referentes à montagem dos conjuntos de dados e controle de qualidade foram relatados nas seções 3.1.1, 3.1.2 e 3.1.3 e resultam nos arquivos SDAT OPA-N e II-NHL-Li descritos na Figura 7. A partir destes, utilizou-se o script tHIERs (constituinte da 3ª versão do DivergenomeTools) para criar os arquivos de entrada para o pacote Hierfstat da plataforma R. O cálculo de AMOVA foi efetuado a partir do script Hierfastat que encapsula a função *varcomp* do pacote e a executa em paralelo, neste estudo, 4 núcleos foram utilizados em cada análise (NAT-EAS, NAT-EUR e NAT-WAFR), sendo estas executadas concomitantemente através do comando xargs do Unix. O script *R table frequencies* une as estimativas de diversidade resultantes do cálculo de AMOVA às frequências alélicas em uma tabela. A partir de cada uma destas tabelas foram selecionados os respectivos polimorfismos com valores de F_{CT} superiores ao 95º percentil e de F_{SC} inferiores à 0,12. Os SNPs diferenciados em NAT para os dois conjuntos, OPA-N e II-NHL-Li, foram unidos em uma única lista não redundante. Esta lista foi submetida ao WS (WebService) SNAP para selecionar polimorfismos em Desequilíbrio de Ligação com os SNPs previamente identificados nas populações CHB/JPT de 1000Genomes. As duas listas foram anotadas utilizando a ferramenta MASSA e análises de enriquecimento foram realizadas a partir dos arquivos de anotação.

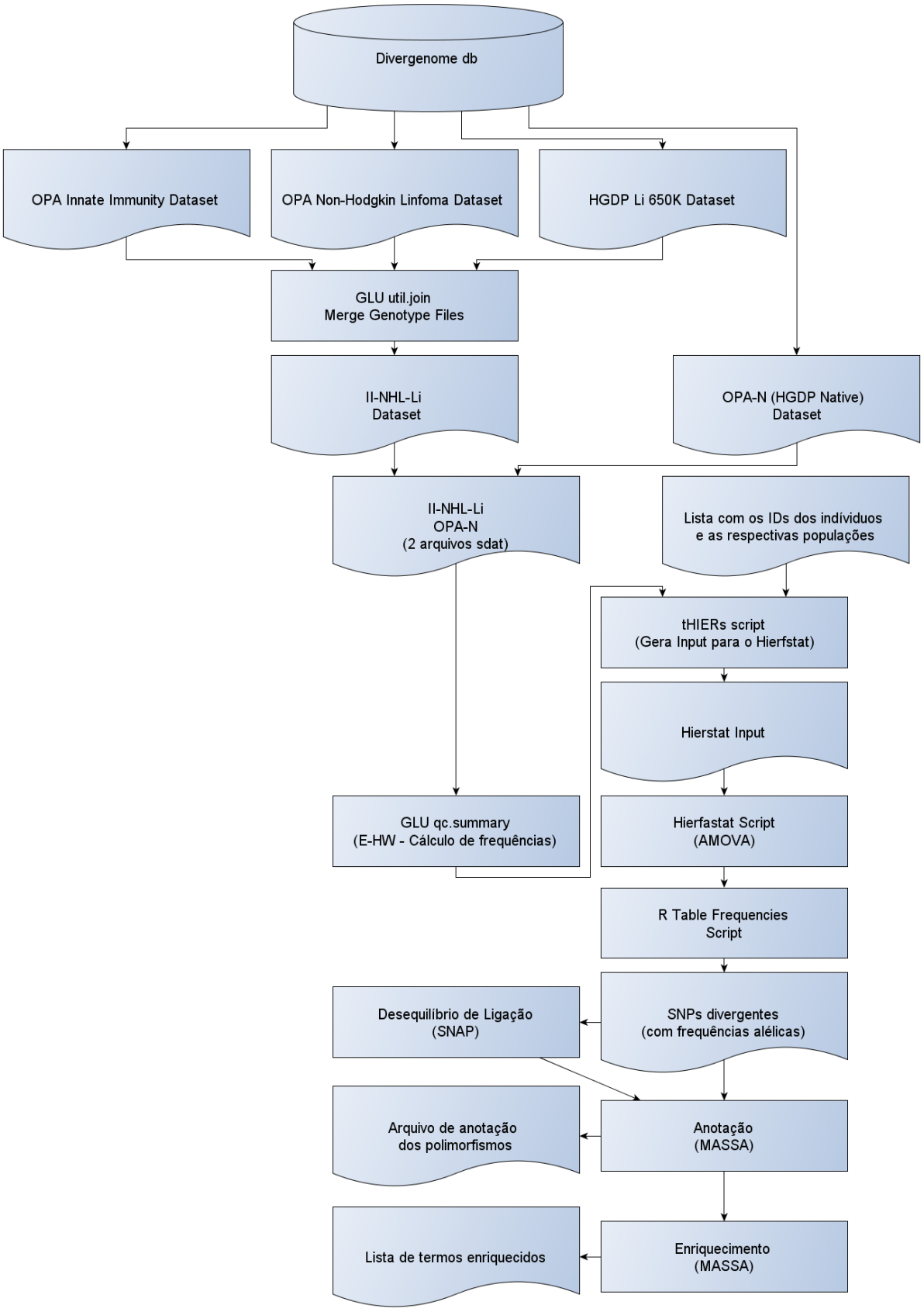


Figura 7: Fluxograma de Análises para seleção de SNPs diferenciados em populações Nativo-Americanas.

3.2 RESULTADOS

Como explicitado anteriormente, para a seleção de SNPs diferenciados em cada teste de diferenciação populacional utilizamos os seguintes critérios: a) Valores de F_{CT} superiores ao nonagésimo quinto percentil; b) Valores de F_{SC} inferiores a 0,12. Essas medidas têm por objetivo selecionar os SNPs cujas frequências alélicas sejam divergentes entre os grupos populacionais, mas convergentes entre as populações dentro de um grupo. Há de se ressaltar que a regra (a) difere da anteriormente proposta em (SOARES-SOUZA, 2010) por individualizar os valores críticos de acordo com as populações envolvidas. Isto é importante devido ao fato de que durante o processo de povoamento das demais regiões a partir da diáspora africana, ocorreu, sistematicamente, o efeito da deriva genética, ou seja, flutuações aleatórias nas frequências alélicas. Esse efeito, também conhecido como *allele surfing*, leva à diferenciação cada vez maior entre as populações no sentido da dispersão dos indivíduos, assim, surge uma correlação entre a distância geográfica e a genética, levando a níveis de divergência diferentes entre as populações humanas. Através da nova metodologia proposta é possível minimizar os efeitos da deriva genética, pois são selecionados os SNPs mais divergentes, independentemente da distância genética média entre os grupos analisados. As distribuições dos valores de F_{CT} para cada análise estão representadas na Figura 8 e os valores críticos de F_{CT} obtidos para cada configuração de AMOVA e diferentes percentis (90, 95 e 99) foram comparados na Figura 7.

Os maiores valores de diferenciação entre grupos populacionais (regra a), demonstrados pela maior massa da distribuição concentrada à direita na Figura 8 e maiores valores críticos na Figura 7, são vistos, respectivamente, nas seguintes configurações: WAFR-NAT, EUR-NAT e EAS-NAT. Tais dados corroboram a origem africana do *Homo sapiens* e estudos prévios sobre a variabilidade genética humana e, em especial, das populações nativo-americanas (BASTOS-RODRIGUES; PIMENTA; PENA, 2006; LI et al., 2008; ROSENBERG et al., 2002, 2005; SCLIAR et al., 2012; WANG et al., 2007).

A maior diferença nos valores críticos entre os conjuntos de dados ocorre entre os grupos Leste da Ásia e Nativo-americanos. Isso ocorre devido à amostragem diferencial do grupo Nativo-americanos no conjunto de dados NAT. Por ter mais populações ameríndias, este grupo representa melhor o *pool* gênico das populações americanas, reflete a proximidade

genética entre nativos e asiáticos e minimiza os efeitos da miscigenação europeia, significativa nas populações Pima e Maia.

É importante notar que não houve grandes alterações na distribuição de F_{CT} com a fusão dos conjuntos de dados II-NHL ao Li e, além disso, as distribuições dos valores de F_{CT} são similares nos conjuntos de dados OPA-N e II-NHL-Li. Tudo isto afasta a possibilidade de enviesamento dos dados devido aos critérios de seleção dos SNPs nos conjuntos OPA-N, OPA-II e OPA-NHL, onde foram incluídos, especialmente, àqueles relacionados à imunidade e, por isto, mais suscetíveis à ação da seleção natural (DAUB et al., 2013; FERRER-ADMETLLA et al., 2008; FUMAGALLI et al., 2009; NIELSEN et al., 2009; NOVEMBRE; HAN, 2012).

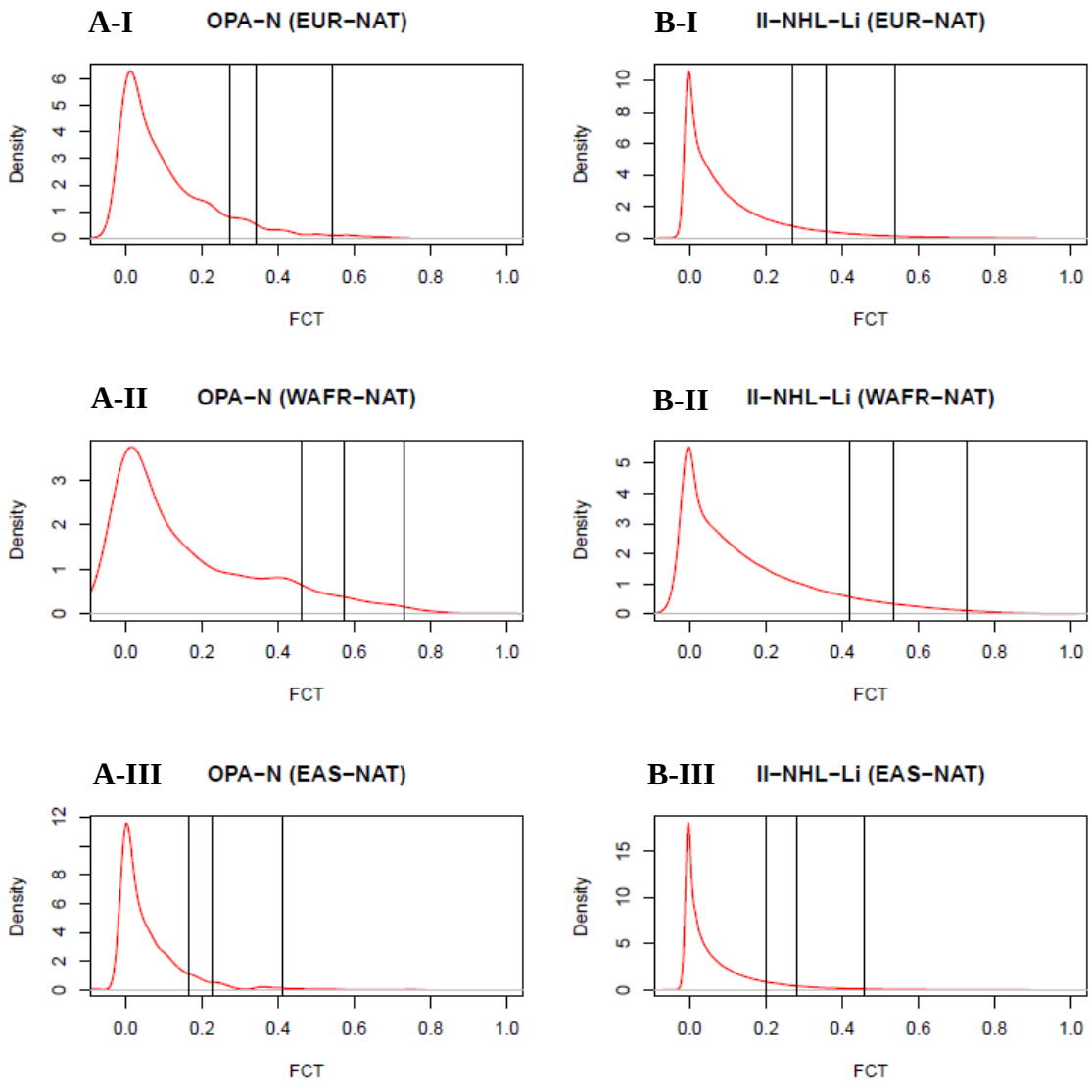


Figura 8: Distribuição dos valores de F_{CT} por conjunto de dados e configuração populacional. Na figura estão representadas as distribuições dos valores de diferenciação

(F_{CT}) para cada conjunto de dados: A) OPA-N, B) II-NHL-Li; e configuração populacional: I) EUR-NAT, II) WAFR-NAT, III) EAS-NAT. Cada reta vertical, a partir da origem, representa, respectivamente, o 90º, 95º e 99º percentis.

Tabela 7: Valores Críticos de F_{CT}

Dataset	Valores críticos de F_{CT} por configuração populacional e percentil								
	0.90			0.95			0.99		
	EUR-NAT	EAS-NAT	WAFR-NAT	EUR-NAT	EAS-NAT	WAFR-NAT	EUR-NAT	EAS-NAT	WAFR-NAT
II-NHL	0.274	0.235	0.465	0.367	0.301	0.587	0.555	0.503	0.770
II-NHL-Li	0.268	0.200	0.418	0.359	0.279	0.553	0.538	0.459	0.725
Native	0.273	0.164	0.461	0.343	0.227	0.571	0.542	0.410	0.730

O critério de seleção (b) visa manter apenas os polimorfismos com baixa estruturação dentro dos grupos. Foram estabelecidos valores de corte para o F_{SC} a partir dos percentis (5, 10 e 25) como demonstrado na Tabela 8 e Figura 9. Entretanto, o valor de 0,12 mostrou-se não apenas suficientemente restritivo, mas também capaz de refletir melhor a atuação da deriva gênica dentro dos grupos, uma vez que este valor corresponde à média de diferenciação estimada para o genoma humano (BARBUJANI et al., 1997; LI et al., 2008; NEI; ROYCHOUDHURY, 1974; ROSENBERG et al., 2005).

Tabela 8: Valores Críticos de F_{SC}

Dataset	Valores críticos de F_{SC} por configuração populacional e percentil								
	0.05			0.10			0.25		
	EUR-NAT	EAS-NAT	WAFR-NAT	EUR-NAT	EAS-NAT	WAFR-NAT	EUR-NAT	EAS-NAT	WAFR-NAT
II-NHL	0.000	0.000	0.000	0.000	0.004	0.000	0.007	0.013	0.005
II-NHL-Li	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.009	0.000
Native	0.000	0.000	0.000	0.000	0.003	0.000	0.010	0.016	0.011

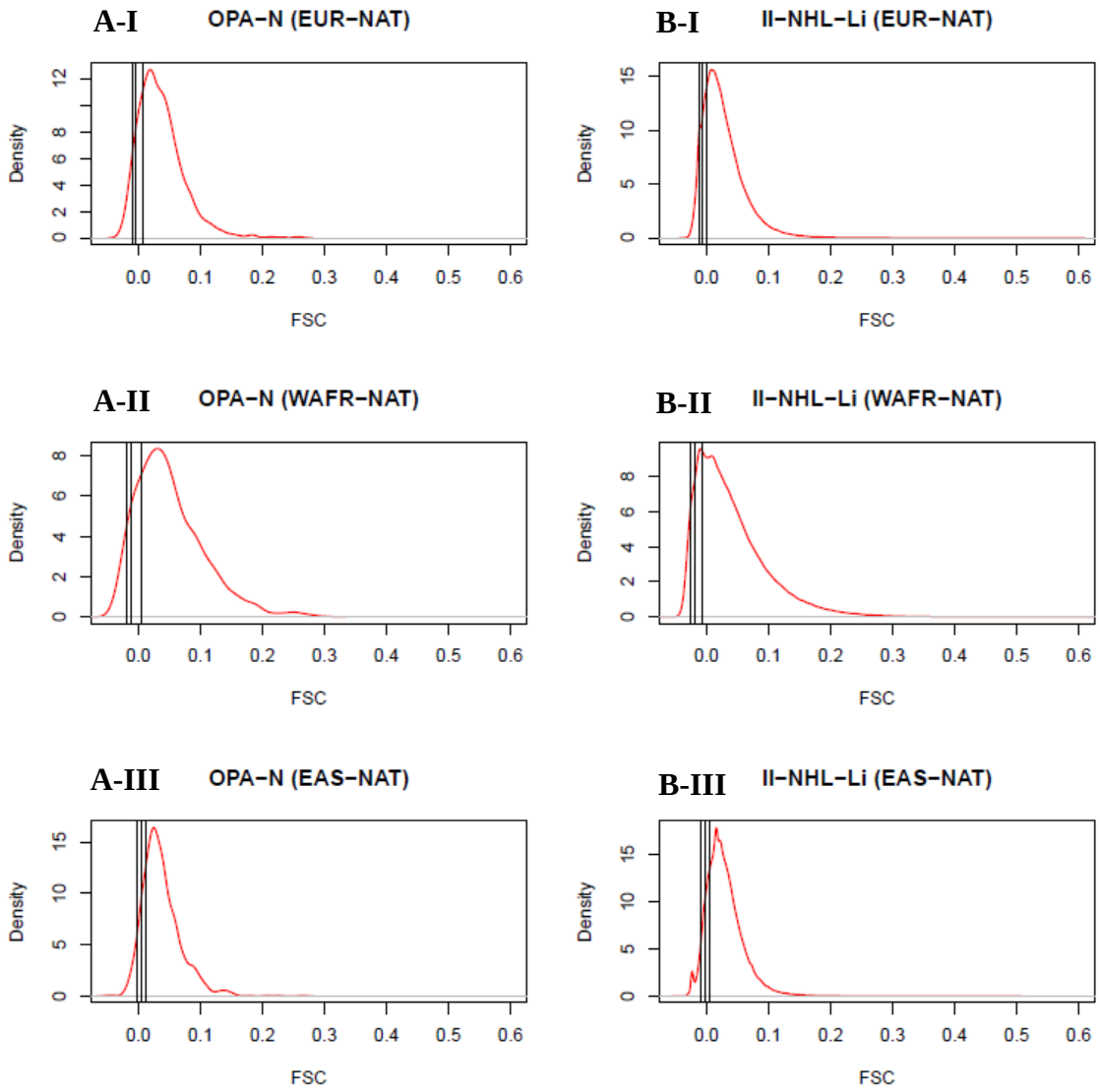


Figura 9: Distribuição dos valores de F_{SC} por conjunto de dados e configuração populacional. Na figura estão representadas as distribuições dos valores de diferenciação (F_{SC}) para cada conjunto de dados: A) OPA-N, B) II-NHL-Li; e configuração populacional: I) EUR-NAT, II) WAFR-NAT, III) EAS-NAT. Cada reta vertical, a partir da origem, representa, respectivamente, o 5º, 10º e 25º percentis.

O número de SNPs divergentes para cada análise e conjunto de dados encontra-se discriminado na Tabela 9. Em média, foram selecionados, entre 25 e 29 mil polimorfismos no conjunto de dados II-NHL-Li e 59 a 61 no conjunto OPA-N. Ainda na Tabela 9 estão explicitados os valores encontrados na união e na interseção desses resultados. No total, foram descritos, respectivamente, 28.819, 28.657 e 25.404 SNPs diferenciados nas análises EUR-NAT, EAS-NAT e WAFR-NAT, totalizando 69.891. A interseção entre os resultados do conjunto de dados II-NHL-Li indica a presença de 1569 polimorfismos divergentes comuns às três análises e para o conjunto de dados OPA-N, não há polimorfismo presente nas três.

Tabela 9: SNPs divergentes por configuração populacional e conjunto de dados

Dataset	Configuração do AMOVA						
	EUR-NAT	EAS-NAT	WAFR-NAT	(EUR ∪ WAFR)-NAT	(EUR ∩ WAFR)-NAT	(EUR ∪ WAFR ∪ EAS)-NAT	(EUR ∩ WAFR ∩ EAS)-NAT
II-NHL-Li	28819	28657	25404	48860[1]	5363[2]	69765[1]	1569[2]
Native	60	59	61	117[1]	4[2]	160[1]	0[2]
II-NHL-Li-NAT	28863[1]	28703[1]	25454[1]	48951[1]	5366[1]/1[2]	69891[1]	1569[1]/0[2]

[1] União entre os conjuntos de resultados excluindo-se as entradas duplicadas. [2] Interseção entre os conjuntos de resultados.

Através do mapeamento SNP-Gene, ou seja, omissão dos identificadores dos polimorfismos em favor do símbolo do gene onde estes se encontram, o número de genes diferenciados foi obtido e está discriminado na Tabela 10. Nota-se a gradiente de diferenciação: 4072 genes na comparação WAFR-NAT, 3930 na EUR-NAT e 3792 na EAS-NAT. No total, são 7503 genes diferenciados, sendo 1135 concordantes entre as três análises do conjunto II-NHL-Li e três entre as configurações do conjunto OPA-N. Destes, 3 estão presentes em todas as análises: *CASR*, *BRIP1* e *TP63*.

Tabela 10: Genes divergentes por configuração populacional e conjunto de dados

Dataset	Configuração do AMOVA						
	EUR-NAT	EAS-NAT	WAFR-NAT	(EUR ∪ WAFR)-NAT	(EUR ∩ WAFR)-NAT	(EUR ∪ WAFR ∪ EAS)-NAT	(EUR ∩ WAFR ∩ EAS)-NAT
II-NHL-Li	3930	3792	4072	6054[1]	1936[2]	7486[1]	1135[2]
Native	43	37	34	70[1]	7[2]	90[1]	3[2]
II-NHL-Li-Nat	3930[1]	3800[1]	4077[1]	6071[1]	1937[1]/6[2]	7503[1]	1135[1]/3[2]

[1] União entre os conjuntos de resultados excluindo-se as entradas duplicadas. [2] Interseção entre os conjuntos de resultados.

A classificação funcional dos polimorfismos divergentes está descrita na Tabela 11. Aproximadamente 54% dos SNPs estão localizados em regiões intergênicas e 42% em regiões intrônicas. Desta forma, 4% dos polimorfismos pertencem a alguma categoria funcional certamente envolvida na expressão gênica. As classes funcionais, de acordo com a posição, mais representadas são: "*cds-reference*"[1] com 947 polimorfismos, mutações na região 3' não traduzida (386 SNPs), mutações não sinônimas (381), polimorfismos próximos à montante do gene (202), SNPs em RNAs não codificantes (196), mutações sinônimas (194), mutações na região 5' não traduzida (62), polimorfismos próximos à jusante do gene (60) e mutações em sítios de *splicing* (3). A severidade das mutações em regiões codificantes foi inquirida a partir das bases de dados provenientes dos softwares PolyPhen2 (ADZHUBEI et al., 2010), SIFT (KUMAR; HENIKOFF; NG, 2009) e PROVEAN (CHOI et al., 2012). A maior parte

das mutações em regiões codificadoras (87-93%) é predita como benigna e, aproximadamente, 10% das variantes são inferidas como possivelmente ou provavelmente deletérias. É interessante notar que há quase o dobro de mutações não sinônimas (381) em relação ao número de sinônimas (194), indicando que, possivelmente, algumas destas áreas ou estão sob seleção positiva em Nativo-Americanos ou são devidos ao fenômeno de *expansion load*.

Quanto à interação com outros elementos genômicos, as classes funcionais mais representadas são: alvos de fatores de transcrição (1577 SNPs), ilhas de CpG (218), alvos de acentuadores (39) e alvos de microRNAs de interferência (7). No que tange às taxas de conservação, entre 4,8% e 8,5% dos sítios são preditos como conservados, de acordo com os softwares PHAST (SIEPEL et al., 2005) e GERP++ (DAVYDOV et al., 2010), respectivamente.

Tabela 11: Mapeamento dos SNPs divergentes em relação à categoria funcional predita

Classe Funcional	Configuração do AMOVA				
	EUR-NAT	**EAS-NAT**	**WAFR-NAT**	**(EUR ∪ WAFR)-NAT**	**(EUR ∪ WAFR ∪ EAS)-NAT**
Intergenic	15592	15432	13789	26448	37843
Intronic	12246	12269	10765	20788	29617
Near Gene 5'	97	81	67	152	202
Near Gene 3'	29	26	18	42	60
5'UTR	26	27	25	42	62
3'UTR	153	170	160	265	386
Splice	1	1	1	2	3
CDS-reference[1]	392	376	351	669	947
Synonymous	84	92	60	128	194
Non-Synonymous	145	151	152	272	381
Missense	147	151	150	269	378
Nonsense	1	0	2	2	2
Frameshift	1	0	0	1	1
Benign[3/4/5]	184/459/437	212/453/426	238/437/417	374/801/763	538/1166/1115
PossiblyDamaging[3]	18	14	12	28	41
ProbablyDamaging[3/4/5]	23/53/86	15/51/84	9/38/74	29/82/143	40/133/205
ncRNA	94	78	66	143	196
miRNA Target[2]	2	2	3	5	7
Enhancer[2]	15	16	13	23	39
CpG Island[2]	88	90	93	159	218
Transcript Factor[2]	652	601	638	1154	1577
Conservation PHAST[2]	1361	1327	1354	2425	3334
Conservation GERP[2]	2376	2304	2417	4271	5897

[1] CDS-reference é descrito no dbSNP como o alelo presente em um contig, podendo sua classificação ser sinônimo ou não-sinônimo de acordo com o contig utilizado como referência. [2] Anotações obtidas a partir do web-server SNP Nexus (DAYEM ULLAH; LEMOINE; CHELALA, 2012) ({HYPERLINK "**http://snp-nexus.org/**"}). Predições de severidade de mutação estimadas pelos seguintes programas: [3]PolyPhen2. [4]Provean. [5]SIFT.

A anotação dos polimorfismos divergentes indica que os mesmos estão distribuídos por 7503 genes, excetuando-se as variantes intergênicas, pertencentes a 669 famílias gênicas, sendo a classificação funcional dos genes descrita na Tabela 12. A maior parte dos genes (85%) está implicada na síntese proteica e apenas 3,8% estão envolvidos em outras funções diversas como: produção de RNAs não-codificantes longos e nucleolares, transcritos multigênicos, segmentos gênicos constitutivos de protocaderinas, loci complexos e pseudogenes.

Tabela 12: Classificação funcional dos genes apresentando SNPs diferenciados em Nativo-Americanos

Classe Funcional	SNPs[1]	Genes
Gene with protein product	28765	6420
RNA, long non-coding	365	104
Pseudogene	123	64
Readthrough	112	27
Unknown	51	15
Protocadherin	16	4
RNA, small nucleolar	2	2
Complex locus constituent	2	1
No info	40455	868

[1]Apenas SNPs localizados em regiões gênicas.

Os genes com maior número de variantes estruturadas são: *CSMD1* (142 SNPs) – *CUB and Sushi multiple domains 1* – possivelmente relacionado ao desenvolvimento e agressividade de neoplasias, regulação do sistema complemento, e à orientação e prolongamento das terminações dos axônios; *CNTNAP2* (118), – *contactin associated protein-like 2* – associado à formação de domínios funcionais críticos para a transmissão do impulso nervoso e à clusterização de canais de potássio, atua ainda como supressor de tumor em gliomas; *PTPRD* (117) – *protein tyrosine phosphatase, receptor type, D* – contribui com o desenvolvimento neuronal, regula a orientação dos axônios e o desenvolvimento neuroendócrino; *RBFOX1* (105) – *RNA binding protein, fox-1 homolog (C. elegans) 1* – atua na distribuição e *splicing* do RNA no complexo de Golgi e atua na mediação da via neuronal da calcitonina através do processamento alternativo do RNA; e *FHIT* (98) – *fragile histidine triad gene* – opera no balanceamento dos sinais de apoptose, proliferação e sobrevivência celular, exerce a função de supressão de tumor e regulação do ciclo celular ao interagir com

microtúbulos e tubulinas induzindo a apoptose. Os loci gênicos apresentando as maiores contagens de polimorfismos divergentes estão discriminados na Tabela 13.

Tabela 13: Genes com maior número de loci divergentes em Nativo-Americanos por configuração populacional

EUR-NAT		EAS-NAT		WAFR-NAT		(EUR ∪ WAFR)-NAT		(EUR ∪ WAFR ∪ EAS)-NAT	
Gene	N	Gene	N	Gene	N	Gene	N	Gene	N
PTPRD	83	NRG1	71	CSMD1	68	PTPRD	109	CSMD1	142
CNTNAP2	57	PDE4D	63	RBFOX1	56	CSMD1	100	CNTNAP2	118
FHIT	45	CNTNAP2	60	NRXN3	44	RBFOX1	90	PTPRD	117
ITPR2	45	CSMD1	57	CNTNAP2	42	CNTNAP2	84	RBFOX1	105
NRG3	44	PARK2	55	PDE11A	38	FHIT	70	FHIT	98
RBFOX1	43	PRKG1	50	DLG2	35	DLG2	63	PARK2	82
KANK1	42	NRXN3	40	ATRNL1	35	ODZ4	56	NRG1	80
NRXN1	41	CTNND2	39	PTPRD	33	WWOX	53	NRXN3	76
CSDM1	39	NPAS3	39	PARD3B	30	OPCML	51	DAB1	74

De acordo com a base de dados *NHGRI GWAS Catalog* há 717 polimorfismos associados (*GWAS hits*) a 321 fenótipos presentes entre os SNPs selecionados. Dentre os fenótipos com maior número de SNPs associados estão: subfenótipos relacionados à obesidade (28 SNPs), altura (28), doença de Crohn (20), glicolisação de IgG (18), performance cognitiva (17), câncer de próstata (14), lúpus eritematoso sistêmico (15), miopia (15), esclerose múltipla (13), doença celíaca (10), diabetes tipo 2 (9), transtorno depressivo maior (9), cárie dental (9); inteligência (8), doença de Alzheimer (8), triglicérides (8), índice de massa corporal (8), dentre outros. Alguns destes fenótipos apresentam alta prevalência em populações Nativo-americanas: obesidade e alto índice de massa corporal (HEARST et al., 2013; TRAURIG et al., 2012), glicolisação de IgG e desfechos autoimunes (NEWKIRK et al., 1998; TAI; NEWKIRK, 2000); fenótipos cognitivos (FERGENBAUM et al., 2009; JERVIS; MANSON, 2007; JERVIS et al., 2010; MITCHELL et al., 2011); lúpus eritematoso sistêmico (BARNABE et al., 2012; SANCHEZ et al., 2010); diabetes tipo 2 (BENYSHEK; MARTIN; JOHNSTON, 2001; BOGARDUS; TATARANNI, 2002; CAMPBELL et al., 2012; CARBONETTO; STEPHENS, 2013). Entretanto, os Nativo-Americanos possuem menor incidência de câncer de próstata (HENDERSON et al., 2008; MAHONEY et al., 2009) e esclerose múltipla (SVENSON et al., 2007; WARREN et al., 2007).

A incidência de diabetes tipo 2 tem crescido nas populações ameríndias em ritmo intenso nas últimas décadas e tem alcançando o status de epidemia em várias populações ao redor do mundo (BENYSHEK; MARTIN; JOHNSTON, 2001; BOGARDUS; TATARANNI,

2002; HANSON et al., 2007). Uma das particularidades da susceptibilidade genética à esta doença é a alta diversidade e estruturação encontrada nos polimorfismos que conferem risco. No presente trabalho, nove polimorfismos em diferentes genes com altos valores de F_{CT} estão associados à susceptibilidade à diabetes não dependente de insulina. Recentemente, a heterogeneidade das frequências dos alelos de risco em diversas populações foi caracterizada, sendo 12 alelos em oito genes e uma região intergênica reconhecidos como fatores de risco para o desenvolvimento da doença. A diferenciação genética encontrada nestes alelos é a maior entre as 1495 doenças analisadas (CHEN et al., 2012). Apenas um gene – *IGF2BP2* – é coincidente no presente trabalho e no referido estudo, reforçando o caráter heterogêneo da susceptibilidade à diabetes. Também é interessante notar que as frequências alélicas que conferem risco são, em geral, baixas na população do Leste Asiático, embora altas nas populações nativo-americanas como visto no presente trabalho e no gene *IGF2* (SOARES-SOUZA, 2010). Todas estas observações indicam uma complexa interação genômica e ambiental no desenvolvimento da diabetes tipo 2.

A anotação dos genes onde há presença de SNPs divergentes indica que os loci estão relacionados à: 6265 processos biológicos, 2306 funções moleculares, 910 componentes celulares; 1504 vias metabólicas segundo o Reactome e 1014 vias de acordo com PharmGKB; metabolização de 904 fármacos ou xenobióticos; além de, 224 fenótipos conforme o PharmGKB e 2096 de acordo com o OMIM.

Com o intuito de identificar regiões com alto desequilíbrio de ligação no genoma Nativo-americano e possíveis associações a fenótipos não identificadas a partir dos dados de 650 mil SNPs, utilizou-se os dados de DL obtidos a partir do WS SNAP, e a partir destes, foram identificados 1.398.714 polimorfismos, sendo 181.760 únicos, ligados àqueles divergentes. A menor distância encontrada entre dois SNPs em DL é 1 pb e a maior, 1.033.228 pb, sendo a distância média igual a 35.002 pb. Cada bloco de ligação é formado, em média, por 21 polimorfismos, sendo que os SNPs com maiores blocos de ligação são: rs10129827 (913 SNPs em DL – gene *GPHN*), rs6573710 (896 – *GPHN*), rs1902644 (842 – intergênico). Os 20 polimorfismos com maior número de variantes correlacionadas são descritos na Tabela 14.

Tabela 14: Descrição dos polimorfismos divergentes em Nativo-Amricanos com altos valores de DL

ID do Polimorfismo	DL[1]	Gene	Cromossomo	Posição
rs10129827	913	*GPHN*	14	67232573
rs6573710	896	*GPHN*	14	67160209
rs1902644[C]	842	*Intergênico*	14	66898156
rs17827790[C]	839	*GPHN*	14	67024964
rs7160476	799	*GPHN*	14	67332377
rs2319185	785	Intergênico	14	66895566
rs8014490[B]	785	Intergênico	14	66811161
rs6573670	748	Intergênico	14	66843997
rs7142068[A,B,C]	655	*GPHN*	14	67463012
rs950197	650	Intergênico	14	66736168
rs10483786[C]	644	*GPHN*	14	66733526
rs10148212	620	*GPHN*	14	67095613
rs7151086	605	*GPHN*	14	67501555
rs6573656	565	Intergênico	14	66711083
rs7145240	559	Intergênico	14	66708299
rs8022194	559	Intergênico	14	66707190
rs10146368	542	Intergênico	14	66693126
rs10512509	537	*CEP112*	17	63981626
rs2111413	537	*CEP112*	17	64016313
rs10139752	529	Intergênico	14	66689160

[1] Número de SNPs em DL. [A] Sítio conservado de ancoragem de fatores de transcrição. [B] Sítio conservado segundo PHAST. [C] Sítio Conservado segundo GERP++.

A anotação dos genes com maiores contagens de SNPs correlacionados é apresentada na tabela 15. O gene *GPHN* codifica uma proteína que ancora receptores de neurotransmissores inibitórios ao citoesqueleto pós-sináptico, e está relacionado ao metabolismo de molibdênio atuando como cofator e à hyperplexia (espectro de fenótipos cognitivos). O loci *CEP112* está relacionado ao sistema neuronal, atuando na inibição sináptica. O loci *FREM3* encontra-se associado à adesão e comunicação celular. *EPM2A* apresenta um grande leque de ações funcionais, tais como: desenvolvimento do sistema nervoso e metabolização do glicogênio e de proteínas através da desfosforilação.

Tabela 15: Anotação dos genes contendo SNP divergentes em Nativo-Americanos com maiores contagens de polimorfismos em DL

Gene	Doença	Fármaco	Processo Biológico	Função Molecular	Componente Celular	*CytoLoc*
GPNH		Phenylgermanium, picrolonic acid	establishment of synaptic specificity at neuromuscular junction, molybdopterin cofactor biosynthetic process	protein complex scaffold	inhibitory synapse, extrinsic to plasma membrane	14q23.3
CEP112	-	-	receptor localization to synapse	-	inhibitory synapse; plasma membrane; centrosome; cytoplasm	17q24.1
FREM3	-	-	cell adhesion; cell communication	metal ion binding	extracellular matrix; basement membrane; integral to membrane	4q31.21
EPM2A	Breast Neoplasms; Deafness; Hodgkin Disease; Neoplasms; therapy-related acute myeloid leukemia (t-ML); Epilepsy, progressive myoclonic 2A (Lafora)	s-adenosylmethionine; azathioprine; cisplatin; dacarbazine; mercaptopurine; procarbazine; thioguanine; purine analogues	protein dephosphorylation; peptidyl-tyrosine dephosphorylation; glycogen metabolic process; behavior; nervous system development	carbohydrate binding; protein tyrosine/serine/threonine phosphatase activity; protein tyrosine phosphatase activity; protein binding; protein serine/threonine phosphatase activity; starch binding	nucleus; cytosol; cytoplasm; polysome; plasma membran; endoplasmic reticulum	6q24.3

Testes de enriquecimento foram realizados, utilizando a função de enriquecimento da ferramenta MASSA, com o intuito de selecionar as informações mais relevantes destes conjuntos de anotações. Nas tabelas 16 e 17, a seguir, são apresentadas frações dos termos com os menores valores de significância para cada categoria.

Tabela 16: Enriquecimento de termos do banco de dados Gene Ontology para genes contendo SNPs divergentes em Nativo-Americanos

GO – Atributos para Enriquecimento	Termos Enriquecidos para polimorfismos divergentes em NAT-(EUR ∪ WAFR ∪ EAS)			
	Termo	**p-Value**	**SNP Count**	**Gene Count**
Processo Biológico	*Translation*	6.38E-26	103	55
	Proteolysis	1.83E-17	819	190
	regulation of transcription, DNA-dependent	4.79E-13	1591	415
	antigen processing and presentation	1.68E-11	6	3
	protein folding	1.82E-11	111	47
	Immune response	4.38E-11	324	116
	intracellular protein transport	1.25E-10	219	69
	complement activation	1.08E-09	33	10
	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1.14E-09	8	7
	nucleosome assembly	3.42E-08	27	14
	small molecule metabolic process	4.96E-08	1781	477
	regulation of immune response	1.15E-07	74	30
	axon guidance	4.96E-07	1375	177
	Glycolysis	4.98E-07	39	13
	protein polymerization	7.12E-07	8	5
	synaptic transmission	1.02E-06	1270	184
	negative regulation of transcription from RNA polymerase II promoter	9.88E-06	870	179
	complement activation, classical pathway	2.47E-05	53	17
	microtubule-based movement	2.58E-05	200	37
	antigen processing and presentation of peptide antigen via MHC class I	3.11E-05	56	23
	blood coagulation	3.67E-05	1239	200
	regulation of cyclin-dependent protein kinase activity	4.05E-05	10	6
	nerve growth factor receptor signaling pathway	4.92E-05	700	102
	DNA repair	5.59E-05	247	88
Função Molecular	*protein binding*	1.18E-32	9174	1913
	nucleic acid binding	2.96E-26	847	249
	Zinc ion binding	1.53E-20	2904	690
	Nucleotide binding	4.30E-17	819	164
	olfactory receptor activity	8.09E-11	447	34
	structural constituent of ribosome	1.28E-10	36	26
	ATP binding	1.42E-10	3120	633
	metal ion binding	2.68E-10	1345	460
	GTP binding	3.18E-10	417	117
	sequence-specific DNA binding transcription factor activity	5.61E-10	1435	297
Componente Celular	*Mitochondrion*	1.44E-14	1459	376
	Ribosome	7.38E-14	58	26
	Intracellular	5.49E-12	1886	429
	MHC class II protein complex	1.11E-11	6	3
	MHC class I protein complex	1.41E-11	8	3
	Integral to plasma membrane	3.11E-11	7566	1427
	Nucleus	2.24E-10	7048	1656
	Nucleolus	7.38E-10	2052	536
	Integral to membrane	2.55E-08	2467	415
	Mitochondrial respiratory chain complex I	5.83E-07	7	5

A base de dados e ontologias *Gene Ontology* classifica os genes em três classes: a) processo biológico - operações ou conjuntos de eventos moleculares com início e fim definidos, relacionados ao funcionamento das unidades biológicas; b) função molecular - atividades elementares do produto gênico a um nível molecular; e c) componente celular -

área de atuação dos produtos gênicos quanto à célula ou seu ambiente extracelular. Desta forma, as anotações entre as classes são relacionadas, descrevendo diferentes perspectivas da atuação gênica. Dentre os termos enriquecidos da categoria processo biológico estão funções relacionadas à regulação da expressão gênica, à imunidade, ao metabolismo intracelular de carboidratos e peptídeos, ao ciclo celular e à atividade neuronal (Tabela 16).

Quanto à função molecular encontram-se sobrerrepresentados termos associados à ancoragem de íons, ATP/GTP, nucleotídeos e fatores de transcrição, à constituição do ribossomo e atividade dos receptores olfatórios. Em relação aos componentes celulares enriquecidos encontram-se as estruturas celulares associadas ao núcleo celular, às proteínas de membrana, mitocôndrias e ribossomos.

Estes resultados são consistentes com recente investigação da atuação da seleção natural adaptativa em Nativo-americanos Totonacs e da Bolívia (WATKINS et al., 2012). Quanto à estruturação, dois SNPs, rs12439270 – *FOXB1* e rs470113 – *TNRC6B* e cinco genes, *SLC6A11*, *ADAMTS9*, *ACVR1B*, *TRPC4* e *STARD13* que apresentam divergências nas frequências alélicas e sinais de seleção positiva coincidem entre a investigação e o presente trabalho. Estes genes estão relacionados à: *FOXB1* – regulação da transcrição gênica, liga-se a fatores de transcrição e acentuadores; *TNRC6B* – silenciamento de genes, regulação da expressão gênica, resposta ao stress e sistema imune; *SLC6A11* – transporte através da membrana e atividade sináptica; *ADAMTS9* – proteólise, organização da matriz celular e desenvolvimento dos melanócitos; *ACVR1B* – transdução de sinal, desenvolvimento do sistema nervoso e do folículo piloso; *TRPC4* – orientação do axônio e transporte de cálcio; *STARD13* – transdução e metabolismo de GTP. Ainda que a preponderância da seleção natural seja questionada como principal fator na diferenciação populacional (HOFER et al., 2009), a identificação de regiões diferenciadas é importante no entendimento da prevalência de fenótipos (MYLES et al., 2008).

Na Tabela 17 estão discriminados os termos com valores mais significativos quanto ao teste de enriquecimento para atributos das bases de dados HGNC, OMIM, PGKB e Reactome. Dentre as vias metabólicas destacam-se as relacionadas à regulação gênica, transdução de sinal, metabolismo, ciclo celular, sistema imune, transporte de moléculas e desenvolvimento do sistema nervoso. Também se encontram enriquecidas as vias de metabolização do fármaco Celecoxib e de susceptibilidade ao câncer de mama. Conforme as anotações da base

PhamrGKB há uma sobre representação dos genes envolvidos na metabolização do lítio e de anti-inflamatórios. Das 5 bandas citogenéticas enriquecidas, duas encontram-se no cromossomo 19 e duas no cromossomo 17. As famílias gênicas enriquecidas apresentadas na Tabela 15 são: LNCRNA – *Long non-coding RNAs*, SLC – *solute carriers*, RPL – *L ribosomal proteins*, EFHAND – *EF- hand domain containing* e PLEKH - *Pleckstrin homology (PH) domain containing*.

Quanto às doenças, o maior sinal de enriquecimento está presente no termo câncer gástrico somático. Outras disfunções com tendência de viés incluem: Síndrome de Stevens-Johnson, hipersensibilidade a fármacos, transtorno bipolar, necrólise epidérmica tóxica, síndrome coronária aguda, câncer de mama somático, tetralogia de Fallot, deficiência do complexo I mitocondrial e susceptibilidade à asma.

Tabela 17: Enriquecimento de termos das bases de dados Reactome, PharmGKB, OMIM e HGNC para genes contendo SNPs divergentes em Nativo-Americanos

Atributos - Enriquecimento	Termos Enriquecidos para polimorfismos divergentes em NAT-(EUR ∪ WAFR ∪ EAS)			
	Termo	**p-Value**	**SNPs**	**Genes**
Reactome Vias Metabólicas	*Disease*	6.64E-43	358	77
	Olfactory Signaling Pathway	1.67E-16	76	35
	Signaling by the B Cell Receptor (BCR)	2.09E-14	31	8
	Downstream signaling of activated FGFR	2.40E-14	1	1
	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	5.74E-14	2	1
	Influenza Infection	2.69E-11	6	3
	Signaling by FGFR in disease	5.42E-11	29	8
	Metabolism of RNA	6.81E-11	40	17
	HIV Infection	9.32E-11	77	13
	DAP12 signaling	1.00E-10	2	2
PGKB Vias Metabólicas	*Celecoxib Pathway*	3.48E-06	2545	329
	Role of BRCA1, BRCA2 and ATR in cancer susceptibility	8.16E-06	19	6
	BRCA1 dependent ub ligase activity	7.15E-05	15	4
	Thromboxane A2 receptor signaling	1.95E-04	2478	312
	BARD1 signaling events	2.11E-04	37	13
PGKB Fármacos	*Lithium*	2.95E-09	2509	309
	Antiinflammatoryagents	3.43E-09	2406	292
	Celecoxib	0.003	238	35
	Brivanib	0.010	1	1
	Tipifarnib	0.010	5	1
PGKB Doenças	*Bipolar Disorder*	0.002	2513	311
	Stevens-Johnson Syndrome	0.003	57	7
	HIV	0.003	43	12
	Acute coronary syndrome	0.005	2380	294
	Epidermal Necrolysis, Toxic	0.006	7	4
OMIM Doenças	*Gastric cancer, somatic*	1.51E-04	12	2
	Breast cancer, susceptibility to	0.007	2	1
	Breast cancer, somatic	0.007	1	1
	Pheochromocytoma	0.018	1	1
	Myasthenic syndrome, slow-channel congenital	0.022	3	2
OMIM Banda	*19p13.3*	8.98E-21	8	6

	17q12	3.36E-17	15	3
	1q21.3	1.12E-16	8	5
	19p13.2	6.99E-16	32	11
	16p13.3	7.10E-16	15	10
Famílias Gênicas	*LNCRNA, ANTISENSE*	2.78E-11	112	39
	EFHAND	4.14E-10	412	79
	LNCRNA, INTRONIC	7.35E-10	6	2
	RPL	1.15E-09	12	10
	SLC	1.81E-09	644	155

Ainda que a análise de enriquecimento singular seja conveniente para detectar termos sobrerrepresentados em diferentes bases de dados, dependendo do tamanho do conjunto de resultados obtidos torna-se difícil identificar as relações entre os genes e as anotações nos diferentes atributos. Afora as restrições estatísticas (ver seção 4.3), esta é uma das principais limitações na interpretação dos resultados de enriquecimento utilizando SEA e GSEA. Espera-se, em breve, através da aplicação de algoritmos de descoberta de redes e métodos de clusterização adicionar novos níveis de extração do conhecimento na ferramenta MASSA. Como exemplo da potencialidade destas novas funções, utilizou-se o WS DAVID (HUANG et al., 2007) para identificar agrupamentos de genes com anotações semelhantes. Foram submetidos 2430 genes contendo anotações para dois atributos – processo biológico de GO e vias metabólicas de Reactome, sendo necessária esta seleção devido ao número máximo de genes (3000) aceito em DAVID. A partir de um método heurístico de determinação dos representantes de cada cluster, DAVID identificou 84 agrupamentos de genes com funções semelhantes, sendo os 10 clusters com maiores escores de enriquecimento representados na Tabela 18.

Tabela 18: Agrupamentos de genes com maiores escores de enriquecimento

Cluster	Número de genes	Escore de Enriquecimento	Função Biológica
C1	413	29,54	Receptores de membrana, sinalização celular.
C2	256	27,34	Receptores de membrana ionotrópicos.
C3	97	21,34	Metabolismo de carboidratos.
C4	14	21,29	Proteínas transmembrana.
C5	257	19,66	Fosforilação, constituição do músculo.
C6	7	18,35	Produção da Vitamina K, coagulação.
C7	54	17,36	Metabolismo de nucleosídeos trifosfatos.
C8	12	14,63	Transmissão sináptica.
C9	8	14,55	Transporte proteico.
C10	19	14,61	Metabolismo de lipídeos.

A partir dos agrupamentos de genes é possível empreender relações entre os termos enriquecidos, sendo que algumas destas relações não são previamente conhecidas em

populações Nativo-Americanas. Dois clusters foram selecionados para exemplificar as relações crípticas entre os termos de enriquecimento, apontando visões interessantes da biologia dos Nativo-Americanos.

O agrupamento C2, composto por genes que atuam no transporte iônico transmembrana, possui 16 genes associados ao transtorno bipolar, à síndrome coronariana aguda e à farmacocinética do lítio, utilizado no tratamento de desordens de humor, tais como depressão e transtorno bipolar. O processo farmacodinâmico de atuação do lítio na estabilização do humor não é plenamente conhecido. Supõe-se que o efeito seja obtido através da interação do fármaco com receptores ionotrópicos nos neurônios. Estes genes estão presentes, ainda, nas vias de metabolismo de anti-inflamatórios e de celecoxib, um inibidor da ciclo-oxigenase-2 utilizado no tratamento da artrite reumatoide, doença com alta incidência em Nativo-americanos. A ciclo-oxigenase-2 está envolvida em diversas funções fisiológicas como angiogênese, resposta ao estresse, alergias, dor e febre.

O agrupamento C53, essencialmente composto por nucleoporinas, atua no transporte intracelular. Genes deste cluster estão associados ao transporte de riboproteínas, monossacarídeos e pequenas moléculas pelo envoltório do núcleo celular. Atuam ainda na regulação da expressão gênica e resistência viral. Genes do agrupamento C53 estão, especialmente, relacionados à infecção por gripe e HIV. As populações Nativo-americanas são historicamente suscetíveis às doenças virais, apresentando baixa resistência frente a doenças como gripe, varíola, catapora e caxumba. Desta maneira, os constituintes deste grupamento constituem potenciais alvos para o entendimento da susceptibilidade viral, em especial, nas populações Nativo-Americanas.

Os resultados destas análises de estrutura populacional são, essencialmente, exploratórios. A utilização sistemática de métodos de extração do conhecimento biológico permite apontar alvos potenciais para explicar a diversidade fenotípica observada nas populações humanas. A partir dos resultados obtidos pelas análises de diferenciação populacional e enriquecimento de termos biológicos é possível selecionar conjuntos de regiões e genes com escopos mais restritos para diferentes funções biológicas. E, então, técnicas mais robustas podem ser utilizadas para testar a ação da seleção natural nestas regiões genômicas e a associação destas variantes com doenças ou fenótipos de interesse.

Além disso, a restrição do escopo torna viável o desenvolvimento de estudos funcionais para testar no âmbito molecular e celular, os achados bioinformáticos.

As perspectivas deste estudo no âmbito biológico incluem o sequenciamento de genes e regiões com indícios de seleção natural e o desenvolvimento de estudos funcionais para testar o impacto da variabilidade genética no desfecho fenotípico. Além disso, a mesma metodologia será aplicada aos dados de três populações Nativo-Americanas – 44 indivíduos Ashaninkas, 45 Shimaas e 16 Aymaras – genotipadas para o arranjo de 2.5M da Illumina, também utilizado nas amostras do Projeto EPIGEN-Brasil (ver seção 1.4).

CAPÍTULO 4 - INTEGRAÇÃO DE BASES DE DADOS PARA ANOTAÇÃO E ENRIQUECIMENTO

A anotação e o enriquecimento de dados visam aumentar a robustez dos resultados obtidos através das análises genético-populacionais. Desta forma, informações são acrescidas aos resultados previamente obtidos nas análises evolutivas e nos estudos de associação, buscando corroborar ou refutar o significado biológico do achado. Apesar da grande quantidade de dados disponíveis sobre variabilidade genética, o estudo das variantes genéticas com interesse evolutivo ou biomédico requer a associação desta informação a outras informações biológicas tais como vias metabólicas, fenótipos associados, severidade das mutações, dentre outras. Uma vez que toda esta informação heterogênea está fragmentada em diferentes bancos de dados, faz-se necessário desenvolver ferramentas que permitam integrar as diferentes fontes de dados biológicos no intuito de avançar nos estudos de epidemiologia genética e genética de populações. Ainda que existam ferramentas lidando com este problema, o gerenciamento de dados de larga escala ainda é problemático, principalmente quanto ao tempo de execução dos processos de integração e enriquecimento. Tecnologias computacionais capazes de implementar a execução paralela do código são uma solução promissora no processamento de grandes volumes de dados – problema também conhecido como *Big Data* – uma vez que elas são capazes de reduzir o tempo de execução através da divisão de tarefas em unidades de processamento independentes. Uma destas tecnologias é a baseada em sistemas multiagentes, onde cada agente é um programa autônomo que realiza uma tarefa particular e pode se comunicar com outros agentes do sistema multiagente para delegar tarefas ou compartilhar resultados. Agentes em um sistema multiagente podem ser executados concomitantemente tirando vantagem do poder de processamento de computadores com múltiplos processadores. De forma geral, duas características dos sistemas multiagente são relevantes para a implementação de processos de integração e enriquecimento: a modularidade e independência do conceito de agente, apropriadas para encapsular fontes heterogêneas e distribuídas de dados biológicos; e a natureza paralela do sistema multiagente, adequada para otimizar o processamento de dados em larga escala. Nesse contexto, a tecnologia multiagente foi escolhida como paradigma para o desenvolvimento do sistema de anotação e enriquecimento em larga-escala denominado MASSA (*MultiAgent System for Snp Annotation*).

Este capítulo está dividido em 11 seções, sendo que as seções 4.1 à 4.9 descrevem as funcionalidades e os componentes da plataforma MASSA. Em resumo, estas nove seções referem-se, respectivamente, à: i) descrição genérica do sistema; ii) descrição da implementação do sistema; iii) delineamento do processo de enriquecimento; iv) caracterização do I/O da ferramenta, ou seja, os arquivos de entrada requeridos e os de saída gerados; v) delineamento dos modos de anotação presentes; vi) descrição das informações e bancos de dados utilizadas no processo de anotação; vii) detalhamento da arquitetura e do fluxo de informações do sistema; viii) caracterização da comunicação entre os agentes; ix) exposição dos algoritmos e do fluxo de ação dos agentes. A seção 4.10 apresenta estudos piloto para determinar a eficiência e escalabilidade da ferramenta. E o estudo de caso apresentado na seção 4.11 ilustra uma das possibilidades de uso da ferramenta e descreve os resultados gerados por MASSA. Além disto, o capítulo 3 apresenta um exemplo de utilização massiva dos resultados obtidos a partir da plataforma.

4.1 VISÃO GERAL DO MASSA

A plataforma MASSA, *Multi-Agent System for SNP Annotation*, tem por objetivo coletar, integrar e fazer o enriquecimento de informações sobre SNPs para gerar uma robusta e confiável anotação destas variantes. Para alcançar esse objetivo, o MASSA foi projetado e implementado com três tipos de agentes, cada um com atividades particulares: Agente de Interface, Agente Coordenador e o Agente de Banco de dados (agente DB). O Agente de Interface gerencia as atividades de entrada e saída de dados, que consistem em ler uma lista de SNPs e gerar os relatórios finais de anotação e enriquecimento, e dá início aos processos de anotação e enriquecimento. O agente Coordenador é responsável por coordenar o processo de anotação de SNPs e genes, delegando tarefas aos agentes DB e combinando os resultados retornados. Os agentes DB encapsulam o acesso aos bancos de dados, inquirindo estes sobre atributos de SNPs e genes, e informando ao agente Coordenador o resultado das consultas. Ao longo da execução do sistema, esses agentes se comunicam através de troca de mensagens para concluir os processos de anotação e enriquecimento (ver Seção 4.8 para detalhes da comunicação dos agentes). O sistema conta ainda com uma opção de anotação simples, onde alguns atributos básicos são retornados.

4.2 IMPLEMENTAÇÃO

MASSA é implementada utilizando o framework, baseado em JAVA, JADE (BELLIFEMINE; POGGI; RIMASSA, 1999), onde cada agente é executado como um thread (processo) individual no sistema, e os comportamentos paralelos dos agentes permitem a distribuição das tarefas, permitindo, assim a paralelização do sistema. Além disso, todas as funcionalidades relacionadas à comunicação entre agentes estão implementadas através dos métodos disponibilizados pelo ambiente JADE.

Três camadas de paralelização foram implementadas no sistema: i) o Agente Interface divide a lista de SNPs em *pR* sublistas (*pR* – taxa de paralelização, ver seção 4.9) e as envia ao agente coordenador para serem anotadas concorrentemente; ii) O agente coordenador delega a tarefa de anotação aos agentes DB conjuntamente, iniciando num primeiro momento todos os agentes BD baseados em informações de SNPs, e logo após, todos os agentes DB baseados em genes; iii) os agentes DB recebem um subconjunto de SNPs ou genes e retornam as informações para cada um destes SNPs e genes simultaneamente – isto é possível devido ao fato de que o sistema possui *pR* cópias de cada agente DB. Com esta abordagem tira-se proveito de máquinas com múltiplos núcleos e na seção 4.10 demonstra-se que estas camadas melhoram significativamente a performance do sistema de anotação, ou seja, quanto maior a taxa de paralelização (pR), menor é o tempo de anotação. Entretanto, cada máquina pode dispor de um limite de saturação dos processadores, considerando a configuração do servidor onde realizamos os testes, Dell Intel Xeon E7 4x8 núcleos 16x8 Gb memória, este limite é de 26 pRs. Além disso, com esta implementação, é possível executar o agente coordenador e os agentes BD em diferentes máquinas, o que provavelmente melhorará ainda mais o tempo de anotação. Espera-se, em breve, testar essa funcionalidade.

4.3 ENRIQUECIMENTO

Outro foco desse sistema é a utilização do resultado da anotação de dados para o enriquecimento de termos associados a informações biológicas relevantes, tais como vias metabólicas, termos de Gene Ontology, famílias gênicas, regiões genômicas, dentre outros. Os testes de enriquecimento consistem no cálculo da probabilidade de se amostrar um número igual ou superior de termos relacionados a um determinado conjunto de SNPs ou genes em

comparação a uma amostragem aleatória do conjunto total de SNPs ou genes (BAUER et al., 2008). Dentre as metodologias estatísticas utilizadas nos cálculos de enriquecimento encontram-se o teste Qui-Quadrado, Exato de Fisher, Binomial, e o teste Z. A partir destes resultados, as metodologias de enriquecimento podem não só aumentar a confiança dos estudos como permitir a identificação dos processos biológicos subjacentes a determinados conjuntos gênicos (HUANG; SHERMAN; LEMPICKI, 2009). Tanto a anotação quanto o enriquecimento de dados são extremamente relevantes para agregar conhecimento biológico aos dados genômicos "crus" que são gerados por genotipagem e sequenciamento, e, dessa forma, facilitar estudos de genética de populações e epidemiologia genética de alto nível.

Entretanto, as análises de enriquecimento possuem algumas limitações: problemas ou omissões na anotação dos genes, redundância de informações, influência do tamanho amostral e populacional nos valores de significância e ausência de padrão ouro para comparação e validação de métodos. SEA apresenta restrições na visualização dos resultados, na diluição das relações entre os termos e genes e na definição do valor de corte – problema usual nos testes que envolvem comparações múltiplas. GSEA requer um valor biológico sumário (por exemplo, aumento da expressão), que dependendo da aplicação, pode não ser fácil de obter. Além disso, os valores são ranqueados de acordo com o valor de p, entretanto, nem sempre este valor corresponde à realidade biológica, por exemplo, uma pequena alteração no nível de expressão de um gene, pode desencadear grandes mudanças em cascata. A limitação de MEA está relacionada aos genes e termos órfãos, sem relações fortes entre as anotações, desta maneira, genes e termos podem ser excluídos das análises (HUANG; SHERMAN; LEMPICKI, 2009).

A princípio, foram considerados seis enfoques estatísticos para avaliar o nível de enriquecimento dos termos biológicos: i) Teste Exato de Fisher, ii) Qui-Quadrado, iii) Distribuição hipergeométrica, iv) Distribuição binomial, v) Teste-t e vi) Teste z. Dentre as restrições de cada enfoque estão: v) e vi) exigem que os dados tenham distribuição normal, o que nem sempre é verdadeiro para as anotações; iii) e iv) podem ser aproximadas pelo teste Exato de Fisher que é computacionalmente mais rápido de ser testado; ii) o teste do Qui-Quadrado exige amostras superiores a 20 e frequências esperadas superiores a 5 quando o teste é 2x2, desta forma, seria necessária a correção de Yates que aumenta a estatística do teste e superestima o p-valor; e i) é um teste conservador. Entretanto, as vantagens do uso do teste Exato de Fisher são: funciona bem para amostras pequenas, não exige suposições

paramétricas por ser um teste exato e não apresenta restrições quanto à quantidade de observações por cela. Desta forma, optou-se pela utilização do teste Exato de Fisher bilateral, usualmente calculado para determinar se duas proporções são diferentes.

Os valores críticos de significância são calculados a partir da correção de Bonferroni, onde o p-valor é estabelecido por $\alpha/n° \, de \, termos$. Devido ao caráter conservador deste teste, para conjuntos de testes superiores a 1000 termos, utiliza-se o valor de 10^{-5} como ponto de corte para a identificação de termos enriquecidos na amostra.

Atualmente, encontra-se implementado no sistema MASSA o Cálculo Exato de Fisher disponível na biblioteca estatística JSC (*Java Statistical Classes*). A função *FisherExactTest* realiza o teste a partir de uma tabela de contingência encapsulada no objeto *ContingencyTable2x2* dessa mesma biblioteca. A tabela é criada a partir das contagens dos termos presentes no resultado da anotação e estas são então comparadas aos cômputos dos termos presentes na população total, ou seja, do total de termos presentes na base de dados. A contagem dos termos no conjunto de referência, ou populacional, está armazenada na classe *BioEnrichment* do sistema e é dependente do banco de dados de origem. Isto implica que um termo só pode ser testado em relação às contagens do próprio banco de origem. No futuro, o desenvolvimento contínuo de ontologias e sua utilização poderão minimizar os efeitos da heterogeneidade dos dados.

MASSA realiza os testes de enriquecimento para os seguintes atributos biológicos: doenças (OMIM e PharmGKB); vias metabólicas (PharmGKB e Reactome); fármacos (PharmGKB); processo biológico, função molecular e componente celular (GO); banda citogenética (OMIM) e família gênica (HGNC). A função de enriquecimento é realizada pelo Agente de Interface, após a geração do arquivo de sumário das anotações. Exemplos dos resultados de enriquecimento podem ser vistos no capítulo 3 e na seção 4.11.

4.4 ENTRADA E SAÍDA DE DADOS

O sistema recebe como input uma lista de identificadores (dbSNP rs ID) e gera quatro arquivos de saída: o relatório da anotação, um arquivo de log, um arquivo sumário e o relatório de enriquecimento (opcional). O relatório de anotação apresenta ao usuário 66

informações sobre o polimorfismo extraídas de 11 bancos de dados distintos. Para a opção de anotação rápida são retornados 16 atributos da base dbSNP. O arquivo de log apresenta informações sobre a proveniência de cada atributo, indicando o nome e a versão de cada banco de dados. O sumário apresenta um resumo dos dados indicando o número de ocorrências dos valores de 14 atributos. O relatório de enriquecimento retorna os resultados dos testes de enriquecimento para 9 atributos. Exemplos destes arquivos são encontrados nas Figura 14 e Figura 15 e nas Tabelas Tabela 31 e Tabela 32 da seção 4.11.

4.5 MODOS DE ANOTAÇÃO

O sistema possui dois modos de anotação, local e remota, que diferem no método de acesso às bases de dados incorporadas ao sistema, o qual pode ser através de consultas tanto a bancos remotos quanto a bancos instalados localmente. A escolha pelo modo de anotação depende dos requerimentos do usuário, como descrito a seguir. No modo de acesso remoto, a coleta de dados é realizada diretamente a partir dos bancos de dados on-line utilizando APIs ou webservices. Este modo é indicado para anotar pequenos conjuntos de dados e possui dois benefícios: a informação está sempre atualizada, uma vez que é requerida em tempo real, a partir do próprio banco de dados on-line, e não há necessidade de se instalar bancos de dados locais, tornando mais simples o processo de configuração e manutenção; a principal desvantagem deste modelo é o limite da banda de acesso aos bancos remotos, restringindo, assim, o tamanho dos conjuntos de polimorfismos a serem anotados em poucas centenas de SNPs. No modo de acesso local, as informações são consultadas a partir de versões locais dos bancos de dados. Esta versão é indicada para a anotação de grandes conjuntos de dados devido à maior velocidade de anotação e à restrição de banda da anotação remota, entretanto, exige esforço na implementação e atualização dos bancos locais.

4.6 BASES DE DADOS

A plataforma MASSA integra informações de 11 bancos de dados de interesse às áreas de genética de populações, farmacogenômica e epidemiologia genética: dbSNP (SHERRY et al., 2001), UCSC (KAROLCHIK et al., 2014), Gene Ontology (ASHBURNER et al., 2000), PharmGKB (WHIRL-CARRILLO et al., 2012), OMIM (“Online Mendelian Inheritance in

Man, OMIM®," 2014), Reactome (CROFT et al., 2014; JOSHI-TOPE et al., 2003; MATTHEWS et al., 2007, 2009; VASTRIK et al., 2007), HGNC (GRAY et al., 2013), NHGRI GWAS Catalog (WELTER et al., 2014), PolyPhen2 (ADZHUBEI et al., 2010), Provean (CHOI et al., 2012), SIFT (KUMAR; HENIKOFF; NG, 2009). O acesso remoto depende da tecnologia disponibilizada por cada base, sendo utilizadas ferramentas de acesso baseadas em MySQL, APIs em Perl e Webservices. Os bancos locais estão armazenados em um SGBD MySQL implementado em um servidor Linux. Na Tabela 19 há uma breve descrição das bases de dados; as formas de acesso e o processo de construção dos bancos de dados locais são descritos no Apêndice A.

Tabela 19: Fonte dos dados de anotação

Base de dados	Descrição
dbSNP	Armazena informações básicas sobre variantes genéticas e é o principal repositório de SNPs.
UCSC	Disponibiliza grande quantidade de anotações genômicas. O MASSA retorna deste, informações sobre SNPs e conservação nucleotídica.
Gene Ontology	Além da ontologia, o GO dispõe de anotações para genes e produtos gênicos. As anotações são classificadas em três áreas: processo biológico, função molecular e componente celular.
PharmGKB	Contém informações farmacogenéticas, tais como, vias metabólicas, fármacos e doenças associadas.
OMIM	Armazena informações sobre doenças, em particular, as mendelianas.
Reactome	Base de dados curada sobre vias metabólicas.
HGNC	Repositório de nomenclaturas e famílias gênicas.
PolyPhen2	Ferramentas de predição da severidade de mutações. Os dados utilizados para a construção das bases de dados locais consistem dos exemplos de anotação disponibilizados pelos autores dos softwares.
SIFT	
Provean	
NHGRI GWAS Catalog	Base de dados sobre *GWASs*.

O encapsulamento dos métodos de acesso remoto é dividido em três grandes grupos: a) Acesso MySQL, neste caso faz-se necessária a instalação e uso da biblioteca JDBC (*mysql-connector* – versão 5.1.21), pois esta realiza a mediação entre o aplicativo em JAVA e o banco de dados relacional; b) Acesso via WebService, neste caso os mantenedores das bases de dados disponibilizam scripts para acesso aos dados, essas APIs podem ser baseadas em SOAP ou WSDL, por exemplo; c) Acesso remoto simulado, nos casos onde não é possível o acesso remoto aos dados, por exemplo, GWAS Catalog, arquivos de texto para consulta são disponibilizados em conjunto com a ferramenta. O acesso via MySQL é utilizado para interrogar todas as bases de dados no modo de anotação local.

4.6.1 DBSNP

O banco de dados dbSNP é um dos principais repositórios de variantes genéticas existentes, em especial, dos polimorfismos de uma única base (SNPs) e, em menor escala, de pequenas inserções, deleções e microssatélites. Atualmente, armazena dados para aproximadamente 63 milhões de variantes, sendo 44 milhões destas validadas. Os principais objetivos desta base de dados consistem em validar e caracterizar as variantes genéticas, permitindo a utilização destes polimorfismos pela comunidade científica. dbSNP disponibiliza ainda anotações e links com informações sobre estes polimorfismos em outras bases de dados. Os métodos de acesso incluem APIs através do conjunto de ferramentas Eutils, acesso via página web (http://www.ncbi.nlm.nih.gov/snp/) e download dos dados via FTP. Para a anotação remota utilizamos as APIs do Eutils que acessam os dados em bases de dados no formato FLT (*Flat file*) e XML (*Extend Markup Language*), os dados recuperados da base dbSNP estão descritos na Tabela 20.

Tabela 20: Mapa de recuperação de informações - agente DB dbSNP

Informação	Campo FLT	Campo XML	Campo SQL
Código do polimorfismo	First Line, first column	<rs>rsID	SNP (snp_id)
Tipo do polimorfismo	-	<rs>snpClass	b137_MapLinkHGVS (snp_type)
Região do transcrito	LOC - fxn-class	<FxnSet> fxnClass	b137_SNPContigLocusId (fxn_class)
Genótipo observado	SNP - alleles	<Observed>	SubPop (Observed)
Nucleotide Numbering coding DNA	-	<hgvs> (NM_)	SNP_HGVS (hgvs_name)
Cromossomo	CTG - chr	<Component> chromosome	b137_SNPChrPosOnRef (chr)
Nucleotide numbering genomic Ref Seq	-	<hgvs> (NC_)	b137_SNPChrPosOnRef (pos)
Versão do Assembly	CTG - assembly	<Assembly> genomeBuild	b137_SNPContigLocusId (build_id)
Coordenada inicial	CTG - chr_pos	<Component Type> accession	b137_SNPContigLocusId (asn_from)
Coordenada final	CTG - chr_pos	<Component Type> orientation	b137_SNPContigLocusId (asn_to)
Orientação	-	<Assembly> genomeBuild	b137_MapLinkHGVS (orientation)
Gene ID	LOC - locus_id	<FxnSet> geneId	b137_SNPContigLocusId (locus_id)
Símbolo do gene	LOC - (second attribute)	<FxnSet> symbol	b137_SNPContigLocusId (locus_symbol)

Alelo Ancestral	-	<Sequence> ancestral_allele	SNPAncestralAllele (ancestral_allele_id)
Identificador do RNAm	-	<hgvs> (NM_)	b137_SNPContigLocusId (mrna_acc)
Versão do RNAm	-	<hgvs> (NM_)	b137_SNPContigLocusId (mrna_ver)
Frequência dos alelos	-	<hgvs> (freq)	SNPAlleleFreq (freq)
Alelo referência	-	<Frequency> allele	SNPAlleleFreq (allele_id)

Os campos FLT e XML são utilizados na pesquisa remota e os campos SQL, na pesquisa local. São indicados acima os campos e os atributos de onde as informações são retiradas. Tomando como exemplo a informação código do gene (Gene ID): no campo FLT o primeiro código, LOC, indica o campo e o segundo código, locus_id, indica o atributo onde a informação está localizada, o mesmo ocorre nos campos XML (<FxnSet> tag-XML correspondente ao campo, geneId indica o atributo a ser retornado) e SQL (b137_SNPContigLocusId refere-se à tabela de consulta e locus_id ao atributo desejado).

4.6.2 UCSC

A plataforma *UCSC Genome Browser* permite a visualização de diferentes níveis de anotação genômica em diversos organismos, atualmente, há dados para 100 espécies (Novembro, 2013). Disponibiliza para humanos dados de conservação e análises comparativas evolucionárias; modelos gênicos; dados de regulação, expressão e epigenéticos; diferenciação tecidual, variação, fenótipos e associação a doenças. Como coordenadora de dados do projeto ENCODE disponibiliza dados sobre sítios de hipersensibilidade à DNAse, marcadores de histona e níveis de expressão. Os dados e as respectivas fontes consultadas pelo sistema MASSA estão descritos na Tabela 21. As versões local e remota compartilham a metodologia de acesso através de instruções MySQL. E a única diferença entre os dados obtidos entre as formas de acesso local e remota é a de que no acesso remoto não dispomos dos valores de conservação para cada sítio, ao invés disso, retorna-se os valores de conservação médios para um conjunto de 1024 sítios. No acesso local são disponibilizados tanto a informação por sítio quanto a por conjunto de 1024 sítios. Isto se deve ao fato de que os valores de conservação por sítio não se encontram na base de dados relacionais, sendo disponibilizados através dos arquivos *wiggle* – formato para compactação de dados de conservação utilizado pelo UCSC.

Tabela 21: Mapa de recuperação de informações - agente DB UCSC

Informação	Campo	Fonte
Strand	Strand	snp137
Alelo referência UCSC	refUCSC	snp137
Genótipo referência UCSC	observed	snp137
Tipo de polimorfismo	class	snp137
Classificação funcional	func	snp137
Conservação Primatas (phastCons46)[a]	score	phastCons46wayPrimates
Conservação Mamíferos (phastCons46) [a]	score	phastCons46wayPlacental
Conservação Vertebrados (phastCons46) [a]	score	phastCons46way
Conservação Primatas (phyloP46) [a]	score	phyloP46wayPrimates
Conservação Mamíferos (phyloP46) [a]	score	phyloP46wayPlacental
Conservação Vertebrados (phyloP46) [a]	score	phyloP46wayAll
Conservação Primatas (phastCons46)[b]	score	phastCons46wayPrimatesOneStep
Conservação Mamíferos (phastCons46) [b]	score	phastCons46wayPlacentalOneStep
Conservação Vertebrados (phastCons46) [b]	score	phastCons46wayOneStep
Conservação Primatas (phyloP46) [b]	score	phyloP46wayPrimatesOneStep
Conservação Mamíferos (phyloP46) [b]	score	phyloP46wayPlacentalOneStep
Conservação Vertebrados (phyloP46) [b]	score	phyloP46wayAllOneStep

[a] Disponível nas versões remota e local do MASSA. [B] Disponível apenas na versão local do MASSA.

4.6.3 GENE ONTOLOGY

A base de dados *Gene Ontology* tem como objetivo criar padrões na nomenclatura dos atributos dos produtos gênicos e dos genes entre diferentes espécies e bancos de dados. Através de um vocabulário controlado, as ontologias, é possível descrever os genes de acordo com sua função molecular, processo biológico e componentes celulares. A função molecular corresponde às atividades, funções, do produto gênico a um nível molecular; processos biológicos são operações ou conjuntos de eventos moleculares com início e fim definidos, relacionados ao funcionamento das unidades biológicas; componentes celulares definem a área de atuação dos produtos gênicos quanto à célula ou seu ambiente extracelular. Os dados acessados são descritos na Tabela 22.

O banco de dados Gene Ontology pode ser acessado por diversas ferramentas, sejam elas do consórcio Gene Ontology ou desenvolvidas por terceiros. Dentre os métodos de acesso disponibilizados pelos desenvolvedores do banco estão: a plataforma AmiGO, GOOSE, módulos em Perl, acesso MySQL e download dos dados. Destes selecionamos a conexão via MySQL para recuperar os dados no modo de anotação remoto.

Tabela 22: Mapa de recuperação de informações - agente DB Gene Ontology

Informação	Campo	Fonte
Símbolo do Gene [a]	gp_symbol	term_J_association_J_evidence_J_gene_product
Símbolo do Gene [b]	symbol	gene_product
Especificação do termo[a1]	term_type	term_J_association_J_evidence_J_gene_product
Especificação do termo[b1]	term_type	Term
Nome do Termo[a]	term_name	term_J_association_J_evidence_J_gene_product
Nome do Termo[b]	name	Term
ID de acesso do Termo[a]	term_id	term_J_association_J_evidence_J_gene_product
ID de acesso do Termo[b]	acc	Term
Função Molecular [a2]	term_name	Ver observação [2]
Função Molecular [b2]	name	Ver observação [2]
Componente Celular [a2]	term_name	Ver observação [2]
Componente Celular [b2]	name	Ver observação [2]
Processo Biológico [a2]	term_name	Ver observação [2]
Processo Biológico [b2]	name	Ver observação [2]

[a] Implementação remota do MASSA. [B] Implementação local do MASSA. [1] Especifica a ontologia a ser utilizada (*molecular function*, *cellular component*, *biological process*). [2] Os termos são discriminados de acordo com o tipo da ontologia descrito no campo especificação do termo (term_type).

4.6.4 PHARMGKB

A plataforma *The Pharmacogenetics and Pharmacogenomics Knowledge Base (*PharmGKB*)* tem por objetivo criar um repositório relacionando as respostas individuais ao uso de fármacos à variantes genéticas. O sítio da plataforma (www.pharmgkb.org) permite a visualização de diferentes informações integradas, tais como genótipos, moléculas, dados clínicos, bibliografia e representação de vias metabólicas, as informações utilizadas por MASSA estão descritas na Tabela 23. O acesso aos dados disponibilizados pela plataforma PharmGKB pode ser realizado através da interface do sítio, download dos dados e APIs implementadas em Perl, sendo esta última a utilizada na conexão remota do MASSA.

Tabela 23: Mapa de recuperação de informações - agente DB PGKB

Informação	Campo	Fonte
Nome abreviado do Gene[a]	Pgkbkey	Gene
Código de Identificação do Gene[a]	pgkbId	Gene
Genes Relacionados[a]	relatedGenes	Gene
Localização genômica do polimorfismo[a]	name	goldenPath

Código de Identificação do Polimorfismo[a]	pgkbkey	goldenPath
Código de Identificação do Haplótipo[a]	name	Haplotype
Nome do Fármaco[a]	name	Drug
Nome da doença[a]	name	Disease
Nome da Via Metabólica[a]	name	Pathway

[a] Disponível nas versões remota e local do MASSA.

4.6.5 OMIM

A plataforma Online Mendelian Inheritance in Man (OMIM) consiste num repositório de dados livre destinado a descrever as doenças mendelianas, focando na relação entre fenótipos e variantes genéticas, as informações requeridas por MASSA estão discriminadas na Tabela 24. O acesso à base de dados OMIM pode ser realizado através da interface de busca do sítio (www.omim.org), de APIs disponibilizadas pelo desenvolvedor ou do download dos dados. As APIs permitem o retorno das informações em diferentes estruturas de dados como: Ruby, HTML, JSON, Python e XML, sendo este último o método de acesso remoto do agente DB OMIM.

Tabela 24: Mapa de recuperação de informações - agente DB OMIM

Informação	Campo	Fonte
Sistema de numeração OMIM[a]	Id	genemap
Localização Citogenética[a]	cytoLoc	genemap
Nome abreviado do Gene[a]	geneSymbol	genemap
Status do Gene[a]	geneStatus	genemap
Título do estudo[a]	title1/2	genemap
Código de Identificação OMIM[a]	mimId	genemap
Método[a]	method	genemap
Comentários[a]	comments1/2	genemap
Desordem/doença[a]	disorder	genemap
Herdabilidade[b]	inheritance	genemap
Referências[a]	Reference	genemap

[a] Disponível nas versões remota e local do MASSA. [B] Disponível apenas na versão remota do MASSA.

4.6.6 REACTOME

Reactome é um banco de dados sobre vias metabólicas e tem como objetivo disponibilizar dados e ferramentas para a visualização, interpretação e análise de vias. Uma das vantagens deste banco é o processo de curadoria e revisão dos dados submetidos que proporciona confiabilidade à anotação disponibilizada. Atualmente, Reactome se encontra na versão 47 disponibilizando anotações para 7200 dos 20744 genes codificadores presentes no Ensembl, 15107 referências bibliográficas e 1421 moléculas organizadas em 6849 reações provenientes de 1491 vias metabólicas, sendo o mapa de recuperação de informações apresentado na Tabela 25. As formas de acesso ao banco incluem: acesso via web, módulo de conexão via BioMart, plug-in para o software Cytoscape e download dos dados. Atualmente, o agente DB Reactome remoto utiliza o acesso remoto simulado.

Tabela 25: Mapa de recuperação de informações - agente DB Reactome

Informação	Campo	Fonte
Via Metabólica	pathway	reactome_pathway

4.6.7 HGNC

O Comitê de Nomenclatura Gênica da HUGO (*Human Genome Organisation*) é responsável por definir e padronizar a nomenclatura de genes. A partir deste comitê são definidos os nomes e símbolos oficiais dos genes humanos. Cada nome e símbolo é único e deve representar as famílias de genes e ser facilmente aplicável à genes de outras espécies (por exemplo: camundongo). O acesso aos dados do HGNC pode ser realizado via interface web através do sítio (http://www.genenames.org/hgnc-searches). Outra forma de acesso remoto, sendo a utilizada por MASSA, se dá através de um script CGI que acessa diretamente o banco de dados em MySQL. As informações retornadas são descritas na Tabela 26.

Tabela 26: Mapa de recuperação de informações - agente DB HGNC

Informação	Campo	Fonte
Código de Identificação do HGNC	hgncId	hugoDB
Símbolo oficial do gene	approvedSymbol	hugoDB

Nome oficial do gene	approvedName	hugoDB
Tipo do locus (ex. microRNA)	locusType	hugoDB
Agrupamento de tipos de locus	locusGroup	hugoDB
Outros símbolos utilizados	synonyms	hugoDB
Etiqueta para designar uma família gênica	geneFamilyTag	hugoDB
Nome da família gênica	geneFamilyDesc	hugoDB

4.6.8 POLYPHEN-2

PolyPhen-2 é a segunda versão de um software criado para predizer a severidade das substituições aminoacídicas na estrutura e função das proteínas. A segunda versão difere da primeira no conjunto de atributos preditivos, no fluxograma de alinhamento das sequencias e no método de classificação. Dois conjuntos de dados foram utilizados no treinamento e validação da ferramenta: a) HumVar, um conjunto de 13032 mutações causadoras de doenças segundo o banco de dados UniProt e 8496 SNPs não sinônimos não deletérios; b) HumDiv, representado por 3155 alelos nocivos em doenças mendelianas segundo o UniProt e 6321 variantes presentes em sequências de mamíferos filogeneticamente próximas, preditas como não danosas. As predições HumVar são utilizadas como repositório de severidade das mutações, tanto no acesso local quanto remoto, sendo utilizado, neste caso, o acesso remoto simulado. As informações retornadas estão descritas na Tabela 27.

Tabela 27: Mapa de recuperação de informações - agente DB PolyPhen

Informação	Campo	Fonte
Identificador da proteína	o_acc	polyphen2
Substituição	nt1/nt2	polyphen2
Posição na sequência polipeptídica	o_pos	polyphen2
Predição PolyPhen-2	prediction	polyphen2
PolyPhen-2 Probabilty	pph2_prob	polyphen2
PolyPhen-2 FDR	pph2_fdr	polyphen2
Predição PolyPhen-1	effect	polyphen2

4.6.9 PROVEAN/SIFT

Provean e SIFT são ferramentas criadas pelo mesmo grupo de pesquisa para estimar o impacto das alterações na cadeia polipeptídica. SIFT foi a primeira a ser criada e calcula o escore de severidade a partir da distribuição dos resíduos aminoacídicos observados em uma

dada posição durante o alinhamento das sequências e das frequências não observadas na distribuição de aminoácidos calculada a partir da distribuição de Dirichlet. Provean calcula a severidade das mutações para todos os tipos de variantes na sequência proteica e o escore final estima a alteração na similaridade de uma sequência em relação a uma homóloga, antes e depois da introdução de uma substituição na sequência da proteína. Os bancos de dados local e remoto – acesso remoto simulado – foram construídos a partir de predições genômicas realizadas para SIFT e Provean, sendo as informações requeridas por MASSA descritas na Tabela 28.

Tabela 28: Mapa de recuperação de informações - agente DB Provean/SIFT

Informação	Campo	Fonte
Identificador da proteína	protein_id	protein sequence change
Substituição	residue_ref/residue_alt	protein sequence change
Posição na sequência polipeptídica	position	protein sequence change
Predição Provean	prediction	Provean prediction
Escore Provean	score	Provean prediction
Predição SIFT	prediction	SIFT prediction
Escore SIFT	score	SIFT prediction

4.6.10 NHGRI GWAS CATALOG

O catálogo do NHGRI é uma coleção de resultados de estudos de associação genômicos obtidos através da revisão da literatura e da comparação a outros bancos de dados de GWAS. São admitidos apenas estudos com pelo menos 100.000 marcadores e valores de significância inferiores a 10^{-5}. Atualmente, o catálogo conta com 11.912 SNPs discriminados em 1751 publicações curadas. O acesso à base pode ser realizado através da interface web, do download dos dados em formato TSV, da base em OWL e da visualização dinâmica a partir dos cariótipos presentes no sítio (http://www.genome.gov/gwastudies/). Os dados pesquisados por MASSA estão discriminados na Tabela 29 e o modo de anotação remoto destes se dá por acesso remoto simulado.

Tabela 29: Mapa de recuperação de informações - agente DB GWAS

Informação	Campo	Fonte
SNP com risco mais elevado	strongest_SNP_risk_allele	gwas_catalog
Região genômica	context	gwas_catalog

Genes	reported_genes	gwas_catalog
Valor de significância	p-value	gwas_catalog
Fenótipo	disease/trait	gwas_catalog
Tamanho amostral (inclui classificação étnica)	initial_sample_size	gwas_catalog
Código de acesso PubMed	pubmedID	gwas_catalog

4.7 ARQUITETURA E FLUXO DO SISTEMA

O sistema MASSA é composto por 24 agentes: 1 Agente de Interface, 1 Agente Coordenador e 22 Agentes de Banco de dados (agente DB), um para cada banco de dados incorporado ao sistema, ambos na versão local e remota. A arquitetura do sistema pode ser vista na Figura 10.

O sistema se inicia através da leitura do agente Interface ao registro de configuração (*masconfig*) e a partir deste arquivo são indicados: o arquivo contendo a lista de SNPs a serem anotados (*snp_list_file*), o modo de acesso (local ou remoto), o tipo de anotação (resumida ou completa) e o grau de paralelização do sistema (*pquery* – pR, daqui em diante). A lista de SNPs indicada no arquivo de configuração é lida pelo agente de Interface, checada em busca de inconsistências (identificadores diferentes do previsto) e dividida pela taxa de paralelização (pR) em N sublistas (onde N = pR), as quais são enviadas separadamente para o agente coordenador em N mensagens. O controle de cada uma das listas é feito através de um identificador de pesquisa (*searchID*) e o coordenador também é informado sobre o modelo de anotação, se local ou remoto, e se há opção de anotação simples.

Após receber as sublistas de SNPs, o agente coordenador as envia aos agentes DB para anotação. Há dois passos sequenciais a serem seguidos pelo Coordenador para gerenciar o processo de anotação: a) a anotação de informações sobre polimorfismos, e b) a anotação de informações sobre genes. No passo (a), cada sublista é enviada aos agentes DB que disponibilizam as informações sobre polimorfismos, como dbSNP, UCSC, GWAS Catalog, PolyPhen2, Provean e SIFT. Dentre os atributos retornados neste primeiro momento está o nome do gene onde o SNP se encontra. Com isso, o Coordenador segue à próxima fase, passo (b), onde filtra a lista de genes e a envia aos demais agentes DB: OMIM, GO, PGKB, HGNC

e Reactome. Entretanto, o passo b) só é executado na anotação completa, na anotação rápida o Coordenador requisita apenas a anotação do agente dbSNP.

Figura 10: Arquitetura do sistema MASSA

Para acomodar as requisições de todos os agentes coordenadores são criadas pR instâncias (cópias) dos agentes DB. Todos os agentes bancos de dados atuam de forma sincrônica, em um design similar. Eles recebem do coordenador a lista de SNPs ou de genes a serem anotados, consultam as bases de dados de forma paralela e retornam um conjunto de atributos ao Coordenador. Ou seja, cada consulta por SNP ou gene resulta em uma inquirição concorrente ao banco de dados correspondente. Após a anotação de toda a lista, o agente DB retorna a lista e as informações ao agente Coordenador, que irá combinar as listas recebidas e enviar a lista completa dos polimorfismos anotados ao agente Interface.

O agente Interface espera até que todos o agente Coordenador retorne as listas de informações baseadas em genes e polimorfismos e, após isto, gera 4 arquivos de saída: o arquivo de anotação, em formato TSV (*tab separated values*); o arquivo de log, com as informações de tempo de execução, proveniência dos dados e quantidade de dados anotados (e não anotados, caso existam); o arquivo de sumário, que informa ao usuário o número de ocorrências dos valores relacionados a alguns atributos; e o relatório de enriquecimento (opcional), onde são informados os resultados dos testes de enriquecimento para alguns atributos de interesse (ver exemplos na seção 4.11). Após a criação dos arquivos de saída, o agente Interface desliga o sistema. A descrição detalhada do funcionamento de cada agente pode ser vista na Seção 4.9.

O sistema lida com as anotações intermediárias, isto é, aquelas enviadas dos agentes DB aos coordenadores e ao agente interface de maneira diferente em cada modo de anotação. Na versão remota, as anotações intermediárias são armazenadas em estruturas de dados especiais, mantidas na memória. Já na versão local, devido ao tamanho das anotações intermediárias, as informações são armazenadas em um banco de dados temporário, nomeado *annotation*. Neste caso, os agentes transmitem entre si apenas o identificador de pesquisa (*search_id*) para suas respectivas anotações, evitando que a memória do computador fique sobrecarregada com os dados da anotação intermediária.

4.8 COMUNICAÇÃO ENTRE OS AGENTES

Os agentes se comunicam trocando mensagens. As partes básicas de informação que uma mensagem pode conter são: a) o remetente da mensagem, que pode ser o nome ou identificador do agente; b) o destinatário da mensagem, identificado por nome ou ID; c) performativo ou o propósito da mensagem; que pode ser, por exemplo, uma requisição (*REQUEST*) ou resposta (*REPLY*) a uma requisição; e d) o conteúdo da mensagem. A representação da troca de mensagens se dá por duas funções: *SendMsg(receiver, PERFORMATIVE, content*) – remetente implícito; e *CheckMsgPool(PERFORMATIVE*), onde o agente checa a própria caixa de mensagens em busca de mensagens com um performativo específico. Funções foram implementadas para preparar o conteúdo da mensagem, recebê-lo e ao identificador da mensagem, como mostrado nos algoritmos 1, 2 e 3 (seção 4.9).

Uma vez que MASSA é um sistema multiagente dinâmico, isto é, os tipos e cômputos de agentes variam entre uma execução e outra dependendo do tipo de anotação requerida, um recurso similar ao de páginas amarelas é utilizado, dessa forma um agente pode procurar um serviço que necessita ao invés de especificar antecipadamente o nome do agente que o disponibiliza. Esta funcionalidade está implementada para que todos os agentes façam registros no serviço de páginas amarelas (*DFAgent* no ambiente JADE) ao se iniciarem, e interroguem este mesmo registro para buscar por outros serviços ou agentes. No ato do registro, os agentes especificam os próprios nomes, o tipo de informação que aceitam (*infotype*) e o tipo de informação que disponibilizam (*servicetype*). Para os agentes DB que encapsulam bases de dados locais ou remotas, o tipo de serviço é o nome do banco de dados. Nesta aplicação específica, o sistema MASSA, *infotype* pode ser "snp" ou "gene", o que quer dizer que o agente pode processar uma lista de SNPs ou de genes, já o *servicetype* pode ser, por exemplo, "*annotation*", "*coordination*", "*omim*" ou "*ucsc*". O registro da função é implementada como *Register(agentName, infotype, servicetype)*, como demonstrado na linha 2 dos algoritmos 1, 2 e 3. Quando um agente necessita achar um agente que disponibiliza um serviço específico, como a busca no banco de dados dbSNP, ou processar um tipo específico de informação, o agente deve interrogar o serviço de páginas amarelas em relação a estes parâmetros. Isto é implementado como a função *SearchAgentID(DFAgent, infotype, servicetype)*, exemplificado nas linhas 9,15 e 35 do Algoritmo 2. Assim, se o Coordenador necessita encontrar todos os agentes que processam a informação do tipo "*snp*" (*infotype=snp*), ele utiliza a função *SearchAgentID(DFAgent, snp, null)*. Isto irá retornar todos os agentes que podem processar uma lista de SNPs, independentemente do tipo de serviço. No sistema MASSA, isto irá retornar, por exemplo, os identificadores (Ids) dos agentes dbSNP, UCSC, GWAS, PolyPhen2 e Provean/SIFT.

4.9 ALGORITMOS DOS AGENTES

4.9.1 AGENTE INTERFACE

O algoritmo 1 descreve como o agente Interface funciona. Ele recebe como entrada um arquivo de configuração (*masconfig.txt*) e um contendo uma lista de SNPs (*snplist.txt*). O primeiro é processado para definir dois parâmetros, a taxa de paralelização (*pR*) e o tipo de

anotação (*anntype*), e o último para obter o conjunto de polimorfismos a ser anotado ($S = \{< snp_1, \ldots, snp_n >\}$). Depois disto, o conjunto de SNPs é dividido em *pR* subconjuntos e armazenado em S_{sub}.

Algorithm 1 Interface Agent.

```
1: input: masconfig.txt,snplist.txt
2: Register("Interface", null, "annotation")
3: (pR, anntype) = Parse(masconfig.txt)
4: S = Read(snplist.txt) {S = {< snp_1, ..., snp_n >}}
5: S_sub = {< S_1, ..., S_n >: S|n = pR ∧ S_n ⊆ S}
6: for all S_i ∈ S_sub do
7:    content = PrepareMsgContent(S_i, anntype)
8:    requestid_i = SendMsg("Coordinator", REQUEST, content)
9:    R = R ∪ {requestid_i}
10: end for
11: A = {∅}
12: while |R| > 0 or Timeout() do
13:    reply = CheckMsgPool(REPLY)
14:    if reply not null then
15:       replyid_i = GetMsgId(reply)
16:       A_i = GetMsgContent(reply) {A_i = {< a_1, ..., a_n > |a_n = (snp, gene, at, value)}}
17:       A = A ∪ A_i
18:       R = R \ {replyid_i}
19:       i++
20:    end if
21: end while
22: WriteAnnotationFile(snplist.annotation, A)
23: WriteSummaryFile(snplist.sum, A)
24: WriteEnrichFile(snplist.enrich, A)
25: WriteLogFile(snplist.log, S, A)
26: ShutDownAgentPlatform()
27: output: snplist.annotation, snplist.log, snplist.sum, snplist.enrich
```

Cada um destes subconjuntos de SNPs ($\{< S_1, \ldots, S_n >\} \in S_{sub}$) são enviados ao agente Coordenador em mensagens separadas. Todas as mensagens contêm um subconjunto de SNPs (S_n) e o tipo de anotação (*anntype*) como conteúdo, a agente Interface mantem um conjunto com todos os Ids das mensagens (|R|), para que ele gerencie as respostas recebidas. O objetivo de dividir o conjunto de SNPs em subconjuntos menores é o de melhorar a velocidade de anotação, uma vez que cada um dos subconjuntos será anotado em paralelo ao longo do processo de anotação (obviamente, partindo do princípio de que o processador disponha de múltiplos núcleos). Além disso, com esta implementação, há a possibilidade de se executar o agente Coordenador e os agentes DB em máquinas diferentes, o que, provavelmente, pode melhorar ainda mais a velocidade de anotação.

Depois de enviar todos as requisições ao agente Coordenador, o agente Interface aguarda as respostas constantemente verificando a sua caixa de mensagens (*CheckMsgPool (REPLY*)) – para implementar esta função foram utilizados os métodos fornecidos pelo ambiente JADE. Se houver uma resposta (*replynotnull*), em seguida, o agente de interface recebe a identificação da mensagem resposta (*replyid*) e lê o conteúdo da mensagem resposta para o conjunto da anotação $A_i = \{< a_1, \dots, a_n > | a_n = (snp, gene, at, value)\}$, de forma que cada elemento do conjunto é uma tupla $a_n = (snp, gene, at, value)$, onde *snp* é o SNP a ser anotado, *gene* é o gene onde está localizado o SNP (caso exista), *at* é o atributo da anotação (tal como cromossomo, posição, etc.), e *value* é o valor do atributo (tal como, 3 para o atributo cromossomo, ou 12345 para o atributo posição). Assim que as mensagens de resposta vão sendo recebidas, os subconjuntos de anotação vão sendo unidos em um único conjunto *A*, e os IDs das respostas vão sendo removidos do conjunto *R* ($R = R/\{replyid_i\}$). Quando todas as mensagens de resposta são recebidas ($|R| = 0$) ou o tempo limite se esgota (*Timeout()*), as saídas são geradas. A função de tempo limite leva em consideração o tempo gasto pela resposta anterior e o multiplica por 5 visando gerar um tempo de espera aceitável. Todas as saídas são produzidas com base nos conjuntos de anotação recebidos (*A*). No algoritmo o processo de geração de outputs é representado como funções genéricas (linhas 22 a 25 do Algoritmo 1) , uma vez que a sua implementação contem principalmente a formatação de saída. Finalmente, depois de gerar todos os relatórios, o agente de interface desliga a plataforma de agentes e o sistema é encerrado.

4.9.2 AGENTE COORDENADOR

A implementação do agente Coordenador é descrita no Algoritmo 2. Uma vez que este agente precisa de instruções do agente Interface para começar a operar, ele inicia-se vasculhando a caixa de mensagens buscando uma mensagem *REQUEST* (linha 4). Após a chegada da requisição do agente Interface contendo uma lista de SNPs (*S*) a ser anotada e o tipo de anotação (*anntype*), o Coordenador verifica se a instrução se refere à anotação simples ou à completa. As implementações das anotações simples e completa são tratadas em partes diferentes do algoritmo, como detalhado abaixo.

A anotação simples requer apenas a anotação de SNPs, e só faz uso do banco de dados dbSNP . Portanto, se este for o caso, o coordenador prepara uma mensagem *REQUEST* ao agente dbSNP contendo a lista recebida de SNPs (*S*) como seu conteúdo (linhas 9 a 11). Note-se que para encontrar o identificador do agente dbSNP, o Coordenador consulta as páginas amarelas por um agente com o tipo de informação "*snp*" e tipo de serviço "*dbSNP*". Assim como o agente Interface, o Coordenador também mantém um conjunto com os IDs de todas as mensagens requisitadas (|*R*|), podendo desta forma controlar as respostas recebidas. Com a chegada da resposta para a anotação simples (linhas 27 a 31 do Algoritmo 2), as respectivas anotações para os subconjuntos de SNPs são armazenadas em $A_j^{snp} = as_1, \dots, as_n$, onde $as_n = (snp, at, value)$, e unidas em um único conjunto A^{snp}. Quando todas as mensagens de resposta para a anotação simples forem recebidas ($|R| = 0$) ou o tempo limite se esgotar (*Timeout()*), uma mensagem é enviada de volta para o agente Interface com o conjunto da anotação completa (*A*) (linhas 48 a 52 do Algoritmo 2) . Importante notar que para a anotação simples $A = A^{snp}$, uma vez que não haverá anotação para genes $A^{gene} = \{\emptyset\}$. Neste ponto, a tarefa do Agente Coordenador na anotação simples é finalizada.

A anotação completa requer a anotação de SNPs e dos genes a eles associados, utilizando todas as bases de dados disponíveis no sistema. Por conseguinte, se este for o caso, o agente coordenador tem de dividir o processo de anotação em duas etapas: primeiro a anotação dos SNPs e, em seguida, a anotação dos genes, uma vez que o gene no qual está localizado um SNP é um dos atributos retornados pela primeira anotação. Para a anotação de SNPs, o agente Coordenador consulta o serviço de páginas amarelas buscando os agentes que processam a informação "*snp*" (*infotype = snp* e *servicetype = null*), e armazena a lista de agentes resultante no conjunto $P = < receiverid_1, \dots, receiverid_n >$ (linha 15). Então, para cada agente presente no conjunto P, o Coordenador envia uma mensagem *REQUEST* contendo a lista original de SNPs a ser anotada (*S*) (linhas 14 a 21 do Algoritmo 2). De acordo com as respostas que vão sendo retornados, os respectivos conjuntos anotados são armazenados em A_k^{snp} (como para a anotação simples). Na anotação completa, entretanto, é necessária a retornar as informações dos genes, dessa forma, passos adicionais foram incluídos como descrito a seguir. Um subconjunto $G_k = \{< g_{1,\dots,} g_n >\}$ do conjunto de SNPs anotados A_k^{snp} é criado apenas com essas anotações em que o atributo é igual ao gene e, consequentemente, o seu valor contem o nome do gene onde o SNP está localizado ($g_n = (snp, at, value) \Lambda\ at = gene$). Isto fornece um mapeamento de SNP para gene. Em seguida,

o agente Coordenador inquire o serviço de páginas amarelas em busca de todos os agentes que processam a informação "*gene*" (*infotype = gene* e *servicetype = null*), e armazena a lista de agentes resultante no conjunto P. Então, para cada agente em P, o Coordenador envia uma mensagem *REQUEST* contendo o conjunto de genes a ser anotado (G_k) (linhas 32 a 42 do Algoritmo 2). Assim que as respostas para a anotação completam começam a chegar, seus respectivos conjuntos de genes anotados são armazenados em A_0^{gene} e unidos em um único conjunto $A^{gene} = \{< ag_1, \dots, ag_n >\}$ onde $ag_n = (snp, gene, at, value)$ (linhas 43 a 47). Quando todas as mensagens de resposta para a anotação completa de SNP e respectivos genes são recebidas ($|R| = 0$) ou o tempo limite é esgotado (*Timeout()*), um conjunto de anotações completo $A = \{< a_1, \dots, a_n >\}$ é criado unindo o conjunto de anotações de SNP ao conjunto de anotações de genes $A^{snp} \cup A^{gene}$, onde $a_n = (snp, gene, at, value)$. Este conjunto A é enviado de volta ao agente Interface através de uma mensagem *REPLY* (linhas 48 a 52 do Algoritmo 2). Neste ponto, a tarefa de anotação completa do Agente Coordenador está completa.

É importante notar que esta implementação do Agente Coordenador é salutar, uma vez que todas as anotações de SNPs são independentes umas das outras, e também as anotações de genes são independentes entre si, permitindo a realização da anotação destes grupos em paralelo e melhorando o desempenho do sistema. Isto é, todas as anotações de SNPs são enviadas aos agentes DB processando a informação "*snp*", ao mesmo tempo, e todas as anotações de genes são enviados para agentes BD que processam a informação "*gene*" simultaneamente. A única dependência entre as tarefas existente é aquela entre a anotação do SNP e a recuperação do atributo gene a ser utilizado no mapeamento de SNP para gene.

Algorithm 2 Coordinator Agent.

input: -
$Register(\text{"}Coordinator\text{"}, \text{"}snp\text{"}, \text{"}coordination\text{"})$
$infotype = snp$
loop
  $request = CheckMsgPool(REQUEST)$
  if $request$ **not** $null$ **then**
    $(S, anntype) = GetMsgContent(request)$
    if $anntype = simple$ **then**
      $receiverid = SearchAgentID(\text{"}DFAgent\text{"}, infotype, dbsnp)$
      $content = PrepareMsgContent(S)$
      $requestid = SendMsg(receiverid, REQUEST, content)$
      $R = R \cup \{requestid\}$
    end if
    if $anntype = complete$ **then**
      $P = SearchAgentID(DFAgent, infotype, null)$ $\{P = < receiverid_1, ..., receiverid_n >\}$
      for all $receiverid_i \in P$ **do**
        $content = PrepareMsgContent(S)$
        $requestid_i = SendMsg(receiverid_i, REQUEST, content)$
        $R = R \cup \{requestid_i\}$
      end for
    end if
    $A^{snp} = \{\emptyset\}, A^{gene} = \{\emptyset\}$
  end if
  $reply = CheckMsgPool(REPLY)$
  if $reply$ **not** $null$ **then**
    $replyid_j = GetMsgId(reply)$
    if $infotype = snp$ **and** $anntype = simple$ **then**
      $A_j^{snp} = GetMsgContent(reply)$ $\{A_j^{snp} = \{< as_1, ..., as_n > | as_n = (snp, at, value)\}\}$
      $A^{snp} = A^{snp} \cup A_i^{snp}$
      $R = R \setminus \{replyid_j\}$
    end if
    if $infotype = snp$ **and** $anntype = complete$ **then**
      $A_k^{snp} = GetMsgContent(reply)$
      $infotype = gene$
      $G_k = \{< g_1, ..., g_n >: A_k^{snp} | g_n = (snp, at, value) \wedge at = gene\}$
      $P = SearchAgentID(DFAgent, infotype, null)$
      for all $receiverid_l \in P$ **do**
        $content = PrepareMsgContent(G_k)$
        $requestid_l = SendMsg(receiverid_l, REQUEST, content)$
        $R = R \cup \{requestid_l\}$
      end for
    end if
    if $infotype = gene$ **and** $anntype = complete$ **then**
      $A_o^{gene} = GetMsgContent(reply)$ $\{A_o^{gene} = \{< ag_1, ..., ag_n > | ag_n = (snp, gene, at, value)\}\}$
      $A^{gene} = A^{gene} \cup A_o^{gene}$
      $R = R \setminus \{replyid_j\}$
    end if
    if $|R| = 0$ **or** $Timeout()$ **then**
      $A = \{< a_1, ..., a_n >: A^{snp} \cup A^{gene} | a_n = (snp, gene, at, value) \forall as_p = (snp_p, at, value) \in A^{snp} \wedge \forall ag_q = (snp_q, gene, at, value) \in A^{gene}, snp_p = snp_q\}$
      $content = PrepareMsgContent(A)$
      $SendMsg(Interface, REPLY, content)$
    end if
  end if
end loop
output: A

4.9.3 AGENTE BANCO DE DADOS

A implementação dos agentes Banco de dados (agentes DB) é descrita em termos gerais no Algoritmo 3. As informações sobre o acesso a cada base de dados são explicitadas na seção 4.6 e no Apêndice A, entretanto, devido às limitações de espaço, os algoritmos de conexão serão omitidos neste trabalho, uma vez que cada banco conta com sua forma articular de acesso (API, WS, SQL, etc). Ao invés disso, o acesso específico a uma base de dados será representado pela função *SearchDatabase*($< snp|gene >, attribute$), onde um banco de dados é consultado sobre um "*snp*" ou o nome de "*gene*" em relação a algum atributo, retornando o valor armazenado para aquele atributo. Os agentes DB se distinguem no algoritmo por 3 parâmetros: i) o tipo de informação que processam (*infotype* que pode ser "*snp*" ou "*gene*"); ii) o tipo de serviço que disponibilizam (o parâmetro *servicetype*, que é o nome do banco de dados que encapsulam); e iii) a lista de atributos que recuperam do banco de dados $T = < at_1, \dots, at_n >$.

Agentes BD precisam de instruções do Agente Coordenador para começar a funcionar. Assim, eles iniciam a execução verificando a caixa de mensagens em busca de um *REQUEST* (linha 4). Quando uma resposta chega, o agente DB lê a lista de SNPs ou genes do conteúdo da mensagem e a armazena no conjunto *S*, para SNPs, ou G, para os genes. Logo após, o agente DB recupera as informações de todos os atributos no conjunto *T*, para todos os SNPs ou genes, e as armazena nos conjuntos A^{snp} ou A^{gene}.

No final, tem-se $A^{snp} = \{< as_{11,\dots,} as_{ij} > | as_{ij} = (snp_i, at_j, value)\}$ ou $A^{gene} = \{< ag_{11,\dots,} ag_{ij} > | ag_{ij} = (snp_i, gene_i, at_j, value)\}$. O agente DB completa a tarefa enviando os conjuntos anotadas de SNPs ou genes de volta ao agente Coordenador.

Algorithm 3 Database Agents.

1: input: $infotype, servicetype, T = < at_1, ..., at_n >$
2: $Register(DBAgentName, infotype, servicetype)$
3: **loop**
4: $request = CheckMsgPool(REQUEST)$
5: **if** $request$ **not** $null$ **then**
6: $senderid = GetMsgSenderId(request)$
7: **if** $infotype = snp$ **then**
8: $S = GetMsgContent(request)$
9: $A^{snp} = \{\emptyset\}$
10: **for all** $snp_i \in S$ **do**
11: **for all** $at_j \in T$ **do**
12: $value = SearchDatabase(snp_i, at_j)$
13: $A^{snp} = A^{snp} \cup \{as_{ij} = (snp_i, at_j, value)\}$
14: **end for**
15: **end for**
16: $content = PrepareMsgContent(A^{snp})$
17: **end if**
18: **if** $infotype = gene$ **then**
19: $G = GetMsgContent(request)$
20: $A^{gene} = \{\emptyset\}$
21: **for all** $(snp, gene)_i \in G$ **do**
22: **for all** $at_j \in T$ **do**
23: $value = SearchDatabase(gene_i, at_j)$
24: $A^{gene} = A^{gene} \cup \{ag_{ij} = (snp_i, gene_i, at_j, value)\}$
25: **end for**
26: **end for**
27: $content = PrepareMsgContent(A^{gene})$
28: **end if**
29: $SendMsg(senderid, REPLY, content)$
30: **end if**
31: **end loop**
32: output: $A^{snp}|A^{gene}$

4.10 ESTUDO EXPERIMENTAL

O sistema de anotação MASSA encontra-se em funcional. Atualmente, o sistema dispõe de 24 agentes, sendo 22 agentes DB de acesso às bases de dados, além do agente Interface e do agente Coordenador, todos plenamente funcionais. Para testar o desempenho do sistema foram realizados dois experimentos: o primeiro objetiva determinar o número ótimo de paralelizações do sistema (pR), e o segundo objetiva testar a performance do MASSA em relação a um sistema similar, o ANNOVAR.

Como descrito anteriormente, o MASSA utiliza a estratégia de divisão dos dados de entrada em N subconjuntos menores, os quais são anotados em paralelo. Para encontrar a taxa ótima de paralelizações para anotação de grandes conjuntos de dados foi realizado um experimento para analisar o comportamento do sistema quanto ao tempo de anotação, em relação à taxa de paralelização no sistema. Para isso, utilizamos os conjuntos de 2.5 milhões e 5 milhões de SNPs dos arranjos de genotipagem da Illumina (seção 1.4.1) como dados de entrada. Os resultados estão apresentados na Figura 11, e mostram que, para conjuntos grandes de SNPs, o número ótimo de threads está em torno de 26 para a anotação simples.

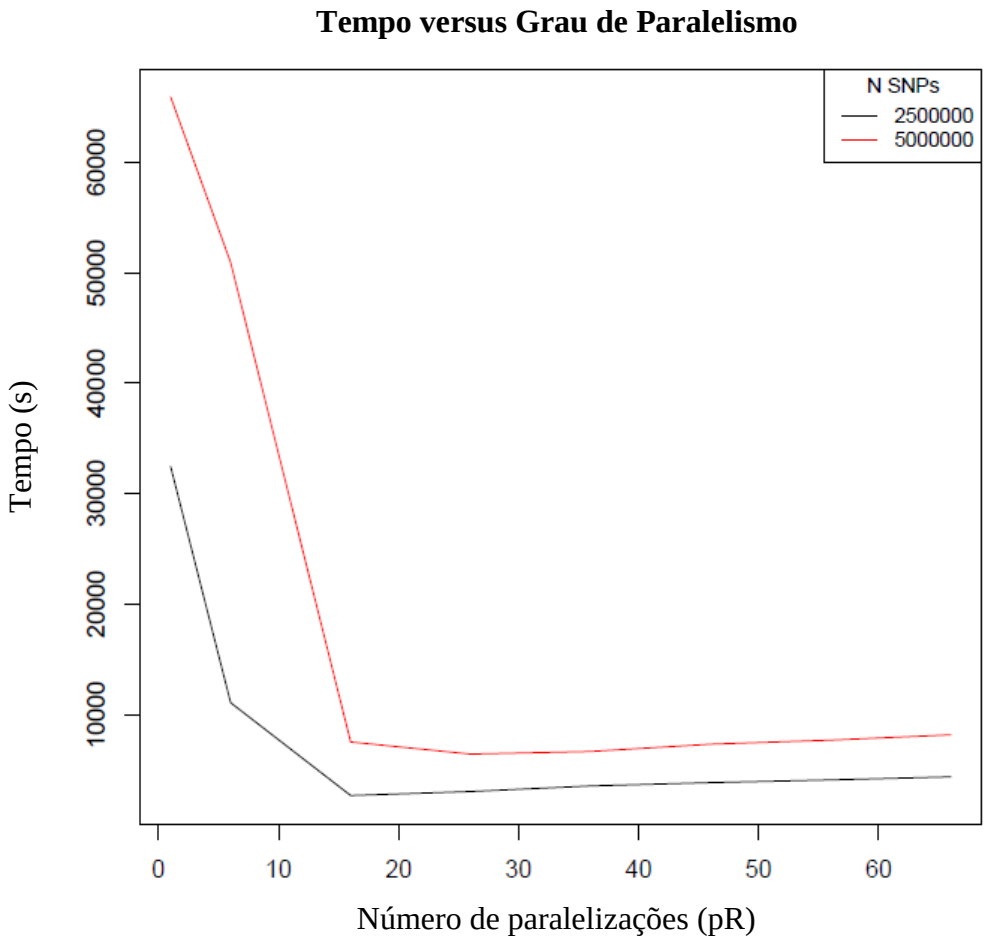


Figura 11: Relação entre o número de paralelizações e o tempo de anotação.

Com o número ótimo de paralelizações determinado, a performance do sistema foi testada em relação ao tempo de anotação para conjuntos de SNPs com tamanhos variados (de 100.000 à 5.000.000 SNPs) em comparação a um sistema similar, o ANNOVAR (WANG; LI; HAKONARSON, 2010). Os resultados mostrados na Figura 12 sugerem que os tempos de

anotação na versão de anotação resumida do MASSA, principalmente para os grandes conjuntos de variantes, é bastante satisfatório, menos de 1h para o conjunto de 2.5M polimorfismos e até 2.5 horas para o conjunto de 5M. Em comparação com a ferramenta ANNOVAR, nota-se que a eficiência de MASSA, em relação ao tempo de anotação, é superior para conjuntos de dados de até 2 M. Entretanto, como o tempo de anotação de MASSA mostrou-se proporcional ao número de polimorfismos anotados, a velocidade desta foi superada pela ANNOVAR nos maiores conjuntos de dados.

Esses experimentos foram realizados em um servidor de 32 núcleos. Para uma visão mais completa do comportamento do sistema, pretendemos repetir os testes em um computador com menor número de núcleos e verificar a variação do número ótimo de threads. Além disso, otimizações na recuperação e escrita das anotações estão sendo implementadas e espera-se que os tempos diminuam sensivelmente.

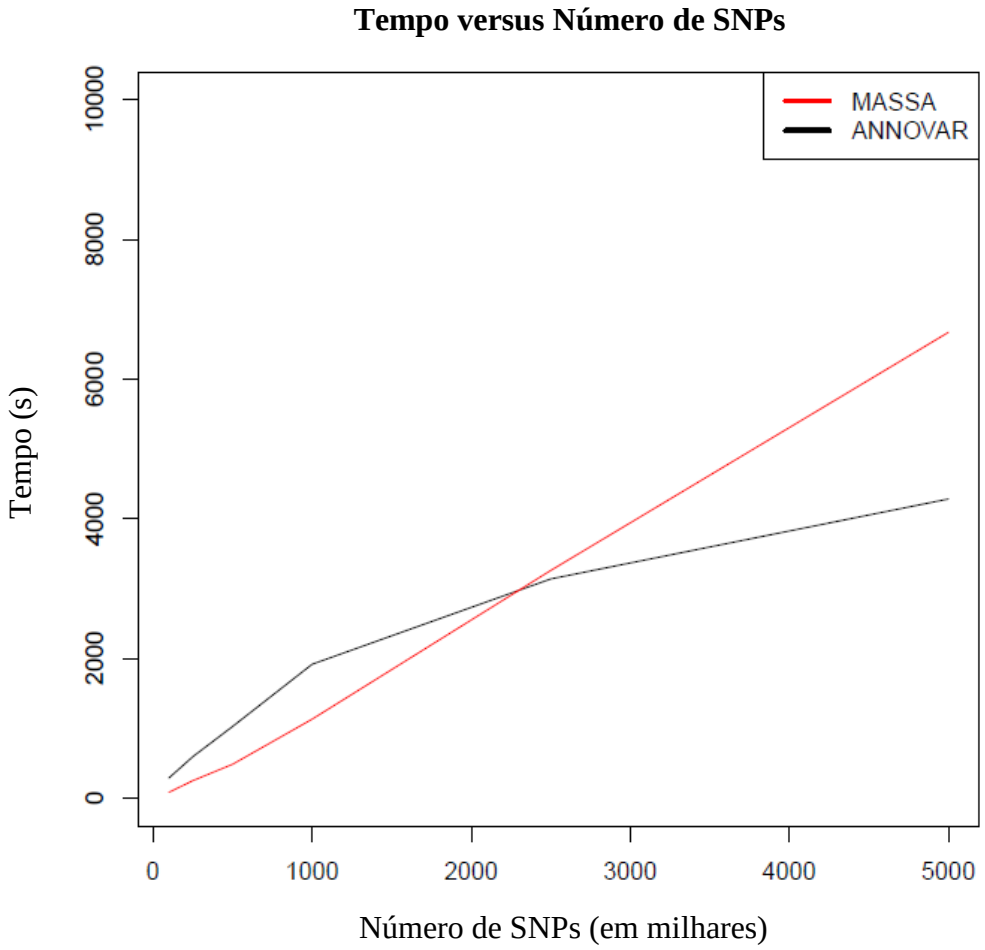


Figura 12: Tempo de anotação em função do número de SNPs submetidos.

Também comparamos MASSA a 14 ferramentas de anotação (Apêndice B) e 14 ferramentas de enriquecimento (Apêndice C). As anotações foram classificadas em 17 grupos baseados no mapeamento ontológico descrito no Apêndice D. Estes grupos envolvem, por exemplo, anotações relacionadas a genes, proteínas, fenótipos e polimorfismos. Na comparação com as ferramentas de anotação, MASSA é a segunda com mais anotações (66), sendo estas distribuídas por 9 grupos de classificação. O WS SNP Nexus (CHELALA; KHAN; LEMOINE, 2009; DAYEM ULLAH; LEMOINE; CHELALA, 2012, 2013) com 194 anotações é a ferramenta mais compreensiva tanto em número de anotações, quanto na distribuição das anotações (11 grupos). Merecem destaque as ferramentas SNP-Info-WebServer (55 anotações) (XU; TAYLOR, 2009), PLINK (47) (PURCELL et al., 2007) SNPdbe (29) (SCHAEFER et al., 2012) e Annovar (25). Entretanto, é importante ressaltar que o número de anotações em algumas ferramentas está enviesado, como no caso de SNP Nexus, devido à repetição das anotações para bancos de dados diferentes, mas tal fato não é necessariamente ruim já que garante maior liberdade e confiança às anotações. Outras ferramentas testadas tem números de anotações inferiores à 20, tais como GLU, CandiSNPer (SCHMITT et al., 2010), MutaGeneSys (STOYANOVICH; PE'ER, 2008), PolySearch (CHENG et al., 2008), SNAP (JOHNSON et al., 2008), SCAN (GAMAZON et al., 2010), MedRefSNP (RHEE; LEE, 2009), Varietas (PAANANEN; CISZEK; WONG, 2010) e WGAViewer (GE et al., 2008). Na comparação com as demais ferramentas, MASSA apresenta duas limitações, baixa flexibilidade na escolha das anotações a serem retornadas e a ausência de anotações baseadas em populações.

Na comparação com as ferramentas de enriquecimento (Apêndice B), MASSA ocupa a quarta posição no número de atributos enriquecidos (10 atributos) e na distribuição dos atributos de acordo com os grupos de anotação (distribui-se por cinco grupos). Foram analisadas cinco ferramentas SEA, quatro compostas (duas SEA/GSEA e duas SEA/MEA), três GSEA, uma MEA, uma PSEA e uma SSEA. Dentre as ferramentas exclusivamente da categoria SEA, MASSA é a mais abrangente. As ferramentas com maior número de atributos enriquecidos são DAVID com 63 atributos distribuídos por sete grupos, ToppGene (17 atributos – 9 grupos) e GSEA (16 – 6). As demais ferramentas: Gowinda (KOFLER; SCHLÖTTERER, 2012), GWAS PathwayIdentifier , i-GSEA4GWAS (ZHANG et al., 2010), SNP-PRAGE (LEE et al., 2011), clusterProfile (YU et al., 2012), MBRole (CHAGOYEN; PAZOS, 2011), GeneCodis3 (NOGALES-CADENAS et al., 2009; TABAS-MADRID; NOGALES-CADENAS; PASCUAL-MONTANO, 2012), GOrilla (EDEN et al., 2009),

Ontologizer 2.0 (BAUER et al., 2008) e GOstat (BEISSBARTH; SPEED, 2004) realizam testes de enriquecimento para menos de 10 atributos. Na confrontação com outras ferramentas, as principais restrições de MASSA são: inflexibilidade na escolha dos atributos para enriquecimento, ausência de uma interface amigável ao usuário, especialmente, para apresentação e visualização dos dados e necessidade de implementar algoritmos de descoberta de redes, clusterização e relacionamento termo-a-termo.

4.11 ESTUDO DE CASO

Nesta seção será descrito um exemplo de uso da ferramenta MASSA. Selecionaram-se como alvo para a anotação os resultados obtidos a partir da análise de miscigenação efetuada pela equipe de ancestralidade do projeto EPIGEN-Brasil (Figura 13). A metodologia utilizada para descrição da estruturação populacional pode ser vista na seção 1.4. O objetivo primário desta anotação é descrever os atributos biológicos relacionados a um dos clusters de ancestralidade. Para isso, seguiram-se os seguintes passos: i) identificação dos SNPs diferenciados nos clusters populacionais; ii) geração de listas de SNPs diferenciados em cada cluster, baseado no resultado i); e iii) execução do MASSA para anotação e enriquecimento dos SNPs de cada cluster.

Para identificação dos SNPs diferenciados, após a inferência dos clusters de ancestralidade, os polimorfismos com maior diferença nas frequências alélicas entre um dado grupo e os demais foram selecionados como AIMs. O delta das frequências alélicas foi calculado da seguinte maneira:

$$abs\big(Cluster1\ fa - mean(Cluster2 \dots 7fa)\big) = \delta > 0.3$$

Ou seja, foram considerados alelos divergentes aqueles em que a diferença absoluta de frequência do alelo referência de um dado cluster em relação à média de frequência dos demais divergia em mais de 0.3.

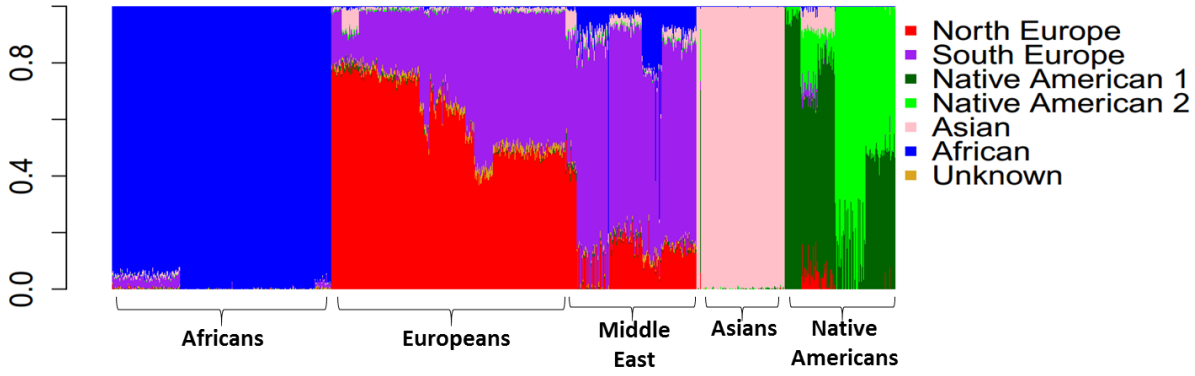


Figura 13: Cluster de ancestralidade para os dados do projeto EPIGEN-Brasil.

O número de SNPs selecionados a partir dos valores de delta e a ancestralidade inferida para cada cluster estão discriminados na Tabela 30. Cada conjunto de SNPs, associado a um cluster, foi anotado com MASSA e o número de genes que apresentam ao menos um polimorfismo entre os mais diferenciados é descrito.

Tabela 30: Número de SNPs selecionados por Cluster

Cluster	Ancestralidade	Número de SNPs	Número de genes
Cluster1	Europa Meridional	7.409	1872
Cluster2	Nativo-americanos1	27.821	4449
Cluster3	Nativo-americanos2	15.748	3159
Cluster4	Leste asiático	10.655	2369
Cluster5	África Ocidental	7.473	1993
Cluster6	África Oriental	6.321	1757
Cluster7	Europa Setentrional	6.538	1688

O cluster 3, referente à ancestralidade Nativo-Americana, foi selecionado para exemplificar as principais funções e resultados da plataforma de anotação. O sistema MASSA gera quatro arquivos: um arquivo de log, o arquivo de anotação, o arquivo sumário e o relatório de enriquecimento. Cada um destes será descrito e exemplificado nos parágrafos seguintes.

A Figura 14 representa um exemplo resumido do arquivo de log da anotação. Neste documento são indicados a data e o horário do início da anotação, o nome do arquivo anotado, a quantidade de SNPs submetidos e devidamente anotados e tempo de execução do sistema. A

segunda parte do arquivo de log – *Data Provenance* – indica as fontes e as respectivas versões dos bancos de dados consultados para cada atributo retornado pela anotação. De acordo com o arquivo de log da Figura 1, os 15.748 SNPs do cluster 3 foram anotados em 1,6 minutos.

```
LOG FILE FOR MAS ANNOTATION SYSTEM
----------------------------------
Date and time: 2/12/2013 at 8:53:95
Data file: Admix.C3.snps.out
SNPs in file: 15748
SNPs annotated: 15748
Execution time: 99 seconds

DATA PROVENANCE
---------------
Attribute    Source      Version
Chromosome  dbsnp 137_04-10-2013
LocusType   hgnc  04-09-2013
```

Figura 14: Exemplo de arquivo de log

PolymorphismId(1)	PolymorphismType(2)	GeneSymbol(3)	GeneId(4)	TranscriptRegion(5)	NucleotideNumberingCodingDNA(6)	Chromosome(7)	ChromosomePosition(8)	AncestralAllele(9)	Orientation(10)
rs3811647	null	TF	7018	intron	c.1330+278	3	1.33E+08	G,G,G,G,G	null
AssemblyBuildVersion(11)	AssemblyCoordStart(12)	AssemblyCoordEnd(13)	mRNAaccession(14)	mRNAversion(15)	Alleles(16)	Frequency(17)	Strand(18)	RefUCSC(19)	ObservedUCSC(20)
37_3	4E+08	4E+08	NM_001063	3	G,A	0.660161,0.339839	+	G	A/G
PolymorphismClass(21)	FunctionalClass(22)	ReactomePathways(23)	PGKBPathways(24)	PGKBDrugs(25)	PGKBDisease(26)	RelatedGenes(27)	GOMolecularFunction(28)	GOCellularComponent(29)	GOBiologicalProcess(30)
single	intron	Transmembrane transport of small molecules;Platelet degranulation;Response to elevated platelet cytosolic Ca2+;Iron uptake and transport;Transferrin endocytosis and recycling;Hemostasis;Platelet activation, signaling and aggregation	Regulation of Androgen receptor activity;extrinsic prothrombin activation pathway;mechanism of gene regulation by peroxisome proliferators via ppara;Further platelet releasate - ;Extrinsic Pathway - ;EPHB forward signaling;HIF-1-alpha transcription factor network	null	null	null	ubiquitin protein ligase binding;ferric iron binding;protein binding	extracellular region;mitochondrion;secretory granule;endosome membrane;endocytic vesicle;basal part of cell;basal plasma membrane; t	cellular iron ion homeostasis;iron ion transport;transmembrane transport;transferrin transport;platelet activation;blood coagulation;platelet degranulation

Cytoloc(31)	GeneStatus(32)	GeneMapMethods(33)	Disorders(34)	MIMids(35)	Inheritance(36)	PhenoMapMethods(37)	Comments(38)	Hgncld(39)	GeneSymbol(40)
3q22.1	null	3	Atransferrinemia	209300	null	null	null	HGNC:11740	TF
GeneName(41)	GeneSynonyms(42)	LocusType(43)	LocusGroup(44)	GeneFamilyTag(45)	GeneFamily(46)	Pubmed(47)	Reported_Genes(48)	Strongest_SNP_Risk_Allele(49)	Context(50)
transferrin	PRO1557, PRO2086	gene with protein product	protein-coding gene	null	null	21785125;21665994;21483845;21208937;19084217;	TF;	rs3811647-?;rs3811647-A;	intron;

P-Value(51)	Disease_Trait(52)	Sample_and_Population(53)	PolyphenProteinID(54)	PolyphenSubstitution(55)	Polyphen2Prediction(56)	PolyPhen2Prob(57)	Polyphen2FDR(58)	Polyphen1Prediction(59)	ProveanProteinID(60)
2E-16;1E-35;5E-10;8E-6;3E-47;3	Hepcidin levels;Alcohol consumption;Iron status biomarkers;Iron levels;	1,657 individuals;5,181 European ancestry individuals;336 European ancestry iron deficiency cases, 343 European ancestry controls;Up to 5,633 Caucasian individuals;459 twin pairs;	null	null	null	null	null	null	null

ProveanSubstitution(61)	ProveanProteinPos(62)	ProveanPrediction(63)	ProveanScore(64)	SIFTPrediction(65)	SIFTScore(66)
null	null	null	null	null	Null

Figura 15: Exemplo de arquivo de anotação

O arquivo de anotação apresenta ao usuário 66 informações para cada um dos polimorfismos submetidos à anotação, entretanto, estas informações são dependentes da disponibilidade da informação nas bases de dados consultadas. Sempre que uma informação não puder ser obtida pelo MASSA o campo do atributo é preenchido com o valor *null*. Uma amostra do relatório de anotação pode ser visualizada na amostra da Figura 15. O resultado da anotação mostra que o polimorfismo rs3811647, por exemplo, está localizado em uma região intrônica do gene *TF*, no cromossomo 3. Este gene está associado à homeostase de íons de ferro, e um estudo de GWAS descreveu a associação entre o consumo de álcool e o polimorfismo na população europeia.

A anotação está disponível para o usuário em um intuitivo arquivo TSV, onde o cabeçalho indica não só o nome de cada atributo, mas também a coluna onde este se encontra, além disso, naqueles atributos onde mais de um valor pode ser retornando, os diferentes valores são separados por ponto e vírgula(“;”). Estas características tornam o processamento do arquivo de anotação por softwares externos bastante prática e pouco propensa a erros. Isto é essencial devido ao fato de que os relatórios de anotação podem ter alguns gigabytes de tamanho dependendo do tamanho da submissão.

O sumário indica ao usuário do sistema MASSA as contagens dos valores de 15 campos específicos, sendo eles: Gene (dbSNP); Região do transcrito (dbSNP e UCSC); Vias Metabólicas (Reactome e PharmGKB); fármacos e xenobióticos (PharmGKB); doenças e fenótipos (OMIM e PharmGKB); Famílias Gênicas (HGNC); Tipo/Função do gene (HGNC); Localização citogenética (OMIM); Processo Biológico (GO); Função Molecular (GO); e Componente Celular (GO). Para cada um destes atributos é indicado o número de polimorfismos e genes associados. Uma amostra do arquivo pode ser visualizada na Tabela 31. Neste exemplo, pode-se visualizar os maiores cômputos dos termos associados a 5 atributos: famílias gênicas (banco HGNC), fenótipo (OMIM), processo biológico (GO), função molecular (GO) e componente celular (GO).

Tabela 31: Exemplo de arquivo sumário

Database Attribute	*Term*	*SNP Count*	*Gene Count*
HGNC Gene Family	SLC	169	81
	ZNF	82	36
	EFHAND	84	33
OMIM Disorder	*Colorectal cancer*	13	5
	Leukemia, acute myeloid	4	3
	Usher syndrome, type 1D/F	11	2
GO Biological Process	*transcription, DNA-dependent*	513	231
	small molecule metabolic process	377	214
	signal transduction	536	205
GO Molecular Function	*protein binding*	1931	801
	zinc ion binding	618	294
	ATP binding	599	265
GO Cellular Component	*cytoplasm*	1733	686
	integral to membrane	1766	663
	nucleus	1521	662

A última função da plataforma MASSA é o enriquecimento de termos, onde é avaliada a sobrerrepresentação dos termos retornados pela anotação em relação a frequências destes nas bases de dados. O enriquecimento é realizado, atualmente, para os seguintes atributos: Vias metabólicas (Reactome e PharmGKB); fenótipos (OMIM e PhamrGKB); xenobióticos (PharmGKB); Localização citogenética (OMIM); famílias gênicas (HGNC); processo biológico (GO); função molecular (GO) e componente celular (GO). É interessante notar que entre os termos enriquecidos encontram-se alguns previamente descritos como sob seleção natural em humanos, tais como, resposta imune (QUINTANA-MURCI; CLARK, 2013), entre os processos biológicos; atividade dos receptores olfatórios em vias metabólicas (GILAD et al., 2003); e pertencentes à família gênica SLC (PICKRELL et al., 2009), envolvida no transporte de íons (Tabela 32).

Tabela 32: Exemplo do relatório de enriquecimento

Database Attribute	*Term*	*p-Value*	*Gene Population Count*	*SNP Count*	*Gene Sample Count*	*Gene List*
HGNC Gene Family	*LNCRNA, ANTISENSE*	1.92E-13	749	24	13	[ADAMTS9-AS2,..., NPPA-AS1]
HGNC Gene Family	*SLC*	1.07E-09	440	169	81	[SLC1A7,..., SLC44A5]
HGNC Gene Family	*ISET, IGD, FN3*	8.46E-09	23	73	16	[IGSF9,..., CNTN4]
Reactome Pathway	*Disease*	3.13E-21	1041	71	32	[ABCA4,..., NCF2]

Reactome Pathway	*Olfactory Signaling Pathway*	1.32E-12	423	9	8	[OR2G2,…, OR9Q1]
Reactome Pathway	*Mitotic M-M/G1 phases*	1.47E-08	258	8	4	[AHCTF1,…, NSL1]
GO Biological Process	*regulation of transcription, DNA-dependent*	4.72E-12	682	5	5	[SARS, SSR1, EIF3H, IGF2BP3, MRPS33]
GO Biological Process	*immune response*	4.03E-10	2159	142	86	[NR5A2,…, ZFP112, ZNF584]
GO Biological Process	*axon guidance*	2.37E-09	900	27	24	[NTRK1, TNR,…, RELN, GLI3]
GO Molecular Function	*protein binding*	2.86E-19	11040	1931	801	[CASP9, C1QB,…, MAST2]
GO Molecular Function	*nucleotide binding*	1.75E-14	1363	179	53	[QDPR, NAV1,…, NOX4]
GO Molecular Function	*nucleic acid binding*	2.12E-14	2095	208	103	[PRDM16,…, CHD5, XRN1]
GO Cellular Component	*Nucleolus*	9.59E-12	2771	448	197	[CASZ1,…, KAZN]
GO Cellular Component	*Nucleus*	8.17E-11	9313	660	408	[CASP9,…, PPT1]
GO Cellular Component	*Mitochondrion*	1.09E-10	2191	312	147	[H6PD,..., TF, SCP2]

A partir da anotação e da análise de enriquecimento de três conjuntos de dados contendo SNPs diferenciados em populações Nativo-americanas – NAT (descrito no capítulo 3) – C2 e C3 – clusters obtidos através da diferença de frequências alélicas nesta seção – foi possível comparar os resultados obtidos e calcular o grau de sobreposição dos resultados.

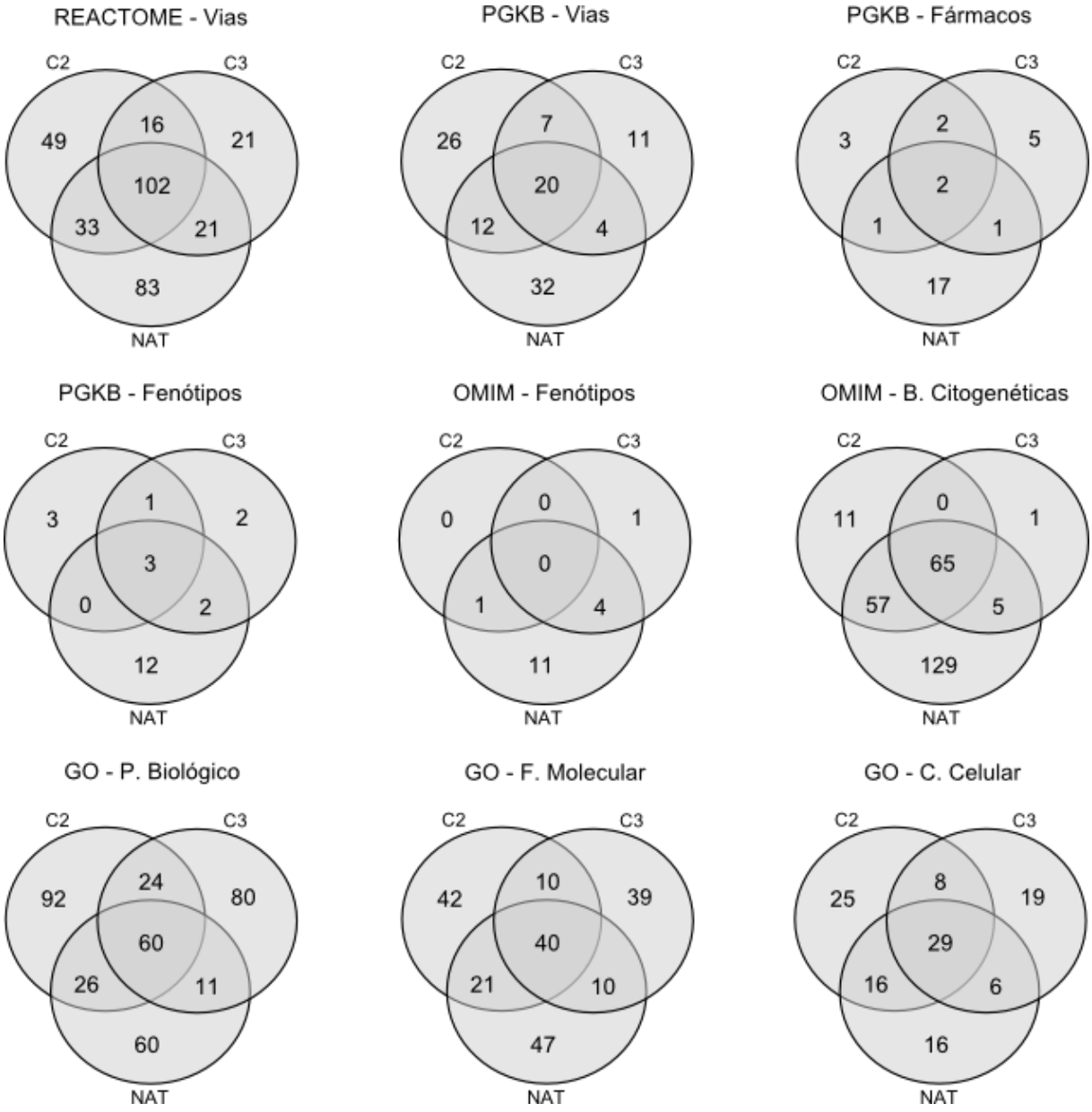


Figura 16: Interseção entre os resultados das análises de enriquecimento para genes divergentes em Nativo-Americanos. NAT – Conjunto de SNPs diferenciados obtido através da AMOVA (ver Capítulo 3); C2 e C3 – Conjuntos de SNPs diferenciados obtidos através da diferença de frequências alélicas.

A Figura 16 indica as interseções entre os resultados das análises de enriquecimento para nove atributos. Foram incluídos todos os termos com valores de significância inferiores à 5%. É importante ressaltar que a sobreposição entre os grupos não foi absoluta devido à: i) utilização de painéis de marcadores diferentes entre os conjuntos NAT e C2-C3; ii) métodos diferentes de seleção de marcadores divergentes; e iii) estruturação observada nos Nativo-Americanos, resultando em dois agrupamentos (C2 e C3). No primeiro caso i) as diferenças podem surgir devido à amostragem diferencial dos painéis de marcadores, assim, em alguns

arranjos de genotipagem, certas áreas do genoma podem estar pouco representadas. As divergências nos resultados podem ter sido causadas pelo uso de estratégias diferentes para a seleção de SNPs em NAT em relação à C2 e C3, caso ii). Na primeira estratégia, apenas os polimorfismos na cauda da distribuição são selecionados, deixando de fora outros SNPs potencialmente divergentes, mas com estimativas menores de distância genética. Entretanto, para os conjuntos C2 e C3, a seleção de polimorfismos ocorre para todos àqueles que possuam diferenças nas frequências alélicas superiores à 0,3. Porém, este método é mais suscetível a selecionar variantes neutras, sob deriva genética. Em iii) a restrição ocorre devido à estruturação dentro do grupo Nativo-Americano, isto faz com que a sobreposição não seja possível para as variantes com diferenças nas frequências alélicas entre os dois grupos e, também, em relação ao grupo NAT. A comparação dos cinco termos com valores mais significativos de enriquecimento é descrita na Tabela 33. Nota-se que a há redundância nas informações, não havendo maior repetição dos termos em nenhuma combinação dois-a-dois (NAT-C2, NAT-C3 e C2-C3).

Tabela 33: Comparação entre os resultados das análises de enriquecimento para três conjuntos de SNPs diferenciados em Nativo-Americanos

Atributo	Nativos – NAT	Nativos – C2	Nativos – C3
HGNC – Famílias Gênicas	LNCRNA, ANTISENSE	LNCRNA, ANTISENSE	LNCRNA, ANTISENSE
	EFHAND	COLLAGEN	SLC
	LNCRNA, INTRONIC	SLC	ISET, IGD, FN3
	RPL	RPL	ISET
	SLC	ISET	PLEKH
Reactome – Vias Metabólicas	*Disease*	*Disease*	*Disease*
	Olfactory Signaling Pathway	*Metabolism of RNA*	*Olfactory Signaling Pathway*
	Signaling by the B Cell Receptor (BCR)	*Extracellular matrix organization*	*Mitotic M-M/G1 phases*
	Downstream signaling of activated FGFR	*Signaling by the B Cell Receptor (BCR)*	*Transmission across Chemical Synapses*
	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	*Transmission across Chemical Synapses*	*Signal Transduction*
PGKB – Fármacos	*Lithium*	*Antiinflammatoryagents*	*lithium*
	Antiinflammatoryagents	*lithium*	*Antiinflammatoryagents*
	Celecoxib	*hydrochlorothiazide*	*celecoxib*
	Brivanib	*dasatinib*	*Bisphosphonates*
	Tipifarnib	*Alpha-adrenoreceptorantagonists*	*gemcitabine*
OMIM – Fenótipos (Doenças)	*Gastric cancer, somatic*	*Gastric cancer, somatic*	*Myocardial infarction, susceptibility to*
	Breast cancer, susceptibility to	*Colorectal cancer, somatic*	*Alport syndrome, autosomal recessive*
	Breast cancer, somatic	*Myasthenic syndrome, slow-channel congenital*	*Tetralogy of Fallot*

	Pheochromocytoma	*Alzheimer disease, susceptibility to*	*Arrhythmogenic right ventricular dysplasia 11 with mild palmoplantar keratoderma and woolly hair*
	Myasthenic syndrome, slow-channel congenital	*Fraser syndrome*	*Cardiomyopathy, dilated, 1JJ*
OMIM – Bandas Citogenéticas	*19p13.3*	*16p13.3*	*16p13.3*
	17q12	*19p13.3*	*19p13.3*
	1q21.3	*17q12*	*19p13.2*
	19p13.2	*1q21.3*	*1q21.3*
	16p13.3	*3p21.31*	*3p21.31*
GO – Processos Biológicos	*Translation*	*translation*	*proteolysis*
	Proteolysis	*regulation of transcription, DNA-dependent*	*regulation of transcription, DNA-dependent*
	regulation of transcription, DNA-dependent	*proteolysis*	*axon guidance*
	antigen processing and presentation	*immune response*	*translation*
	protein folding	*synaptic transmission*	*protein folding*
GO – Função Molecular	*protein binding*	*nucleic acid binding*	*protein binding*
	nucleic acid binding	*structural constituent of ribosome*	*nucleotide binding*
	Zinc ion binding	*nucleotide binding*	*nucleic acid binding*
	Nucleotide binding	*metal ion binding*	*structural constituent of ribosome*
	olfactory receptor activity	*zinc ion binding*	*metal ion binding*
GO – Componente Celular	*Mitochondrion*	*nucleus*	*nucleolus*
	Ribosome	*MHC class I protein complex*	*nucleus*
	Intracellular	*nucleolus*	*mitochondrion*
	MHC class II protein complex	*ribosome*	*cell junction*
	MHC class I protein complex	*integral to plasma membrane*	*ribosome*

NAT – Conjunto de SNPs diferenciados obtido através da AMOVA (ver Capítulo 3); C2 e C3 – Conjuntos de SNPs diferenciados obtidos através da diferença de frequências alélicas.

CONSIDERAÇÕES FINAIS

Durante séculos a ciência analítica assentou-se sobre a especialização e o reducionismo como forma de ratificar ou refutar hipóteses. Entretanto, nos últimos anos, alguns paradigmas das ciências biológicas têm sido alterados, com esta disciplina passando de um campo descritivo a um orientado à análise de dados. E o surgimento e aumento da importância da bioinformática estão intrinsicamente ligados aos desafios propostos pela quantidade e heterogeneidade dos dados biológicos. Esta revolução requer a aplicação de conceitos e metodologias que contribuam para a elucidação dos mecanismos genéticos responsáveis pela variação fenotípica, seja ela adaptativa ou deletéria. Entretanto, esta tarefa permanece árdua e complexa. A partir deste contexto, pode-se situar o escopo do presente trabalho como uma contribuição à extração de conhecimento relevante da variabilidade genética de populações miscigenadas e autóctones latino-americanas. Para tal, quatro desafios foram abordados: i) a otimização do processo analítico; ii) descrição da variabilidade populacional de populações americanas; iii) identificação de variantes divergentes na população nativo-americana; iv) aplicação de novos métodos computacionais para integrar e sumarizar os dados biológicos.

A otimização dos processos analíticos corresponde à formalização de fluxogramas de análises e à construção de ferramentas e *pipelines* de suporte às análises. Os fluxogramas e *pipelines* das seções 1.1 a 1.3 foram utilizados com sucesso nos artigos de diversidade genética do capítulo 2 e nas análises de estruturação do capítulo 3, além de outros estudos do LDGH. Entretanto, a grande quantidade de dados disponibilizada pelo projeto EPIGEN-Brasil levou à necessidade de se ampliar o número de softwares, em especial, aqueles relacionados às estimativas de ancestralidade (seção 1.4). Com a inclusão de novos programas e análises aos fluxogramas, novas ferramentas de conversão têm sido criadas e integradas ao *pipeline* de conversão de dados. Espera-se que a extensão e a aplicação de metodologias de computação intensiva para lidar com dados de larga escala permitam criar uma terceira versão do *Divergenome tools* ainda mais eficiente e compreensiva. Apesar do esforço contínuo demandado para manter atualizado este conjunto metodológico, os ganhos são evidentes por diminuírem o tempo de execução das análises e aumentarem a confiabilidade e reprodutibilidade dos dados.

O estudo da diversidade genética das populações latino-americanas é essencial para a epidemiologia genética e reconstrução da história demográfica destas populações. Ainda que estudos genético-populacionais enfoquem mais na descrição da variabilidade, importantes visões foram obtidas nos estudos do capítulo 2. Quanto à variabilidade de populações autóctones, o estudo da diversidade da população Quechua não apenas reforça o modelo de história populacional proposto por (TARAZONA-SANTOS et al., 2001) como indica a população Quechua como um alvo potencial para estudos de larga-escala, uma vez que esta população retém um grande reservatório de variabilidade genética em relação às populações nativo-americanas. Os estudos envolvendo as populações miscigenadas de Minas Gerais e de Lima no Peru, também oferecem perspectivas interessantes como a associação de fenótipos deletérios à ancestralidade ameríndia e à ampla distribuição da ancestralidade europeia e africana em brasileiros portadores de doença falciforme. Estes resultados ressaltam a oportunidade oferecida por estas populações nos estudos de associação por mapeamento de miscigenação.

A identificação de polimorfismos com alta divergência populacional apresentada no capítulo 3 representa um grande avanço em comparação com os resultados obtidos anteriormente, houve um aumento significativo, aproximadamente 700 vezes, no número de variantes selecionadas. Faz-se necessário utilizar métodos de detecção da seleção natural para evidenciar os sinais de seleção positiva. Além disso, os dados gerados permitem a identificação de regiões genômicas e funções biológicas candidatas a serem caracterizadas funcionalmente *in vivo*.

O último propósito do presente trabalho é aplicação da metodologia de sistemas multiagentes na integração e enriquecimento de dados biológicos. Ainda que os resultados sejam bastante promissores demonstrando a escalabilidade e a extensibilidade da plataforma, ainda faz-se necessário aperfeiçoar o acesso e a recuperação dos dados visando a diminuição do tempo de execução, ampliar o escopo das análises de enriquecimento e implementar novas metodologias, como as análises de redes complexas, para refinar ainda mais a geração do conhecimento.

REFERÊNCIAS

ABECASIS, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. **Nature**, v. 491, n. 7422, p. 56–65, 1 nov. 2012.

ADZHUBEI, I. A. et al. A method and server for predicting damaging missense mutations. **Nature methods**, v. 7, n. 4, p. 248–9, abr. 2010.

AKASHI, H.; OSADA, N.; OHTA, T. Weak selection and protein evolution. **Genetics**, v. 192, n. 1, p. 15–31, set. 2012.

ALEXANDER, D. H.; NOVEMBRE, J.; LANGE, K. Fast model-based estimation of ancestry in unrelated individuals. **Genome research**, v. 19, n. 9, p. 1655–64, set. 2009.

ALTSHULER, D. M. et al. Integrating common and rare genetic variation in diverse human populations. **Nature**, v. 467, n. 7311, p. 52–8, 2 set. 2010.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature genetics**, v. 25, n. 1, p. 25–9, maio 2000.

BALARESQUE, P. L.; BALLEREAU, S. J.; JOBLING, M. A. Challenges in human genetic diversity: demographic history and adaptation. **Human molecular genetics**, v. 16 Spec No, p. R134–9, 15 out. 2007.

BALDING, D. J. IEW A tutorial on statistical methods for population association studies. **Group**, v. 7, p. 781–791, 2006.

BARBUJANI, G. et al. An apportionment of human DNA diversity. **Proceedings of the National Academy of Sciences of the United States of America**, v. 94, n. 9, p. 4516–9, 29 abr. 1997.

BARNABE, C. et al. Prevalence of systemic lupus erythematosus and systemic sclerosis in the First Nations population of Alberta, Canada. **Arthritis care & research**, v. 64, n. 1, p. 138–43, jan. 2012.

BASTOS-RODRIGUES, L.; PIMENTA, J. R.; PENA, S. D. J. The genetic structure of human populations studied through short insertion-deletion polymorphisms. **Annals of human genetics**, v. 70, n. Pt 5, p. 658–65, set. 2006.

BAUER, S. et al. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. **Bioinformatics (Oxford, England)**, v. 24, n. 14, p. 1650–1, 15 jul. 2008.

BEISSBARTH, T.; SPEED, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. **Bioinformatics (Oxford, England)**, v. 20, n. 9, p. 1464–5, 12 jun. 2004.

BELLIFEMINE, F.; POGGI, A.; RIMASSA, G. JADE–A FIPA-compliant agent framework. **Proceedings of PAAM**, 1999.

BENYSHEK, D. C.; MARTIN, J. F.; JOHNSTON, C. S. A reconsideration of the origins of the type 2 diabetes epidemic among Native Americans and the implications for intervention policy. **Medical anthropology**, v. 20, n. 1, p. 25–64, jan. 2001.

BHATIA, G. et al. Estimating and interpreting FST: the impact of rare variants. **Genome research**, v. 23, n. 9, p. 1514–21, set. 2013.

BOGARDUS, C.; TATARANNI, P. A. Reduced early insulin secretion in the etiology of type 2 diabetes mellitus in Pima Indians. **Diabetes**, v. 51 Suppl 1, p. S262–4, fev. 2002.

BOHL, M.; RYNN, M. **Tools For Structured and Object-Oriented Design:United States Edition**. 7th Editio ed. [s.l.] Prentice Hall, 2007. p. 400

BROMBERG, Y. Building a genome analysis pipeline to predict disease risk and prevent disease. **Journal of molecular biology**, v. 425, n. 21, p. 3993–4005, 1 nov. 2013.

BROMBERG, Y.; CAPRIOTTI, E. SNP-SIG Meeting 2011: identification and annotation of SNPs in the context of structure, function, and disease. **BMC genomics**, v. 13 Suppl 4, p. S1, jan. 2012.

BROMBERG, Y.; CAPRIOTTI, E. Thoughts from SNP-SIG 2012: future challenges in the annotation of genetic variations. **BMC genomics**, v. 14 Suppl 3, p. S1, jan. 2013.

BROWNING, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. **Human genetics**, v. 124, n. 5, p. 439–50, dez. 2008.

BROWNING, S. R.; BROWNING, B. L. Haplotype phasing: existing methods and new developments. **Nature reviews. Genetics**, v. 12, n. 10, p. 703–14, out. 2011.

BROWNING, S. R.; BROWNING, B. L. Identity by descent between distant relatives: detection and applications. **Annual review of genetics**, v. 46, p. 617–33, jan. 2012.

BUSTAMANTE, C. D.; BURCHARD, E. G.; DE LA VEGA, F. M. Genomics for the world. **Nature**, v. 475, n. 7355, p. 163–5, 14 jul. 2011.

CAMPBELL, D. D. et al. Amerind ancestry, socioeconomic status and the genetics of type 2 diabetes in a Colombian population. **PloS one**, v. 7, n. 4, p. e33570, jan. 2012.

CANN, H. M. Human genome diversity. **Comptes rendus de l'Académie des sciences. Série III, Sciences de la vie**, v. 321, n. 6, p. 443–6, jun. 1998.

CARBONETTO, P.; STEPHENS, M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. **PLoS genetics**, v. 9, n. 10, p. e1003770, jan. 2013.

CAVALLI-SFORZA, L. L. The Human Genome Diversity Project: past, present and future. **Nature Reviews Genetics**, v. 6, n. 4, p. 333–343, 2005.

CHAGOYEN, M.; PAZOS, F. MBRole: enrichment analysis of metabolomic data. **Bioinformatics (Oxford, England)**, v. 27, n. 5, p. 730–1, 1 mar. 2011.

CHELALA, C.; KHAN, A.; LEMOINE, N. R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. **Bioinformatics (Oxford, England)**, v. 25, n. 5, p. 655–61, 1 mar. 2009.

CHEN, R. et al. Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. **PLoS genetics**, v. 8, n. 4, p. e1002621, jan. 2012.

CHENG, D. et al. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. **Nucleic acids research**, v. 36, n. Web Server issue, p. W399–405, 1 jul. 2008.

CHESSEL, D.; DUFOUR, A. B.; THIOULOUSE, J. The ade4 package-I- One-table methods. **R News**, v. 4, p. 5–10, 2004.

CHEVITARESE, J. **Determinação da estrutura genética das populações humanas e inferência dos fatores evolutivos que contribuíram para sua formação.** [s.l.] UFMG, 2009.

CHOI, Y. et al. Predicting the functional effect of amino acid substitutions and indels. **PloS one**, v. 7, n. 10, p. e46688, jan. 2012.

CHURCHHOUSE, C.; MARCHINI, J. Multiway admixture deconvolution using phased or unphased ancestral panels. **Genetic epidemiology**, v. 37, n. 1, p. 1–12, jan. 2013.

CLARK, A. G. The role of haplotypes in candidate gene studies. **Genetic epidemiology**, v. 27, n. 4, p. 321–33, dez. 2004.

CLARK, A. G. et al. Ascertainment bias in studies of human genome-wide polymorphism. **Genome Research**, p. 1496–1502, 2005.

CORELLA, A. et al. Mitochondrial DNA diversity of the Amerindian populations living in the Andean Piedmont of Bolivia: Chimane, Moseten, Aymara and Quechua. **Annals of human biology**, v. 34, n. 1, p. 34–55, 2007.

CROFT, D. et al. The Reactome pathway knowledgebase. **Nucleic acids research**, v. 42, n. 1, p. D472–7, 1 jan. 2014.

D'ANTONIO, M. et al. WEP: a high-performance analysis pipeline for whole-exome data. **BMC bioinformatics**, v. 14 Suppl 7, p. S11, jan. 2013.

DA SILVA, M. C. F. et al. Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers. **Blood**, v. 118, n. 16, p. 4493–5; author reply 4495, 20 out. 2011.

DAUB, J. T. et al. Evidence for polygenic adaptation to pathogens in the human genome. **Molecular biology and evolution**, v. 30, n. 7, p. 1544–58, jul. 2013.

DAVYDOV, E. V et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. **PLoS computational biology**, v. 6, n. 12, p. e1001025, jan. 2010.

DAYEM ULLAH, A. Z.; LEMOINE, N. R.; CHELALA, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). **Nucleic acids research**, v. 40, n. Web Server issue, p. W65–70, jul. 2012.

DAYEM ULLAH, A. Z.; LEMOINE, N. R.; CHELALA, C. A practical guide for the functional annotation of genetic variations using SNPnexus. **Briefings in bioinformatics**, v. 14, n. 4, p. 437–47, jul. 2013.

DE FILIPPO, C. et al. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. **Briefings in bioinformatics**, v. 13, n. 6, p. 696–710, nov. 2012.

DRAY, S.; DUFOUR, A. B. The ade4 package: implementing the duality diagram for ecologists. **Journal of Statistical Software**, v. 22, n. 4, p. 1–20, 2007.

DRAY, S.; DUFOUR, A. B.; CHESSEL, D. The ade4 package-II: Two-table and K-table methods. **R News**, v. 7, n. 2, p. 47–52, 2007.

EDEN, E. et al. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. **BMC bioinformatics**, v. 10, p. 48, jan. 2009.

EXCOFFIER, L.; LAVAL, G.; SCHNEIDER, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. **Evolutionary bioinformatics online**, v. 1, p. 47–50, jan. 2005.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application. **Network**, v. 491, p. 479–491, 1992.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. **Genetics**, v. 164, n. 4, p. 1567–87, ago. 2003.

FERGENBAUM, J. H. et al. Obesity and lowered cognitive performance in a Canadian First Nations population. **Obesity (Silver Spring, Md.)**, v. 17, n. 10, p. 1957–63, out. 2009.

FERRER-ADMETLLA, A. et al. Balancing selection is the main force shaping the evolution of innate immunity genes. **Journal of immunology (Baltimore, Md. : 1950)**, v. 181, n. 2, p. 1315–22, 15 jul. 2008.

FRANÇOIS, O. et al. Principal component analysis under population genetic models of range expansion and admixture. **Molecular biology and evolution**, v. 27, n. 6, p. 1257–68, jun. 2010.

FRIEDLAENDER, J. S. et al. The Genetic Structure of Pacific Islanders. **Structure**, v. 4, n. 1, 2008.

FUMAGALLI, M. et al. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. **Genome research**, v. 19, n. 2, p. 199–212, fev. 2009.

GAMAZON, E. R. et al. SCAN: SNP and copy number annotation. **Bioinformatics (Oxford, England)**, v. 26, n. 2, p. 259–62, 15 jan. 2010.

GARRIGAN, D.; HAMMER, M. F. Reconstructing human origins. **Genetics**, 2006.

GE, D. et al. WGAViewer: software for genomic annotation of whole genome association studies. **Genome research**, v. 18, n. 4, p. 640–3, abr. 2008.

GILAD, Y. et al. Natural selection on the olfactory receptor gene family in humans and chimpanzees. **American journal of human genetics**, v. 73, n. 3, p. 489–501, set. 2003.

GLAAB, E. et al. EnrichNet: network-based gene set enrichment analysis. **Bioinformatics (Oxford, England)**, v. 28, n. 18, p. i451–i457, 15 set. 2012.

GOLDSTEIN, D. B. et al. Sequencing studies in human genetics: design and interpretation. **Nature reviews. Genetics**, v. 14, n. 7, p. 460–70, jul. 2013.

GOSLEE, S. C.; URBAN, D. L. The ecodist package for dissimilarity-based analysis of ecological data. **Journal of Statistical Software**, v. 22, n. 7, p. 1–19, 2007.

GRAY, K. A. et al. Genenames.org: the HGNC resources in 2013. **Nucleic acids research**, v. 41, n. Database issue, p. D545–52, jan. 2013.

HAMMER, M. F. et al. Sex-biased evolutionary forces shape genomic patterns of human diversity. **PLoS genetics**, v. 4, n. 9, p. e1000202, jan. 2008.

HANSON, R. L. et al. A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array. **Diabetes**, v. 56, n. 12, p. 3045–52, dez. 2007.

HEARST, M. O. et al. Trends of overweight and obesity among white and American Indian school children in South Dakota, 1998-2010. **Obesity (Silver Spring, Md.)**, v. 21, n. 1, p. E26–32, jan. 2013.

HENDERSON, J. A. et al. Prostate cancer incidence among American Indian and Alaska Native men, US, 1999-2004. **Cancer**, v. 113, n. 5 Suppl, p. 1203–12, 1 set. 2008.

HOFER, T. et al. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. **Annals of human genetics**, v. 73, n. 1, p. 95–108, jan. 2009.

HOLSINGER, K. E.; WEIR, B. S. Genetics in geographically structured populations : defining , estimating and interpreting F ST. v. 10, n. SePTeMBer, 2009.

HUANG, D. W. et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. **Nucleic acids research**, v. 35, n. Web Server issue, p. W169–75, jul. 2007.

HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. **Nucleic acids research**, v. 37, n. 1, p. 1–13, jan. 2009.

HUGHES, A. L. Near neutrality: leading edge of the neutral theory of molecular evolution. **Annals of the New York Academy of Sciences**, v. 1133, p. 162–79, jan. 2008.

HUNG, J.-H. et al. Gene set enrichment analysis: performance evaluation and usage guidelines. **Briefings in bioinformatics**, v. 13, n. 3, p. 281–91, maio 2012.

HURST, L. D. Fundamental concepts in genetics: genetics and the understanding of selection. **Nature reviews. Genetics**, v. 10, n. 2, p. 83–93, fev. 2009.

JAKOBSSON, M. et al. Genotype , haplotype and copy-number variation in worldwide human populations. **Nature**, v. 451, n. February, 2008.

JENSEN, L. J.; BORK, P. Ontologies in quantitative biology: a basis for comparison, integration, and discovery. **PLoS biology**, v. 8, n. 5, p. e1000374, maio 2010.

JERVIS, L. L. et al. Predictors of performance on the MMSE and the DRS-2 among American Indian elders. **The Journal of neuropsychiatry and clinical neurosciences**, v. 22, n. 4, p. 417–25, jan. 2010.

JERVIS, L. L.; MANSON, S. M. Cognitive impairment, psychiatric disorders, and problematic behaviors in a tribal nursing home. **Journal of aging and health**, v. 19, n. 2, p. 260–74, abr. 2007.

JOHNSON, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. **Bioinformatics (Oxford, England)**, v. 24, n. 24, p. 2938–9, 15 dez. 2008.

JOMBART, T.; AHMED, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. **Bioinformatics** , 16 set. 2011.

JOSHI-TOPE, G. et al. The Genome Knowledgebase: a resource for biologists and bioinformaticists. **Cold Spring Harbor symposia on quantitative biology**, v. 68, p. 237–43, jan. 2003.

KAROLCHIK, D. et al. The UCSC Genome Browser database: 2014 update. **Nucleic acids research**, v. 42, n. 1, p. D764–70, 1 jan. 2014.

KOFLER, R.; SCHLÖTTERER, C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. **Bioinformatics (Oxford, England)**, v. 28, n. 15, p. 2084–5, 1 ago. 2012.

KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature protocols**, v. 4, n. 7, p. 1073–81, jan. 2009.

LACHANCE, J.; TISHKOFF, S. A. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. **BioEssays : news and reviews in molecular, cellular and developmental biology**, v. 35, n. 9, p. 780–6, set. 2013.

LARSON, N. B.; SCHAID, D. J. Regularized rare variant enrichment analysis for case-control exome sequencing data. **Genetic epidemiology**, v. 38, n. 2, p. 104–13, fev. 2014.

LEE, J. et al. SNP-PRAGE: SNP-based parametric robust analysis of gene set enrichment. **BMC systems biology**, v. 5 Suppl 2, p. S11, 14 dez. 2011.

LEWIS, P. O.; ZAYKIN, D. **GDA (Genetic Data Analysis): Computer program for the analysis of allelic data.**Storrs University of Connecticut, , 2002.

LEWONTIN, R. The apportionament of human diversity. **Evolutionary Biology**, v. 6, p. 381–398, 1972.

LI, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. **Science (New York, N.Y.)**, v. 319, n. 5866, p. 1100–4, 22 fev. 2008.

LI, Y.; GRAUBARD, B. I. Testing Hardy – Weinberg Equilibrium and Homogeneity of Hardy – Weinberg Disequilibrium using Complex Survey Data. **Biometrics**, 2009.

LIU, L. et al. Robust methods for population stratification in genome wide association studies. **BMC bioinformatics**, v. 14, p. 132, jan. 2013a.

LIU, N.; ZHAO, H. A non-parametric approach to population structure inference using multilocus genotypes. **Human genomics**, v. 2, n. 6, p. 353–64, jun. 2006.

LIU, Y. et al. Softwares and methods for estimating genetic ancestry in human populations. **Human genomics**, v. 7, p. 1, jan. 2013b.

MACHADO, M. et al. Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies. **Investigative genetics**, v. 2, n. 1, p. 3, jan. 2011.

MAGALHÃES, W. C. S. **DIVERGENOME: uma plataforma bioinformática para o estudo da diversidade genética humana e aplicações na identificação de episódios de seleção natural na evolução humana**. [s.l.] UFMG, 2011.

MAHONEY, M. C. et al. Changes in cancer incidence patterns among a northeastern American Indian population: 1955-1969 versus 1990-2004. **The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association**, v. 25, n. 4, p. 378–83, jan. 2009.

MARX, V. Biology: The big challenges of big data. **Nature**, v. 498, n. 7453, p. 255–60, 13 jun. 2013.

MASEL, J. Genetic drift. **Current biology : CB**, v. 21, n. 20, p. R837–8, 25 out. 2011.

MATTHEWS, L. et al. **An Introduction to the Reactome Knowledgebase of Human Biological Pathways and Processes : Bioinformatics Primer : Pathway Interaction**

Database. Disponível em: <http://pid.nci.nih.gov/PID/2007/071211/full/pid.2007.3.shtml>. Acesso em: 30 jan. 2014.

MATTHEWS, L. et al. Reactome knowledgebase of human biological pathways and processes. **Nucleic acids research**, v. 37, n. Database issue, p. D619–22, jan. 2009.

MCVEAN, G. A genealogical interpretation of principal components analysis. **PLoS genetics**, v. 5, n. 10, p. e1000686, out. 2009.

MITCHELL, C. M. et al. Trajectories of cognitive development among American Indian young children. **Developmental psychology**, v. 47, n. 4, p. 991–9, jul. 2011.

MULLIGAN, C. J. et al. P OPULATION G ENETICS , H ISTORY , AND H EALTH. **Distribution**, p. 295–315, 2004.

MYLES, S. et al. Identification and analysis of genomic regions with large between-population differentiation in humans. **Annals of human genetics**, v. 72, n. Pt 1, p. 99–110, jan. 2008.

NEI, M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Sciences of the United States of America**, v. 70, n. 12, p. 3321–3, dez. 1973.

NEI, M.; ROYCHOUDHURY, A. Sampling variances of heterozygosity and genetic distance. **Genetics**, p. 379–390, 1974.

NEWKIRK, M. M. et al. Advanced glycation endproducts (AGE) on IgG, a target for circulating antibodies in North American Indians with rheumatoid arthritis (RA). **Cellular and molecular biology (Noisy-le-Grand, France)**, v. 44, n. 7, p. 1129–38, nov. 1998.

NIELSEN, R. Population genetic analysis of ascertained SNP data. **Human genomics**, v. 1, n. 3, p. 218–24, mar. 2004.

NIELSEN, R. et al. Darwinian and demographic forces affecting human protein coding genes. **Genome Research**, p. 838–849, 2009.

NOGALES-CADENAS, R. et al. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. **Nucleic acids research**, v. 37, n. Web Server issue, p. W317–22, jul. 2009.

NOVEMBRE, J.; DI RIENZO, A. Spatial patterns of variation due to natural selection in humans. **Nature reviews. Genetics**, v. 10, n. 11, p. 745–55, nov. 2009.

NOVEMBRE, J.; HAN, E. Human population structure and the adaptive response to pathogen-induced selection pressures. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 367, n. 1590, p. 878–86, 19 mar. 2012.

Online Mendelian Inheritance in Man, OMIM®. Disponível em: <http://omim.org/>. Acesso em: 20 fev. 2014.

PAANANEN, J.; CISZEK, R.; WONG, G. Varietas: a functional variation database portal. **Database : the journal of biological databases and curation**, v. 2010, p. baq016, jan. 2010.

PEISCHL, S. et al. On the accumulation of deleterious mutations during range expansions. **Molecular ecology**, v. 22, n. 24, p. 5972–82, dez. 2013.

PEREIRA, L. et al. Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population. **PloS one**, v. 7, n. 8, p. e41200, jan. 2012.

PICKRELL, J. K. et al. Signals of recent positive selection in a worldwide sample of human populations. p. 826–837, 2009.

PRITCHARD, J. K. et al. Association Mapping in Structured Populations. **Society**, p. 170–181, 2000.

PRITCHARD, J. K.; DONNELLY, P. Case – Control Studies of Association in Structured or Admixed Populations. **Theoretical Population Biology**, v. 237, p. 227–237, 2001.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945–959, 2000.

PURCELL, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American journal of human genetics**, v. 81, n. 3, p. 559–75, set. 2007.

QUINTANA-MURCI, L.; CLARK, A. G. Population genetic tools for dissecting innate immunity in humans. **Nature reviews. Immunology**, v. 13, n. 4, p. 280–93, abr. 2013.

RAMACHANDRAN, S. et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. 2005.

RF FLOW 5. **What do the different flowchart shapes mean?** Disponível em: <http://www.rff.com/flowchart_shapes.htm>. Acesso em: 15 fev. 2014.

RHEE, H.; LEE, J.-S. MedRefSNP: a database of medically investigated SNPs. **Human mutation**, v. 30, n. 3, p. E460–6, mar. 2009.

ROBERTSON, A; HILL, W. G. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. **Genetics**, v. 107, n. 4, p. 703–18, ago. 1984.

RODRIGUES, M. R. et al. A graph-based approach for designing extensible pipelines. **BMC bioinformatics**, v. 13, p. 163, jan. 2012.

ROGERS, A. R.; JORDE, L. B. Ascertainment Bias in Estimates of Average Heterozygosity. **New York**, p. 1033–1041, 1996.

ROSENBERG, N. et al. Genetic structure of human populations. **Science**, v. 2381, n. 2002, 2002.

ROSENBERG, N. A. et al. Clines , Clusters , and the Effect of Study Design on the Inference of Human Population Structure. **Structure**, v. 1, n. 6, 2005.

ROSENBERG, N. A. Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel , Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. **Annals of Human Genetics**, p. 841–847, 2006.

ROSENBERG, N. A.; NORDBORG, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. **Nature reviews. Genetics**, v. 3, n. 5, p. 380–90, maio 2002.

SABETI, P. C. et al. Detecting recent positive selection in the human genome from haplotype structure. v. 419, n. October, 2002.

SANCHEZ, E. et al. Genetically determined Amerindian ancestry correlates with increased frequency of risk alleles for systemic lupus erythematosus. **Arthritis and rheumatism**, v. 62, n. 12, p. 3722–9, dez. 2010.

SCHAEFER, C. et al. SNPdbe: constructing an nsSNP functional impacts database. **Bioinformatics (Oxford, England)**, v. 28, n. 4, p. 601–2, 15 fev. 2012.

SCHMITT, A. O. et al. CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. **Bioinformatics (Oxford, England)**, v. 26, n. 7, p. 969–70, 1 abr. 2010.

SCLIAR, M. O. et al. The population genetics of Quechuas, the largest native South American group: autosomal sequences, SNPs, and microsatellites evidence high level of diversity. **American journal of physical anthropology**, v. 147, n. 3, p. 443–51, mar. 2012.

SELDIN, M. Admixture mapping as a tool in gene discovery. **Current opinion in genetics & development**, 2007.

SHERRY, S. T. et al. dbSNP: the NCBI database of genetic variation. **Nucleic acids research**, v. 29, n. 1, p. 308–11, 1 jan. 2001.

SIEPEL, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. **Genome research**, v. 15, n. 8, p. 1034–50, ago. 2005.

SMITH, M. W.; BRIEN, S. J. O. MAPPING BY ADMIXTURE LINKAGE DISEQUILIBRIUM : ADVANCES , LIMITATIONS AND GUIDELINES. **Nature Reviews Genetics**, n. July, p. 1–11, 2005.

SOARES-SOUZA, G. B. **Identificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e aplicações biomédicas.** [s.l.] UFMG, 2010.

STOYANOVICH, J.; PE'ER, I. MutaGeneSys: estimating individual disease susceptibility based on genome-wide SNP array data. **Bioinformatics (Oxford, England)**, v. 24, n. 3, p. 440–2, 1 fev. 2008.

SUN, H. et al. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. **Bioinformatics (Oxford, England)**, 30 out. 2013.

SVENSON, L. W. et al. Prevalence of multiple sclerosis in First Nations people of Alberta. **The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques**, v. 34, n. 2, p. 175–80, maio 2007.

TABAS-MADRID, D.; NOGALES-CADENAS, R.; PASCUAL-MONTANO, A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. **Nucleic acids research**, v. 40, n. Web Server issue, p. W478–83, jul. 2012.

TAI, A. W.; NEWKIRK, M. M. An autoantibody targeting glycated IgG is associated with elevated serum immune complexes in rheumatoid arthritis (RA). **Clinical and experimental immunology**, v. 120, n. 1, p. 188–93, abr. 2000.

TANG, H. et al. Estimation of individual admixture: analytical and study design considerations. **Genetic epidemiology**, v. 28, n. 4, p. 289–301, maio 2005.

TARAZONA-SANTOS, E. et al. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. **American journal of human genetics**, v. 68, n. 6, p. 1485–96, jun. 2001.

THE INTERNATIONAL HAPMAP CONSORTIUM. The International HapMap Project. **Nature**, v. 426, n. 6968, p. 789–96, 18 dez. 2003.

THE INTERNATIONAL HAPMAP CONSORTIUM. A haplotype map of the human genome. **Nature**, v. 437, n. 7063, p. 1299–320, 27 out. 2005.

THE INTERNATIONAL HAPMAP CONSORTIUM. A second generation human haplotype map of over 3.1 million SNPs. **Nature**, v. 449, n. October, p. 851–862, 2007.

THOMAS, D. C.; WITTE, J. S. Point : Population Stratification : A Problem for Case-Control Studies of Candidate-Gene Associations ? 1. **Prevention**, v. 11, n. June, p. 505–512, 2002.

THORISSON, G. A.; MUILU, J.; BROOKES, A. J. Genotype-phenotype databases: challenges and solutions for the post-genomic era. **Nature reviews. Genetics**, v. 10, n. 1, p. 9–18, jan. 2009.

TRAURIG, M. T. et al. Variants in the LEPR gene are nominally associated with higher BMI and lower 24-h energy expenditure in Pima Indians. **Obesity (Silver Spring, Md.)**, v. 20, n. 12, p. 2426–30, dez. 2012.

VASTRIK, I. et al. Reactome: a knowledge base of biologic pathways and processes. **Genome biology**, v. 8, n. 3, p. R39, jan. 2007.

VITEBSKY, P. Yakut. In: SMITH, G. (Ed.). **The nationalities question in the Soviet Union**. London: Longmans, 1990. p. 302–317.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic acids research**, v. 38, n. 16, p. e164, set. 2010.

WANG, S. et al. Genetic Variation and Population Structure in Native Americans. **Americas, The**, v. 3, n. 11, 2007.

WARREN, S. et al. Incidence of multiple sclerosis among First Nations people in Alberta, Canada. **Neuroepidemiology**, v. 28, n. 1, p. 21–7, jan. 2007.

WATKINS, W. S. et al. Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. **BMC genetics**, v. 13, p. 39, jan. 2012.

WEIR, B. S.; COCKERHAM, C. C. Estimating F-Statistics for the Analysis of Population Structure. **Evolution**, v. 38, n. 6, p. 1358–1370, 1984.

WEIR, B. S.; HILL, W. G. E STIMATING F-S TATISTICS. 2002.

WELTER, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. **Nucleic acids research**, v. 42, n. 1, p. D1001–6, 1 jan. 2014.

WHIRL-CARRILLO, M. et al. Pharmacogenomics knowledge for personalized medicine. **Clinical pharmacology and therapeutics**, v. 92, n. 4, p. 414–7, out. 2012.

WILDER, J. A.; MOBASHER, Z.; HAMMER, M. F. Genetic evidence for unequal effective population sizes of human females and males. **Molecular biology and evolution**, v. 21, n. 11, p. 2047–57, nov. 2004.

XU, Z.; TAYLOR, J. A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. **Nucleic acids research**, v. 37, n. Web Server issue, p. W600–5, jul. 2009.

YANG, R. C. Estimating Hierarchical F-Statistics. **Evolution**, v. 52, n. 4, p. 950–956, 1998.

YU, G. et al. clusterProfiler: an R package for comparing biological themes among gene clusters. **Omics : a journal of integrative biology**, v. 16, n. 5, p. 284–7, maio 2012.

ZHANG, F. et al. Genetic studies of human diversity in East Asia. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 362, n. 1482, p. 987–95, 29 jun. 2007.

ZHANG, K. et al. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. **Nucleic acids research**, v. 38, n. Web Server issue, p. W90–5, jul. 2010.

APÊNDICE A – DESCRIÇÃO DAS FORMAS DE ACESSO AOS BANCOS DE DADOS E DA CONSTRUÇÃO DAS BASES LOCAIS

dbSNP

A versão local da base de dados atualmente instalada em nosso servidor é a 137. Os seguintes passos foram realizados a fim de implementar a cópia local: i) download da estrutura do banco (tabelas, relacionamentos e indíces) associada ao *Homo sapiens*; (ii) download dos dados brutos para a mesma espécie; (iii) desenvolvimento execução de scripts para a conversão do formato MSSQL em MySQL; (iv) criação da estrutura do banco de dados local a partir das informações convertidas anteriormente; (v) inserção dos dados do *Homo sapiens* no banco dbSNP local, nomeado dbsnp_137.

O sítio dbSNP disponibiliza um conjunto de ferramentas para acesso remoto denominado Eutils. Deste grupo de ferramentas utilizou-se, especificamente, o script efetch.fcgi, que recebe como entrada o valor numérico do identificador do SNP e o tipo de output desejado, a API então retorna um arquivo baseado em *html* com as informações disponíveis para a variante. O agente dbSNP remoto executa este script para cada polimorfismo e processa o resultado, retornando apenas as informações desejadas.

UCSC

As versões local e remota compartilham a metodologia de acesso através de instruções MySQL. E a única diferença entre os dados obtidos entre as formas de acesso local e remota é a de que no acesso remoto não dispomos dos valores de conservação para cada sítio, ao invés disso, retorna-se os valores de conservação médios para um conjunto de 1024 sítios. No acesso local são disponibilizados tanto a informação por sítio quanto a por conjunto de 1024 sítios. Isto se deve ao fato de que os valores de conservação por sítio não se encontram na base de dados relacionais, sendo disponibilizados através dos arquivos *wiggle*.

A versão local do banco de dados UCSC mantida no servidor é a hg19, baixada em Março de 2013. Os procedimentos para a instalação da base de dados local foram: i) download da estrutura de dados, versão hg19, para 8 tabelas no formato MySQL; (ii) download dos escores de conservação PhastCons46 e PhyloP46 em formato *wiggle* para

vertebrados, mamíferos e primatas; (iii) desenvolvimento e execução do script MyWiggum para conversão do formato *wiggle* em TSV; (iv) preenchimento das bases descritas na Tabela 21; (v) criação de índices para as tabelas de conservação base-a-base.

A conexão remota ao banco de dados UCSC se dá através do acesso MySQL remoto. A base é então requerida e retorna os atributos desejados assim como na anotação local. O endereço host utilizado é *genome-mysql.cse.ucsc.edu*, o usuário é *genome* e não é necessária senha.

GENE ONTOLOGY

Atualmente é mantida a versão local do banco de dados Gene Ontology de Abril de 2013. O processo de criação da base envolve os seguintes passos: i) download dos arquivos *termdb* e *assocdb* em formato MySQL; ii) reconstrução destas tabelas no banco de dados local Gene Ontology.

A base de dados *Gene Ontology* disponibiliza o acesso remoto através da conexão MySQL. A base *go_latest* (mais recente) é inquirida no endereço host mysql.ebi.ac.uk:4085 utilizando o usuário *go_select* e senha *amigo*. A consulta é realizada a partir do símbolo do gene, e as informações retornadas são processadas pelo agente db GO de forma a separá-las de acordo com a classe de anotação: processo biológico, função molecular e componente celular.

PHARMGKB

O aceso aos dados disponibilizados pela plataforma PharmGKB pode ser realizado através da interface do sítio (www.pharmgkb.org), através de APIs implementadas em Perl e através do download dos dados. A versão remota do MASSA acessa o banco de dados através do script *search.pl* disponibilizado pelo desenvolvedores do PharmGKB. O acesso local é realizado através de um banco local em MySQL construído a partir dos dados baixados pelo mesmo script utilizado pelo agente de conexão remoto.

A conexão remota ao servidor do PharmGKB é obtida através da API em Perl search.pl disponibilizada pelo PhamGKB. É necessária a instalação do módulo SOAP::Lite para a correta execução do agente PGKB. O script foi encapsulado pelo MASSA sendo

executado O resultado retornado pelo script é processado e as informações são disponibilizadas de maneira organizada e inteligível ao usuário. A interrogação do banco é realizada a partir do nome do gene.

Para o banco local foi realizado o download dos dados a partir da API em Perl *search.pl*. O script pharmLeaks.pl foi desenvolvido para interrogar a base de dados PharmGKB de maneira sequencial. A partir de uma lista de genes o script *pharmLeaks.pl* executa a API de busca da base PharmGKB e constrói os arquivos tabelados que serão utilizados para popular o banco local. São construídas simultaneamente as relações entre as tabelas através de chaves estrangeiras numéricas. São produzidas, então, oito tabelas: *Gene*, *Disease*, *Drug*, *GoldenPath*, *Haplotype*, *Literature*, *Pathway*, Output sendo esta utilizada como controle das chaves estrangeiras. O banco MySQL foi construído previamente respeitando o formato dos dados retornados pela API *search*. Após o download dos dados as tabelas podem ser adicionadas ao banco.

OMIM

O acesso à base de dados OMIM pode ser realizado através da interface de busca do sítio (www.omim.org), de APIs disponibilizadas pelo desenvolvedor ou do download dos dados. As APIs permitem o retorno dos dados em diferentes estruturas de dados como: XML, HTML, JSON, Python e Ruby. O agente remoto OMIM realiza o acesso aos dados através da API no formato XML. O agente local OMIM realiza o acesso aos dados a partir de uma versão local em MySQL obtida através do download dos dados. A única diferença entre os dois métodos de acesso é que a informação sobre os padrões de herança da doença não estão disponíveis na versão local. Para a construção do banco local realizou-se o download dos dados através do FTP. Para tanto é necessário o registro prévio no OMIM.

O script *omim2sqlBygene.pl* foi desenvolvido para processar o arquivo *genemap.txt* em um arquivo tabulado compatível com o banco local em MySQL desenvolvido para abrigar os dados baixados. A estrutura do banco segue o mesmo padrão do arquivo *genemap*, com 18 atributos. O script separa a sexta coluna (*Gene Symbols*) do arquivo genemap.txt em diferentes linhas (cada linha do arquivo de saída deve estar relacionada à apenas um símbolo gênico). As demais colunas e informações forma mantidas no formato original.

A conexão remota à plataforma OMIM é obtida através da API disponibilizada pelos desenvolvedores. É necessário o registro prévio para obter a autorização de acesso aos dados. A chave utilizada pelo MASSA é 0807D10B6344A56E6C07AC19A6AF5A1957C6B92C, registrada por Giordano B. Soares-Souza. O acesso é feito através do servidor localizado nos EUA, endereço: http://api.omim.org/. O acesso aos dados é configurado através da própria URL e o MASSA não utiliza nenhum filtro além do nome do gene a ser investigado. Todos os campos disponíveis são retornados e processados, de forma a manter o mesmo formato e conteúdo da pesquisa local.

REACTOME

O banco local foi construído a partir do download do arquivo *Reactome Pathways Gene Set* versão da base 46. Este arquivo discrimina as vias metabólicas e os genes associados a cada uma delas. O formato foi convertido para TSV, individualizando as relações entre genes e vias, ou seja, cada linha passa a apresentar associação única gene-via. A estrutura da base local em MySQL foi construída, incluindo o índice da tabela e, então, o arquivo foi armazenado no banco. A versão remota do MASSA, no momento, não dispõe de acesso direto ao Reactome, mas, ao invés disso, a consulta é realizada a partir do arquivo TSV gerado durante a criação do banco local. A anotação é feita através da leitura do arquivo que será disponibilizado em conjunto com a ferramenta até que o acesso remoto seja gerado.

HGNC

O acesso aos dados do HGNC pode ser realizado via interface web através do sítio (http://www.genenames.org/hgnc-searches). Outra forma de acesso é um script CGI que acessa diretamente o banco de dados em MySQL.

A conexão remota ao servidor HGNC é realizada através do script *hugoPerl.pl* que constrói uma URL de acesso ao CGI do HGNC. Este script é encapsulado pelo MASSA e o resultado retornado é processado e apresentado ao usuário de forma organizada e inteligível.

Para a construção do banco local os dados do HGNC foram baixados através do CGI do sítio do HGNC (http://www.genenames.org/cgi-bin/hgnc_downloads). Todos os campos de pesquisa foram marcados e nenhum filtro foi utilizado para garantir a totalidade dos dados.

O banco local foi construído a partir do padrão de colunas retornado pelo HGNC: apenas uma tabela MySQL com todos os campos descritos na Tabela 26.

POLYPHEN-2

Para a construção do banco local de PolyPhen-2, os seguintes passos foram realizados: i) download dos arquivos contendo as predições para os conjuntos de dados HumVar e HumDiv; ii) construção da estrutura da base de dados local; iii) inserção dos dados no banco; iv) construção dos índices das tabelas HumVar e HumDiv. Atualmente, apenas os dados do conjunto HumVar são retornandos para evitar a sobreposição de anotações para um mesmo polimorfismo.

Como não há uma base de dados remota, passível de ser acessada pelo MASSA, construiu-se um arquivo resumido das informações presentes nas anotações de HumVar, contendo apenas os campos retornados pelo agente PolyPhen. As pesquisas na versão remota são, dessa forma, realizadas através da leitura direta deste resumo.

PROVEAN/SIFT

Os passos para a construção do banco local incluem: i) download das anotações disponibilizadas – Provean Scores (v1.1); construção da estrutura da base de dados local; iii) inserção dos dados na base local provean; iv) adição dos índices a tabela.

NHGRI GWAS CATALOG

O banco de dados local foi gerado a partir das seguintes ações: i) download dos dados em formato TSV; ii) extração dos campos mais relevantes do arquivo (ver Tabela 29); iii) montagem da estrutura local da base em MySQL; iv) inserção do arquivo resumido; v) criação dos índices da tabela. O acesso remoto à base de dados ainda não está disponível e a anotação na versão remota é efetuada através da leitura do arquivo resumido produzido para a construção da base local.

APÊNDICE B – COMPARAÇÃO ENTRE 14 FERRAMENTAS DE ANOTAÇÃO

Ferramenta/ Categoria de Anotação	Enriquecimento	Polimorfismo	Genótipos	Haplótipos	Gene	RefSeq	DL	Proteína	Via/Interação	mRNA	miRNA	CNVs/Estrutural	Fenótipo	Farmagogenômica	Seleção Natural	População	Referências	Outros
MASSA	10	11	4	-	17	4	-	13	2	2	-	-	12	1	-	-	-	-
Annovar	-	5	6	-	3	5	-	2	-	1	-	2	3	-	-	B	-	-
GLU	-	3	2	-	-	7	-	2	-	-	-	-	-	-	-	-	-	-
PLINK	-	14	15	-	6	7	-	-	-	-	3	2	-	-	-	-	-	7
SNPnexus	-	54	12	-	40	1	-	30	-	9	13	21	15	-	-	15 B1	-	1
Illumina	-	2	1	-	3	1	-	-	-	-	-	-	-	-	-	-	-	1
CandiSNPer	-	4	-	-	-	1	2	-	-	-	-	-	-	-	-	B	-	-
MutaGeneSys	-	1	2	-	-	-	-	-	-	-	-	-	1	-	-	B	-	1
SNPdbe	-	2	1	-	1	-	-	21	-	-	-	1	1	-	-	1	-	2
PolySearch	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SNAP	-	3	3	-	1	-	7	-	-	-	-	-	-	-	-	B	-	1
SCAN	-	3	1	-	1	3	2	-	-	-	-	5	-	-	-	-	-	-
MedRefSNP	-	1	1	-	1	1	-	-	1	-	-	-	-	-	-	-	2	-
Varietas	-	7	1	-	2	-	-	-	-	-	-	-	6	-	-	-	1	1
WGAViewer	-	3	2	-	1	3	3	-	-	-	-	4	-	-	1	-	-	-
SNPinfo Web Server	-	24	7	-	4	1	6	6	-	-	4	-	1	-	-	-	-	2

B – Alguns atributos da anotação são baseados em populações de escolha. B1 – 15 Anotações para uma população, mas pode-se retornar resultados para as 15 populações do HapMap.

APÊNDICE C – COMPARAÇÃO ENTRE 14 FERRAMENTAS DE ENRIQUECIMENTO

Ferramenta/Categoria de Anotação	Enriquecimento	Método	Polimorfismo	Genótipos	Haplótipos	Gene	RefSeq	DL	Proteína	Via/Interação	mRNA	miRNA	CNVs/Estrutural	Fenótipo	Farmagogenômica	Seleção Natural	População	Referências	Outros
MASSA	10	SEA	11	4	-	17 (4)	4 (1)	-	13	2 (2)	2	-	-	12 (2)	1 (1)	-	-	-	-
Gowinda[1]	Genérico	SEA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GWAS PathwayIndentifier	1	SEA/GSEA	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
i-GSEA4GWAS	4	GSEA	-	-	-	3	-	-	-	1	-	-	-	-	-	-	-	-	-
ToppGene	17	SEA/GSEA	-	-	-	6	1	-	1	2	-	1	-	3	1	-	-	1	1
GSEA-SNP1	14	GSEA	-	-	-	7	-	-	1	-	-	1	-	3	1	-	-	-	1
SNP-PRAGE	1	PSEA	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
clusterProfiler	4	MEA	-	-	-	3	-	-	-	1	-	-	-	-	-	-	-	-	-
MBRole	9	SEA	-	-	-	-	-	-	-	3	-	-	-	2	4	-	-	-	-
GeneCodis3	9	SEA	-	-	-	3	-	-	1	2	-	-	-	1	1	-	-	1	-
GOrilla	3	GSEA	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-
Ontologizer 2.0	3	SEA/MEA	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-
GOstat	3	SEA	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-
SNP-based PEA	1	SSEA	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
DAVID	63	SEA/MEA	-	-	-	10	13	-	16	15	-	-	-	2	-	-	-	3	4

[1]**Gowinda é um software genérico, realiza a análise de enriquecimento para conjuntos de dados disponibilizados pelo usuário.**

APÊNDICE D – MAPEAMENTO ONTOLÓGICO E TERMINOLÓGICO DAS ANOTAÇÕES GENÔMICAS

INFORMAÇÕES	Ontologia	Id do Termo	Variável	Descrição	Description
Polimorfismo					
Polymorphism ID	polymorphic_variant	SO:0001766	polyId	Código de Identificação	Polymorphism Code
Polymorphism Type	polymorphic_variant	SO:0001766	polType	Tipo (SNP, InDel, CNV, STR)	Type (SNP, InDel, CNV, STR)
Single Nucleotide Polymorphism	snp	SO:SO_0000694	snp	Código de Identificação dbSNP	dbSNP Id
Affymetrix ID	affymetrix	edamontology:format_1639	affyId	Código de Identificação Affymetrix	Affymetrix Id
Illumina ID	Illumina data	swo2:SWO_0000233	IlluId	Código de Identificação Illumina	Illumina Id
1000Genomes	-	-	1kgId	Código de Identificação 1000G	1000Genomes Id
Genotyping Center	genotyping_center	-	genCenter	Centro de genotipagem	Genotyping Center
Genotyping Plataform	genotyping_plataform	-	genPlataform	Plataforma de Genotipagem	Genotyping Plataform
SNP Error	-	-	snpError	Indica erro de genotipagem	Genotyping error
SNP Pos Duplication	duplication	SO:SO_1000035	snpPosDup	Posição duplicada do SNP	Duplicated SNP positions
Chromosome	chromosome	chromosome	chr	Cromossomo	Chromosome
Chromosome Orientation	Genomic Orientation	C48673	orient	Orientação 5'->3'	5'->3' Orientation
DNA Strand Orientation	DNA sense specification	data_0853	strand	Fita-molde	DNA-strand
HG17 Position	Genome build identifier	data_2340	hg17pos	Posição Cromossômica HG17	HG17 Chromosome Position
HG18 Position	Genome build identifier	data_2340	hg18pos	Posição Cromossômica HG18	HG18 Chromosome Position
HG19 Position	Genome build identifier	data_2340	chrPos	Posição Cromossômica HG19	HG19 Chromosome Position
Transcript Region	transcript_region	SO:SO_0000833	transReg	Região (Exon, Intron, 5'UTR, 3'UTR)	Regions (Exon, Intron, 5'UTR, 3'UTR)
Nucleotide Numbering coding DNA	-	-	coordRelGene	Posição de uma sequência contígua que incia-se e inclui um códon de iniciação e um de parada. De acordo com a referência funcional.	Position in a contiguous sequence which begins

					with, and includes, a start codon and ends with, and includes, a stop codon.According functional reference.
Assembly Build Version	-	-	assmBV	Versão de montagem do Genoma	Genome Assembly Build Version
Assembly Coord Start	-	-	assmCStart	Posição de início do polimorfismo de acordo com a versão do genoma	Start position of variant according to reference human genome assembly
Assembly Coord End	-	-	assmCEnd	Posição de término do polimorfismo de acordo com a versão do genoma	End position of variant according to reference human genome assembly
Phast Conserved Nucleotide - Primate	conserved_region	SO:0000330	phastConsP	Escore de Conservação em primatas (PhastCons)	Residue conservation in primates (phastCons)
Phats Conserved Nucleotide - Mammals	conserved_region	SO:0000330	phastConsM	Escore de Conservação em mamíferos (PhastCons)	Residue conservation in mammals (phastCons)
Phast Conserved Nucleotide - Vertebrate	conserved_region	SO:0000330	phastConsV	Escore de Conservação em vertebrados (PhastCons)	Residue conservation in vertebrate (phastCons)
Phast Conserved Elements - Primate	conserved_region	SO:0000330	phastEConsP	Escore de Conservação em primatas (PhastCons)	Residue conservation in primates (phastCons)
Phats Conserved Elements - Mammals	conserved_region	SO:0000330	phastEConsM	Escore de Conservação em mamíferos (PhastCons)	Residue conservation in mammals (phastCons)

Phast Conserved Elements - Vertebrate	conserved_region	SO:0000330	phastEConsV	Escore de Conservação em vertebrados (PhastCons)	Residue conservation in vertebrate (phastCons)
PhyloP Conserved Nucleotide - Primate	conserved_region	SO:0000330	phyloPP	Escore de Conservação em primatas (phyloP)	Residue conservation in primates (phyloP)
PhyloP Conserved Nucleotide - Mammals	conserved_region	SO:0000330	phyloPM	Escore de Conservação em mamíferos (phyloP)	Residue conservation in mammals (phyloP)
PhyloP Conserved Nucleotide - Vertebrate	conserved_region	SO:0000330	phyloPV	Escore de Conservação em vertebrados (phyloP)	Residue conservation in vertebrate (phyloP)
Non Synonymous	coding_variant_quality	SO:0001814	nSynPol	Variante nucleotídica que altera o aminoácido da sequência proteica	Nucleotide variant that alters amino-acid of a protein sequence
Exon	exon	SO:0000147	exonLoc	Localização - Éxon	Exon location
Consensus splice site	splice_site	SO:0000162	spliceLoc	Localização - Sítio de splice	Splice Site Location
5' UTR	5_prime_UTR_variant	SO:0001623	5Loc	Localização - 5' UTR	5' UTR location
3' UTR	3_prime_UTR_variant	SO:0001624	3Loc	Localização - 3' UTR	3' UTR location
SNPedia ID	-	-	snpediaId	Código de identificação SNPedia db	SNPedia db Id
SNPedia Entry	-	-	snpediaLink	Código de entrada para SNPedia	SNPedia Entry Code
CDNA Position	cDNA	SO:0000756	cdnaPos	Posição no DNA sintetizado a partir da transcriptase reversa, RNA molde	Position in DNA synthesized by reverse transcriptase using RNA as a template
Informaçã do Genótipo					
Genotype Value	genotype	SO:SO_0001027	genoValue	Valor da variante no genoma	Value of a variant in genome
Reference Allele	reference_variant	reference_variant	refAllele	Alelo referência. Utilizado para normalização dos dados	Reference Allele. Used to normalize allele calls. Methods to normalize can vary: functional

					relevance, frequency, strand
Other Allele	no_reference_variant	no_reference_variant	compAllele	Alelo complementar. Utilizado para normalização dos dados	Complementar Allele. Used to normalize allele calls. Methods to normalize can vary: same as above
Reference Homozygote Genotype	reference_homozygote	reference_homozygote	refGenotype	Genótipo Homozigoto referência	Homozygote for reference allele
Heterozygote	heterozygote	heterozygote	hetGenotype	Genótipo Heterozigoto	Usually two different alleles for a variant
Other Homozygote Genotype	no_reference_homozygote	no_reference_homozygote	compGenotype	Genótipo Homozigoto complementar	Homozygote for the other allele
NCBI reference allele	reference_variant	reference_variant	refAlleledbSnp	Alelo referência dbSNP	Reference allele at NCBI database
UCSC reference allele	reference_variant	reference_variant	refAlleleUcsc	Alelo referência UCSC	Reference allele at UCSC database
Allele Frequency	variant_frequency	SO:SO_0001763	freqAllele	Frequência alélica	Alleles Frequency
Genotype Frequency	-	-	freqGenotype	Frequência genotípica	Genotypic Frequency
Ancestral Allele	ancestral allele	transmed:TMO_0166	ancAllele	Alelo Ancestral	Ancestral Allele
Human alleles	-	-	hsAlleles	Alelos observados em humanos	Alleles observed in humam
Predominant human allele	-	-	predHsAlleles	Alelo predominante em humanos	Most frequent allele in humans
Heterozygosity	-	-	het	Valor de Heterozigosidade	Heterozigosity Value
Chimp allele	-	-	chimpAllele	Alelos observados em chimpanzés	Alleles Observed in Chimps
Macaque allele	-	-	macAllele	Alelos observados em Macaca	Alleles Observed in Macaca
dbSNP MAF	maf dbsnp	maf dbsnp	mafdbsnp	Alelo de Menor Frequência dbSNP	CEU Minor Allele Frequency
HapMap CEU MAF	maf	-	mafCeu	Alelo de Menor Frequência CEU	ASI Minor Allele Frequency
HapMap ASI MAF	maf	-	mafAsi	Alelo de Menor Frequência ASI	YRI Minor Allele

					Frequency
HapMap YRI MAF	maf	-	mafYri	Alelo de Menor Frequência YRI	dbSNP Minor Allele Frequency
HapMap Strand	strand hapmap	-	strand HapMap	Fita-molde para a genotipagem ou sequenciamento em HapMap.	Strand to sequencing or genotyping in HapMap
Damaging Allele	variant_phenotype	SO:SO_0001769	damageAllele	Alelo danoso de perda ou ganho de função	Dangerous allele by loss or gain of function
Damaging Allele Polyphen	variant_phenotype	SO:SO_0001769	damageAlPphen	Alelo danoso de perda ou ganho de função	Dangerous allele by loss or gain of function
Damaging Allele SIFT	variant_phenotype	SO:SO_0001769	damageAlSift	Alelo danoso de perda ou ganho de função	Dangerous allele by loss or gain of function
Haplótipo					
Haplotype ID	haplotype	haplotype	haploId	Códido do haplótipo	Haplotype Code
Diplotype	diplotype	SO:SO_0001028	diplotype	Par de haplótipos	Pair of haplotype
Haplotype Value	-	-	haploVal	Valor do haplótipo	Haplotype Value
Software Method	software	swo:SWO_0000001	haploMethod	Método de Inferência (Software)	Inference Method (Software)
Gene					
Name	Gene name	data_2299	geneName	Nome completo do gene	Gene Name
HGNC Gene Name	Gene name (HGNC)	data_1791	geneName	Nome oficial do gene (HGNC)	Official gene name (HGNC)
Symbol	Gene symbol	data_1026	geneSymbol	Símbolo do gene (HGNC)	Gene Symbol (HGNC)
Gene Synonyms	-	-	geneSyns	Nomes Sinônimos	Nomes Sinônimos
Alternative Symbol	-	-	altGeneSymbols	Símbolos Alternativos	Alternative Symbols
Previous Names	-	-	prevGeneNames	Nomes Anteriores	Previous Names
Gene ID	Gene ID	data_2295	geneId	Código do gene dbSNP	dbSNP Gene Code
HGNC Gene Id	Gene ID (HGNC)	data_2298	hugoId	Código do gene HGNC	HGNC Gene Code
Entrez Id	Gene ID (NCBI)	data_1027	entrezGeneId	Código do gene NCBI	NCBI Gene Code
Ensembl Id	Gene ID (Ensembl)	data_1033	ensemblId	Código do gene Ensembl	Ensembl Gene Code
UniProt Gene Id	UniProt accession	data_3021	uniProtGeneId	Código do gene UniProt	UniProt Gene Code
CCDS Id	-	-	ccdsId	Código do gene Consensus CDS	Consensus CDS Gene

					Code
UCSC Id	-	-	ucscGeneId	Código do gene UCSC	UCSC Gene Code
Vega Id	-	-	vegaId	Código do gene Vega	Vega Gene Code
Mouse Genome DB Id	-	-	mouseGdbId	Código do gene Mouse Genome db	Mouse Genome db Gene Id
Rat Genome DB Id	-	-	ratGdbId	Código do gene Rat Genome db	Rat Genome db Gene Id
Gene Type	Gene type	37717001	geneType	Tipo (codificante, pseudogene, RNAs, elementos transponíveis, retrovirus, complexo, etc)	Type (coding, pseudogene, RNAs, transposable elements, retrovirus, complex, etc)
Gene Group	gene_group	SO:SO_0005855	geneGroup	Grupo (codificante, pseudogene, RNA não codificante)	Group (coding, pseudogene, Non-Coding RNA)
Gene Tag	-	-	geneTag	Etiqueta da família gênica	Gene Family Tag
Gene Family	Gene Family	C26004	geneFamily	Família Gênica	A set of genes coding for diverse proteins which, by virtue of their high degree of sequence similarity, are believed to have evolved from a single ancestral gene.
Gene Sequence	Nucleic acid features (coding sequence)	data_1313	geneSeq	Sequência Gênica	Gene Sequence
Gene Status	-	-	geneStatus	Grau de confiabilidade no qual o assinalamento de loci ao cromossomo ou entre loci foi estabelecido	Certainty with which assignment of loci to chromosome or to linkage between 2 loci has been established

Assembly Build Version	-	-	assmBV	Versão de montagem do Genoma	Genome Assembly Build Version
Assembly Coord Start	-	-	assmCStart	Posição de início do polimorfismo de acordo com a versão do genoma	Start position of variant according to reference human genome assembly
Assembly Coord End	-	-	assmCEnd	Posição de término do polimorfismo de acordo com a versão do genoma	End position of variant according to reference human genome assembly
Translational Regulator	translation_regulatory_region	SO:0001680	translReg	Elementos reguladores da tradução	Elements involved in the control of the process of translation.
Transcription Regulator	transcription_regulatory_region	SO:0001679	transcReg	Elementos reguladores da transcrição	Elements involved in the control of the process of transcription.
Transcription Factor	transcription_factor	-	transcFactor	Fatores de Transcrição (associam-se à regiões regulatórias envolvidads no controle da transcrição)	Factors that bind a regulatory region involved in the control of the process of transcription.
Transcript Factor Binding Sites	TF_binding_site	SO:SO_0000235	tfbsSite	Sítios de ligação de fatores de transcrição.	Transcript Factor Binding Sites
Molecular Function	molecular_function	GO:0003674	molfunction	Função do produto gênico, habilidades	The functions of a gene product are the jobs that it does or the "abilities" that it has.
Cellular Component	cellular_component	GO:0005575	celComp	Localização efetora ou constitutiva do produto gênico	The cellular component ontology describes locations,

					at the levels of subcellular structures and macromolecular complexes.
Biological Process	biological_process	GO:0008150	bioProcess	Série de eventos ou funções moleculares	A biological process is a recognized series of events or molecular functions. A process is a collection of molecular events with a defined beginning and end.
Promoter region	promotor	promotor	promoter	Região promotora	Região promotora
Frameshift	frameshift	SO:0000865	frameshift	Mudança da janela de leitura (Predito)	Mudança da janela de leitura (Predito)
Expression	expression profiling	GRO:ExpressionProfiling	expression	Nível de expressão gênica	Nível de expressão gênica
Gene Map Method	-	-	geneMapMet	Método de mapeamento do gene	Gene Mapping Method
Sequência referência					
Reference Sequence ID	RefSeq accession	data_1098	refSeqId	Código da sequência referência	Código da sequência referência
Reference Sequence Name	EMBL/GenBank/DDBJ ID	data_1103	refSeqName	Nome da sequência referência	Nome da sequência referência
Cytogenetic Band	Cytogenetic map	data_1283	cytoLoc	Localização (Banda Citogenética)	Localização (Banda Citogenética)
Contig	contig	SO:SO_0000149	contig	Código do contig (conjunto de segmentos sobrepostos de DNA que unidos representam uma região consenso de DNA)	Código do contig (conjunto de segmentos sobrepostos de DNA que unidos representam uma região consenso de DNA)
Assembly Build Version	sequence_assembly	SO:SO_000053	assmBV	Versão de montagem do Genoma	Genome Assembly

					Build Version
Assembly Coord Start	-	-	assmCStart	Posição de início do polimorfismo de acordo com a versão do genoma	Start position of variant according to reference human genome assembly
Assembly Coord End	-	-	assmCEnd	Posição de término do polimorfismo de acordo com a versão do genoma	End position of variant according to reference human genome assembly
Nucleotide numbering genomic Ref Seq	Sequence position	data_1016	coorRelSeq	Posição de uma sequência contígua que incia-se e inclui um códon de iniciação e um de parada. De acordo com a posição física.	Position in a contiguous sequence which begins with, and includes, a start codon and ends with, and includes, a stop codon.According physical reference.
Gene transcript	transcript	SO:0000673	mRNA	Código do transcrito	Transcript Id
Enhancer	enhancer	SO:0000165	enhancer	Predição de amplificadores	Enhancers Prediction
Enhancer Binding Sites	enhancer_binding_site	SO:SO_0001461	enhBinSites	Sítios de ligação de amplificadores de expressão.	Expression enhnacer binding sites
Nearby Genes(KB distance)	-	-	linkDistGenes	Genes próximos ou ligados fisicamente (distância em Kb)	Nearby or Linked genes (Kb distance)
Disease-causing region?	disease_causing_variant	SO:SO_0001772	damageReg	Regiões relacionadas ao desenvolvimento de doenças	Disease-causing region
CpG Islands	CpG_island	SO:0000307	cpgIslands	Localização de ilhas de CpG	CpG Islands location
População					
Population ID	Population Group	C17005	popId	Código da população	Population Code
Population Name	Ethnic Group	C16564	popName	Nome da população	Population Name
Geographic Origin	Geographic Area	C16632	popOrigin	Origem Biogeográfica da população	Biogeographic Origin of

					Population
Country	Country	C25464	popLoc	Localização da população	Population Location
Coordinates	Coordinate	C44465	popCoords	Coordenadas da população amostrada	Sample population coordinates
Fenótipo					
Phenotype Name	phenotypic variability	HP:0003812	phenotype	Nome do fenótipo expresso	Name of the phenotypic expression
Affected Status	-	-	phenoStatus	Caracterização do indivíduo quanto à expressão do fenótipo	Caracterização do indivíduo quanto à expressão do fenótipo
Inheritance	Mode of inheritance	HP:0000005		Tipo de herança genética	Genetic Inheritance type
Quantitative	quantitative_variant	-	phenoValue	Valor quantitativo do fenótipo	Quantitative Value of phenotype
Qualitative	-	-	phenoValue	Valor qualitativo do fenótipo	Qualitative Value of phenotype
Disease	Disease or Disorder	C2991	disease	Nome da doença (fenótipos malignos)	Disease Name (malign phenotypes)
Disease Class	-	-	diseaseClass	Classificação da doença	Disease Classification
PharmGKB Disease	Disease ID (PharmGKB)	data_2651	pgkbDid	Identificador de doenças do PGKB	Identifier of a disease from the pharmacogenetics and pharmacogenomics knowledge base (PharmGKB).
OMIM	OMIM ID	data_1153	mimId	Código do fenótipo no banco OMIM	
MIM Morbid	-	-	mimMorbid	Classificação gênica da doença segundo o OMIM	OMIM gene clasifictation of disease
Phenotype Mapping Method	-	-	phenoMapMet	Método de mapeamento do fenótipo	Phenotyping Mapping Code
Mesh Terms	Disease	D004194	meshTerms	Termos MeSH (Medical Subject Headings - controle de vocabulário para indexação de artigos no PubMed) para doença.	MeSH Terms (Medical Subject Headings -

					controlled vocabulary thesaurus used for indexing articles for PubMed)
GAD db	-	-	gadId	Fenótipos associados à uma variante no banco de dados GAD (Genetic Association Database)	Phenotypes associated with a variant in GAD database (Genetic Association Database)
COSMIC db	-	-	cosmicId	Fenótipos associados à uma variante no banco de dados COSMIC (Catologue of Somatic Mutations in Cancer Database)	Phenotypes associated with a variant in COSMIC database (Catalogue Of Somatic Mutations In Cancer)
NHGRI db	National Human Genome Research Institute	C82617	nhgriId	Fenótipos associados à uma variante no banco de dados NHGRI (National Humam Genome Research Insitute)	Phenotypes associated with a variant in NHGRI gwas database
GWASdb	gwas_trait	-	gwasdbId	Fenótipos associados à uma variante no banco de dados GWASdb	Phenotypes associated with a variant in GWASdb database (GWAS database)
Fármacos e xenobióticos					
Drug	drug	SOPHARM:SOPHARM_20000	drug	Drogas metabolizadas pela variante	Drugs metabolized by variant
PharmGKB Drug Id	Drug ID (PharmGKB)	data_2652	drug	ID de drogas metabolizadas pela variante	Drug identifier
Vias Metabólicas					
Metabolic Pathway	classic_metabolic_pathway	PW:0000002	pathway	Nome da via metabólica	Metabolic Pathway name
PharmGKB Pathway Id	Pathway ID (PharmGKB)	data_2650	pathPid	Identificador da via no banco PGKB	Pathway identifier in PGKB
mRNA					
mRNA	mRNA	SO:0000234	mRnaId	Código de Identificação do mRNA	mRNA Accession

					Id
mRNA ID	mRNA_contig	SO:SO_0001829	mRnaId	Código de Identificação do mRNA	mRNA Accession Id
mRNA Version	alternatively_spliced_transcript	SO:1001187	mRnaVer	Versão do mRNA (splice alternativo)	Version of mRNA (alternative splice)
Micro RNA					
miRNA	miRNA	SO:0000276	miRnaId	Micro RNA	Micro RNA
miRNA target	miRNA Target Site	obo:SO_0000934	miRnaTarget	Alvo de Micro RNA	Micro RNA Targets
snoRNAs	snoRNA	SO:0000275	snoRnas	Pequeno RNA nucleolar	Small Nucleolar RNA
scaRNAs	Cajal body	GO:0015030	scaRnas	Pequeno RNA do tipo Cajal	Small Cajal RNA
CNVs e Variantes Estruturais					
CNV ID	copy_number_variation	SO:0001019	cnvId	Código do CNV	CNV code
CNV name	copy_number_variation	SO:0001019	cnvName	Variante relacionada ao aumento ou diminuição do número de cópias em dada região.	A variation that increases or decreases the copy number of a given region.
Number of copies	copy_number_change	SO:0001563	cnvNumber	Variante de sequência onde cópias de trechos estão ou aumentadas ou diminuídas.	A sequence variant where copies of a feature (CNV) are either increased or decreased.
Assembly Build Version	assembly	SO:0001248	assmBV	Versão de montagem do Genoma	Genome Assembly Build Version
Assembly Coord Start	-	-	assmCStart	Posição de início do polimorfismo de acordo com a versão do genoma	Start position of variant according to reference human genome assembly
Assembly Coord End	-	-	assmCEnd	Posição de término do polimorfismo de acordo com a versão do genoma	End position of variant according to reference human genome assembly

CNV Gene Expression	expression profiling	GRO:ExpressionProfiling	cnvGeneExp	Uso de métodos de alto desempenho para avaliar o nível e o momento da expressão gênica em uma amostra biológica.	The use of high-throughput methods (e.g. DNA microarrays) for evaluating the level and timing of gene expression in a biologic sample (a cell or tissue).
Segmental duplication	Segmental Duplications, Genomic	D056916	segDup	Duplicação segmental	Low-copy (2-50) repetitive DNA elements that are highly homologous and range in size from 1000 to 400,000 base pairs.
Proteína					
Protein Name			protein Name	Nome da proteína	Protein Name
Protein ID			protein Id	Código de identificação da proteína	Protein Id
Ensembl Id	Protein ID (Ensembl)	data_2398	protein EnsId	Código de identificação da proteína	Protein Id
Enzyme Name			enzyme Name	Nome da enzima	Enzyme Name
Enzyme Id			enzyme Id	Código da enzima no Enzyme db	Enzyme db Id
AminoAcid Position	amino acid sequence position	transmed:TMO_0123	aaPos	Posição na cadeia de aminoácidos	Position in Amino Acid chain
AminoAcid Change	amino_acid_substitution	SO:SO_0001606	aaChange	Posição de troca na cadeia de aa	Change of AminoAcid caused by mutation
Splice Distance	splicing_variant	SO:0001568	spliceDist	Distância da mutação em relação ao sítio de splice	Distance of mutation relative to splice site
Reference AminoAcid			aaReference	Aminoácido referência. Usado para normalizar o resíduo aminoacídico (frequência, relevância funcional)	Reference aminoacid. Used to normalize aminoacid residue. Methods to normalize can vary: functional relevance,

					frequency.
Variant AminoAcid			aaVariant	Aminoácido variante. Usado para normalizar o resíduo aminoacídico (frequência, relevância funcional)	Variant aminoacid. Used to normalize aminoacid residue. Methods to normalize can vary: functional relevance, frequency.
Codon	codon	SO:0000360	codon	Códon	Codon
Peptide	polypeptide	SO:0000104	peptide	Sequência Peptídica	Peptide
Evolução Biológica					
Fst	Population genetics	topic_3056	fstVal	Valor do cálculo de fixação entre populações quanto à amostragem total.	Statistics of population differentiation due to genetic structure.
XP-EHH	Population genetics	topic_3056	xpEhhVal	Valor da estatística XP-EHH	Cross Population Extended Haplotype Homozogysity (XP-EHH) Value
iHS	Population genetics	topic_3056	ihsVal	Valor da estatística iHS	iHS Value
Haplogroups	-	-	haploG		
Outros					
Out Link	Links List	C19470	link	Links para outros endereços	Out links
Description/Comments	Comment	C25393	description	Descrição	Description
Source database	Database name	data_1056		Banco de dados fonte	Source Database
Source database Id	Database identifier	data_1048	extdb	Códigos para bancos de dados fonte	Source Databases Ids
Referências Bibliográficas					
Title of article	Publication Name	C93639	litTitle	Título do artigo	Article Title
Pubmed ID	PubMed ID	data_1187	pmid	Código de identificação do artigo no PubMed	PubMed Id
Bibliographic References	Literature	D008091	literature	Referências bibliográficas	Bibliographic References
Desequilíbrio de Ligação (DL)					
r2	linkage_group	-	ldCorrelation	Valor do coeficiente de correlação entre duas entidades genômicas	Value of correlation coeffcient within distinct

					genomic entities.
D'	-	-	ldprima	Valor do desequilíbrio de ligação entre duas entidades genômicas	Value of linkage disequilibrium within distinct genomic entities.
Proxy SNP	snp_captured_by_proxy	snp_captured_by_proxy	snpProxy	Polimorfismo ligado à variante interrogada	Linked polymorphisms in respect to query variant
Proxy Gene	-	-	geneProxy	Gene ligado à variante interrogada	Linked genes in respect to query variant
Distance	-	-	ldDist	Distância entre duas entidades genômicas ligadas (correlacionadas)	Distance within two genomic entities (correlated)
Recombination Rate	-	-	recombRate	Taxa de recombinação entre duas entidades genômicas	Recombination Rate within genomic entities
Genetic Map Distance	-	-	geneMapDist	Distância calcula através de mapas gênicos	Distance calculated by genomic maps
Genetic Map Position	-	-	geneMapPos	Distância calcula através da distância física	Distance calculated by physical distance
Tag SNP	tag_snp	tag_snp	tagSnp	SNP representativo de outros em uma região de alto desequilíbrio de ligação.	Representative single nucleotide polymorphism (SNP) in a region of the genome with high linkage disequilibrium
Bases de dados					
dbsnp	dbsnp	dbsnp	dbsnp	dbSNP é um banco de dados do NCBI.	dbSNP is the SNP database of the NCBI. Any individual is a particular version of dbSNP.

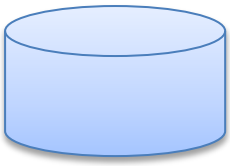
Hapmap	hapmap	hapmap	hapmap	Rescurso público que irá auxiliar pesquisadores a encontrar genes associados a doenças humanas e a resposta farmacêutica através da disponibilização de mapas de haplótipos.	A public resource that will help researchers find genes associated with human disease and response to pharmaceuticals by providing data about human haplotypes.
HGVBase	hgvbase	hgvbase	hgvBase	Catálogo de genes humanos e variação genômica. Útil como ferramenta de pesquisa para auxiliar na definição dos componentes genéticos de variação fenotípica.	A catalog of normal human gene and genome variation, useful as a research tool to help define the genetic component of human phenotypic variation.
OMIM	omim	omim	omim	Catálogo de genes humanos e desordens genéticas. Incorpora variações alélicas envolvidas em doenças.	This database is a catalog of human genes and genetic disorders. It embeds information about allelic variations involved in this disorders.
PharmGKB	pharmgkb	pharmgkb	pgkb	PharmGKB cura informações que estabelecem conhecimento sobre as relaçõese entre fármacos, doenças e genes, incluindo suas variações e produtos gênicos.	PharmGKB curates information that establishes knowledge about the relationships among drugs, diseases and genes, including their

					variations and gene products.

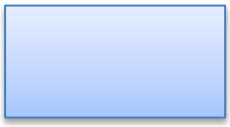
Ontologias Utilizadas no mapeamento
SNP Ontology
Suggested Ontology for Pharmacogenomics
Human Disease Ontology
Pathway Ontology
Gene Regulation Ontology
Human Phenotypic Ontology
Gene Ontology
NCI Thesaurus
Sequence Types and Features
Bioinformatics operations, types of data, formats, and topics

ANEXO A – SIMBOLOGIA DOS FLUXOGRAMAS

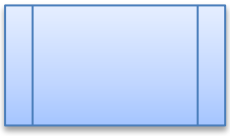
Alguns símbolos utilizados neste trabalho são detalhadas abaixo para melhor compreensão dos fluxogramas deste estudo.



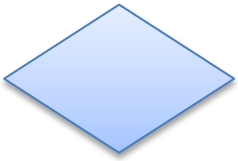
Cilindros representam arquivos ou bases de dados, ou seja, as estruturas de armazenamento de dados.



Utilizado para informar uma ação, tarefa, processo ou operação. Indica a ação que será praticada.



Indica um conjunto de processos ou ações pré-determinado, geralmente indicado em outro local. Em algoritmos, corresponde à uma subrotina.



Processo decisório. Indica as possibilidades de seguimento de um fluxo a partir de uma decisão a ser tomada, sendo que cada seta corresponderá a uma ação.

Documento ou relatório.

Indica o fluxo das ações ou tarefas.

Indica fluxos de ações ou tarefas alternativos.

Fontes: (BOHL; RYNN, 2007; RF FLOW 5, 2014)

ANEXO B – DISTRIBUIÇÃO DAS POPULAÇÕES E SUBPOPULAÇÕES DO HGDP POR: GRUPO POPULACIONAL, REGIÃO GEOGRÁFICA E GRUPO LINGUÍSTICOS.

Grupo Populacional	População	Localização	Grupo Linguístico[a]	N de Indivíduos
África Ocidental	Bantu NE.	Kênia	Nigero-Congolesa	11
	Bantu SE. Pedi	África do Sul	Nigero-Congolesa	1
	Bantu SE. Sotho	África do Sul	Nigero-Congolesa	1
	Bantu SE. Tswana	África do Sul	Nigero-Congolesa	2
	Bantu SE. Zulu	África do Sul	Nigero-Congolesa	1
	Bantu SO. Herero	África do Sul	Nigero-Congolesa	2
	Bantu SO. Ovambo	África do Sul	Nigero-Congolesa	1
	Iorubá	Nigéria	Nigero-Congolesa	25
	Mandenka	Senegal	Nigero-Congolesa	24
África Oriental	Pigmeus Biaka	África Central	Nigero-Congolesa	31
	Pigmeus Mbuti	República Dem. Congo	Nilo-Saariana	13
	San	Namíbia	Coisã	7
América Central	Maia	México	Maia	25
	Pima	México	Azteca-Tanoano	24
América do Sul	Cayapa	Equador	Barbacoano	7
	Karitiana	Brasil	Equatorial-Tucano	23
	Matsiguenga	Peru	Equatorial-Tucano	21
	Piapoco e Curripaco	Colômbia	Equatorial-Tucano	13
	Quechua	Andes Centrais	Andino	22
	San Martin	Peru	Andino	17
	Suruí	Brasil	Equatorial-Tucano	21
Centro Sul Asiático	Balochi	Paquistão	Indo-Europeu	25
	Brahui	Paquistão	Dravídica	25
	Burusho	Paquistão	Isolado	25
	Hazara	Paquistão	Indo-Europeu	25
	Kalash	Paquistão	Indo-Europeu	25
	Makrani	Paquistão	Indo-Europeu	25
	Pathan	Paquistão	Indo-Europeu	24
	Sindhi	Paquistão	Indo-Europeu	24
Europa	Adygei	Cáucaso Russo	Circassiana	15
	Bergamo	Itália	Indo-Europeu	13
	Franceses	França	Indo-Europeu	29
	Bascos Franceses	França	Basco	24
	Orcadiana	Ilhas Orkney	Indo-Europeu	16
	Russos	Rússia	Indo-Europeu	25
	Sardenha	Itália	Indo-Europeu	28
	Toscanos	Itália	Indo-Europeu	8

Leste Asiático	Cambojanos	Camboja	Austro-Asiático	10
	Dai	China	Tai-Kadai	10
	Daur	China	Altaico	10
	Han	China	Sino-tibetano	39
	Hezhen	China	Altaico	9
	Japoneses	Japão	Japônico	30
	Lahu	China	Sino-tibetano	10
	Miaozu	China	Austro-Asiático	10
	Mongóis	China	Altaico	10
	Naxi	China	Sino-tibetano	10
	Oroqen	China	Altaico	10
	She	China	Austro-Asiático	10
	Tu	China	Altaico	10
	Tujia	China	Sino-tibetano	10
	Uigur	China	Altaico	10
	Xibo	China	Altaico	9
	Yakut	Sibéria	Altaico (?)	25
	Yizu	China	Sino-tibetano	10
Oceania	NAN Melanésia	Bougainville	Proto-Oceanico	15
	Papua	Nova Guiné	Bougainville Sul	17
Oriente Médio	Beduínos	Israel (Negev)	Afro-asiático	48
	Drusos	Israel (Carmel)	Afro-asiático	47
	Mozabite	Algéria (Mzab)	Afro-asiático	30
	Palestinos	Israel (Central)	Afro-asiático	49

[a] Os grupos linguísticos listados estão representados por filos ou famílias linguísticas. Foi escolhido o maior nível de classificação visando melhor representar descontinuidades linguísticas nos grupos populacionais.
(?) Classificação amplamente discutida.
Classificação linguística de acordo com Ethnologue (www.ethnologue.com).
Modificada de (SOARES-SOUZA, 2010)

ANEXO C – DISTRIBUIÇÃO GEOGRÁFICA APROXIMADA DAS POPULAÇÕES DO HGDP-CEPH E DAS QUATRO POPULAÇÕES NATIVO AMERICANAS DO PERU E EQUADOR DE NOSSO LABORATÓRIO.

Figura 17: Distribuição geográfica aproximada das populações do HGDP-CEPH e das quatro populações Nativo Americanas do Peru e Equador de nosso laboratório e dos grupos populacionais, cujos dados estão disponíveis para esse estudo.

1 Bantu NE e SE/SO; **2** Mandenka; **3** Ioruba; **4** San; **5** Pigmeu Mbuti; **6** Pigmeu Biaka; **7** Mozabite; **8** Orcadiana; **9** Adygei; **10** Russa NO; **11** Francesa Basca; **12** Francesa; **13** Bérgamo; **14** Sardenha; **15** Toscana; **16** Beduína; **17** Drusa; **18** Palestina; **19** Balochi; **20** Brahui; **21** Makrani; **22** Sindhi; **23** Pathan; **24** Burusho; **25** Hazara; **26** Uigur; **27** Kalash; **28** Han (China S); **29** Han (China N); **30** Dai. **31** Daur; **32** Hezhen; **33** Lahu; **34** Miaozu (Miao); **35** Oroqen; **36** She; **37** Tujia; **38** Tu; **39** Xibo; **40** Yizu (Yi); **41** Mongólia; **42** Naxi; **43** Camboja; **44** Japonesa; **45** Yakut; **46** Melanésia; **47** Papua; **48** Karitiana; **49** Suruí; **50** Piapoco e Curripaco; **51** Maia; **52** Pima; **53*** Populações Nativas do Peru e Equador (**53-a** Cayapa, **53-b** Quechua, **53-c** San Martín e **53-d** Matsiguenga) **A** Populações do *SNP500Cancer* (Ascendência Caucasiana; Ascendência Asiática; Hispânicos e Afro-americanos.

Grupos Populacionais: África Ocidental (vermelho); África Oriental (amarelo); Oriente Médio (azul escuro); Europa (rosa); Centro Sul Asiático (azul claro); Leste Asiático (roxo); Oceania (laranja); América Central (preto); América do Sul (marrom).

Nativo-americanos correspondem ao conjunto das populações dos grupos América Central (círculos pretos) e América do Sul (círculos marrons). Nordeste Asiático corresponde ao conjunto formado pelas populações Daur (31), Hezhen (32) e Oroqen (35).