UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Genômica e miscigenação nos contextos biomédico e evolutivo

AUTOR: Rennan Garcias Moreira

Belo Horizonte

2019

Rennan Garcias Moreira

Genômica e miscigenação

nos contextos biomédico e evolutivo

Versão final

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Doutor em Bioinformática.

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos

BELO HORIZONTE
2019

043 Moreira, Rennan Garcias.
Genômica e miscigenação nos contextos biomédico e evolutivo [manuscrito] / Rennan Garcias Moreira. – 2019.

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

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Bioinformática - Teses. 2. Genômica. 3. Leucemia. 4. Recidiva. 5. Miscigenação. I. Santos, Eduardo Martín Tarazona. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada por Cleide Vieira de Faria – CRB 6/2314 – BC/UFMG

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

ATA DA DEFESA DE TESE

Rennan Garcias Moreira

114/2019 entrada 1°/2015 CPF: 059.401.386-08

Às oito horas e trinta minutos do dia **03 de dezembro de 2019**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"GENOMICA E MISCIGENAÇÃO NOS CONTEXTOS BIOMÉDICO E EVOLUTIVO"**, 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	CPF	Indicação
Dr. Eduardo Martin Tarazona Santos	UFMG	012949056 02	APROVADO
Dr. Dani Gamerman	UFMG	43124917715	APROVADO
Dra. Glória Regina Franco	UFMG	623.387.496-34	APROVADO
Dr. Leandro Machado Colli	USP	306.691.698-70	APROVADO
Dra. Maria Cátira Bortolini	UFRGS	387044550 53	Aprovado
Dr. Marcelo Rizzatti Luizon	UFMG	27730818892	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, 03 de dezembro de 2019.

Dr. Eduardo Martin Tarazona Santos - Orientador _____

Dr. Dani Gamerman _____

Dra. Glória Regina Franco _____

Dr. Leandro Machado Colli _____

Dra. Maria Cátira Bortolini _____

Dr. Marcelo Rizzatti Luizon _____

Agradecimentos

Especiais agradecimentos,

Primeiramente a Deus, pela saúde, pela vida, enfim, por tudo.

À minha esposa, pelo amor, cumplicidade e suporte emocional. Pelo privilégio da convivência, pelo apoio incondicional, e especialmente pelos exemplos de caráter, profissionalismo e luta contínua e irrestrita em busca dos sonhos. Essa tese também é resultado do seu esforço.

A minha família: pai, mãe, irmãs, cunhado, pela compreensão, carinho, afeto diário e apoio emocional em todos os sentidos, essenciais para o alcance de todos objetivos. Também à querida família que compartilho com minha esposa: Antonio, Mary, Leandro, Lília, Pedro, Denise, Lauro.

Ao prof. Dr. Eduardo Tarazona Santos, não apenas pela orientação, mas pelo suporte contínuo e apoio incondicional ao meu crescimento profissional e pessoal durante todos os anos de convivência. Agradeço pelo excepcional ambiente de trabalho, consequência da sua dedicação profissional exemplar.

À prof. Dra. Glória Franco, também pelo apoio incondicional ao meu crescimento profissional, colaborações, parcerias e total disponibilidade em me receber sempre.

Particularmente a ambos, Eduardo Tarazona e Glória Franco, pelos exemplos de excelência profissional e de caráter, e pelo aprendizado constante especialmente em valores que transcendem o ambiente de trabalho.

Aos amigos dos laboratórios do Instituto, em especial do LDGH e LGB. Referência destacada à querida amiga Moara Machado. Também a profa. Fernanda Soares, quem contribuiu bastante para um dos capítulos da tese.

Aos demais professores e funcionários do Instituto, com especial apreço aos professores Marcelo Luizon, Francisco Lobo, Carolina Gomes e Ricardo Gomes, pela disponibilidade, parcerias profissionais e especialmente pela gentileza permanente na convivência diária. Levo-os também como exemplos no campo pessoal e profissional.

Ao Programa de Pós-Graduação em Bioinformática, por oferecer oportunidades e ambiente acadêmico de excelência para o desenvolvimento dos alunos.

Ao Prof. Jeffrey Kidd, de Michigan/EUA, pela oportunidade disponibilizada e excepcional aprendizado proporcionado. Também aos amigos do laboratório, Amanda, Feichen, Sarah, Jing.

Aos familiares e demais amigos externos à UFMG, pela convivência e apoio em todos os momentos.

Enfim, a todos que contribuíram durante esse projeto profissional e de vida.

Resumo

Essa tese traz contribuições ao progresso de diferentes áreas da Genômica e da Bioinformática em diversas dimensões. Estudos que compreendem desde a geração e análise do dado genético até a sua aplicação com propósitos biomédicos e evolutivos são discutidos. São apresentadas contribuições para a geração de dados genômicos de qualidade, as consequências sofridas por populações negligenciadas no campo biomédico, tendo em vista a desconsideração da diversidade genética humana populacional, além de perspectivas para a compreensão dos eventos de adaptação poligênica utilizando populações miscigenadas como modelo.

Os resultados apresentados no primeiro capítulo demonstram a importância do processo de geração de dados genéticos, com enfoque em diferentes etapas, desde a manipulação do material biológico até a determinação das variantes. A relevância da escolha dos métodos e a necessidade da utilização de ferramentas computacionais adequadas para lidar com as particularidades do ensaio biológico são abordadas. As análises indicam divergências de acordo com os métodos de chamada de variantes utilizados e destacam os principais aspectos a serem observados no estabelecimento de processos customizados de geração e análise de dados.

Parte dos dados gerados com os métodos do primeiro capítulo contribuiu para o estudo genético-populacional realizado no capítulo seguinte. Considerando evidências reportadas na literatura científica que suportam a existência de associação entre marcadores moleculares dos genes *PDE4B* e *MYT1L*, a ancestralidade nativo-americanana e a recidiva de leucemia linfoblástica aguda (LLA), a diversidade molecular desses dois genes foi analisada em populações miscigenadas e predominantemente nativo-americanas da América Latina. Um dos principais resultados indica que em várias populações há marcadores em desequilíbrio de ligação cujos alelos de risco para a recidiva da LLA possuem maior frequência quanto maior for a proporção de ancestralidade nativo-americana. A diversidade e estrutura genética são discutidas também no contexto da regulação gênica, já que podem ter consequências diretas sobre a resposta ao tratamento da LLA. Tendo em vista que os estudos são distorcidos quando a diversidade genética humana global não é apropriadamente considerada, esse capítulo discuti os resultados considerando os protocolos de tratamento da LLA aplicados em países da América Latina.

Os estudos de populações negligenciadas, em especial as populações com histórico de formação por eventos de miscigenação, não se beneficiam dos progressos na área da Genômica apenas no que diz respeito à melhor compreensão e aprofundamento das análises no campo biomédico. O avanço dos métodos e técnicas de análise também permitem que hipóteses relacionadas aos pilares teóricos da evolução das populações humanas sejam testadas no nível Genômico. O último capítulo apresenta perspectivas construídas a partir de um esforço inicial para utilizar as particularidades de populações miscigenadas na busca por sinais de seleção poligênica. As suaves alterações de frequências em locos múltiplos podem revelar a atuação de pressões seletivas antes não identificadas e a análise de populações miscigenadas foi explorada como um modelo para a aplicação de métodos que auxiliem na identificação desses sinais evolutivos.

Portanto, as principais contribuições dessa tese envolvem a geração de dados de qualidade e a aplicação de novas metodologias para o estudo de populações pouco referenciadas nos contextos biomédico e evolutivo. Assim, espera-se fornecer uma visão geral de como a Bioinformática permite lidar com a complexidade dos processos de construção do conhecimento biológico considerando os avanços metodológicos recentes.

Palavras-chave: Nativos-americanos, Leucemia, Recidiva, Seleção poligênica, Bioinformática

Abstract

This dissertation provides contributions to the progress of different areas of the Bioinformatics and Genomics fields in several dimensions. It presents studies lying on the generation and analysis of genetic data and its applicability for biomedical and evolutionary purposes. Readers will find contributions toward generation of good-quality genomic data, discussions on biomedical consequences suffered by neglected populations given underestimation of global human population genetic diversity, and perspectives for a better comprehension of polygenic adaptation in admixed populations.

Results presented in chapter one highlight the importance of genetic data generation, discussing steps ranging from the biological material handling to genetic variant calling. This chapter addresses the importance of choosing correct analytical methods and computational tools suitable to deal with the particularities of the biological assay. Analyses show differences in results according to variant calling strategies and point out issues to be considered when customizing processes for data generation and analyses.

Part of the dataset generated in the first chapter was used for population-genetics analyses performed in the second chapter. Given previously reported evidences supporting association among *PDE4B* and *MYT1L* molecular markers, native-american ancestry and acute lymphoblastic leukemia (ALL) relapse, molecular diversities of such genes were assessed in admixed and native-american populations from Latin America. One of the major findings shows linkage disequilibrium association in several populations between markers whose ALL-relapse risk-alleles frequencies are directly related with proportion of native-american ancestry. The genetic structure and diversities of such genes are analyzed and also discussed regarding potential regulatory functions, since they may directly affect ALL treatment outcome in native-americans. Several reports have highlighted uncompleteness of studies focusing on diseases comprehension and treatment outcome if not considering global human genetic diversity. Thus, this chapter contributes for fixing such distortion by analyzing and discussing results which may have consequences for ALL treatment protocols applied in Latin America countries.

Studies performed with neglected populations, especially those historically arising from admixture processes, do not benefit from advances in Genomics solely due to progresses in knowledge and better-quality analyses for biomedical purposes. Technical advances and methods improvements also allow testing hypotheses about human populations evolution in the broad context of Genomics. The last chapter presents perspectives after an initial effort toward using admixed populations particularities for seeking polygenic adaptative signals. Slight alleles frequencies shifts in multiple loci may reveal selection footprints not yet identified. Hence, admixed populations may be used as a model system for applying new methods suitable to point out natural selection polygenic signals not yet reported.

Thus, major discussions involve the generation of good-quality data and application of new methods to study populations underrepresented in scientific reports. One expect to provide an overview of how Bioinformatics aids in dealing with the high-complexity of building biological knowledge considering modern methods and technologies.

Keywords: Native-american, Leukemia, Relapse, Polygenic selection, Bioinformatics

LISTA DE ABREVIAÇÕES

RNA - *Ribonucleic Acid*

DNA - *Deoxyribonucleic Acid*

PCR – *Polymerase Chain Reaction*

GATK – *Genome Analysis Toolkit*

LDGH – Laboratório de Diversidade Genética Humana

WDL - *Workflow Description Language*

BWA - *Burrows-Wheeler Aligner*

uBAM – *unmapped Binary Alignment/Map*

QD – *Quality by Depth*

FS – *Fisher Strand*

SOR – *Strand Odds Ratio*

MQ – *Mapping Quality*

MQRankSum – *Mapping Quality Rank Sum Test*

QUAL – *Quality*

SNP – *Single Nucleotide Polymorphism*

VQSR - *Variant Quality Score Recalibration*

IGV - *Integrative Genomics Viewer*

MIA - marcadores informativos de ancestralidade

LLA - Leucemia Linfoblástica Aguda

IDH – Índice de Desenvolvimento Humano

BFM - Berlin-Frankfurt-Münster

DRM - doença residual mínima

AMP – *adenosine monophosphate*

CEDEFS – Centro de Documentação Eloy Ferreira da Silva

MAF – *minor allele frequency*

ENCODE - *Encyclopedia of DNA Elements*

AMOVA - *Analysis of Molecular Variance*

LISTA DE FIGURAS

CAPÍTULO 1

Figura 1. Principais etapas e procedimentos realizados na construção das bibliotecas com o kit Agilent Haloplex.----------25

Figura 2. Perfil do Bioanalyzer da distribuição dos tamanhos de fragmentos (pb) esperados da biblioteca Haloplex customizada.----------26

Figura 3. Etapas do pipeline do *Genome Analysis Toolkit* (GATK).----------31

Figura 4. Ilustração adaptada da distribuição da qualidade das bases ordenadas pela posição em todas as *reads* de uma amostra antes de qualquer tratamento, obtida pelo uso do software FastQC. Em verde: >Q28, em amarelo: >Q20, em vermelho: <Q20. Retângulos amarelos: percentis 25º-75º; linhas inferiores e superiores em cada posição: percentis 10° e 90° respectivamente. Linha azul: valor de qualidade médio em cada posição. Linhas em vermelho: mediana.----------35

Figura 5. Distribuição da cobertura das variantes inferidas pelo *SureCall*----------36

Figura 6. Fluxograma geral das etapas do processo de customização baseado no guia de boas práticas do *Genome Analysis Toolkit* (GATK)----------39

Figura 7. Ilustração adaptada da distribuição da qualidade das bases ordenadas pela posição em todas as *reads* de uma amostra após tratamento pelo programa Trimmomatic. Em verde: >Q28, em amarelo: >Q20, em vermelho: <Q20. Retângulos amarelos: percentis 25º-75º; linhas inferiores e superiores em cada posição: percentis 10° e 90° respectivamente. Linha azul: valor de qualidade médio em cada posição. Linhas em vermelho: mediana.----------40

Figura 8. Distribuições dos parâmetros utilizados como indicadores para o processo de filtragem de Indels. SOR: *StrandOddsRatio*, FS: *FisherStrand*, QD: *QualByDepth*, QUAL: *Quality*, ReadPosRankSum: *ReadPosRankSumTest*. Em vermelho, a distribuição obtida com os SNPs do dbSNP, já em azul a distribuição obtida com as variantes inferidas.----------41

Figura 9. Distribuições dos parâmetros utilizados como indicadores para o processo de filtragem de SNPs. SOR: *StrandOddsRatio*, FS: *FisherStrand*, QD: *QualByDepth*, QUAL: *Quality*, ReadPosRankSum: *ReadPosRankSumTest*. Em vermelho, a distribuição obtida com os SNPs do dbSNP, já em azul a distribuição obtida com as variantes inferidas.---------- ---------42

Figura 10. Distribuição da cobertura das variantes inferidas pelo *Genome Analysis Toolkit* (GATK)----------43

Figura 11. Fluxograma do processo de refinamento dos resultados obtidos a partir do *pipeline* customizado. i= número de *reads* independentes que suportam a variante; r = número total de reads que suportam a variante; os números 1, 2 e 3 indicam a primeira, segunda e terceira variante mais abundante.----------44

Figura 12. Diagramas de Venn indicando o número de SNPs concordantes entre os programas SureCall e *Genome Analysis Toolkit* (GATK) antes (A) e após (B) a aplicação da correção dos genótipos. A concordância das variantes do tipo indel inferidas pelo *SureCall* e o *Genome Analysis Toolkit* (C).----------45

CAPÍTULO 2

Figura 1. Ilustração das interações moleculares relacionadas às *PDE4* e o *AMP* cíclico em linfócitos neoplásicos e no microambiente tumoral. Adaptado de Cooney & Aguiar (2016). -68

Figura 2. Mapa das localidades e populações amostradas----------72

Figura 3. Número de amostras exclusivas e compartilhadas entre os bancos de dados TargetSeq, BeadXpress, TaqMan-rs6683977 e TaqMan-rs6683977+2.5M.----------74

Figura 4. Representação gráfica das proporções relativas de miscigenação inferidas pelo programa ADMIXTURE a partir de 88 maracadores informativos de ancestralidade para as populações do estudo: CEU (descendentes de europeus/EUA); YRI (*Yoruba*/Nigéria); ASH (*Ashaninka*/Peru); MAC (*Machiguenga*/Peru); AYM (*Aymara*/Peru); QUE (*Quechua*/Peru); MEX (*Taharumara* e *Huichol*/México); GUA (*Guarani*/Brasil); TUP (*Tupiniquim*/Brasil); AMG (Miscigenados/MG - Brasil). Cada barra vertical representa um indivíduo e indica a proporção de ancestralidade de acordo com as cores das populações parentais----------84

Figura 5. Frequências alélicas dos SNPs de risco para recidiva de LLA *PDE4B*-rs6683977.G (alelo ancestral) (vazado) e *MYT1L*-rs17039396.A (alelo derivado) (sólido) em função da proporção de ancestralidade das populações: AMG: Miscigenados Brasileiros, AMR: Hispânicos Miscigenados dos EUA/Latinos Americanos do 1000GP (populações PUR, CLM, MXL e PEL), ASH: *Ashaninka*, AYM: *Aymara*, GUA: *Guarani*, MAC: *Machiguenga*, MEX: (*Huichol* e *Tarahumara*), QUE: *Quechua*, TUP: *Tupiniquim*.---------- ----------88

Figura 6. Análise de componentes principais realizada com os SNPs genotipados nos genes *PDE4B* e *MYT1L*. PC1 = 12%, PC2 = 10%. NAT_BRA: Nativos Brasileiros; NAT_PERU: Nativos Peruanos; AMG_BRA: Miscigenados Brasileiros, EUR: Europeus (1000GP), AFR: Africanos (1000GP).----------91

Figura 7. Distribuição dos haplótipos (representados pelas cores) dos genes e suas frequências relativas (tamanho dos blocos haplotípicos) de cada gene relacionado à recidiva da LLA em cada grupo estudado. Cada linha do gráfico representa um haplótipo de um indivíduo. a) *PDE4B*; b) *MYT1L*. EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP).----------94

Figura 8. Desequilíbrio de ligação entre os polimorfismos do gene *MYT1L* nas populações peruanas (PERU), mexicanas (MEX), africanas (AFR) e europeias (EUR). Quadros sólidos indicam desequilíbrio (r^2>0,8), quadros em escala de cinza indicam escala gradual de intensidade de desequilíbrio, enquanto que quadros claros indicam alta probabilidade de recombinação. O retângulo em destaque indica a posição do SNP de interesse rs17039396.-- 99

Figura 9. Desequilíbrio de ligação entre os polimorfismos do gene *PDE4B* nas populações peruanas (PERU), mexicanas (MEX), africanas (AFR) e europeias (EUR). Quadros sólidos indicam desequilíbrio (r^2>0,8), quadros em escala de cinza indicam escala gradual de intensidade de desequilíbrio, enquanto que quadros claros indicam alta probabilidade de recombinação. O triângulo, o círculo, o retângulo e o triângulo invertido em destaques indicam a posição dos SNPs de interesse rs546784, rs524770, rs6683977 e rs641262, respectivamente. ----------100

Figura 10. Ilustração da região de 20Kb do gene *PDE4B* em formato UCSC *Genome Browser*. Posicionamento dos SNPs com F_{CT} > 0,12 em relação às regiões de modificação de histonas (**H3KMe1, H3K4Me3, H3K27Ac**) e aos resultados de modelagem para a presença ou ausência de marcas de cromatina (**ChromHMM** - Laranja: acentuador forte, Amarelo: acentuador fraco, Verde claro: transcrição fraca) em células linfoblastóide (**GM12878** - rosa) e de leucemia mielóide (**K562** - roxo); às regiões de sítios de ligação de elementos regulatórios em *cis*-

(clivagem por DNAse) e em *trans*- (ligação de fatores de transcrição - **Txn Factor ChIP**). a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151. As posições dos SNPs de interesse inicial (rs546784, rs524770, rs6683977, rs641262) estão em destaque em linhas verticais em azul claro.----------104

Figura 11. Enfoque na região do gene PDE4B (GRCh37, chr1:66.766.000 – 66.769.000) com evidências mais fortes de funçao reulatória em formato UCSC Genome Browser. Posicionamento dos SNPs com FCT > 0,12 (para com EUR – azul e EUR e AFR – vermelho) em relação às regiões de modificação de histonas (**H3KMe1, H3K4Me3, H3K27Ac**) e aos resultados de modelagem para a presença ou ausência de marcas de cromatina **(ChromHMM** - Laranja: acentuador forte, Amarelo: acentuador fraco, Verde claro: transcrição fraca) em células linfoblastóide (**GM12878** - rosa) e de leucemia mielóide (**K562** - roxo); às regiões de sítios de ligação de elementos regulatórios em cis- (clivagem por DNAse) e em trans- (ligação de fatores de transcrição - **Txn Factor ChIP**). a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151. As posições dos SNPs de interesse inicial (rs524770, rs6683977) estão em linhas verticais em azul claro----------105

Figura 12. Ilustração da região de 20Kb do gene *MYT1L* em formato UCSC *Genome Browser*. Posicionamento dos SNPs com F_{CT} > 0,12 em relação às regiões de modificação de histonas (**H3KMe1, H3K4Me3, H3K27Ac**) e aos resultados de modelagem para a presença ou ausência de marcas de cromatina (**ChromHMM** – não há sinais em evidência) em células linfoblastóide (**GM12878** - rosa) e de leucemia mielóide (**K562** - roxo); às regiões de sítios de ligação de elementos regulatórios em *cis*- (clivagem por DNAse) e em *trans*- (ligação de fatores de transcrição - **Txn Factor ChIP**). a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151. A posição do SNP de interesse inicial rs17039396 está em destaque por uma linha vertical em azul claro.----------106

CAPÍTULO 3

Figura 1. Ilustração de um dos arquivos *.bed* (*browser extensible data*) contendo informação da ancestralidade local por indivíduo. UND – ancestralidade não determinada; AFR – ancestralidade africana, EUR- ancestralidade europeia, NAT – ancestralidade nativo-americana. As ancestralidades são apresentados em genótipos, considerando os dois cromossomos homólogos. Arquivo texto separado por tabulações com quebra das linhas.- -156

Figura 2. Ilustração da estrutura do arquivo com os símbolos dos genes que compõem os grupos gênicos analisados. A primeira coluna traz o nome do grupo e a segunda o sítio eletrônico de origem. Arquivo texto separado por tabulações com quebra das linhas.---------- -157

Figura 3. Fluxograma com a indicação das etapas e processos executados no *pipeline* desenvolvido----------160

Figura 4. Ilustração da estrutura do arquivo após o mapeamento dos genes de acordo com as ancestralidades e posições genômicas. Primeira coluna indica o cromossomo, segunda coluna a posição inicial, a terceira coluna a posição final, a quarta coluna os genótipos de ancestralidade e da quinta coluna em diante os símbolos gênicos. Arquivo texto separado por tabulações com quebra das linhas.----------162

Figura 5. Exemplo de uma distribuição nula da ancestralidade europeia para um grupo gênico específico construída a partir de 10.000 permutações das posições gênicas----------164

Figura 6. Histograma exemplificativo dos p-valores gerado para o cálculo da taxa de falsos-positivos (FDR). População CLM e ancestralidade africana.----------165

Figura 7. Diagrama de Venn apresentando os números de grupos gênicos significativos concordantes e exclusivos entre as populações estudadas. Os números representam apenas a quantidade de grupos gênicos, sem considerar a condição de excesso ou escassez de ancestralidade. AFR: ancestralidade africana; EUR: ancestralidade europeia, NAT: ancestralidade nativo-americana. Populações do 1000GP: CLM: colombianos de Medellin, MXL: mexicanos de Los Angeles/EUA; PUR: porto-riquenhos. SAL: população de Salvador – Projeto EPIGEN----------167

Figura 8. Diagrama de Venn apresentando os números de grupos gênicos significativos concordantes e exclusivos entre as populações estudadas. Os resultados compreendem o número de grupos gênicos considerando a condição de excesso ou escassez de ancestralidade. AFR: ancestralidade africana; EUR: ancestralidade europeia, NAT: ancestralidade nativo-americana. Populações do 1000GP: CLM: colombianos de Medellin, MXL: mexicanos de Los Angeles/EUA; PUR: porto-riquenhos. SAL: população de Salvador – Projeto EPIGEN------168

Figura 9. Relação entre q-valor e o tamanho dos grupos gênicos (número de genes) analisados para a ancestralidade africana na população miscigenada de colombianos (CLM) ----------173

LISTA DE TABELAS

CAPÍTULO 1

Tabela 1. País de origem, origem étnica e número de amostras utilizadas na preparação de bibliotecas.----------23

Tabela 2. Lista dos genes e regiões com as respectivas posições selecionadas para o sequenciamento direcionado----------29

Tabela 3. Números de variantes e suas características obtidos com o *SureCall*----------36

Tabela 4. Números de variantes e suas características obtidos com o *pipeline* customizado do *Genome Analysis Toolkit* (GATK) antes de qualquer correção. Ts/Tv: transições/transversões.----------38

Tabela 5. Números de variantes e suas características obtidos com o *pipeline* customizado do *Genome Analysis Toolkit* (GATK) após correções----------44

CAPÍTULO 2

Tabela 1. Países, populações, ancestralidades, bancos e números de amostras----------73

Tabela 2. Relação das variantes selecionados com a indicação dos genes onde estão localizados e o número de variantes cobertos em razão do desequilíbrio de ligação.----------76

Tabela 3. Proporções de ancestralidade por população obtidas a partir da média das ancestralidades individuais.. CEU (descendentes de europeus/EUA); YRI (*Yoruba*/Nigéria); ASH (*Ashaninka*/Peru); MAC (*Machiguenga*/Peru); AYM (*Aymara*/Peru); QUE (*Quechua*/Peru); TAH (*Taharumara*/México); HUI (*Huichol*/México); GUA (*Guarani*/Brasil); TUP (*Tupiniquim*/Brasil); AMG (Miscigenados/MG - Brasil).---- ----------84

Tabela 4. Frequências dos alelos ancestrais (*PDE4B*- rs546784, rs6683977 e rs641262) e alelos derivados (*PDE4B*- rs524770 e *MYT1L*-rs17039396) reportados como associadas à recidiva de LLA (YANG *et al.* 2011, 2012) em diferentes populações. ASH: *Ashaninka*, MAC: *Machiguenga*, AYM: *Aymara*, QUE: *Quechua*, MEX (*Huicholes* e *Tarahumara*), GUA: *Guarani*, TUP: *Tupiniquim*, 1000GP: AMR: Hispânicos miscigenados/EUA e Latinos Americanos, EUR: Europeus, AFR: Africanos, EAS: Leste-Asiáticos, SAS: Sul-Asiáticos. n: indica o número total de amostras usado para o cálculo da frequência alélica.Valores em negrito

indicam desvio do equilíbrio de Hardy-Weinberg (p<0,05). (-) indica que os SNPs não estavam incluídos nos bancos de dados analisados na população em questão.----------87

Tabela 5. Índices de heterozigosidade observada intrapopulacinal. ASH: *Ashaninka*, AYM: *Aymara*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, MCP: Martinho Campos, CAR: Carmésia, RES: Resplendor, SJM: São João das Missões, Projeto 1000 Genomas: EUR: Europeus, AFR: Africanos, EAS: Leste Asiáticos, SAS: Sul-Asiáticos.---------- --89

Tabela 6 Índices de F_{ST} par-a-par entre as populações. ASH: *Ashaninka*, AYM: *Aymara*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, MCP: Martinho Campos, CAR: Carmésia, RES: Resplendor, SJM: São João das Missões; Projeto 1000 Genomas: EUR: Europeus, AFR: Africanos, EAS: Leste Asiáticos, SAS: Sul-Asiáticos. (* p<0,005)---------- ---90

Tabela 7. Índices de F_{CT} par-a-par entre as populações de nativos (brasileiros e peruanos) e as populações de africanos (AFR) e europeus (EUR) do Projeto 1000 Genomas para cada SNP do banco BeadXpress. Em negrito valores significativos (p<0.05), enquanto que em cinza valores de F_{CT}>0,12. NAN: valores não obtidos pela aplicação de filtros de qualidade.---------- --92

Tabela 8. Valores dos índices de F_{CT} par-a-par entre as populações para as variantes previamente reportadas como associadas à recidiva de LLA genotipadas por sequenciamento direcionado. As significâncias (p-valor) são também apresentados. NAT: populações nativas brasileiras, peruanas e mexicanas; Projeto 1000 Genomas: EUR: Europeus, AFR: Africanos, EAS: Leste-Asiáticos, SAS: Sul-Asiáticos.----------93

Tabela 9. SNPs do banco de dados TargetSeq, BeadXpress e Taqman-rs6683977+2.5M em desequilíbrio de ligação LD (r2>0.80) com *PDE4B*-rs6683977 e *MYT1L*-rs17039396 em diferentes populações. EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP), TAH: *Tarahumara*, HUI:*Huichol*, QUE: *Quechua*, AYM: *Aymara*, ASH: *Ashaninka*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, AMG-BRA: Miscigenados Brasileiros. 0: nenhuma associação encontrada; (-) população sem dados para o banco em questão.---------- -------97

Tabela 10. Haplótipos formados pelas variantes em desequilíbrio de ligação (r^2>0.8) com o SNP *PDE4B*-rs6683977 (destacado em negrito) no banco de dados TargetSeq (em ordem: rs6668516, rs546784, rs6683604, rs12137080, rs12137115, rs495477, rs494735, rs6683977, rs638111, rs641262). Os números de amostras estão apresentados entre parênteses. Populações do 1000GP: AMR= Hispânicos miscigenados dos EUA e Latinos Americanos, SAS= Sul-asiáticos, EAS= Leste-asiáticos, EUR= Europeus. MEX = nativo-americanos do Mexico (*Tarahumaras* e *Huicholes*); populações peruanas de nativos: QUE=*Quechuas*, AYM=*Aymaras*, ASH=*Ashaninkas*, MAC=*Machiguengas*. (-) indica ausência ou haplótipos em baixa frequência.----------101

Tabela 11. Evidências da atuação de elementos funcionais nas regiões sequenciadas (banco de dados TargetSeq) dos genes *PDE4B* e *MYT1L* obtidos pelo acesso ao banco de dados ENCODE. Em negrito os SNPs em que o índice de diferenciação F_{CT} é maior que 0,12 para comparações com os grupos de africanos e/ou europeus.----------107

Tabela 12. Indicação dos scores obtidos pela consulta ao banco de dados RegulomeDB. 2b: ligação de fator de transcrição + qualquer motivo + DNase Footprint + pico de DNAse; 4: fator de transcrição + pico de DNAse; 5: fator de transcrição ou pico de DNAse; 6: outro.---------108

Tabela 13. Documentos e respectivas referências obtidos em buscas para identificação dos protocolos de tratamento da LLA utilizados em países latino-americanos----------109

Tabela Apêndice 1. Frequências alélicas dos SNPs do BeadXpress localizados nos genes *PDE4B* e *MYT1L* com destaque (em cinza) para os reportados em Yang *et al.* (2011). 1000GP: Projeto 1000 Genomas. ASH: *Ashaninka*, AYM: *Aymara*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, MCP: Martinho Campos, CAR: Carmésia, RES: Resplendor, SJM: São João das Missões, EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). Nan: genótipos não obtidos.----------133

Tabela Apêndice 2. Frequências alélicas (relativas ao número de amostras do banco de dados TargetSeq) em diferentes populações das variantes encontradas entre as posições genômicas 2.215.144 e 2.235.144 do gene *MYT1L* (GRCh37). ASH: *Ashaninka*, MAC: *Machiguenga*, AYM: *Aymara*, QUE: *Quechua*, TUP: *Tupiniquim*, HUI:*Huichol*, EUR: Europeans (1000GP), AFR: Africanos (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). Destaque (em cinza) para o SNP rs17039396----------135

Tabela Apêndice 3. Frequências alélicas (relativas ao número de amostras do banco de dados TargetSeq) em diferentes populações das variantes encontradas entre as posições genômicas 66.759.100 e 66.779.100 do gene *PDE4B* (GRCh37). ASH: *Ashaninka*, MAC: *Machiguenga*, AYM: *Aymara*, QUE: *Quechua*, TUP: *Tupiniquim*, HUI:*Huichol*, EUR: Europeans (1000GP), AFR: Africans (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). Destaque (cinza) para os SNPs rs546784, rs524770, rs6683977, rs641262----------138

Tabela Apêndice 4. Valores de F_{CT} para cada uma das variantes do banco de dados TargetSeq do gene *PDE4B* (posições genômicas 66.759.100 a 66.779.100 - GRCh37) calculados par-a-par entre as populações que compõem o banco de dados TargetSeq. Destaque (cinza) para os SNPs rs546784, rs524770, rs6683977, rs641262----------143

Tabela Apêndice 5. Valores de F_{CT} para cada uma das variantes do banco de dados TargetSeq do gene *MYT1L* (2.215.144 a 2.235.144 - GRCh37) calculados par-a-par entre as populações que compõem o banco de dados TargetSeq. Destaque (cinza) para o SNP rs17039396-------145

CAPÍTULO 3

Tabela 1. Populações do Projeto 1000 Genomas e do Projeto EPIGEN e suas respectivas proporções de ancestralidade. MXL – mexicanos de Los Angeles, CLM – colombianos de Medellin, PUR – porto-riquenhos. EUR – ancestralidade europeia, AFR – ancestralidade africana e NAT – ancestralidade nativo-americana.----------156

Tabela 2. Número de grupos gênicos significativos em cada população por ancestralidade (EUR – europeia, AFR – africana, NAT- nativo-americana) e combinações de ancestralidade.----------165

Tabela 3. Nomes dos grupos gênicos, as ancestralidades e suas condições de excesso ou escassez que apresentaram significância em todas as populações analisadas.----------169

SUMÁRIO

APRESENTAÇÃO 13

CAPÍTULO 1. SEQUENCIAMENTO DIRECIONADO E *PIPELINE* CUSTOMIZADO PARA A GERAÇÃO DE DADOS GENÉTICOS POR SEQUENCIAMENTO POR SÍNTESE: PARALELO ENTRE FERRAMENTAS COMERCIAL E DE DESENVOLVIMENTO LIVRE 17

INTRODUÇÃO 17

OBJETIVOS 22

METODOLOGIA 23

RESULTADOS 34

DISCUSSÃO 46

REFERÊNCIAS 52

CAPÍTULO 2. ANCESTRALIDADE NATIVO-AMERICANA E IMPLICAÇÕES CLÍNICAS PARA A LEUCEMIA LINFOBLÁSTICA AGUDA: ANÁLISES EM ESCALA FINA DE VARIANTES COM IMPORTÂNCIA BIOMÉDICA 57

INTRODUÇÃO 57

OBJETIVOS 70

METODOLOGIA 70

RESULTADOS 82

DISCUSSÃO 110

REFERÊNCIAS 124

APÊNDICES 133

CAPÍTULO 3. PERSPECTIVAS SOBRE A CONTRIBUIÇÃO DA INFERÊNCIA DA ANCESTRALIDADE LOCAL PARA A IDENTIFICAÇÃO DE SINAIS DE SELEÇÃO POLIGÊNICA 148

INTRODUÇÃO 148

OBJETIVOS 154

METODOLOGIA 154

RESULTADOS PRELIMINARES 158

DISCUSSÃO PRELIMINAR 170

REFERÊNCIAS 176

CONCLUSÃO GERAL 181

REFERÊNCIAS 183

ANEXOS 184

APRESENTAÇÃO

Os recentes progressos alcançados na área da Genômica foram expressivos em razão dos avanços tecnológicos obtidos nas últimas décadas, mais especificamente na era iniciada após o sequenciamento do primeiro genoma humano (MARDIS 2011). Em razão dos altos investimentos para o desenvolvimento de novas tecnologias de sequenciamento de ácidos nucléicos, atualmente é possível acessar as informações do genoma completo de qualquer organismo em escala de horas. Essa revolução tecnológica permitiu que estudos de diversas áreas do conhecimento biológico incluíssem a Genômica como um dos seus pilares de base teórica e prática (MARDIS 2008). As aplicações no campo da biologia humana foram ainda mais evidentes em razão das inúmeras linhas de pesquisa focadas em medicina, saúde e evolução.

A geração massiva de informação que impulsionou a Genômica, atualmente ao nível de uma das áreas mais importantes para estudos de base biológica, também levou ao surgimento de novos desafios em diferentes campos do conhecimento. O desenvolvimento dessa área permitiu que hipóteses anteriormente consideradas como difíceis de serem testadas por dificuldades metodológicas passassem a ser colocadas em prova em estudos amplos, alicerçados em um volume de dados sem precedentes. Nesse contexto, a manipulação e análise de um universo de dados de grande magnitude, e que ainda se encontra em crescimento expressivo, indicaram limites antes não conhecidos. Atualmente, são reconhecidas as limitações dos métodos e ferramentas de análise frente à taxa de geração de dados (NEKRUTENKO & TAYLOR 2012). O que não é de se surpreender, tendo em vista a superação da lei de Moore quando aplicada à realidade da Genômica (WETTERSTRAND 2019) e a constatação de que estimativas do início da década indicassem que, naquele momento, a cada dois dias era gerada no mundo a mesma quantidade de informação que havia sido acumulada em toda a história até o ano de 2003 (BOTTLES *et al.* 2014). Esse déficit notório de capacidade analítica revelou a necessidade de progressos em áreas que permitissem extrair informações claras e interpretáveis a partir de imensos bancos de dados. Diante desse cenário, o advento de estratégias e mecanismos computacionais voltados à análise de dados biológicos complexos motivou a consolidação e o progresso de um já tradicional campo do conhecimento: a Bioinformática (HAGEN 2000).

Nesse contexto de desenvolvimento vertiginoso, a Genômica e suas aplicações transcenderam limites de áreas do conhecimento e encontraram na Bioinformática o suporte

ideal para o progresso científico. Fazendo uso da riqueza de possibilidades proporcionada pela interface entre a Bioinformática e a Genômica essa tese busca contribuir em diferentes etapas dos processos de geração, interpretação e aplicação dos dados biológicos para a solução de problemas voltados às questões biomédicas e evolutivas.

Por meio da constatação de que lacunas de conhecimento precisam ser preenchidas nos processos de geração de dados de qualidade (TAUB *et al.* 2010, O'RAWE *et al.* 2013), o primeiro capítulo dessa tese discorre sobre resultados encontrados na análise dos dados gerados por sequenciamento massivo de ácidos nucléicos e pelo emprego de métodos de definição de variações genéticas. A compreensão acerca das especificidades técnicas de geração da informação, em conjunto com os conhecimentos biológicos que envolvem os estudos, permitem elevar a qualidade do dado biológico a patamares de maior confiança estatística. A hipótese testada nesse capítulo foi de que a chamada de variantes é um processo que apresenta resultados mais confiáveis à medida em que são consideradas as propriedades da biblioteca sequenciada e de todo o processo de geração do dado. Nesse sentido, o capítulo 1 contribui para a discussão sobre as implicações técnicas das diferentes ferramentas bioinformáticas e processos de análise considerando as particularidades dos dados biológicos a serem utilizados.

A contribuição para a qualidade do dado gerado tem a pretensão de trazer maior confiança aos estudos conduzidos com propósitos amplos, tais como os aplicados na área biomédica. Nesse sentido, o segundo capítulo da tese traz um estudo aprofundado sobre o complexo cenário de interação existente entre a composição genética individual e populacional e suas implicações para o tratamento de enfermidades complexas em populações humanas. A realidade atual de desenvolvimento e progresso técnico no campo científico inevitavelmente espelha as desigualdades de outras áreas revelando o protagonismo de países com economias mais desenvolvidas (O'CONNOR *et al.* 2018). Essa concentração de estudos e avanços tem implicações imediatas sobre a característica do conhecimento gerado, o qual impacta imediatamente populações humanas de diferentes origens. De fato, a diversidade genética populacional humana e seus padrões de distribuição revelam peculiaridades importantes que precisam ser reconhecidas em prol do avanço científico (POPEJOY & FULLERTON 2016, SIRUGO *et al.* 2019). O segundo capítulo dessa tese aprofunda o conhecimento genético-populacional acerca de um caso biomédico de tratamento da leucemia linfoblástica aguda, o qual pode ser tomado como exemplo das distorções ainda existentes no progresso científico, em que hiatos de conhecimento passam a afetar populações humanas de

diferentes localidades (BHATIA 2011). Considerando evidências reportadas na literatura científica que suportam a existência de associação entre marcadores moleculares dos genes *PDE4B* e *MYT1L*, a ancestralidade nativo-americanana e a recidiva de leucemia linfoblástica aguda (LLA), esse capítulo apresenta análises da diversidade molecular desses dois genes em populações miscigenadas e predominantemente nativo-americanas da América Latina. A principal hipótese testada foi de que a frequência dos alelos de risco para a recidiva de LLA acompanharia de forma direta a proporção de ancestralidade nativo-americana das populações. A diversidade e estrutura genética foram analisadas e discutidas também no contexto da regulação gênica, já que podem ter consequências diretas sobre a resposta ao tratamento da LLA em populações com proporção significativa de ancestralidade nativo-americana. Processos como os resultantes do intercâmbio genético entre populações com históricos evolutivos diferentes têm consequências diretas sobre os complexos sistemas biológicos que envolvem seus descendentes. A mistura de perfis genômicos diversos que ocorre em processos de miscigenação populacional gera indivíduos com características particulares. Como já mencionado, tais populações miscigenadas são inegavelmente negligenciadas na construção do conhecimento biológico direcionado ao progresso biomédico (POPEJOY & FULLERTON 2016, BENTLEY *et al*. 2017*,* SIRUGO *et al.* 2019). Por outro lado, representam oportunidades únicas para estudos que buscam compreender as bases dos processos evolutivos que historicamente moldaram a diversidade humana (MCKEIGUE 2005, SELDIN *et al.* 2011). Nesse sentido, o terceiro e último capítulo apresenta perspectivas que buscam utilizar a grande densidade de dados genéticos oferecidos pelos avanços da Genômica e as características peculiares das populações miscigenadas para o estudo de forças evolutivas relacionadas aos processos adaptativos. Pelo uso de novas técnicas de estudo genético-populacional ao nível genômico, associadas ao poder analítico oferecido por processos automatizados em ambiente computacional, esse capítulo apresenta um estudo investigativo em busca de sinais genômicos adaptativos em populações miscigenadas. Parte-se da hipótese de que genes cujos produtos estejam inter-relacionados em redes moleculares complexas estão submetidos a pressões seletivas que atuam concomitantemente sobre suas variantes, levando a suaves alterações não-aleatórias da frequência de seus alelos. As pretensões desse estudo vão além da proposta de apenas identificar sinais adaptativos poligênicos em âmbito genômico, já que representa um dos primeiros esforços para testar hipóteses de atuação da seleção natural em populações humanas das Américas considerando os resultados dos processos de miscigenação.

Portanto, a presente tese discorre sobre assuntos de diferentes áreas do conhecimento relacionados à Genômica e Bioinformática que vêm sofrendo progressos expressivos recentes. Busca-se integrar e discutir resultados de diferentes naturezas, mas que convergem aos campos de aplicação da Genômica. Pela apresentação de diferentes métodos e análises em bioinformática espera-se contribuir para o progresso na área de geração de dados genéticos, e com estudos conduzidos nos campos biomédico e evolutivo.

CAPÍTULO 1. SEQUENCIAMENTO DIRECIONADO E *PIPELINE* CUSTOMIZADO PARA A GERAÇÃO DE DADOS GENÉTICOS POR SEQUENCIAMENTO POR SÍNTESE: PARALELO ENTRE FERRAMENTAS COMERCIAL E DE DESENVOLVIMENTO LIVRE

INTRODUÇÃO

O processo de geração de sequencias de ácidos nucléicos e de determinação de variantes envolve várias etapas, desde procedimentos no campo da Biologia Molecular até análises em programas computacionais disponibilizados especificamente para esse propósito. Porém, um dos maiores desafios atualmente no campo da Bioinformática é o desenvolvimento de processos que contemplem a complexidade do dado biológico a ser gerado permitindo a customização da análise de acordo com as peculiaridades do ensaio desejado (ROY *et al.* 2018). Nos casos de geração de dados de painéis customizados, o estabelecimento de *pipelines* específicos contribui para a eficiência das análises, pois permite concomitantemente a adequação da análise ao ensaio biológico de interesse (DE SUMMA *et al.* 2017) e a automação do processo levando em consideração as características da maquinaria computacional disponível e as dependências de software*s* existentes.

A manipulação da amostra, prévia ao sequenciamento do ácido nucléico, é obrigatória para o sequenciamento das moléculas. Esse processo é conhecido na literatura científica como construção da biblioteca, o qual consiste na preparação das moléculas para estarem adequadas à tecnologia do sequenciador selecionado. A construção da biblioteca também pode ter o papel de isolar o subconjunto de fragmentos de ácidos nucléicos que se tem interesse em sequenciar (VAN DIJK *et al.* 2014). Nesse sentido, considerando a grande diversidade de métodos para a seleção das moléculas de interesse e também as diferentes tecnologias de sequenciamento disponíveis no mercado, o processo de construção das bibliotecas é uma etapa de grande relevância para a obtenção das sequências finais (SOLONENKO *et al.* 2013, BARAN-GALE *et al.* 2013, JONES *et al.* 2015).

Como na maioria dos casos as bibliotecas precisam conter apenas os fragmentos que se tem interesse em sequenciar, o procedimento para a separação das moléculas alvo das demais pode ser um trabalho extensivo tendo em vista o tamanho, a diversidade do genoma e o conjunto de ácidos nucléicos extraídos (VAN DIJK *et al.* 2014). Em razão dessa dificuldade, é notável a tendência de mercado que indica que em um futuro próximo, mesmo que se busque identificar apenas algumas variações específicas no genoma ou a presença de

certos RNAs específicos, o sequenciamento de todo o genoma ou de todo o complexo de RNAs extraídos será economicamente viável (KWONG *et al.* 2015, PARK & KIM 2016, MATTICK *et al.* 2018). Até que essa relação custo/benefício seja alcançada as empresas e grupos de pesquisa buscam desenvolver métodos moleculares de construção de bibliotecas que aumentem a eficiência, reduzam os custos, mas que, concomitantemente, possibilitem a geração de sequências com alta qualidade (JIANG *et al.* 2016).

Um dos campos de desenvolvimento de metodologias de construção de bibliotecas é conhecido como sequenciamento direcionado (*Target Sequencing*) (MAMANOVA *et al.* 2010, SIKKEMA-RADDATZ *et al.* 2013). Os métodos relacionados ao sequenciamento direcionado buscam isolar do restante do genoma apenas aqueles fragmentos de DNA que possuam as sequências de interesse. Um dos propósitos mais comuns da aplicação desse método é o sequenciamento apenas dos éxons dos genes presentes no genoma, ou seja, o sequenciamento do exoma (BAMSHAD *et al.* 2011). Porém, há estratégias no mercado, e até mesmo desenvolvidas em laboratórios sem fins comerciais, que permitem isolar quaisquer outras regiões do genoma de organismos conhecidos (ou até mesmo de organismos pouco estudados) (BODI *et al.* 2013, SAMORODNITSKY *et al.* 2015). Esses são painéis customizados gerados de acordo com a necessidade e o propósito do sequenciamento.

As estratégias moleculares para a construção de bibliotecas de sequenciamento direcionado customizadas são bastante diversas, mas se baseiam principalmente em dois métodos, os quais também possuem variações e que por isso explicam o grande número de opções existentes no mercado e na literatura científica (KOZAREWA *et al.* 2015). O primeiro método é mais simples, baseado em reação em cadeia da polimerase (PCR) envolve apenas a síntese de pares de *primers* que são desenvolvidos com a finalidade de formarem grupos de amplificação do tipo *multiplex* para que em poucas, ou mesmo em uma única reação, promovam a amplificação em conjunto das regiões do genoma que se pretende sequenciar (SAMORODNITSKY *et al.* 2015b). A vantagem dessa estratégia é baseada na facilidade do procedimento, o qual requer pouca quantidade de DNA inicial e a realização de apenas algumas reações de PCR. Isso faz com o que o processo seja relativamente rápido e de baixo custo. Por outro lado há desvantagens que estão relacionadas com as diferenças na eficiência da amplificação das regiões, que pode ocorrer pela heterogeneidade dos *primers* utilizados, pelo conteúdo GC das regiões e pela afinidade da enzima utilizada. A diferença na eficiência terá impacto imediato na no número de cópias geradas de cada sequencia alvo, o que interferirá diretamente na homogeneidade da cobertura de sequenciamento alcançada para

cada região. Há também susceptibilidade à fidelidade da enzima no processo de amplificação, o que pode levar à inserção de erros, induzindo mutações não existentes (para mais detalhes veja QUAIL *et al.* 2012). Outros problemas que podem ocorrer pelo uso da PCR no processo de construção de bibliotecas são principalmente: (i) a geração de quimeras de *primers*, que podem ser formadas quando há excesso dessas moléculas ou quando a reação não alcança o limiar ótimo de eficiência; (ii) a existência de variantes nos sítios de anelamento dos *primers*, inibindo a amplificação (*dropout*); (iii) a dificuldade de se amplificar um número grande de regiões em um mesmo processo *multiplex,* dentre outras (AIRD *et al.* 2011).

As desvantagens que ocorrem em razão do uso da técnica de PCR na construção de bibliotecas são consideráveis, a ponto de estimular o desenvolvimento de alternativas de construção de bibliotecas que não fazem uso da PCR. As estratégias de construção de bibliotecas livres de PCR (*PCR-free*) estão sendo cada vez mais disponibilizadas no mercado e na literatura científica (Meienberg *et al.* 2016) e tendem a ser especialmente importantes em estudos que buscam mutações de baixa frequência, como os que são realizados para a identificação de mutações somáticas em carcinomas (ALIOTO *et al.* 2015). Porém, apesar desses problemas, a estratégia de construçãode bibliotecas pelo uso de PCRs é bastante difundida, sendo o método aplicado na maioria dos produtos no mercado. Especialmente pelas otimizações oferecidas pelas empresas, que investem bastante no desenvolvimento de enzimas de alta fidelidade e com alta eficiência no processo de PCR.

O segundo método de construção de bibliotecas de sequenciamento direcionado customizadas se baseia na fragmentação do DNA genômico para a formação de moléculas mais ou menos homogêneas em tamanho que possam ser capturadas por sondas de complementariedade (MAMANOVA *et al.* 2010, BODI *et al.* 2013, KOZAREWA *et al.* 2015). A construção de oligonucleotídeos específicos para serem complementares às regiões de interesse permite a captura apenas dos fragmentos que contêm as sequências desejadas. As principais vantagens desse método são: (i) maior uniformidade de cobertura das regiões no sequenciamento, (ii) menor susceptibilidade a erros relacionados ao alinhamento de *primers*, (iii) menor susceptibilidade a erros causados por variantes do tipo inserção/deleção. Como desvantagens citam-se o maior custo e complexidade do processo de construção de bibliotecas e a maior quantidade de DNA inicial necessária para a reação (SAMORODNITSKY *et al.* 2015a,b). Além disso, apesar de mais robusta, essa técnica não é totalmente isenta de erros inerentes a PCR, já que alguns ciclos de amplificação podem também ser necessários nesse método. De toda forma, a necessidade de um número menor de ciclos e também a utilização

de *primers* universais complementares a adaptadores conhecidos reduzem consideravelmente a chance de erros quando comparado ao método de construção de bibliotecas que usa apenas a PCR.

Além da etapa de construção da biblioteca, o processo de geração de sequencias também envolve métodos e estratégias computacionais empregados desde nos programas internos do sequenciador até no tratamento posterior dos dados brutos gerados. Dentro do sistema de sequenciamentro, a determinação das bases que formam as sequências é embasada em parâmetros complexos, desde com métricas de sistemas ópticos (BENTLEY *et al.* 2008) até com parâmetros dos sistemas de semicondutores que captam alterações no potencial elétrico do microrreator (ROTHBERG *et al.* 2011). Além disso, nos processos externos ao sequenciador, a identificação de variações nas sequências, as quais podem indicar mutações, é também um processo minucioso de avaliação de parâmetros estatísticos e de métricas desenvolvidas no campo da bioinformática (MAGI *et al.* 2010, NIELSEN *et al.* 2011). Portanto, além do processo químico, o processo computacional de geração de sequências e de determinação de variantes é outro campo em desenvolvimento e que deve ser considerado com atenção para a geração de dados genéticos, já que a confiança na chamada das bases e das variantes deve ser a máxima possível.

Atualmente há diversas plataformas e software*s*, tanto de distribuição livre, quanto comercializados no mercado, que compõem toda a cadeia de análise, desde o processamento do dado bruto gerado no sequenciador até a determinação do painel final de variantes (KUMAR *et al.* 2012, SANDMANN *et al.* 2017). Dependendo do software*,* o resultado pode também incluir anotações sobre as mutações, o que envolve detalhar informações sobre sua natureza, efeito preditivo, funcionalidade molecular, dentre outras (MUDGE & HARROW, 2016).

A grande maioria das empresas atuantes no mercado oferecem soluções que contemplam desde o kit de construção de bibliotecas até o software para o processamento dos dados do sequenciador. Dessa forma, ao adquirir o kit o cliente também passa a ter direito ao uso da plataforma informática que processa o dado gerado. A vantagem desses programas reside no fato de serem desenvolvidos especificamente para os kit*s* de construção de bibliotecas produzidos pelas empresas e que, por isso, levam em consideração as peculiaridades do ensaio molecular executado antes do sequenciamento. Além disso, são programas e plataformas fáceis de utilizar e que não requerem alta capacidade computacional para serem executados ou, quando assim ocorre, disponibilizam servidores remotos para o

processamento dos dados. Porém, por serem programas fechados, a velocidade com que são atualizados e incorporam melhorias é menor do que a que ocorre com os programas abertos. Além disso, outra desvantagem ocorre pelo paradoxo que surge pelo desenvolvimento de softwares de fácil utilização e a possibilidade de se alterar parâmetros. Muitas vezes, os softwares comerciais não permitem ao usuário ajustar os parâmetros da análise já que o objetivo das companhias é deixar o programa mais fácil de ser manuseado (KUMAR *et al.* 2012). Exemplos de softwares fechados são: SureCall™, da Agilent Technologies®, Ion repórter™ da ThermoFisher® e as plataformas GenomeStudio™ e BaseSpace™ da Illumina®.

Já os softwares e plataformas de livre distribuição são ferramentas que, por estarem em constante aprimoramento, são mais completas e muitas vezes incluem os algoritmos mais eficientes e com melhor desempenho na predição de variantes (SANDMANN *et al.* 2017). Essas opções permitem que o usuário tenha liberdade na definição de parâmetros fundamentais e que possa customizar o procedimento de acordo com as necessidades do seu experimento (como por exemplo, por permitirem ao usuário incluir na análise seus próprios painéis de variantes conhecidas). Porém, a maior desvantagem desses softwares se deve às dificuldades para a instalação e total aproveitamento das funcionalidades, visto que há muitas dependências com relação a bibliotecas de dados e ao sistema operacional (KUMAR *et al.* 2012). Além disso, a automatização das análises requer que sejam desenvolvidos *pipelines* específicos de acordo com sistema computacional utilizado, como por exemplo, para se adequar às ferramentas disponíveis e às especificidades das máquinas utilizadas. Como exemplo desses softwares e plataformas de tratamento de dados de sequenciamento e chamada de variantes citam-se a plataforma Genome Analysis Toolkit (GATK) (MCKENNA *et al.* 2010) e os softwares FreeBayes (GARRISON & MARTH 2012) e Samtools (LI *et al.* 2009).

Tendo em vista as diferenças nas eficiências dos softwares existentes (O'RAWE *et al.* 2013, SANDMANN *et al.* 2017), assim como o constante aprimoramento dos softwares de distribuição livre, o presente capítulo trata da geração de sequências no sequenciador Illumina MiSeq a partir de uma biblioteca customizada de sequenciamento direcionado e também do desenvolvimento de um *pipeline* customizado que trabalha as sequencias geradas. Semelhante a outros estudos comparativos já publicados (O´RAWE *et al.* 2013, HWANG *et al.* 2015, HAMPEL *et al.* 2017), o objetivo desse estudo foi de traçar um paralelo entre os resultados obtidos com o programa oferecido pela empresa fabricante do kit de construção da

biblioteca de sequenciamento direcionado e aqueles obtidos com uma das plataformas de distribuição livre atualmente mais utilizadas, o GATK (MCKENNA *et al.* 2010). Parte-se da hipótese de que a chamada de variantes é um processo que apresenta resultados mais confiáveis à medida em que são consideradas as propriedades da biblioteca sequenciada. Além de revelar as diferenças encontradas nos resultados de ambos os programas, também foi proposto como objetivo final do presente estudo o desenvolvimento de um *pipeline* baseado nas boas práticas do GATK (MCKENNA *et al.* 2010) que possa ser utilizado para a chamada das variantes de dados gerados a partir de bibliotecas de sequenciamento direcionado do tipo Haloplex (Agilent Tecnologies).

Por fim, cabe esclarecer que o desenvolvimento do estudo apresentado nesse capítulo esteve embasado no uso de ferramentas de Bioinformática que são amplamente utilizadas em estudos que envolvem a Genômica, tais como VCFTools (DANECEK *et al.* 2011), BCFTools (LI 2011), PLINK (PURCELL *et al.*2007), dentre outras. O prévio conhecimento acerca das estruturas dos dados genômicos, como os armazenados em arquivos do tipos .fastq, .vcf e .bed, bem como de ferramentas para a manipulação desses tipos de arquivos e linguagens de programação, tais como *perl* e *python*, foi necessário para a realização bem sucedida das análises propostas. Nesse sentido, a experiência prévia adquirida ao longo do estudo desenvolvido no âmbito do Projeto EPIGEN (KEHDY *et al.*2015), em que os trinta primeiros genomas completos de indivíduos brasileiros foram analisados, permitiu ao autor dessa tese adquirir a experiência necessária para o uso de ferramentas de Bioinformática que são essenciais para a manipulação de dados genômicos em larga escala, bem como para a realização de análises estatísticas, de genética de populações, dentre outras. Portanto, destaca-se aqui a contribuição ao estudo de Kehdy *et al.* (2015), em que foi possível o desenvolvimento de habilidades e competências para a realização dos estudos propostos não apenas nesse capítulo, mas também nos demais capítulos que compõem essa tese.

OBJETIVOS

Objetivo Geral

Gerar e disponibilizar um conjunto de dados de sequenciamento de regiões específicas por meio da customização de um *pipeline* de análises que contemple as peculiaridades do desenho experimental.

Objetivos específicos

- Gerar e sequenciar bibliotecas customizadas a partir de um kit comercial

- Comparar os resultados de chamada de variantes entre um software comercial e outro de desenvolvimento livre
- Disponibilizar um painel final com variantes de alta confiança
- Disponibilizar um *pipeline* de análises que contemple as peculiaridades do desenho experimental da biblioteca Haloplex

METODOLOGIA

Amostragem

As amostras utilizadas na geração do painel de variantes compreendem tanto indivíduos nativos americanos quanto indivíduos de populações com histórico de miscigenação. Um total de 150 amostras foi selecionado para a construção das bibliotecas de interesse. Como a origem e a caracterização das populações utilizadas no estudo têm maior relevância para os propósitos apresentados no capítulo 2, mais detalhes sobre esses grupos podem ser encontrados na referida seção. Para os propósitos do presente capítulo cumpre apenas informar o número de amostras por população (Tabela 1) e a origem étnica.

Tabela 1. País de origem, origem étnica e número de amostras utilizadas na preparação de bibliotecas.

País	**Origem étnica**	Nº de Amostras
Peru	*Quechua*	20
	Aymara	24
	Machiguenga	22
	Ashaninka	16
Brasil	*Tupiniquim*	22
	Guarani	24
México	*Huichol*	12
	Tarahumara	10
Total		150

Bibliotecas

A estratégia de sequenciamento de regiões alvo escolhida para a construção das bibliotecas foi a oferecida pela empresa Agilent Technologies Inc., por meio do kit Haloplex 500 Kb (comercializado em 2014, Catalog Number 5190-5436, Lot 6248658). Essa estratégia foi selecionada em razão da quantidade suficiente de dados gerados (500 Kb por amostra) que permitiriam sequenciar as regiões alvo dos genes selecionados. Além disso, essa técnica é baseada na estratégia de circularização do DNA e captação por oligonucleotídeos complementares, o que permite alta especificidade das regiões alvo e redução do número de ciclos na amplificação por PCR. O kit é customizado de acordo com as necessidades do

cliente, que por meio da plataforma *web SureDesign*[1], desenvolvida e disponibilizada pelo fabricante, identifica as regiões do DNA humano a serem sequenciadas para a geração dos dados de interesse. Por meio de análise *in-silico* prévia, a plataforma indica os pontos de clivagem enzimática que serão utilizados na construção da biblioteca e a estimativa de abrangência alcançada pelo ensaio com relação às regiões genômicas de interesse do cliente.

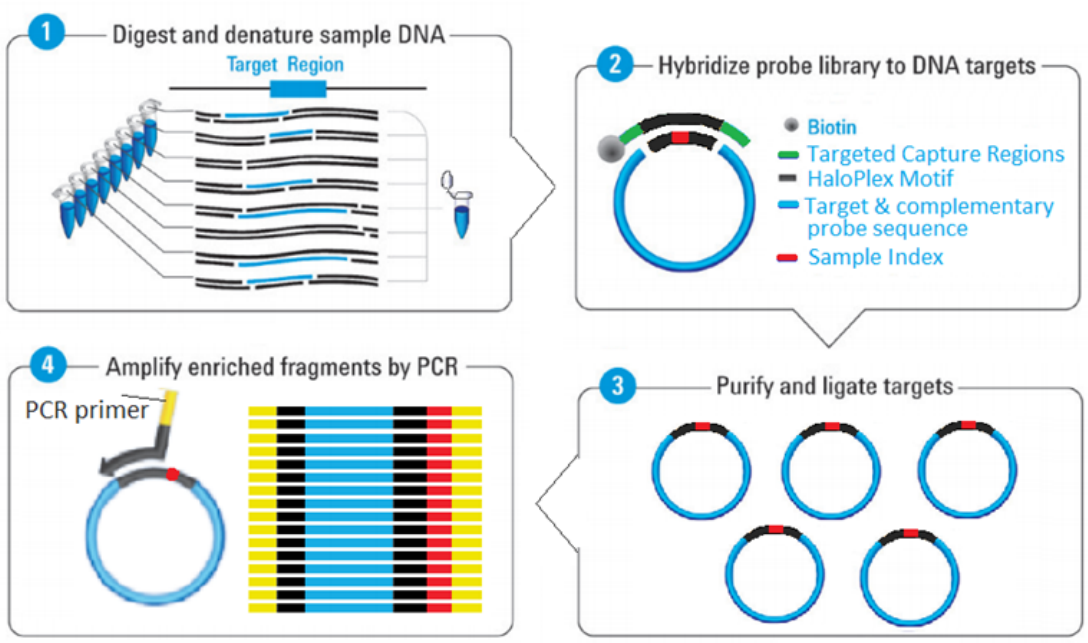
O DNA genômico utilizado como material inicial da construção das bibliotecas foi obtido de amostras de sangue total periférico, por meio de extração com kit próprio (*Gentra® Puregene® Blood Kit* - Qiagen). A quantificação do DNA após a extração foi realizada no equipamento Nanodrop (ThermoFisher). A quantidade inicial de DNA inicial necessária indicada pelo protocolo do Haloplex é de 225 ng de DNA, distribuídos em 45 µL (5ng/µL). Resumidamente, o protocolo do Haloplex estabelece clivagens do DNA genômico com pares de enzimas de restrição e posterior captação das moléculas das regiões alvo pela hibridização com oligonucleotídeos sintetizados especificamente de acordo com o ensaio customizado definido na plataforma *SureDesign*. Após alguns ciclos de amplificação por PCR, que permitem a inserção de sequencias de índices para individualização das amostras, as bibliotecas são purificadas, novamente amplificadas e finalmente analisadas em eletroforese para avaliação do padrão de fragmentos obtidos (Figura 1). Previamente, o fabricante disponibiliza ao cliente a distribuição dos tamanhos dos fragmentos esperada para o kit customizado de acordo com o desenho encomendado (Figura 2), o qual é utilizado para tomada de decisão com relação à inclusão da amostra no grupo de amostras a serem sequenciadas.

Vale ressaltar que a utilização da estratégia de captura do Haloplex de regiões de interesse em conjunto com a realização de ciclos de PCR dificulta o processo de identificação de *reads* únicas (*reads:* são definidas como o produto do sequenciamento de um fragmento de DNA, que no caso de sequenciadores de DNA de 2ª geração são porções curtas de até 300 pb de tamanho), ou seja, aquelas que foram geradas de moléculas de DNA de diferentes células. Como ocorre o uso de enzimas de restrição, os flancos das *reads* possuem as mesmas sequências (sítios de reconhecimento de clivagem das enzimas). Dessa forma, na análise *in silico*, as *reads* geradas, mesmo que sejam produto de amplficação de moléculas de DNA de células diferentes tendem a ser consideradas como duplicatas de PCR, pois possuem sempre as mesmas sequências de início e fim (SAMORODNITSKY *et al.* 2015). Isso não ocorre com

[1] Disponível em: <https://earray.chem.agilent.com/suredesign/>. Acesso em: 03/04/2015

os processos de clivagem aleatória (como procedimentos de sonicação por exemplo), já que o DNA de cada célula é fragmentado em pontos únicos. Assim, no momento da análise identifica-se que as *reads* de mesmo tamanho, posição de início e fim são idênticas, permitindo então marcá-las como cópias, resultantes de reações de PCR. O índice de duplicatas de PCR é uma característica importante, especialmente para o processo de chamada de variantes, e por isso deve ser considerado nos processos posteriores.

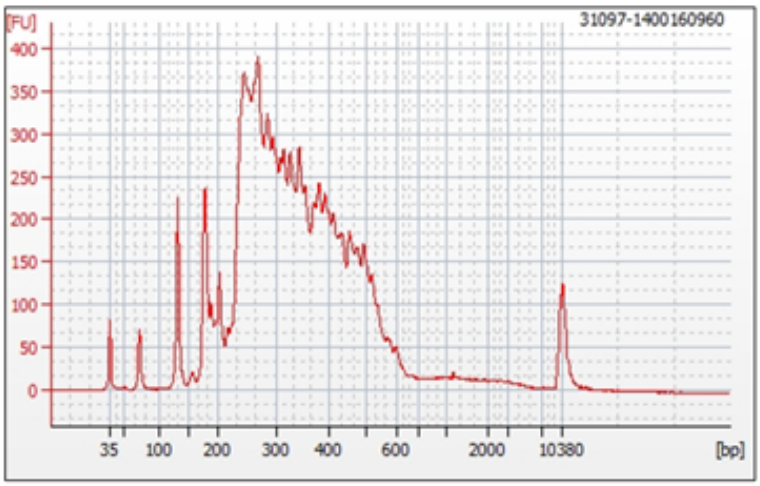
Figura 1. Principais etapas e procedimentos realizados na construção das bibliotecas com o kit Agilent Haloplex.



Adaptado de: HaloPlex Target Enrichment System (Agilent, BW, GE)[2].

[2] Disponível em: < https://www.agilent.com/cs/library/flyers/public/HaloPlex_TechFlier_final_v2.pdf>. Acesso em: 10/09/2018

Figura 2. Perfil do Bioanalyzer da distribuição dos tamanhos de fragmentos (pb) esperados da biblioteca Haloplex customizada.



As regiões selecionadas como de interesse para o sequenciamento incluíram posições em genes previamente indicados como diferenciados em nativos americanos (SOUZA, 2010), genes com interesse farmacogenético (*PDE4B* e *MYT1L*) e também regiões neutras para a realização de estudos evolutivos. No total, regiões de 38 genes diferentes foram sequenciadas, bem como 13 regiões neutras intergênicas (Tabela 2), compreendendo cerca de 500.000 pares de bases em cada amostra. De acordo com o resultado esperado apresentado pelo *SureDesign*, os genes e as regiões totalizariam 945 alvos, os quais gerariam 17.153 amplicons que cobririam 96,82% das bases previamente selecionadas para o sequenciamento.

Sequenciamento

O sequenciamento foi realizado no Laboratório de Genômica do Centro de Laboratórios Multiusuários (CELAM) da UFMG, localizado do Instituto de Ciências Biológicas. Foi utilizado o equipamento MiSeq da fabricante Illumina, o qual emprega a metodologia de sequenciamento massivo em paralelo por síntese (mais detalhes em SHENDURE & JI 2008). Essa tecnologia permite que no equipamento MiSeq sejam gerados até 15 bilhões de pares de bases sequenciadas, que podem representar cerca de 22 a 25 milhões de pares de *reads* que possuem até 300 pares de bases cada.

Considerando que a metodologia de sequenciamento em larga escala e os processos de preparação de bibliotecas seriam implementados pela primeira vez no Laboratório de Diversidade Genética Humana (LDGH) e que as amostras a serem sequenciadas seriam de populações nativo-americanas pouco estudadas (vide capítulo 2 para mais detalhes sobre as populações), a chance de se encontrar variantes novas foi levada em conta para a determinação da cobertura média esperada no sequenciamento de cada amostra. Dessa forma,

as 150 amostras foram divididas em três grupos que variaram de 47 a 53 amostras cada um, os quais foram sequenciados separadamente com o kit Illumina v3 de 600 ciclos. Assim, a razão da capacidade de sequenciamento empregada e do total de 500.000 pares de bases por amostra permitiu que a cobertura média projetada fosse superior a 200x. O valor de alta cobertura contribui para maior confiança no processo de chamada de variantes reduzindo a probabilidade de falsos positivos em casos de identificação de heterozigotos e alelos ainda não descritos.

Pipeline de análises

O equipamento MiSeq gera as sequências já em arquivos no formato *fastq* individualizados por amostras, os quais são utilizados como arquivos de entrada para *pipelines* de chamada de variantes. Como descrito anteriormente, o objetivo desse capítulo foi traçado considerando um paralelo entre ferramentas comercial e de desenvolvimento livre tendo em vista a reconhecida divergência existente entre as estratégias disponíveis para o procedimento de chamada de variantes (O'RAWE *et al.* 2013, HWANG *et al.* 2015, LAURIE *et al.* 2016). Assim, o presente estudo se propôs a executar os algoritmos de duas estratégias distintas, disponibilizadas nos programas *SureCall* e *GATK*.

Surecall

O fabricante do kit de construção das bibliotecas oferece também um programa para a análise das sequencias geradas, determinação dos genótipos e anotação das variantes inferidas. A Agilent Technologies disponibiliza aos seus clientes a plataforma *SureCall v3.5*, desenvolvida para a manipulação dos dados gerados pela estratégia Haloplex e demais produtos da empresa. Esse programa requer o uso do sistema operacional Windows e capacidade computacional ao menos superior a 2 Ghz de processamento, memória RAM acima de 12Gb e 470Gb de espaço disponível de armazenamento. Dessa forma, foi possível utilizá-lo em um dos computadores de mesa disponíveis no LDGH.

O SureCall permite a customização das análises pela alteração de alguns parâmetros a critério do usuário. Para a chamada das variantes o primeiro procedimento realizado é a aparagem das bases (tradução livre do termo “trimming”, bastante utilizado no campo da bioinformática). Os arquivos .fastq gerados pelo sequenciador são utilizados diretamente no SureCall sem a indicação de necessidade de qualquer manipulação anterior. Para a aparagem das bases de baixa qualidade o parâmetro “Quality threshold for trimming” foi ajustado para o valor 5 e as *reads* com extensão de pelo menos 30% o tamanho total (300 pb), ou seja, 90 pb,

foram mantidas. Esse valores são sugeridos como ideais pelo manual do programa e por isso são indicados como *default*.

Em seguida ao processo de aparagem o programa realiza o alinhamento das sequências utilizando o algoritmo MEM ("maximal exact matches") do programa Burrows-Wheeler Aligner (BWA, LI & DURBIN 2009) desenvolvido para tratar com mais acurácia sequencias longas (>100 pb). Os parâmetros utilizados foram os disponíveis como *default*. Para a chamada das variantes foi utilizado o algoritmo do SNPPET, desenvolvido pela própria Agilent Technlogies, Inc., o qual faz uso de uma estratégia probabilística para as chamadas de variantes. Em materiais de divulgação da empresa testes mostraram melhor performance do SNPPET frente ao programa GATK v.1.1(apresentado a seguir) utilizando o "Unified Genotyper" como algoritmo de chamada de variantes (ASHUTOSH *et al.* 2014). Em seguida ao SNPPET, as variantes encontradas foram filtradas para eliminação de possíveis falsos positivos. A anotação das variantes foi também realizada pelo SureCall v3.5 e considerou os seguintes bancos de dados disponibilizados pelo programa: ClinVarAnnotations, cosmic_100715, gwasV3_ucsc_281215, Hs_hg19_Gene_20151215. Ao final do fluxo de análises o programa gerou uma planilha apresentando as variantes encontradas em cada amostra, um valor de confiança estatística (p-value) para a variante, o gene impactado, outras anotações (como ID do dbSNP, OMIM ID, etc.), informações de frequência alélica, possíveis efeitos, dentre outros dados. A chamada de variantes foi realizada em cada amostra separadamente, portanto, é importante destacar que não se considerou as amostras de forma conjunta para a determinação dos genótipos e nem para a chamada de variantes.

Tabela 2. Lista dos genes e regiões, cromossomos (Crom) com as respectivas posições selecionadas para o sequenciamento direcionado

Gene/Região	Crom	posição	Gene/Região	Crom	posição	Gene/Região	Crom.	posição
ACAN	15	89346664 - 89418595	*GLI3*	7	42000538 - 42277479	*PDE4B*	1	66759100 - 66779100
ADCY1	7	45613729 - 45762725	*GSK3B*	3	119540160 - 119813274	NT_006576.16	5	4308910 - 4310969
ADCY2	5	7396311 - 7830204	*ITPR2*	12	26488275 - 26986141	NT_006576.16	5	9965864 - 9968303
ADCY3	2	25042028 - 25142718	*MUC4*	3	195473626 - 195539158	NT_007592.15	6	14701312 - 14704126
APOB	2	21224291 - 21266955	*NRG1*	8	31496892 - 32622568	NT_007819.17	7	8287819 - 8291201
APP	21	27251851 - 27252851	*PARK2*	6	161768442 - 163148844	NT_009952.14	13	21786357 - 21788159
APP	21	27252851 - 27543456	*PDE4D*	5	58264855 - 59817957	NT_009952.14	13	23286884 - 23288808
BRIP1	17	59756537 - 59940930	*PLCB1*	20	8112814 - 8949013	NT_010393.16	16	17909619 - 17912456
CALR	19	13048404 - 13049404	*PRKCA*	17	64298916 - 64806872	NT_010859.14	18	7450207 - 7451781
CALR	19	13049404 - 13055314	*PRKCE*	2	45878474 - 46415139	NT_011109.16	19	3549641 - 3552069
CASR	3	121901520 - 121902520	*PTPRD*	9	8314236 - 10612733	NT_011387.8	20	7604013 - 7606701
CASR	3	121902520 - 122005354	*RAF1*	3	12625090 - 12705735	NT_016354.19	4	29319625 - 29322099
CDH13	16	82660389 - 83830225	*RBFOX1*	16	6069122 - 7763350	NT_016354.19	4	105450162 - 105451914
CDH9	5	26880699 - 27121267	*RELA*	11	65421057 - 65430575	NT_021937.19	1	1021368 - 1022843
CNTNAP2	7	145813443 - 148118100	*SMAD3*	15	67358173 - 67487543	NT_023736.17	8	5094941 - 5096742
CSMD1	8	2792865 - 4852504	*SRC*	20	35973078 - 36034463	NT_025028.14	18	23004514 - 23006360
CTNNB1	3	41236318 - 41301597	*STAT3*	17	40465332 - 40540523	NT_025741.15	6	68545632 - 68547631
EGFR	7	55086704 - 55324323	*TP63*	3	189349195 - 189615078	NT_029419.12	12	9923356 - 9926079
ERBB4	2	212240432 - 213403575	*ZNF280A*	22	22868050 - 22874623	NT_030059.13	10	70198161 - 70200298
FHIT	3	59735026 - 61237143	*ESR1*	6	151938366 - 151958366	NT_030059.13	10	79271368 - 79272941
FHIT	3	61237143 - 61238343	*ESR1*	6	152337857 - 152357858	NT_034772.6	5	36518682 - 36521560
FYN	6	111981525 - 112194665	*MYT1L*	2	2215144 - 2235144			

Genome Analysis Toolkit

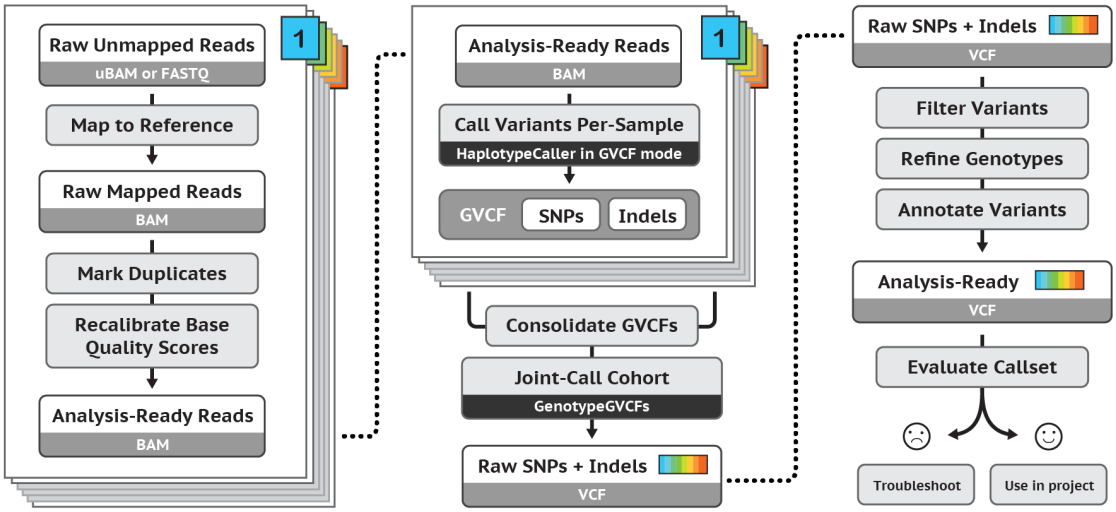
A plataforma de análises GATK ("Genome Analysis Toolkit") (MCKENNA *et al.* 2010), desenvolvida pelo Broad Institute, Cambridge, MA, EUA, é hoje reconhecidamente a que disponibiliza os melhores algoritmos e estratégias de análise de sequenciamento em larga escala (PIROOZNIA *et al.* 2014, LAURIE *et al.* 2016). O *pipeline* indicado para a análise de sequenciamento direcionado - como é o caso das bibliotecas construídas com o kit Haloplex - e identificação de variantes pode ser encontrado entre as opções disponíveis dentro do rol de alternativas do "GATK Best Practices" (VAN DER AUWERA *et al.* 2013).

Esse guia de boas práticas é bastante dinâmico, sendo atualizado constantemente de acordo com a divulgação de melhorias nos algoritmos do GATK e especialmente quando há alteração da versão do programa. Atualmente o programa oferece o algoritmo de chamada de variantes "HaplotyperCaller", o qual apresenta avanços de performance (especialmente com relação às variantes do tipo inserção-deleção) quando comparado ao algoritmo anteriormente oferecido, chamado "UnifiedGenotyper". No início dos trabalhos desse capítulo a versão do GATK disponível e utilizada para as análises era a 3.7. Porém, após alguns meses o programa sofreu alterações mais significativas e a versão 4.0 foi divulgada. Dessa forma, o *pipeline* inicialmente adaptado para o processamento dos dados gerados com o kit Haloplex foi baseado na versão 3.7 do GATK e em seguida readaptado para a versão 4.0.

De toda forma, os procedimentos de implantação do *pipeline* são facilitados tendo em vista que os desenvolvedores do GATK disponibilizam amplo material de apresentação da plataforma, com exemplificações, explicações e tutoriais de trabalho[3]. Nos últimos anos a equipe vem divulgando diferentes plataformas que facilitam bastante a utilização do programa, dentre elas a disponibilização dos *pipelines* na plataforma WDL (*Workflow Description Language* - < http://www.openwdl.org/#three>), a qual, interativamente e com linguagem computacional simples, permite o armazenamento e compartilhamento de *pipelines* de análises. Em conjunto com o WDL, utiliza-se o Cromwell, que é um sistema de gerenciamento de fluxos de trabalho que permite a conexão de diferentes plataformas para a execução dos *pipelines* localmente ou remotamente. O *pipeline* do GATK utiliza outros algoritmos além dos desenvolvidos pelo Broad Institute e programas independentes, como: PICARD, SamTools e BWA-MEM (Figura 3). Vale ressaltar que os desenvolvedores alertam os usuários para a possível necessidade de se adaptar o *pipeline* sugerido de acordo com as especificidades dos dados utilizados.

[3] Disponível em: <https://software.broadinstitute.org/gatk/>. Acesso em 09/08/2017

Figura 3. Etapas do *pipeline* do *Genome Analysis Toolkit* (GATK).



Fonte: Figura retirada do sítio eletrônico do GATK[4].

O *pipeline* "Best-Practices" do GATK (VAN DER AUWERA *et al.* 2013) é dividido em três etapas principais: pré-processamento, descoberta de variantes e refinamento dos dados. Atualmente, a versão 4.0 do programa utiliza arquivos do tipo uBAM (*Binary Alignment/Map* não-mapeado) como entrada inicial. Portanto, o *pipeline* adaptado a partir do GATK implantado para uso do LDGH, incluiu também algumas etapas de pré-processamento além das indicadas no GATK "Best-Practices". As primeiras etapas indicadas no guia do GATK correspondem ao mapeamento das *reads*, marcação das duplicatas, recalibração das bases (Figura 3). São utilizados os programas BWA com o algoritmo MEM para o alinhamento das sequencias ao genoma de referência e também o programa PICARD para a marcação das duplicatas. Em seguida à etapa de pré-processamento há procedimentos que utilizam algoritmos desenvolvidos pelo Broad Institute para a chamada de variantes e genotipagem de cada posição variável de cada amostra. O algoritmo que realiza a chamada das variantes e genotipagem é o "Haplotype Caller", o qual apresenta boa eficiência de acordo com estudos recentes (NI *et al.* 2015, LAURIE *et al.* 2016). Nessa etapa as amostras são tomadas em conjunto, o que permite que as variantes sejam inferidas a partir da análise de todas as amostras concomitantemente, levando à maior acurácia das inferências (POPLIN *et al.* 2017). Ao final dessa etapa são gerados arquivos com a indicação de variantes SNPs e inserções-deleções (indels). Por fim, a última etapa do processo compreende o refinamento das análises para a eliminação de possíveis variantes falso-positivas e também a anotação das

[4] Disponível em: <https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145>. Acesso em 09/06/2017

variantes mantidas. O *pipeline* gera dois arquivos de variantes (.vcf) ao final, um para os SNPs e outro para os indels, que incluem nas colunas os genótipos de cada uma das amostras analisadas.

A utilização do GATK não possui requerimento mínimo de configurações de *hardware* para que seja executado, porém uma boa capacidade de memória RAM deve ser considerada (>8Gb). Por outro lado, há dependências quanto aos sistemas operacionais e aos softwares necessários para as análises. O servidor Sagarana (CELAM/UFMG) é aberto para uso à comunidade científica e foi utilizado para o desenvolvimento do *pipeline* do GATK adaptado. O Sagarana atende às necessidades do processo tendo em vista a alta capacidade computacional do servidor e as dependências já instaladas e disponíveis para uso, como por exemplo R, Python 2.6, Java8. O Sagarana possui vários nós com 64 núcleos e 512 Gb de memória RAM. Além disso, possui um nó especial, para montagem de genomas, com 256 núcleos e 2048 Gb de memória RAM. No total, são quase 300 Tb de HDs. Para a utilização nesse estudo, os nós com 512 Gb de RAM foram utilizados.

Controle de qualidade dos dados

Os dados gerados pelo sequenciador MiSeq (Illumina, Inc.) foram avaliados no programa *MiSeq Control Software* quanto ao percentual de bases sequenciadas acima do índice de qualidade Q30 (BROCKMAN *et al.* 2008). Esse índice de qualidade é resultado do processamento do sinal de fluorescência detectado pelo sistema ótico do equipamento, considerando diversas co-variáveis (como intensidade, contraste, etc.) para a determinação da base. É indicado em escala logarítmica, sendo que o valor 30 representa a probabilidade de 0,001 de haver erro na determinação da base. Outro fator de qualidade avaliado foi o número de *reads* gerados, o qual se espera ser próximo ou superior a 25 milhões de acordo com as especificações do fabricante.

Em seguida à geração dos arquivos *.fastq*, os dados foram avaliados para outras métricas de qualidade pelo uso do programa FASTQC (ANDREWS 2010). Nesse programa foi possível avaliar o percentual de conteúdo GC para cada amostra, a existência de contaminação de sequencias por adaptadores e o tamanho médio das *reads* obtidas.

Com o auxílio e colaboração do PhD. Balast Zsolt, quem esteve por curto período de tempo durante o ano de 2017 no LDGH, a qualidade das variantes geradas utilizando o *pipeline* desenvolvido foi controlada pela comparação com um banco de dados de genotipagem gerado pelo *array* Illumina HumanOmni2.5-8v1.1, o qual permite a

genptpagem de cerca de 2,5 milhões de variantes. No banco de dados do LDGH, das 150 amostras utilizadas nesse estudo, 24 também possuem dados do Illumina HumanOmni2.5-8v1.1, o que permitiu controlar a qualidade dos genótipos com relação às variantes existentes no *array* e que foram também sequenciadas.

Os dados gerados pelo método de genotipagem por *array* são inicialmente manipulados pelo programa *Genome Studio*, disponibilizado pela Illumina, Inc.. Os dados foram analisados pelo então estudante de doutorado Victor Borda, do nosso grupo de pesquisa, para que fossem tratados para retirada de inconsistências e disponibilizados em formatos adequados para as análises seguintes. Inicialmente, os dados foram exportados para os formatos do programa PLINK (*.ped* e *.map*) sendo que os genótipos foram padronizados para os determinados apenas pela fita direta *(forward)* (uso do parâmetro "UseForwardStrand"). Em seguida, já pelo uso do programa PLINK, os dados foram tratados para que os SNPs repetidos fossem retirados e os identificadores "kgp" mantidos pela Illumina para cada variante fossem todos padronizados para o formato com identificador "rs" do dbSNP (Database for Short Genetic Variations – NCBI). Os SNPs com a mesma posição física, mas que possuíam identificadores diferentes foram considerados como a mesma variante, sendo mantidos no grupo de dados final apenas aqueles que possuíam maior índice de qualidade de chamada de base (valor de *Call Rate*). Por fim, para a finalização do controle de qualidade as opções *geno* e *mind* do programa PLINK foram utilizadas para a remoção de todos os SNPs e indivíduos que apresentaram índice de dados faltantes superior a 10% (PURCELL *et al.* 2007).

Exemplo de validação dos resultados

Além do controle dos dados gerados pela validação com os genótipos obtidos pelo uso do *array* Illumina HumanOmni2.5-8v1.1, especificamente o polimorfismo de nucleotídeo único *PDE4B*-rs6683977 (o qual tem especial importância nos estudos reportados no capítulo 2 dessa tese) foi genotipado utilizando-se o método TaqMan (*Assay ID* C___1270960_10 ThermoFicher™) em 58 das 82 amostras de indivíduos de populações peruanas que também foram utilizadas para o sequenciamento direcionado. A análise das concordâncias dos genótipos obtidos com aqueles encontrados pelos métodos de chamada de variante do GATK e *SureCall* são apresentados como indicadores da qualidade dos dados obtidos.

Criação de um pipeline *específico*

Um *script* em linguagem *perl* foi desenvolvido para contemplar todo o processo de tratamento dos dados brutos em formato *.fastq* gerados pelo equipamento MiSeq. Todas as etapas, desde o pré-processamento até a filtragem customizada dos arquivos de variantes finais, foram automatizadas. O *script* levou em consideração as ferramentas instaladas no servidor Sagarana, mas pode ser facilmente adaptado a outras máquinas. O colaborador Balastz Zsolt participou do processo de adaptação do *pipeline* para a versão 4.0 do GATK e do processo de controle de qualidade dos dados pela comparação com os dados de *array* disponíveis para algumas amostras.

Análises descritivas das variantes

A manipulação dos arquivos *.vcf* bem como as análises descritivas dos resultados obtidos após a chamada das variantes foram realizadas nos programas VCFTOOLS (DANECEK *et al.* 2011) e BCFTOOLS (LI 2011).

RESULTADOS

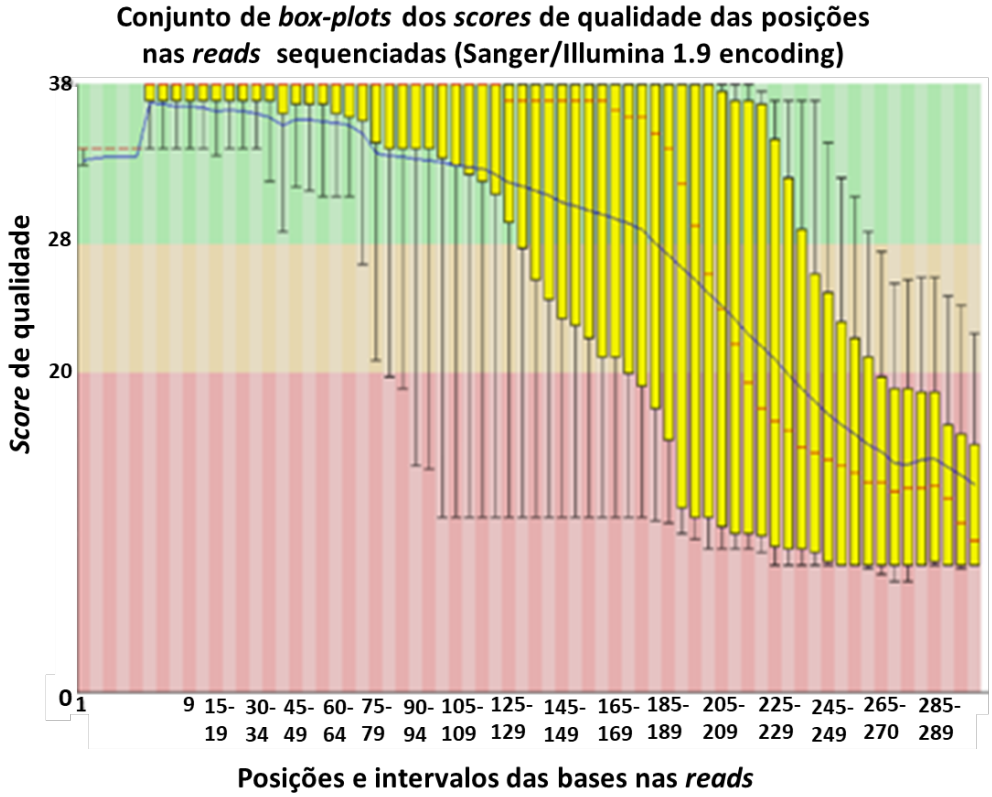
Construção de bibliotecas customizadas

O kit de sequenciamento direcionado Haloplex se mostrou bastante trabalhoso durante a construção das bibliotecas, e também revelou alguns percalços relacionados ao tempo de reação da etapa de hibridização dos oligonucletídeos. Porém, a distribuição do tamanho dos fragmentos das amostras foi na maioria das vezes bastante próxima à esperada (Figura 2).

Controle de qualidade das sequencias

Os dados obtidos do sequenciador são no formato *.fastq*, o qual é adequado para a entrada no programa FASTQC (Andrews 2010). Para cada amostra essa ferramenta disponibiliza informações sobre métricas de qualidade das sequencias, como por exemplo, o percentual de conteúdo GC para cada amostra, a existência de contaminação de sequencias por adaptadores e o tamanho médio das *reads* obtidas. Em um primeiro momento foram identificadas algumas *reads* que apresentaram contaminação por adaptadores e outras sequencias que foram eliminadas do processo antes do prosseguimento das análises. No geral, em todas as amostras a qualidade da maioria das bases sequenciadas ficou acima de Q20, o que indica probabilidade igual ou menor que 1% de erro (Figura 4).

Figura 4. Ilustração adaptada da distribuição da qualidade das bases ordenadas pela posição em todas as *reads* de uma amostra antes de qualquer tratamento, obtida pelo uso do software FastQC. Em verde: >Q28, em amarelo: >Q20, em vermelho: <Q20. Retângulos amarelos: percentis 25º-75º; linhas inferiores e superiores em cada posição: percentis 10° e 90° respectivamente. Linha azul: valor de qualidade médio em cada posição. Linhas em vermelho: mediana.



Surecall

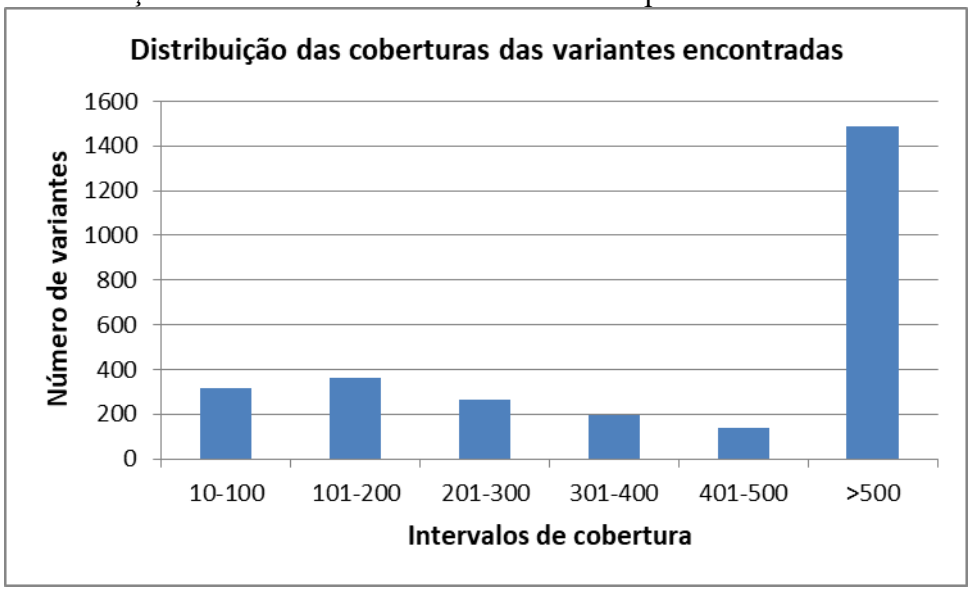
As análises no *SureCall* foram realizadas de forma individual, já que o programa trata cada amostra por vez. Porém, os resultados apresentados a seguir mostram métricas calculadas ao nível global das amostras (n=150) (Tabela 3). Como o próprio programa possui funcionalidades que permitem o pré-processamento das sequências (por exemplo, aparagem das bases finais e iniciais das leituras), não há necessidade de se realizar essa etapa antes de utilizar o programa. Após as análises, o programa também gera um relatório com as métricas de cada etapa como, por exemplo, o número de *reads* aparadas, o número de *reads* descartadas, além de indicar para cada amostra as distribuições da profundidade das bases de acordo com o total de *reads* analisadas. Porém, não há um gráfico como o ilustrado na Figura 4, o que impediu comparações entre as sequências antes e após o processamento. Há outras métricas importantes de qualidade reportadas como a cobertura média das *reads* de cada amostra.

Tabela 3. Números de variantes e suas características obtidos com o *SureCall*

Valores Globais (n=150)							
SNPs	Indels	Sitios multialélicos	SNPs multialélicos	Transições	Transversões	Ts/Tv	Singletons
2354	355	157	4	1607	751	2,14	784

A maioria das variantes encontradas apresentou cobertura alta como mostrado na Figura 5. Das 2760 variantes, 1485 tiveram cobertura superior a 500.

Figura 5. Distribuição da cobertura das variantes inferidas pelo SureCall



Genome Analysis Toolkit (GATK)

Para a execução do GATK foi preciso estabelecer um *pipeline* de análises que contemplasse não apenas o programa em si, mas também as ferramentas acessórias que são necessárias para a realização de todo o processo até a etapa final de chamada das variantes. Dessa forma, em conjunto com o Ph.D. Wagner Magalhães, foi desenvolvido um *masterscript* em *perl* que permitiu integrar diversas etapas do processo, desde a fase de pré-processamento das leituras em *.fastq* obtidas do sequenciador, até a geração final dos arquivos de variantes em formato *.vcf*. O *pipeline* desenvolvido teve como base o sistema do servidor Sagarana, da UFMG e está disponível em domínio público através da iniciativa LDGH-Brazil Scientific Workflow (<http://ldgh.com.br/scientificworkflow/flowcharts.php>), de onde se tem acesso ao repositório github (https://github.com/ldgh/Targeted_Sequencing_Pipeline). No github o pipeline pode ser obtido e posteriormente adaptado para a execução em outros ambientes computacionais. O fluxograma geral das etapas do processo com a especificação dos produtos intermediários de cada etapa pode ser consultado na Figura 6. Para o caso específico do sequenciamento realizado, os parâmetros utilizados em cada etapa podem ser encontrados no *pipeline*.

Na primeira etapa do processo ocorre a aparagem de pares de bases de acordo com a qualidade do sequenciamento para o aumento da qualidade das sequências a serem utilizadas. Para isso foi utilizado o programa Trimmomatic (BOLGER *et al.* 2014). O resultado obtido está ilustrado na Figura 7 em que se percebe expressiva melhora quando comparado ao resultado apresentado na Figura 4, a qual mostra a qualidade das sequencias antes de qualquer processamento.

Após a aparagem das bases e o ajuste de qualidade das sequências o *pipeline* prossegue, sendo que em várias etapas há a geração de gráficos e métricas que auxiliam o usuário a ajustar os parâmetros do programa para a geração de resultados mais robustos. A etapa final do *pipeline* chamada "HardFiltering" permite que o usuário altere parâmetros que influenciam na retenção ou não de variantes no painel final gerado. Nessa etapa os indels e os SNPs são tratados de forma distinta tendo em vista as diferentes características dessas variantes. Os parâmetros a serem ajustados são *Qual By Depth* (QD), *Fisher Strand* (FS), *Strand Odds Ratio* (SOR), *Mapping Quality* (MQ), *Mapping Quality Rank Sum Test* (MQRankSum), *Read Pos Rank Sum Test* (ReadPosRankSum). Brevemente, o índice QUAL (*Quality*) indica o valor de qualidade daquela base. Já o índice QD (*QualByDepth*) representa a normalização do índice QUAL, obtida pela divisão dos índices de QUAL pelos valores de profundidade não-filtrados das amostras não-homozigotas-referência. O índice FS (*FisherStrand*) indica a probabilidade em escala *Phred* (EWING & GREEN 1998) de haver um viés de fita naquela posição (a variante ocorre apenas na fita *forward* ou apenas na *reverse*). O índice SOR (*StrandOddsRatio*) é outra forma de indicar o viés de fita, porém utilizando um teste de *odds-ratio* como base. Por fim, o *ReadPosRankSum* (ReadPosRankSumTest) é um parâmetro que avalia as posições nas *reads* dos alelos referência e alternativo, e com isso permite, por exemplo, identificar variantes que estejam sempre no final das *reads* (o que seria um indicativo de erro para o sequenciamento por síntese da Illumina). Portanto, esses são indicadores gerados para cada variante que de certa forma revelam propriedades, como por exemplo, qual é o contexto das sequencias que circundam a variante, quantas *reads* cobriram essa variante, a proporção de *reads forward* e *reverse* que incluem essas variantes, dentre outros (DE SUMMA *et al.* 2017). Para a tomada de decisão quanto ao ajuste dos parâmetros utiliza-se um banco de dados com variantes de qualidade conhecida que servem como referência.

No caso específico aqui apresentado, variantes do banco dbSNP foram utilizadas para esse propósito. A descrição detalhada desses parâmetros e a importância para os ajustes de filtragem não serão detalhadas nessa tese, tendo em vista a limitação de espaço para

exposição, porém podem ser encontrados no guia de boas práticas do GATK (VAN DER AUWERA *et al.* 2013). As figuras 8 e 9 apresentam os gráficos de alguns desses parâmetros obtidos a partir das amostras analisadas, sendo que apenas os parâmetros cujos valores foram ajustados no *pipeline* desenvolvido para os SNPs e os indels foram incluídos nas figuras. Os gráficos mostram de forma comparativa a distribuição dos parâmetros para variantes do dbSNP e as variantes que estão em análise. Ao final do processo o *pipeline* do GATK apresentou mais variantes do que o obtido com o uso do SureCall (Tabelas 3 e 4).

Tabela 4. Números de variantes e suas características obtidos com o *pipeline* customizado do *Genome Analysis Toolkit* (GATK) antes de qualquer correção. Ts/Tv: transições/transversões.

Valores Globais (n=150)							
SNPs	**Indels**	**Sitios multialélicos**	**SNPs multialélicos**	**Transições**	**Transversões**	**Ts/Tv**	**Singletons**
7258	1613	311	96	4648	2631	1,77	2812

Assim como o resultado obtido com o Surecall, a maioria das variantes encontradas apresentou cobertura alta como mostrado na Figura 10. Porém, é importante notar que a razão de transições por transversões (Ts/Tv) está abaixo de 2. O valor de Ts/Tv próximo a 2 é o esperado quando se considera o genoma como um todo (GATK 2018), tendo em vista a probabilidade duas vezes maior de ocorrer transições do que transversões. Dessa forma, esse foi um indicativo de que as variantes encontradas pelo GATK deveriam passar por um processo de reanálise.

Refinamento dos resultados

Considerando a grande diferença entre os resultados obtidos pelo SureCall e o GATK, buscou-se reanalisar os dados gerados pelo GATK a fim de se obter um painel de variantes com valores de maior confiança. Para tanto, os painéis de indels e de SNPs foram trabalhados de forma separada, tendo em vista a diferença dos dois tipos de variantes. No caso dos indels, como a estratégia de sequenciamento foi de geração de *reads* curtas, a inferência é naturalmente mais difícil, sendo muito sensível ao tamanho do indel e à presença de repetições *in tandem* (GRIMM *et al.* 2013) . Nesse caso, os indels identificados em regiões de homopolímeros foram removidos do painel final de variantes (FANG *et al.* 2014). Da mesma forma, indels que foram identificados em outras regiões *in tandem* que não fossem homopolímeros foram separados em um arquivo diferente dos demais. Por fim, em razão das complexidades que envolvem as variantes do tipo inserção/deleção, essas variantes não foram trabalhadas com mais detalhes e carecem ainda de checagens e reanálises.

Figura 6. Fluxograma geral das etapas do processo de customização baseado no guia de boas práticas do *Genome Analysis Toolkit* (GATK)

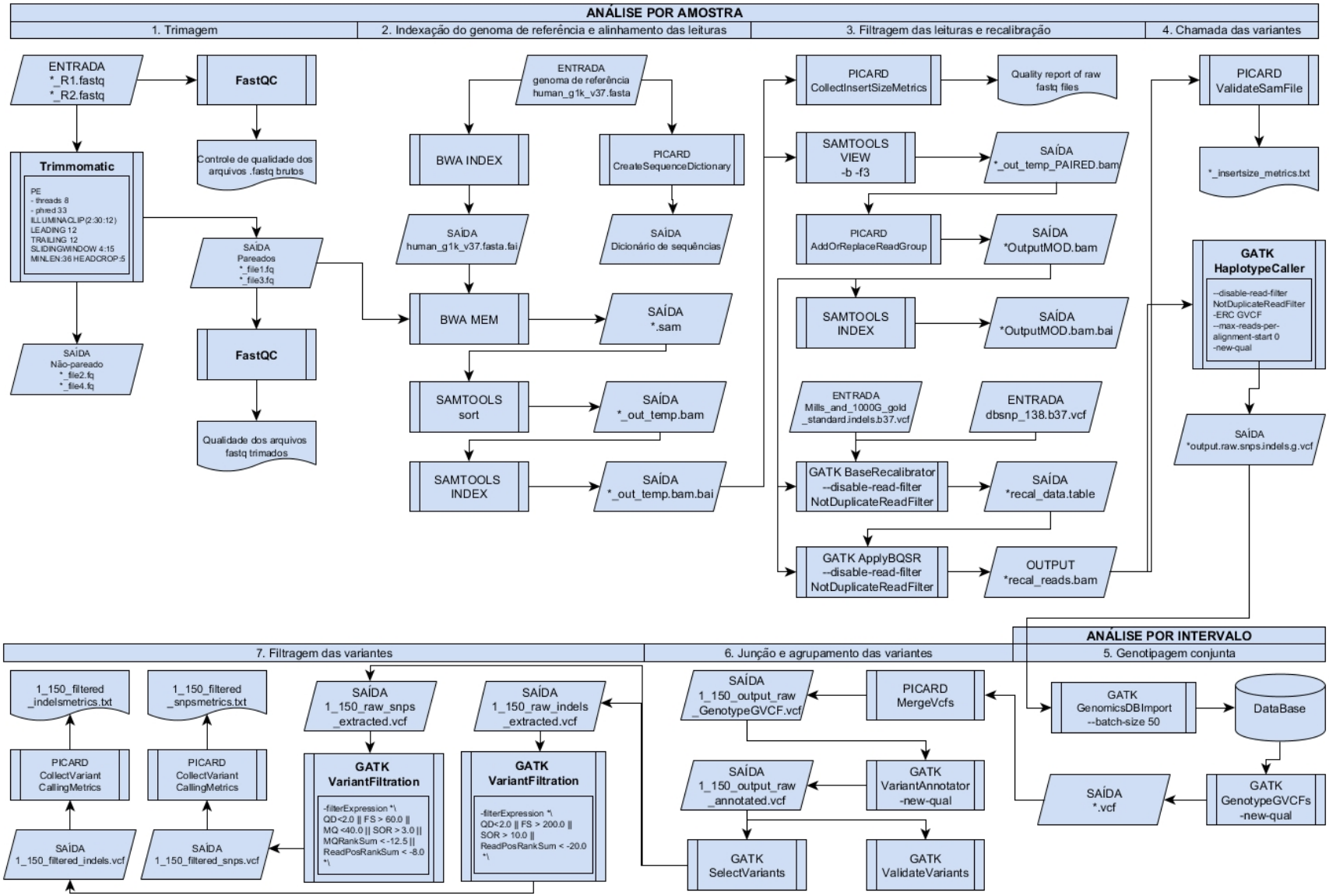


Figura 7. Ilustração adaptada da distribuição da qualidade das bases ordenadas pela posição em todas as *reads* de uma amostra após tratamento pelo programa Trimmomatic. Em verde: >Q28, em amarelo: >Q20, em vermelho: <Q20. Retângulos amarelos: percentis 25º-75º; linhas inferiores e superiores em cada posição: percentis 10° e 90° respectivamente. Linha azul: valor de qualidade médio em cada posição. Linhas em vermelho: mediana.

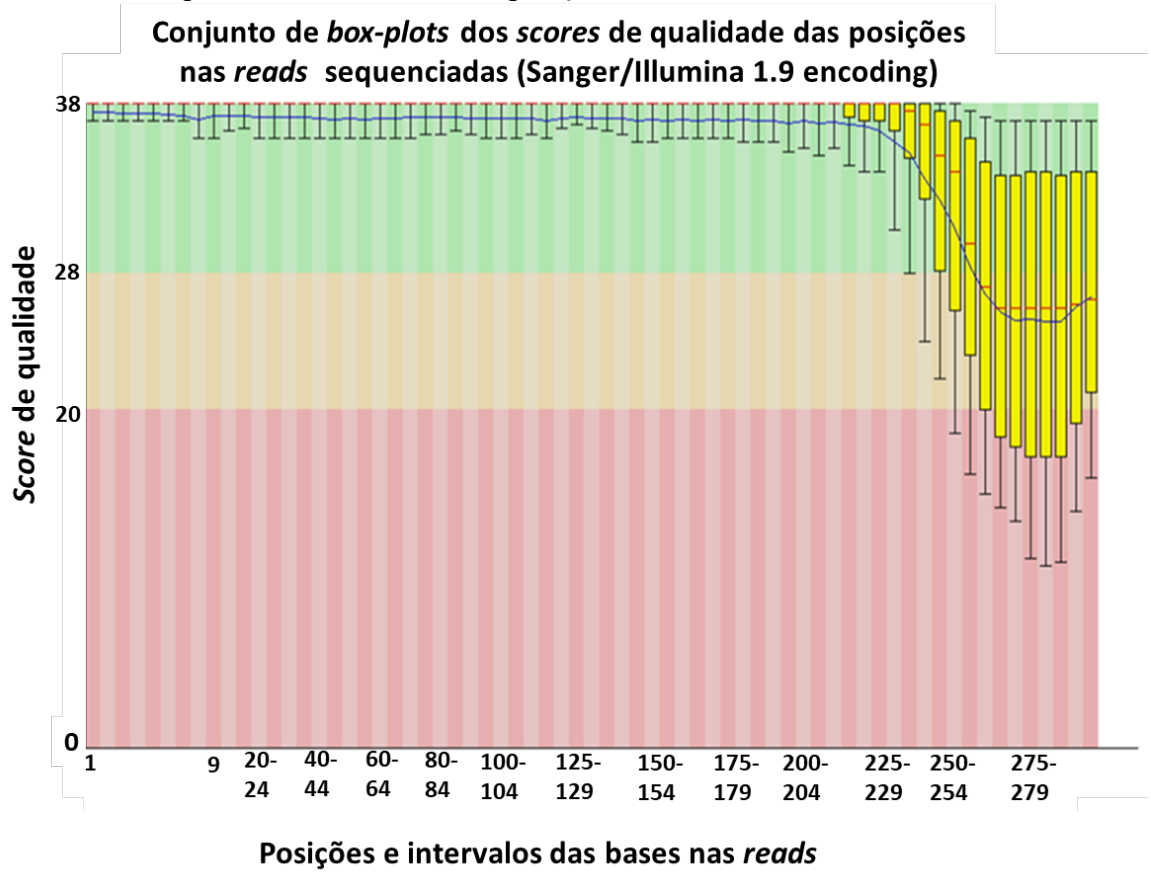


Figura 8. Distribuições dos parâmetros utilizados como indicadores para o processo de filtragem de Indels. SOR: *StrandOddsRatio*, FS: *FisherStrand*, QD: *QualByDepth*, QUAL: Quality, ReadPosRankSum:*ReadPosRankSumTest*. Em vermelho, a distribuição obtida com os SNPs do dbSNP, já em azul a distribuição obtida com as variantes inferidas. As linhas verticais em vermelho indicam os valores utilizados para ajuste das métricas.

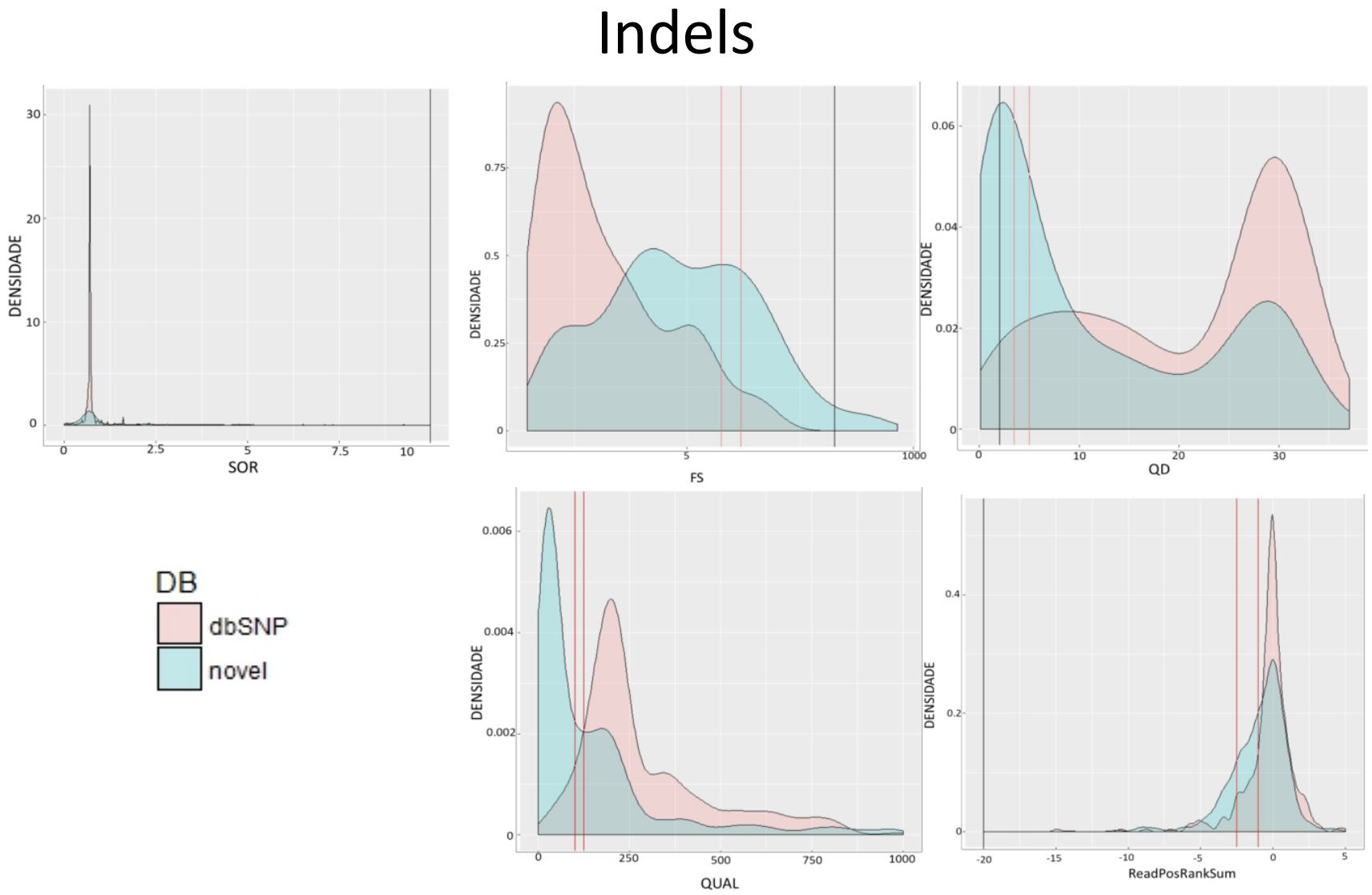


Figura 9. Distribuições dos parâmetros utilizados como indicadores para o processo de filtragem de SNPs. SOR: *StrandOddsRatio*, FS: *FisherStrand*, QD: *QualByDepth*, QUAL: Quality, ReadPosRankSum:*ReadPosRankSumTest*. Em vermelho, a distribuição obtida com os SNPs do dbSNP, já em azul a distribuição obtida com as variantes inferidas. As linhas verticais em vermelho indicam os valores utilizados para ajuste das métricas.

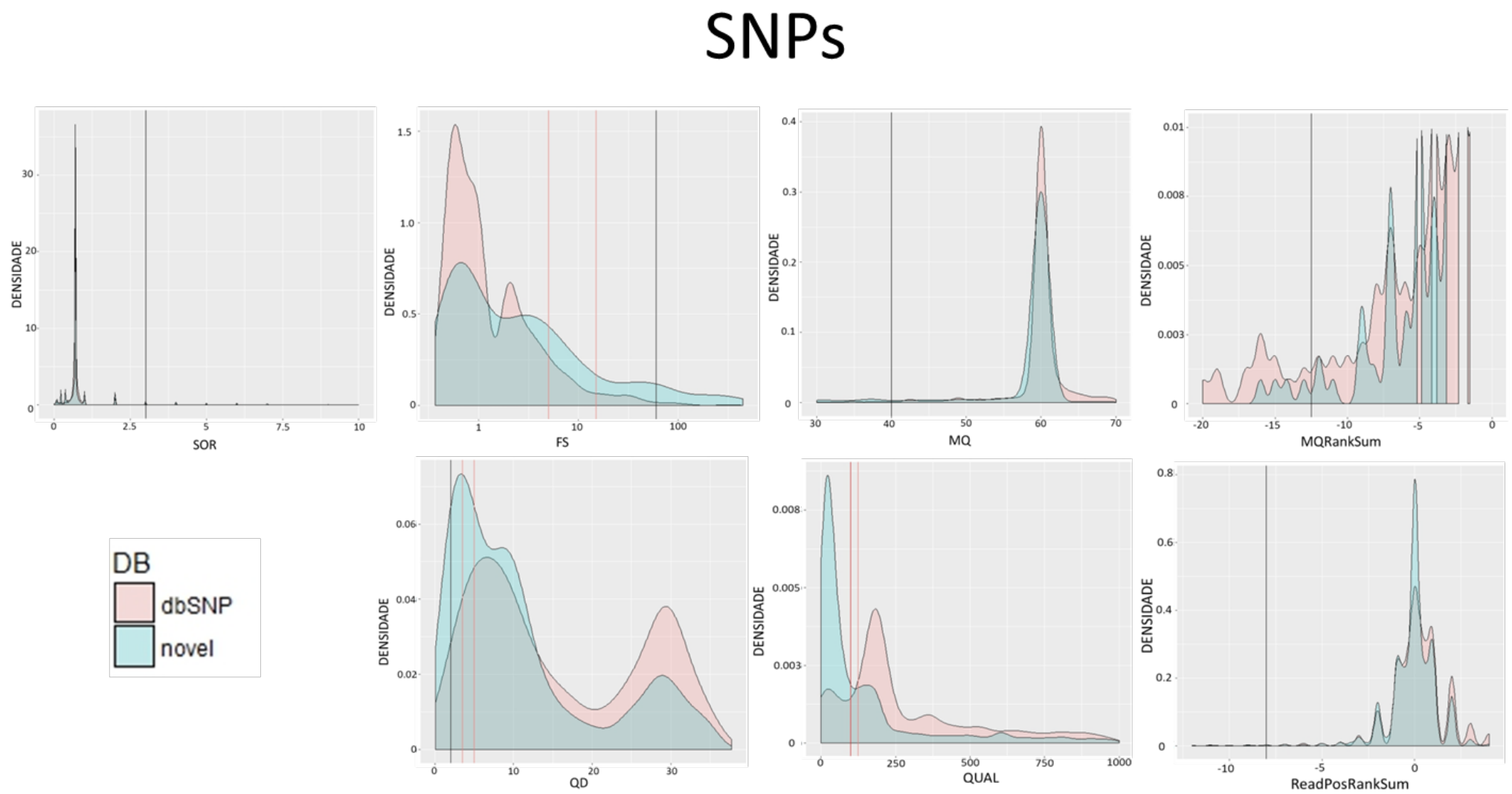
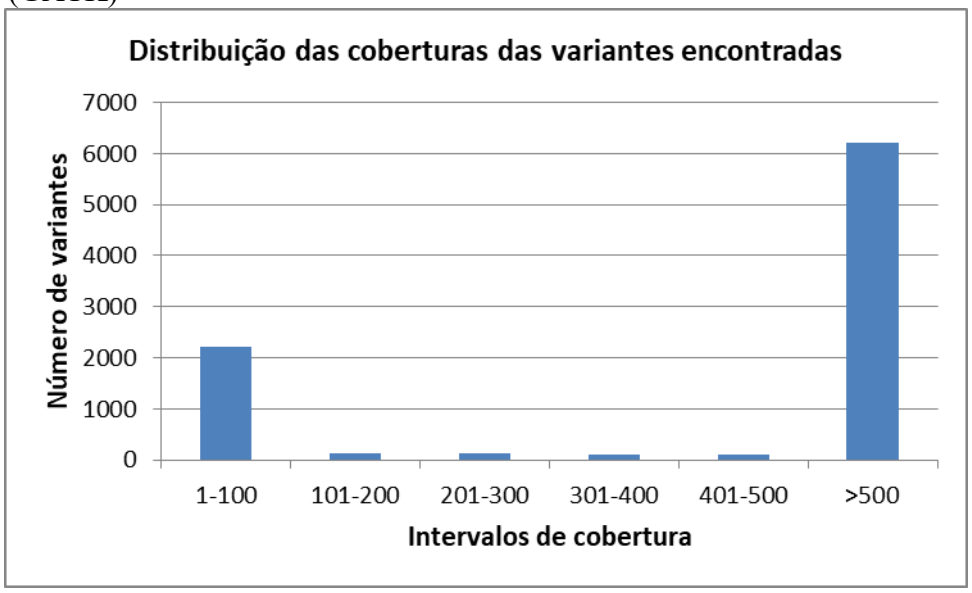


Figura 10. Distribuição da cobertura das variantes inferidas pelo *Genome Analysis Toolkit* (GATK)



As variantes do tipo SNP, por sua vez, foram reanalisadas com maior profundidade tendo em vista a existência de dados de tipagem por *array* para 24 amostras. Para essas análises de refinamento foi desenvolvido em colaboração com o Ph.D. Balast Zsolt um *script* em linguagem *python* que também está disponibilizado no repositório github do LDGH (https://github.com/ldgh/Genotype_Correction). Brevemente, o referido *script* utiliza os arquivos no formato *pileup* gerados pelo programa SamTools como entrada para que os genótipos chamados pelo GATK sejam reanalisados. O *script* é mais restritivo à chamada dos genótipos e considera o número de *reads* em que a variante está presente e também o número de *reads* com diferentes coordenadas de início. Apenas as bases com qualidade >9 foram consideradas. Os parâmetros do *script* também foram ajustados para otimizar a detecção dos genótipos usando como base a genotipagem obtida por *array* disponível para 24 das 150 amostras. O fluxograma com as etapas do *script* pode ser consultado na Figura 11.

Após a execução do *script* sobre o conjunto de amostras, o número de SNPs reduziu de 7258 para 5139 (Tabela 5). A interseção dessas variantes com as existentes no *array* de genotipagem indicou 674 variantes em comum, as quais totalizaram 11.108 genótipos considerando todas as 150 amostras. A checagem da concordância dessas variantes com o *array* de genotipagem indicou que 88, dos 11.108 genótipos chamados no total estavam incorretos, o que representa apenas 0.8% de erro. Notou-se que essas variantes estavam em regiões com particularmente baixa cobertura. A maioria dos genótipos incorretos (84%) foi em razão da chamada de heterozigotos pelo GATK,

quando na verdade o *array* indicou que eram homozigotos. Portanto, com a aplicação do *script* e a correção das variantes, o número total de genótipos passou a 9.702, o que indicou a perda de 1.406 genótipos. Isso se refletiu no aumento de genótipos não chamados (*missing data*). Assim, antes da correção havia 3.423 SNPs com menos de 10% de *missing data*, sendo que após a correção, 2.686 passaram a ter menos de 10% de *missing data*. De toda forma, vale ressaltar o aumento da relação Ts/Tv (transição/transversão) para 1,96, próximo ao esperado teórico (Tabela 5). Assim como com o painel inferido pelo SureCall, do total de 5.139 SNPs a maioria (=3615) apresentou cobertura acima de 100x.

Figura 11. Fluxograma do processo de refinamento dos resultados obtidos a partir do *pipeline* customizado. i= número de *reads* independentes que suportam a variante; r = número total de *reads* que suportam a variante; os números 1, 2 e 3 indicam a primeira, segunda e terceira variante mais abundante.

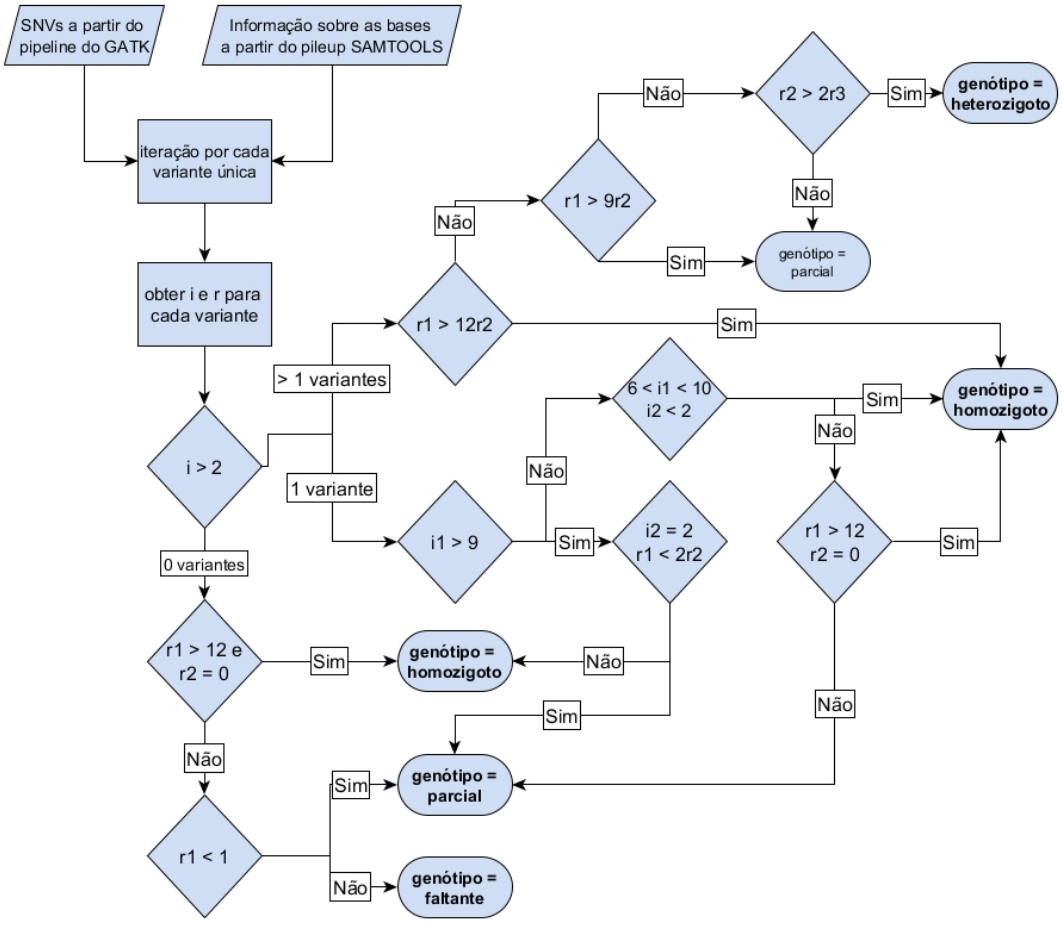


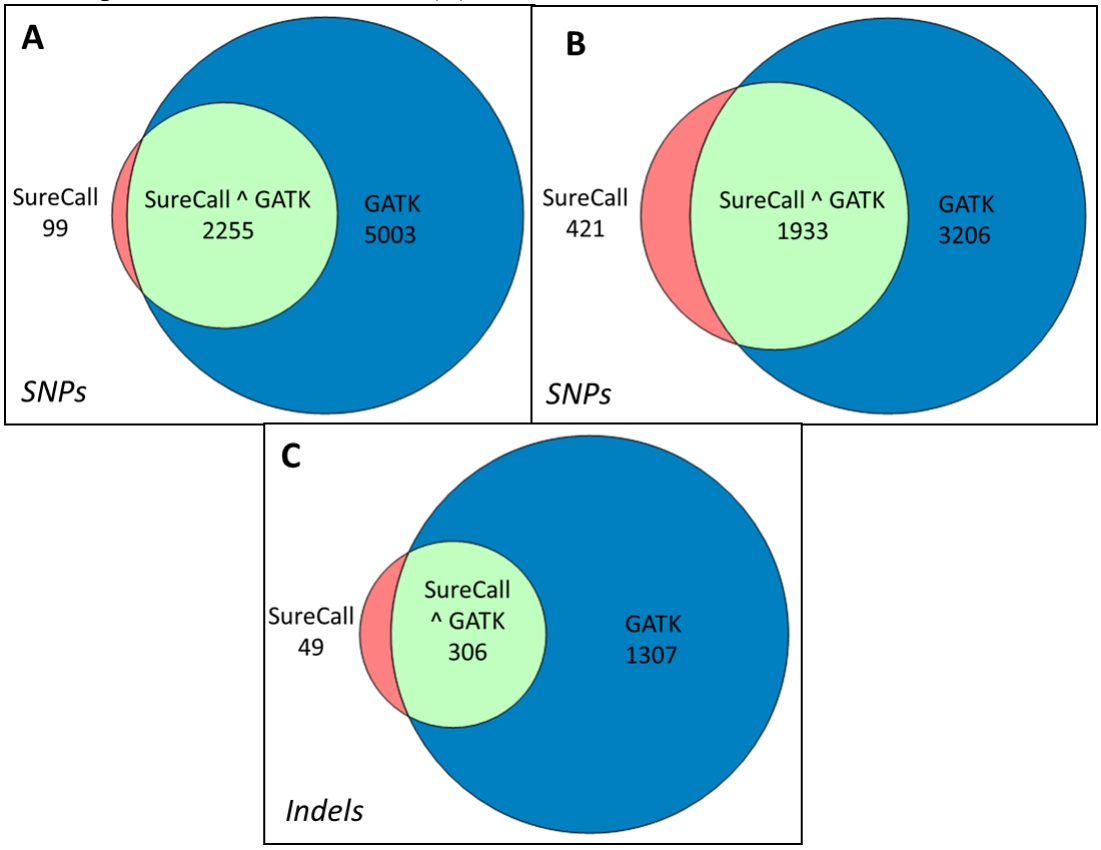
Tabela 5. Números de variantes e suas características obtidos com o *pipeline* customizado do *Genome Analysis Toolkit* (GATK) após correções

Valores Globais (n=150)					
SNPs	**SNPs multialélicos**	**Transições**	**Transversões**	**Ts/Tv**	**Singletons**
5139	29	3411	1743	1,96	1011

Comparação dos resultados obtidos pelo SURECALL e pelo GATK

Os diagramas de Venn apresentam as concordâncias e discordâncias das variantes inferidas por ambos os métodos. As comparações foram feitas de três formas: (i) concordância das variantes SNPs entre os programas antes da correção de genótipos, (ii) concordância dos SNPs após a realização da correção e (iii) concordância dos *indels* (já que não foi aplicada correção aos *indels*) (Figura 12).

Figura 12. Diagramas de Venn indicando o número de SNPs concordantes entre os programas SureCall e *Genome Analysis Toolkit* (GATK) antes (A) e após (B) a aplicação da correção dos genótipos. A concordância das variantes do tipo indel inferidas pelo SureCall e o GATK (C).



Validação dos genótipos obtidos para o SNP rs6683977

Considerando as 58 amostras em que o processo de genotipagem do SNP rs6683977 também foi realizado pelo método TaqMan (Assay ID: C___1270960_10 - Thermo Fisher™), apenas um genótipo foi discordante entre os encontrados pelos métodos de chamada de variantes por sequenciamento. Os métodos TaqMan e a chamada de variantes do GATK indicaram apenas uma amostra (SH247 – *Machiguenga*) como homozigota para o alelo C desse SNP. Já o método de chamadas de variantes do SureCall indicou que esse SNP seria heterozigoto para a mesma amostra.

DISCUSSÃO

A análise dos resultados apresentados indica diferenças expressivas entre os programas utilizados para a chamada de variantes. Além disso, também ressalta a importância da construção de procedimentos customizados para a geração de sequencias de ácidos nucléicos que contemplem as especificidades do ensaio biológico proposto.

Apesar de trabalhoso quando comparado a outros procedimentos de construção de bibliotecas de sequenciamento direcionado, o Haloplex se mostrou uma estratégia eficiente. Por se tratar de um método que utiliza oligonucleotídeos para hibridização e captura dos fragmentos de interesse, o procedimento como um todo foi realizado em maior tempo do que caso fosse utilizada a estratégia de uso de PCR para amplificação das regiões alvo. Porém, como bastante discutido na literatura, a utilização da técnica de captura por sondas possui vantagens frente à estratégia de PCR multiplex para o sequenciamento direcionado (BODI *et al.* 2013, KOZAREWA *et al.* 2015, SAMORODNITSKY *et al.* 2015).

As análises dos gráficos gerados pelo programa FASTQC (ANDREWS 2010) mostram que as *reads* tiveram, no geral, bases com boa qualidade, sendo que a maioria esteve acima do índice de qualidade Q30, indicando probabilidade de erro de cerca de 0,1%. Porém, a Figura 4 revelou que as bases de pior qualidade de sequenciamento estavam ao final das *reads* de leitura. Esse é um padrão esperado para o tipo de sequenciamento por síntese, como o usado pela Illumina no MiSeq. Como o sinal da base é lido a partir da síntese de várias cópias do mesmo fragmento de DNA, o avanço dos ciclos de síntese acumulam erros nas diferentes cópias do mesmo fragmento à medida que as fitas são geradas, o que dificulta a leitura das bases próximas à extremidade 3’ de fragmentos longos (BOLGER *et al.* 2014). O padrão de geração de erros do sequenciamento Illumina é bem conhecido e por isso é tratado de forma adequada nas etapas de pré-processamento dos programas, incluindo o SureCall e o *pipeline* adaptado do GATK desenvolvido nesse estudo.

SureCall

O software disponibilizado pela fabricante do kit Haloplex (Agilent Tech.) para a geração do painel de variantes apresentou facilidades no processo de instalação e de utilização. Apesar de ser um programa que executa algoritmos complexos, os requerimentos de máquina permitiram que fosse instalado em um computador de mesa. A escolha dos parâmetros para a geração do painel de variantes não se mostrou complexa e revelou uma grande variedade de métricas abertas aos ajustes do usuário

caso necessário. Como exemplo, cita-se a possibilidade de ajuste dos parâmetros de alinhamento e de aparagem.

Por outro lado, os procedimentos de análise do programa não são claramente discutidos, impossibilitando compreender com detalhes como as etapas de chamada de variantes são realizadas. Apesar da divulgação pela empresa do funcionamento geral do algoritmo (ASHUTOSH *et al.* 2014), pouco é esclarecido quanto ao tratamento dado a variantes em regiões repetitivas e também para a chamada de indels no caso de erros de alinhamento (comuns em regiões hipervariáveis - veja NIELSEN *et al.* 2011) ou mesmo aos filtros usados antes da geração do painel final de mutações.

Há dois aspectos que precisam ser ressaltados nesse estudo e que são de grande relevância para as análises realizadas no SureCall. O primeiro deles diz respeito à forma como o programa trata as *reads* que possuem início e término idênticos, o que indica que sejam duplicatas de PCR de um mesmo fragmento (CASBON *et al.* 2011). Porém, tendo em vista que o processo de geração das bibliotecas Haloplex utiliza enzimas de restrição para fragmentação do DNA, existirão *reads* idênticas que serão provenientes de fragmentos de DNA de células diferentes; assim como *reads* que são cópias idênticas de um único fragmento de DNA geradas pelos ciclos de PCR no processo de construção da biblioteca. Diante disto, o tratamento dessas *reads* influencia diretamente sobre o processo de chamada de variantes, porém não há informação clara sobre a melhor forma de tratar essa questão no caso do Haloplex. Após buscas em fóruns e por haver menção em um estudo científico (SAMORODNITSKY *et al.* 2015), entende-se que a empresa sugere que sejam mantidas as duplicatas no processo de análise. Essa questão é de fato importante e merece atenção nos procedimentos de construção de bibliotecas, ao ponto de atualmente a solução para esse impasse ser um diferencial de mercado. A própria tecnologia Haloplex já passou a ser adaptada para, ao nível molecular, promover a marcação das *reads* que são duplicatas e assim permitir que o procedimento de chamadas de variantes seja realizado da melhor forma possível. Atualmente pode-se adquirir o kit Haloplex^{HS}, o qual possui em uma das fases iniciais do protocolo de construção das bibliotecas uma etapa de inserção de sequências específicas nos fragmentos de interesse. Essa identificação fragmento-específica permite que todas as cópias provenientes de cada fragmento sejam identificadas nas análises bioinformáticas e possam assim ser tratadas devidamente. Esse avanço na tecnologia, geralmente conhecido como “barcodes moleculares”, contribui bastante para a redução dos erros de

chamada de variantes e especialmente para identificação de variantes que tenham baixa frequência na amostra, como em casos de tecidos somáticos (ZOBECK *et al.* 2016).

O outro aspecto que merece ser discutido diz respeito ao processo de chamada de variantes, o qual pode ser realizado de forma individual, ou seja, amostra por amostra ou pela análise das amostras em conjunto. O SureCall promove a chamada das variantes em cada amostra de forma separada. As análises em amostras múltiplas são apenas possíveis para amostras de trio ou amostras em pares (como, por exemplo, amostras de tecido normal e tecido tumoral) (ASHUTOSH *et al.* 2014). A estratégia de chamadas de variantes em cada amostra em separado não leva em conta importantes informações que podem ser consistentes e concordantes entre todas as amostras. Sabe-se que as taxas de erros variam ao longo do genoma, sendo dependentes da sequência local, porém, esse padrão tende a se repetir em todas as amostras (MURALIDHARAN *et al.* 2011). Ao se promover a chamada de variantes quando todas as amostras estão reunidas em um mesmo grupo a análise é mais robusta, já que o suporte estatístico tende a ser maior quando *reads* diferentes, mesmo que em pouca quantidade, apontam a existência de certa variante consistentemente em várias amostras (DE PRISTO *et al.* 2011, LI 2011, NIELSEN *et al.* 2011). Portanto, tendo em vista o caráter negligenciado das populações utilizadas nesse estudo (como será discutido no capítulo 2 dessa Tese), a existência de variantes ainda não conhecidas, bem como de variantes com baixa frequência, ressalta-se a importância de se promover análises considerando a estratégia que agrupa as amostras para a chamada de variantes.

Os resultados obtidos pela execução do SureCall e apresentados na Tabela 3 mostram valores dentro do esperado para análises das regiões do ensaio. Em especial é importante destacar a razão transição/transversão (Ts/Tv) próxima ao valor teórico de 2.0 e o número de variantes únicas (*singletons*) sendo representativo da maioria das variantes. Além disso, a maior parte das variantes apresentou alta cobertura, o que é esperado tendo em vista a influência do número de *reads* para a confiança na chamada das variantes.

Genome Analysis Toolkit – GATK

Como o software GATK é uma ferramenta de livre distribuição, sua instalação e execução não são desenvolvidas com a finalidade de apresentar ao usuário uma interface amigável de utilização. Porém, apesar de requerer conhecimentos em sistema Unix e também ser necessário instalar outras ferramentas acessórias, o processo de

instalação e uso do GATK não apresentou obstáculos. Os fóruns, tutoriais e artigos disponibilizados pelo Broad Institute foram extremamente importantes para a familiarização com o processo de execução do programa. O guia de boas práticas divulgado e constantemente atualizado (VAN DER AUWERA *et al.* 2013) é uma referência importante para qualquer usuário iniciante, pois permite que todo o processo de análise do programa seja compreendido e, caso necessário, seja customizado de acordo com as peculiaridades do ensaio em análise. Esse é um diferencial do GATK frente a vários programas de distribuição livre, e supera também muitos dos programas comercializados no mercado. Tendo em vista o grande número de usuários do programa e a dedicação de muitos dos profissionais desenvolvedores em contribuir para a divulgação e facilitação de uso, a existência de canais de comunicação permite que quaisquer dúvidas sejam expostas e muitas vezes prontamente esclarecidas.

A necessidade de se instalar outras ferramentas para o uso do GATK não é um impedimento a sua execução já que o programa está adaptado para fazer uso dos arquivos de saída dessas ferramentas. Porém, para a automação do processo é importante que um *pipeline* de análises que integre todo o processo seja desenvolvido. Assim, o *script* disponibilizado na plataforma *github* favorecerá projetos futuros que necessitem de customização para a geração de dados. Com o uso de ferramentas acessórias como o Trimmomatic (BOLGER *et al.* 2014), o resultado após a aparagem das bases de baixa qualidade pôde ser reavaliado, o que permitiu identificar o aumento da qualidade da maioria das bases das sequências e traçar um comparativo (Figuras 4 e 7), destacando o ganho obtido com essa etapa de pré-processamento.

Quanto à questão da remoção das *reads* idênticas (duplicatas) discutidas anteriormente, o mesmo se aplica ao procedimento no GATK. Sendo assim, por ser uma característica do procedimento de construção de biblioteca utilizado, as duplicatas foram mantidas ao longo de todo o processo de análise. Por sua vez, o GATK permite que as análises sejam realizadas em amostras múltiplas, ou seja, todas as amostras são tomadas em conjunto para a inferência das variantes. Dessa forma, entende-se que esse seja um ganho considerável do uso do GATK (Mielczarek *et al.* 2016) quando comparado ao Surecall.

Infelizmente, considerando que o Haloplex é uma estratégia de sequenciamento direcionado que não gera sequencias de grandes porções do genoma, o processo de chamada de variantes sugerido pelo GATK para esses casos é baseado na etapa de "HardFiltering" e não na etapa de VQSR (*Variant Quality Score Recalibration*). O

processo VQSR é um método avançado disponibilizado pelo GATK que permite um melhor balanço entre especificidade e sensibilidade. Isso é possível em razão da estratégia de aprendizado de máquina do programa, a qual utiliza os próprios dados em análise para identificar a melhor forma de filtragem das variantes. Como o painel do Haloplex é relativamente pequeno, não é adequado ao processo de VQSR, pois não disponibiliza informação suficiente para o processo de aprendizagem de máquina. Assim, o guia de boas práticas do GATK sugere o procedimento de "Hard Filtering" (DE SUMMA *et al.* 2017), o qual é bastante explicado em artigos de instrução do BroadInstitute, nos quais também encontram-se sugestões de valores para os parâmetros de filtragem. Além disso, pelo uso do programa estatístico R (R CORE TEAM 2013), o GATK permite que a filtragem das variantes seja ajustada de acordo com plotagens que facilitam ao usuário estabelecer valores de corte para os parâmetros. Dessa forma, apesar da impossibilidade de uso do VQSR, os valores de diversas métricas de filtragem foram ajustados considerando a distribuição de variantes conhecidas (dbSNP), como mostrado nas Figuras 8 e 9, o que eleva a confiança nas variantes novas reportadas pelo processo.

Refinamento das análises do GATK

Como os valores da Tabela 4 foram expressivos, especialmente pela razão Ts/Tv, o procedimento de refinamento das variantes foi realizado tanto pelo uso da estratégia de *mpileup* do SamTools quanto pela comparação com as variantes obtidas pelo método de genotipagem por *array*. Após o refinamento das análises, os valores finais mostrados na Tabela 5 revelam que a razão Ts/Tv foi próxima a 2, como esperado no campo teórico[5]. Ademais, a redução do número de variantes de aproximadamente 7.258 (Tabela 4) para 5.129 (Tabela 5), pode indicar a existência de variantes falso-positivas no painel final. Mas essa interpretação é relativizada quando se avalia os resultados comparativos entre o GATK e o SureCall, os quais serão discutidos em momento oportuno a seguir. Com relação às variantes do tipo indel, o presente estudo sugere que análises mais profundas sejam realizadas sobre essas variantes para que um painel final seja estabelecido.

Comparativo entre SureCall e GATK

Tendo em vista os resultados obtidos e as concordâncias e discordâncias das variantes encontradas em ambos os programas (Figura 12), pode-se notar que o GATK

[5] Disponível em: <http://rosalind.info/glossary/transitiontransversion-ratio/>. Acesso em: 31/04/2017

revelou um número bastante superior de variantes, sendo a análise realizada antes (Figura 12A) ou após (Figura 12B) o refinamento das variantes inferidas pelo GATK. Porém, é importante notar que após o refinamento das análises do GATK, grande parte das variantes não mais consideradas (322 variantes de um total de 2.129 filtradas) foram também inferidas pelo SureCall, o que conferiria a essas variantes grande grau de confiança. Sabe-se que o método escolhido para o refinamento das análises do GATK é bastante restritivo, o que pode ter levado à eliminação de variantes verdadeiras do painel final. Porém, é importante ressaltar que a validação do SNP rs6683977 foi concordante para quase todas as amostras, com exceção do único homozigoto para o alelo C encontrado. Nesse caso, o SureCall indicou que essa amostra seria heterozigota, o que reforça a importância do método de chamada de variantes em amostras múltiplas implantado no GATK, o qual pode ser mais preciso em casos de chamada de genótipos do que quando a chamada ocorre em cada amostra individualmente. Portanto, dois conjuntos de dados de SNPs inferidos pelo GATK estão disponíveis como resultado desse estudo: (i) conjunto inicial com 7.258 SNPs, sendo 3.423 SNPs com menos de 10% de *missing data* e (ii) conjunto após a correção com 5.129 SNPs sendo 2.686 SNPs com menos de 10% de *missing data*.

A análise do compartilhamento de resultados relativos às variantes do tipo indel mostrou alta concordância (Figura 12C), sendo que das 355 variantes inferidas pelo SureCall apenas 49 não foram inferidas pelo GATK. Porém, o GATK inferiu um numero quase cinco vezes superior de variantes. É provável que esse resultado esteja relacionado com a maior eficiência do GATK, que promove o realinhamento local com o algoritmo *HaplotypeCaller* quando há identificação de variações na região genômica em análise. De toda forma, mesmo levando em conta que o GATK possui hoje um dos melhores algoritmos para a inferência de variações do tipo indel, sugere-se que o painel reportado seja utilizado com cautela e que as variantes de interesse específico sejam confirmadas por métodos manuais, como por meio da visualização no programa IGV (*Integrative Genomics Viewer*) (ROBINSON *et al.* 2017).

Por fim, os números expressivos de variantes identificadas pelo GATK e em menor escala também identificados pelo *Surecall* podem ser explicados pela natureza das amostras utilizadas nesse estudo. Como as amostras são provenientes de populações negligenciadas, e que, portanto, não são devidamente representadas em bancos de dados genômicos, espera-se que muitas variantes encontradas não tenham ainda sido

reportadas, especialmente as consideradas raras (baixa frequência) (NEEL 1978, GRAVEL *et al.* 2011).

Conclui-se que o processo de geração de dados genômicos por sequenciamento massivo em paralelo é ainda um campo em desenvolvimento e que constantemente vem apresentando estratégias mais confiáveis e eficazes para a determinação de variações e genótipos. Os resultados apresentados aqui corroboraram a hipótese inicial que indica a importância de se incorporar ajustes ao processo de geração de dados e chamada de variantes que contemplem as especificidades dos dados a serem gerados, incluindo os tipos de amostras e parâmetros de qualidade dos sequenciamentos realizados. O desenvolvimento de *pipelines* específicos auxilia na automação dos processos e permite que os melhores algoritmos e estratégias de análises sejam incorporados e atualizados à medida que avanços na área sejam alcançados. Os painéis de variantes gerados nesse estudo possuem alto grau de confiabilidade e são adequados ao uso em estudos posteriores tendo em vista os parâmetros restritos de análise aplicados considerando duas estratégias diferentes disponíveis para análises de dados. Porém, muitas variantes verdadeiras podem ter sido eliminadas em razão do uso de parâmetros de filtragem rigorosos. Dessa forma, reanálises futuras devem ser promovidas sempre que métodos e algoritmos mais eficientes e com reconhecido ganho de confiança nos dados gerados sejam desenvolvidos.

REFERÊNCIAS

AIRD, Daniel *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. **Genome biology**, v. 12, n. 2, p. R18, 2011.

ALIOTO, Tyler S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. **Nature communications**, v. 6, p. 10001, 2015.

ANDREWS, Simon *et al.* FastQC: a quality control tool for high throughput sequence data. 2010.

ASHUTOSH, D.J. *et al.* 2014, SNPPET: A Fast and Sensitive Algorithm for Variant Detection and Confirmation From Targeted Sequencing Data. Disponível em: <https://cdn.technologynetworks.com/ep/pdfs/snppet-a-fast-and-sensitive-algorithm-for-variant-detection-and-confirmation-from-targeted.pdf>. Acesso em: 20 de jan. de 2019.

BAMSHAD, Michael J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. **Nature Reviews Genetics**, v. 12, n. 11, p. 745, 2011.

BARAN-GALE, Jeanette *et al.* Massively differential bias between two widely used Illumina library preparation methods for small RNA sequencing. **bioRxiv**, p. 001479, 2013.

BENTLEY, David R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. **Nature**, v. 456, n. 7218, p. 53, 2008.

BODI, Kip *et al.* Comparison of commercially available target enrichment methods for next-generation sequencing. Journal of biomolecular techniques: **JBT**, v. 24, n. 2, p. 73, 2013.

BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014..

BROCKMAN, William *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems**. Genome research**, v. 18, n. 5, p. 763-770, 2008.

CASBON, James A. *et al.* A method for counting PCR template molecules with application to next-generation sequencing. Nucleic acids research, v. 39, n. 12, p. e81-e81, 2011.

DANECEK, Petr *et al.* The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156-2158, 2011.

DE SUMMA, Simona *et al.* GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. **BMC bioinformatics**, v. 18, n. 5, p. 119, 2017.

DEPRISTO, Mark A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nature genetics**, v. 43, n. 5, p. 491, 2011.

EWING, Brent; GREEN, Phil. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome research**, v. 8, n. 3, p. 186-194, 1998.

FANG, Han *et al.* Reducing INDEL calling errors in whole genome and exome sequencing data. **Genome medicine**, v. 6, n. 10, p. 89, 2014.

GARRISON, Erik; MARTH, Gabor. Haplotype-based variant detection from short-read sequencing. **arXiv preprint arXiv**:1207.3907, 2012.

GENOME ANALYSIS TOOLKIT - Evaluating the quality of a variant callset. Disponível em <https://software.broadinstitute.org/gatk/documentation/article?id=6308>. Acesso em 10 out 2018.

GRAVEL, Simon *et al.* Demographic history and rare allele sharing among human populations. **Proceedings of the National Academy of Sciences**, v. 108, n. 29, p. 11983-11988, 2011.

GRIMM, Dominik *et al.* Accurate indel prediction using paired-end short reads. **BMC genomics**, v. 14, n. 1, p. 132, 2013.

HAMPEL, Ken J. *et al.* Variant call concordance between two laboratory-developed, solid tumor targeted genomic profiling assays using distinct workflows and sequencing instruments. **Experimental and molecular pathology**, v. 102, n. 2, p. 215-218, 2017.

HWANG, Sohyun *et al.* Systematic comparison of variant calling pipelines using gold standard personal exome variants. **Scientific reports**, v. 5, p. 17875, 2015.

JIANG, Zhihua *et al.* Genome wide sampling sequencing for SNP genotyping: methods, challenges and future development. **International journal of biological sciences**, v. 12, n. 1, p. 100, 2016.

JONES, Marcus B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. **Proceedings of the National Academy of Sciences,** v. 112, n. 45, p. 14024-14029, 2015.

KEHDY, Fernanda SG *et al.* Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. **Proceedings of the National Academy of Science**s, v. 112, n. 28, p. 8696-8701, 2015.

KOZAREWA, Iwanka *et al.* Overview of target enrichment strategies. **Current protocols in molecular biology,** v. 112, n. 1, p. 7.21. 1-7.21. 23, 2015.

KUMAR, Santosh; BANKS, Travis W.; CLOUTIER, Sylvie. SNP discovery through next-generation sequencing and its applications**. International journal of plant genomics**, v. 2012, 2012.

KWONG, Jason C. *et al.* Whole genome sequencing in clinical and public health microbiology. **Pathology**, v. 47, n. 3, p. 199-210, 2015.

LAURIE, Steve *et al.* From wet-lab to variations: Concordance and speed of bioinformatics pipelines for whole genome and whole exome sequencing. **Human mutation**, v. 37, n. 12, p. 1263-1271, 2016.

LI, Heng. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, v. 27, n. 21, p. 2987-2993, 2011.

LI, Heng *et al.* The sequence alignment/map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, 2009.

MAGI, Alberto *et al.* Bioinformatics for next generation sequencing data. **Genes**, v. 1, n. 2, p. 294-307, 2010.

MAMANOVA, Lira *et al.* Target-enrichment strategies for next-generation sequencing. **Nature methods,** v. 7, n. 2, p. 111, 2010.

MATTICK, John S. *et al.* Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. **The Medical Journal of Australia**, v. 209, n. 5, p. 197-199, 2018.

MCKENNA, Aaron *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome research**, v. 20, n. 9, p. 1297-1303, 2010.

MEIENBERG, Janine *et al.* Clinical sequencing: is WGS the better WES?. **Human genetics**, v. 135, n. 3, p. 359-362, 2016.

MIELCZAREK, M.; SZYDA, J. Review of alignment and SNP calling algorithms for next-generation sequencing data. **Journal of applied genetics**, v. 57, n. 1, p. 71-79, 2016.

MUDGE, Jonathan M.; HARROW, Jennifer. The state of play in higher eukaryote gene annotation. **Nature Reviews Genetics**, v. 17, n. 12, p. 758, 2016.

MURALIDHARAN, Omkar *et al.* A cross-sample statistical model for SNP detection in short-read sequencing data. **Nucleic acids research**, v. 40, n. 1, p. e5-e5, 2011.

NEEL, James V. Rare variants, private polymorphisms, and locus heterozygosity in Amerindian populations. **American journal of human genetics**, v. 30, n. 5, p. 465, 1978.

NIELSEN, Rasmus *et al.* Genotype and SNP calling from next-generation sequencing data. **Nature Reviews Genetics**, v. 12, n. 6, p. 443, 2011.

O'RAWE, Jason *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. **Genome medicine,** v. 5, n. 3, p. 28, 2013.

PARK, Sang Tae; KIM, Jayoung. Trends in next-generation sequencing and a new era for whole genome sequencing. **International neurourology journal**, v. 20, n. Suppl 2, p. S76, 2016.

POPLIN, Ryan *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. **BioRxiv**, p. 201178, 2018.

PURCELL, Shaun *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American journal of human genetics**, v. 81, n. 3, p. 559-575, 2007.

QUAIL, Michael A. *et al.* Optimal enzymes for amplifying sequencing libraries. **Nature methods**, v. 9, n. 1, p. 10, 2012.

TEAM, R. Core *et al.* R: A language and environment for statistical computing. 2013. Vienna, Austria. URL <http://www.R-project.org/.>.

ROBINSON, James T. *et al.* Variant review with the integrative genomics viewer. **Cancer research**, v. 77, n. 21, p. e31-e34, 2017.

ROTHBERG, Jonathan M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. **Nature**, v. 475, n. 7356, p. 348, 2011.

ROY, Somak *et al.* Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. **The Journal of Molecular Diagnostics**, v. 20, n. 1, p. 4-27, 2018.

SAMORODNITSKY, Eric *et al.* Comparison of custom capture for targeted next-generation DNA sequencing. **The Journal of Molecular Diagnostics**, v. 17, n. 1, p. 64-75, 2015. **a**

SAMORODNITSKY, Eric *et al.* Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. **Human mutation**, v. 36, n. 9, p. 903-914, 2015.**b**

SANDMANN, Sarah *et al.* Evaluating variant calling tools for non-matched next-generation sequencing data. **Scientific reports**, v. 7, p. 43169, 2017.

SHENDURE, Jay; JI, Hanlee. Next-generation DNA sequencing. **Nature biotechnology**, v. 26, n. 10, p. 1135, 2008.

SIKKEMA-RADDATZ, Birgit *et al.* Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. **Human mutation**, v. 34, n. 7, p. 1035-1042, 2013.

SOLONENKO, Sergei A. *et al.* Sequencing platform and library preparation choices impact viral metagenomes. **BMC genomics**, v. 14, n. 1, p. 320, 2013.

SOUZA, G. Indentificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e implicações biomédicas. Dissertação (Mestrado em Genética) – Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, p. 107. 2010.

VAN DER AUWERA, Geraldine A. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. **Current protocols in bioinformatics,** v. 43, n. 1, p. 11.10. 1-11.10. 33, 2013.

VAN DIJK, Erwin L.; JASZCZYSZYN, Yan; THERMES, Claude. Library preparation methods for next-generation sequencing: tone down the bias. **Experimental cell research**, v. 322, n. 1, p. 12-20, 2014.

ZOBECK, Katie L. *et al.* HaloPlexHS Utilizes Molecular Barcodes to Improve Low Frequency Allele Detection. **Cancer Genetics**, v. 209, n. 6, p. 296, 2016.

CAPÍTULO 2. ANCESTRALIDADE NATIVO-AMERICANA E IMPLICAÇÕES CLÍNICAS PARA A LEUCEMIA LINFOBLÁSTICA AGUDA: ANÁLISES EM ESCALA FINA DE VARIANTES COM IMPORTÂNCIA BIOMÉDICA

INTRODUÇÃO

Avanços nas últimas décadas nas tecnologias de genotipagem e sequenciamento de ácidos nucléicos criaram as bases para o desenvolvimento da medicina personalizada (RABBANI *et al.* 2016). Atualmente vivenciamos a transição do que há alguns anos era apenas uma perspectiva para uma nova realidade na área médica, em que o perfil genético dos pacientes revela informações precisas sobre susceptibilidade a doenças, permite melhores prognósticos e até mesmo a personalização de tratamentos (SCHORK 2015). Apesar de algumas vezes serem considerados paradoxais (BURKE *et al.* 2010) e alvos de ceticismo (BAYER & GALEA 2015), esses avanços contribuem também para progressos no campo da saúde pública, criando perspectivas para o que se propõe ser a "saúde pública de precisão" (KHOURY *et al.* 2016).

Dados genético-populacionais são fontes de informação que contribuem para a melhor compreensão sobre a diversidade observada na prevalência de enfermidades e nas respostas aos tratamentos disponíveis em diferentes populações do globo (NORDLING 2017). A diversidade genética humana revela as bases para essa heterogeneidade observada no campo médico, tendo em vista que as distintas histórias evolutivas dos diferentes grupos humanos podem guardar as respostas para muitas questões epidemiológicas ainda em aberto (LU *et al.* 2014, QUINTANA-MURCI 2016, NORDLING 2017). Não obstante, vários estudos têm demonstrado associações entre aspectos clínicos e a ancestralidade populacional, tanto relacionados à susceptibilidade à doenças quanto à resposta a tratamentos (CSERVENKA *et al.* 2017, MORENO *et al.* 2017, DING *et al.* 2018). Portanto, o progresso no campo da medicina genômica permite que essas associações possam ser estudadas refinadamente (FIGUEROA *et al.* 2016, QUINTANA-MURCI 2016, SUN *et al.* 2016). Mesmo que ainda distante de todas as respostas para a reconhecida complexidade das enfermidades e da melhor forma de tratá-las, a contribuição dos estudos genético-populacionais para o avanço médico é notória. O detalhamento dos estudos para escalas moleculares irá

inegavelmente contribuir para a melhor compreensão das bases genéticas das doenças que acometem as diferentes populações humanas.

Ancestralidade nativo-americana e miscigenação nas Américas

As populações nativo-americanas têm história evolutiva particular quando comparadas às demais populações de outros continentes. As hipóteses mais aceitas para a origem dessas populações sugerem que sejam resultado de ondas migratórias de relativamente poucos indivíduos em eventos de migração isolados provenientes da Ásia, pela região da Beríngia, que posteriormente se espalharam pelo continente americano (REICH *et al.* 2012, SKOGLUND *et al.* 2015). A dinâmica de formação dessas populações compreendeu especialmente grupos de tamanho reduzido, mas também eventos de expansão populacional, além de momentos e circunstancias tanto de fluxo gênico restrito quanto de interações intensas entre populações (TARAZONA-SANTOS *et al.* 2001; WANG *et al.* 2007; 2008). Um dos exemplos que melhor ilustra a existência de uma estrutura genética e cultural característica foi a formação do Império Inca nas regiões dos Andes da América do Sul nos últimos 4.000 anos. A formação desse grupo foi resultado de vários eventos de contração e expansão populacionais que marcaram períodos de surgimento e declínio de culturas diversificadas e a ocorrência de fluxo gênico entre populações anteriormente isoladas, inclusive com populações do leste do continente estabelecidas na floresta amazônica (STANISH 2001, BARBIERI *et al.* 2014). Não obstante, a complexidade genética das populações do continente Americano foi posteriormente incrementada pela interação das populações nativas com indivíduos de origem europeia e africana. As chegadas de conquistadores europeus e escravos africanos nos últimos séculos resultaram em eventos de miscigenação que contribuíram para a formação de populações geneticamente bastante diversas, que atualmente estão espalhadas por todo o continente Americano (WANG *et al.* 2008; KEHDY *et al.* 2015; BOLNICK *et al.* 2016).

Há várias populações no continente Americano que são resultantes da miscigenação de indivíduos de diferentes ancestralidades, como ocorre, por exemplo, com comunidades de descendentes nativo-americanos nos Estados Unidos, que são denominados ameríndios, hispânicos ou latinos. Esses indivíduos têm, na maioria das vezes, grande parte de seu genoma de origem nativo-americana, mas também compartilham material genético de origem europeia e africana (PRICE *et al.* 2007, BRYC *et al.* 2015). Cada uma dessas porções do genoma guarda informações sobre a

história evolutiva do grupo ancestral, o qual conviveu com pressões seletivas peculiares e sofreu de forma particular os efeitos de condições demográficas únicas, como os causados pela deriva genética em populações de tamanho reduzido. Portanto, as características das variantes genéticas dessas populações são particulares, devendo-se levar em conta a existência de variantes exclusivas, diferenças nas frequências alélicas de vários locos quando comparadas às demais populações do globo e, consequentemente, padrões de desequilíbrio de ligação próprios desses grupos (GONZÁLEZ-BURCHARD *et al.* 2005).

Mapeamento por miscigenação em populações com ancestralidade nativo-americana

Estudos de mapeamento por miscigenação usando painéis de marcadores informativos de ancestralidade (MIA), seguidos de procedimentos de genotipagem em escala mais densa ou sequenciamento para o mapeamento-fino, têm sido utilizados com sucesso na identificação e localização de locos genéticos que estejam relacionados à susceptibilidade a doenças e a outras características fenotípicas humanas (SELDIN *et al.* 2011). Grande parte desses estudos considera variantes que são raras ou monomórficas em uma ou mais das populações ancestrais. Vale ressaltar que a importância das variantes raras vem ganhando destaque na literatura no que diz respeito ao papel que desempenham na causa de doenças complexas (IYENGAR & ELSTON 2007). A maioria dos estudos focados em doenças complexas mostra que as causas têm contribuição de uma única mutação rara, recente e de amplo efeito ou de um conjunto de várias mutações raras, recentes e de efeito restrito (MITCHELL 2012). Dessa forma, a análise dessas variantes se mostra mais importante em populações ainda pouco estudadas, como as de origem nativo-americana e as miscigenadas. A presença nessas populações de possíveis alelos comuns exclusivos ainda não conhecidos e de mutações raras ainda não reportadas representa uma oportunidade única para o avanço no conhecimento acerca da base genética de algumas doenças humanas (GONZÁLEZ-BURCHARD *et al.* 2005).

Especificamente para as populações latinas e hispânicas do continente Americano, estudos de mapeamento por mistura demonstram haver alelos em locos com importância biomédica com frequências diferenciadas quando comparadas a outras populações. Galanter e colaboradores (2014) estudaram a incidência de asma em populações de hispânicos de grandes metrópoles americanas buscando encontrar

associação entre a ancestralidade e os genes que estariam envolvidos com a doença. Os autores realizaram estudo de associação por todo o genoma e posteriormente promoveram o mapeamento por mistura. Os resultados indicaram um sinal ainda não reportado de associação com a doença em certa região do cromossomo 6, e mostraram que a ancestralidade nativo-americana apresentava efeito protetivo sobre a manifestação da asma. Além disso, os autores confirmaram associação previamente reportada na literatura para uma região no cromossomo 17. Drake e colaboradores (2014) realizaram estudo de mapeamento por mistura em hispânicos dos Estados Unidos e Porto Rico para avaliar a associação com a resposta a drogas broncodilatadoras. Os autores promoveram um estudo de associação por todo o genoma e posteriormente o mapeamento por mistura e o mapeamento fino. Os resultados indicaram a existência de variantes raras nos genes da família *SLC* (proteínas de membrana envolvidas no transporte de metabólitos endógenos e xenobióticos) que se mostraram associadas a diferentes respostas à droga broncodilatadora albuterol em latinos.

Há certa variedade de estudos que mostram resultados semelhantes para outras características, especialmente as relacionadas à asma e aos tratamentos empregados para essa doença (HERNANDEZ-PACHECO *et al.* 2016). Porém, há também evidências de diferenciação em genes que desempenham papéis em outros tipos de doenças. Em razão da grande importância médica, estudos que buscam associação entre variantes genéticas e doenças como o câncer também têm demonstrado haver relação com a ancestralidade (WALSH *et al.* 2013, ZHAO *et al.* 2016, BERMEJO *et al.* 2017, KAROL *et al.* 2017, KAUR *et al.* 2018).

Leucemia Linfoblástica Aguda (LLA)

A Leucemia Linfoblástica Aguda (LLA) é uma enfermidade que compreende um grupo heterogêneo de condições em sua maioria relacionadas a lesões genéticas conhecidas nas células precursoras dos linfócitos B e T. Essas condições incluem tanto a inibição do desenvolvimento das células hematopoiéticas em estágio inicial, quanto alterações no mecanismo molecular de regulação da divisão celular, causando a proliferação descontrolada dessas células na medula óssea (BIOINDI *et al.* 2016). As alterações genéticas que caracterizam a reconhecida heterogeneidade da LLA incluem mutações pontuais em genes conhecidos (*ARID5B*, *CEBPE*, *GATA3*, e *IKZF1*), assim como alterações estruturais, como translocações e também aneuploidias (HARRISON 2009). Os perfis moleculares específicos das células neoplásicas são características que

influenciam diretamente no tratamento aplicado, na resposta ao tratamento ministrado e no risco de recidiva (SOUSA *et al.* 2015). Dentre as alterações genéticas mais estudadas da LLA, a translocação cromossômica recíproca entre os cromossomos 9 e 22, t(9,22), que causa a fusão dos genes *BCR-ABL1*, conhecida como cromossomo Philadelphia, é a que possui maior impacto sobre o tratamento e o prognóstico, com baixos índices de resposta ao tratamento e baixas taxas de sobrevivência (KURZROCK *et al.* 1988). No caso da LLA em crianças, a translocação mais comum ocorre entre regiões dos cromossomos 12 e 21, causando a fusão dos genes *ETV6-RUNX1* (QUIROZ *et al*. 2019). Porém, recentemente vêm sendo identificadas novas alterações que, apesar de não apresentarem a translocação t(9,22), são tão agressivas e de prognóstico tão desfavorável quanto a cromossomo Philadelphia; e que por isso passaram a ser chamadas *Philadelphia-like* (TASIAN *et al.* 2017).

A LLA é o tipo de leucemia mais comum em crianças, sendo responsável por aproximadamente 80% dos casos. Já em adultos corresponde a apenas cerca de 20% dos casos de leucemia (JABBOUR *et al.* 2015). Em crianças a LLA de linfócitos B (80-85%) é mais comum do que a de linfócitos T (10-15%). Em adultos jovens (16-21 anos) a frequência da LLA de linfócitos T sobe para cerca de 25% dos casos (LEVINE *et al.* 2016). Dados de um estudo revisional recente mostram números relativamente atualizados a nível global para diversos subtipos de leucemia. Segundo Miranda-Filho *et al.* (2018), a incidência da LLA varia de acordo com diversos fatores, mas principalmente com a idade, sendo mais comum em crianças com menos de 5 anos e em adultos com mais de 65 anos; com o gênero, sendo mais comum no sexo masculino; e com o país e a origem. Com relação a essas duas últimas características, há uma maior prevalência em países da América Latina e indivíduos com ancestralidade nativo-americana. De acordo com dados levantados por Miranda-Filho e colaboradores (2018) a prevalência de LLA no Equador, na Costa Rica, e na Colômbia foram as três maiores em comparação a todos os demais países dos cinco continentes pesquisados.

A maior prevalência de LLA em países da América Latina está em concordância com resultados encontrados por estudos que indicam a importância da origem étnica dos acometidos. Para a LLA a ancestralidade nativo-americana é um fator reconhecido de risco (LIM *et al.* 2014, WALSH *et al.* 2013, QUIROZ *et al*. 2019) e está possivelmente associada a diversos mecanismos moleculares de susceptibilidade a esse tipo de leucemia. Algumas das evidências encontradas mostraram que certas mutações

germinativas pontuais, como nos genes *GATA3* (PEREZ-ANDREU *et al.* 2013), *CYP1A1* (SWINNEY *et al.* 2011) e *ARID5B* (XU *et al.* 2012) estão associadas à maior prevalência de LLA em populações com ancestralidade nativo-americana.

Prognóstico, recidiva e tratamentos para a LLA

Apesar de ser uma doença de elevada incidência em várias populações, a LLA tem taxa de sobrevivência relativamente alta, sendo de aproximadamente 80% em casos pediátricos (COOPER *et al.* 2015). Por outro lado, a taxa de sobrevivência em adultos cai para cerca de 20 a 40% (SIVE *et al.* 2012). Mas deve-se ressaltar que esses números refletem a média mundial, sendo que em países com menor índice de desenvolvimento humano (IDH), as taxas de sobrevivência são menores. Em estudos recentes conduzidos no México as taxas de sobrevida de LLA pediátrica reportadas foram em torno de 55% a 67% (JIMÉNEZ-HERNÁNDEZ *et al.* 2015, JAIME-PÉREZ *et al.* 2016). Esses valores são concordantes com o que se encontra em outros países da América Central, como na Guatemala (64%) (ANTILLÓN *et al.* 2017). Ademais, considerando não apenas a LLA, mas outros tipos de neoplasias que acometem crianças, há registros de taxas ainda mais baixas e discrepâncias expressivas quando se traça um comparativo entre países levando-se em conta o nível de desenvolvimento (RODRIGUEZ-GALINDO *et al.* 2015).

A taxa de sobrevivência também está relacionada de maneira direta e expressiva à taxa de recidiva (NGUYEN *et al*. 2008, HUNGER & MULLIGHAN 2015). A tendência de índices mais alarmantes nos países de menor desenvolvimento econômico e social se repete quando se avaliam as taxas de recidiva e as taxas de sobrevivência após a recidiva (JAIME-PÉREZ *et al.* 2018). Infelizmente, os prognósticos para aqueles que são acometidos pela recidiva da LLA são desfavoráveis. A dificuldade no tratamento da neoplasia recidivada ocorre em razão das características das células tumorais que se tornam resistentes ao tratamento. Após a exposição ao tratamento inicial, as células resistentes tendem a permanecer e se proliferar especialmente em razão de alterações em genes que desempenham papel na diferenciação celular, como *IKZF1*, *EBF1* e *CDKN2A*/*2B* (BHOJWANI *et al.* 2013). Além dessas alterações, outras já foram também identificadas e revelam a grande diversidade de células que são elevadas à alta frequência com a recidiva. Segundo já reportado (HUNGER & MULLIGHAN 2015), na média, os exomas de células tumorais possuem cerca de 10 a 20 mutações codificantes não silenciosas no momento do diagnóstico, mas após o

evento da recidiva, pode-se identificar um número duas vezes maior desse tipo de mutações.

As taxas de sobrevivência geral e as taxas de recidiva mencionadas anteriormente são também consequência direta dos tratamentos ministrados aos pacientes, os quais muitas vezes, quando são oferecidos, não são adaptados à realidade biológica e socioeconômica local (MAGRATH *et al.* 2013). Como já exposto anteriormente, a LLA é uma enfermidade heterogênea, que engloba diversos subtipos de neoplasias as quais acometem diferentemente pacientes pediátricos, jovens-adultos e adultos, o que interfere diretamente nos prognósticos e nos tratamentos ministrados. O rol de opções de tratamento inclui diversos protocolos de quimioterapia disponíveis, radioterapia, transplante de medula óssea e imunoterapia.

A radioterapia direta do sistema nervoso central é uma opção que vem sendo abandonada em razão de seus diversos efeitos colaterais, enquanto que a opção de transplante de medula óssea é na grande parte das vezes escolhida nos casos em que ocorre recidiva (HUNGER & MULLIGHAN 2015). Apesar de ser uma das opções mais antigas no tratamento da LLA, a imunoterapia é a alternativa que mais apresenta avanços recentes e grandes perspectivas para o desenvolvimento de novas drogas. A partir de 2014, com o lançamento do fármaco *blinatumomab*, resultados favoráveis vêm sendo reportados para o tratamento da LLA. Essa droga é constituída por anticorpos modificados que permitem o reconhecimento das células tipo B leucêmicas pelas células T (BATLEVI *et al.* 2016). Já a quimioterapia é a alternativa com mais difusão global dentre os tratamentos disponíveis, e por isso também apresenta diversas opções de protocolos (PUI & EVANS 2006). A base dos protocolos utilizados no tratamento da LLA pediátrica foi desenvolvida na década de 1970, pelo grupo conhecido como Berlin-Frankfurt-Münster (BFM). No geral a base protocolar para o tratamento da LLA envolve fases distintas: indução, consolidação e manutenção a longo prazo. Também inclui a profilaxia do sistema nervoso central após certos intervalos periódicos ao longo do tratamento (TERWILLIGER & ABDUL-HAY 2017).

A fase de indução tem como objetivo promover a remissão completa e assim restaurar a hematopoiese. Essa fase dura geralmente 4 a 6 semanas e compreende a prescrição de glicocorticoides, vincristina, uma preparação com asparaginase, e em alguns casos antraciclina (HUNGER & MULLIGHAN 2015, TERWILLIGER & ABDUL-HAY 2017). Essa fase geralmente leva à remissão, porém as chances de

recidiva são bastante expressivas caso o tratamento não avance às fases seguintes. As drogas ministradas após a indução têm o papel de consolidar a remissão e por isso dão nome à fase seguinte de 'consolidação'. Essa etapa dura cerca de 6 a 8 meses de uma combinação intensiva de quimioterápicos que variam bastante de acordo com o protocolo seguido. Após a etapa de consolidação, protocolos baseados no BFM sugerem uma fase de intensificação tardia, também chamada de reindução. Essa fase compreende a administração de glicocorticoides por 8 semanas. É reconhecidamente uma etapa com alta citotoxidade e que por isso sugere que seja intercalada a administração do glicocorticoide com ácido folínico para recuperar os tecidos não-afetados dos efeitos citotóxicos. Vale ressaltar que a maior preocupação nos tratamentos quimioterápicos ministrados se dá em razão do balanço eficácia-citotoxidade (HIJIYA & VAN DER SLUIS 2016, SCHMIEGELOW 2016). A forma como os fármacos são administrados, incluindo a frequência e as doses prescritas, interfere diretamente na eficácia do tratamento. Por essa razão há diversas discussões na literatura médica que indicam a necessidade de se reavaliar protocolos para a supressão ou redução de algumas dessas fases descritas, em especial a fase de intensificação tardia (PUI 2013, VORA *et al.* 2013, SCHRAPPE *et al.* 2016). Por fim, a fase posterior à consolidação ou intensificação tardia (quando é o caso) é conhecida como a fase de manutenção, na qual os pacientes recebem terapia de baixa intensidade por 18 a 30 meses. Geralmente são ministrados glicocorticoides, mercaptopurina ou tioguanina, e vincristina de forma espaçada.

Os recentes e expressivos avanços no tratamento da LLA vêm ocorrendo não apenas pelo desenvolvimento de novas drogas, mas também pelo ajuste dos protocolos de quimioterapia levando em conta o perfil genético-molecular do paciente e da neoplasia (HUNGER & MULLIGHAN 2015). A era da medicina personalizada já teve início e seus benefícios se mostram evidentes no tratamento da LLA (ARIËS *et al.* 2015, JACKSON *et al.* 2016, KAROL *et al.* 2017). Porém, essa realidade ainda não está disponível àqueles que buscam tratamentos na maioria dos centros de oncologia nos países subdesenvolvidos ou em desenvolvimento. E para que esses avanços sejam aplicáveis, e benéficos às populações desses países, será também necessário compreender melhor os aspectos biológicos peculiares às populações desses locais. Os estudos clínicos que são estabelecidos para o desenvolvimento de novas drogas são na maioria realizados em países desenvolvidos, cujo perfil genético dos pacientes é primariamente de origem europeia (POPEJOY & FULLERTON 2016, SIRUGO *et al.*

2019). Portanto, isso indica que novos quimioterápicos e protocolos são desenvolvidos tendo como base genomas cujas variantes e suas frequências alélicas são próprias de populações europeias.

As consequências dessa disparidade acerca do conhecimento genético-molecular dos perfis das células neoplasicas e dos pacientes entre diferentes populações faz com que muitas pessoas não se beneficiem dos progressos dos tratamentos da LLA. Em alguns casos, inclusive, essa diferença de resposta aos tratamentos de acordo com a ancestralidade predominante do paciente pode causar prejuízos notáveis. Como exemplo, Yang e colaboradores mostraram que a ancestralidade nativo-americana está associada à maior probabilidade de recidiva em crianças com LLA (YANG *et al.* 2011). Ao estudarem mais de 2.500 amostras de indivíduos (dos quais 405 se autodeclararam hispânicos) submetidos ao tratamento da LLA, os autores encontraram fortes sinais nos genes *PDE4B* e *MYT1L* que indicaram associação da ancestralidade nativo-americana (indivíduos com mais de 10% de ancestralidade nativo-americana) com a recidiva da doença. Os SNPs *PDE4B*-rs6683977 e *MYT1L*-rs17039396 foram os mais fortemente associados à recidiva de LLA. Os autores detalham os resultados especialmente para o SNP *PDE4B*-rs6683977 e demonstram que o alelo de risco em homozigose alcança probabilidades significativamente superiores de recidiva dos que as observadas nos indivíduos com genótipos heterozigotos ou homozigotos para o outro alelo.

É importante ressaltar que nesse estudo a associação observada se manteve mesmo ajustando para conhecidos fatores de risco de recidiva, tais como: contagem de leucócitos, idade no momento do diagnóstico, subtipo molecular das células tumorais, entre outros. Segundo o estudo, a associação se perdia nos casos em que era ministrada uma fase de intensificação tardia. Porém, vale notar que após a fase de indução um importante indicador da probabilidade de recidiva é o índice DRM (doença residual mínima); e no estudo de Yang e colaboradores mesmo entre os indivíduos com baixo DRM a associação da ancestralidade nativo-americana com a recidiva foi observada (YANG *et al.* 2011). Como o índice de DRM após a fase de indução é um indicativo importante para as etapas seguintes do tratamento (VORA *et al.* 2013, SCHRAPPE *et al.* 2016), possivelmente os indivíduos com alta ancestralidade nativa têm maior probabilidade de redução da intensidade da fase de intensificação ou até mesmo serem liberados dessa fase (PUI 2013). Como mostrado nesse estudo seria importante que a fase de intensificação tardia fosse mantida naqueles em que o componente de

ancestralidade nativo-americana fosse superior a 10% pois, apesar de poderem apresentar baixos níveis de DRM, a chance de recidiva é alta nesses indivíduos.

PDE4B *e* MYT1L – *mecanismos moleculares*

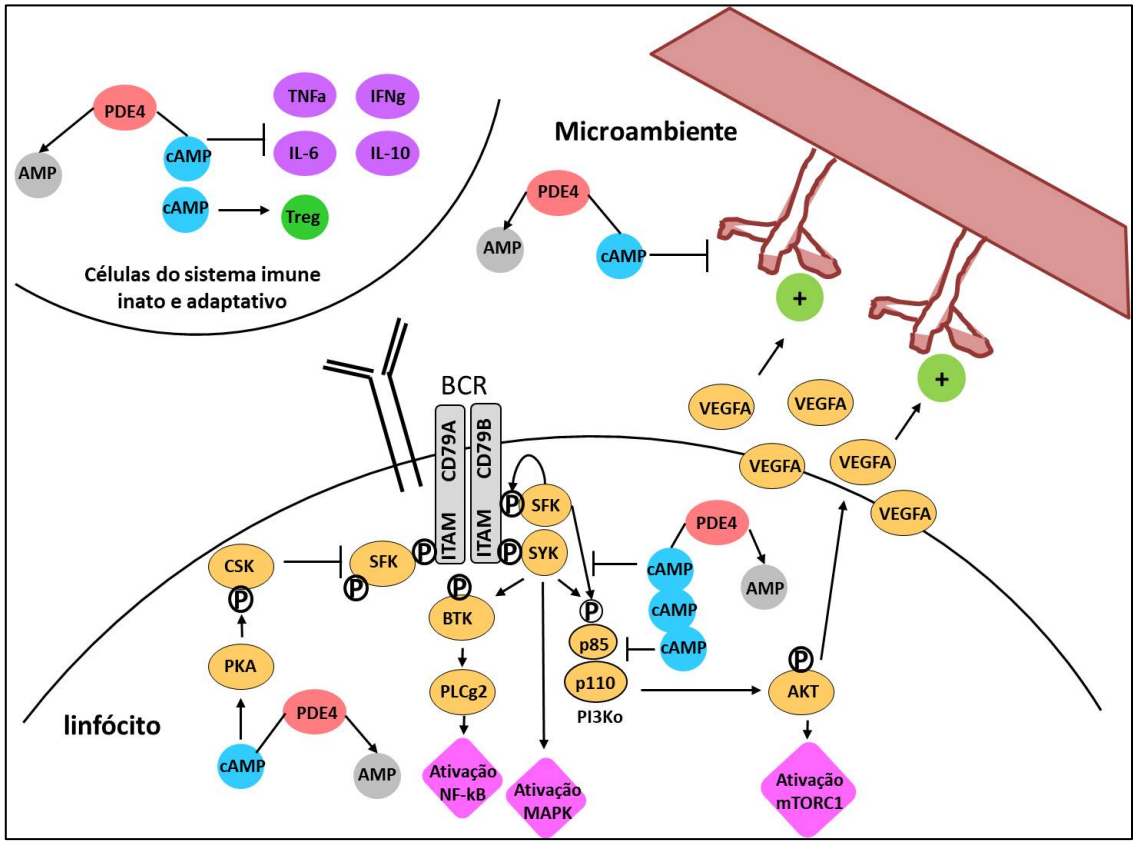
O mecanismo molecular que pode estar ligado aos resultados do tratamento da LLA mostrados nos estudos de Yang e colaboradores (2011, 2012) é descrito com base no gene *PDE4B*. Esse gene codifica a enzima fosfodiesterase 4B, a qual pertence à família das enzimas hidrolíticas responsáveis pela degradação do AMP cíclico, que atua como mensageiro secundário em vários tipos celulares. A fosfodiesterase 4B é uma das isoformas dentre outras existentes (4A, 4C e 4D). De acordo com o "The Human Protein Atlas", estudos de RNA-Seq mostram que esse gene é principalmente expresso no cérebro e na medula óssea (UHLÉN *et al.* 2015). Outros estudos indicam que o *PDE4* é mais expresso em células inflamatórias e que por isso seus inibidores possuem potencial terapêutico para doenças de base inflamatória como asma e psoríase (PAGE & SPINA 2011). Segundo reportado no estudo de Yang e colaboradores (2011), as células com alta expressão do *PDE4B* também se mostraram mais resistentes à prednisona, que é um glicocorticoide frequentemente usado nos protocolos de quimioterapia para o tratamento da LLA. Esse resultado corroborou outros estudos que indicaram que inibidores do gene *PDE4B* aumentaram os níveis de receptores de glicocorticoides em células B leucêmicas (OGAWA *et al.* 2002, MEYERS *et al.* 2007). Em outro estudo posterior, em que vários autores são comuns ao estudo de Yang *et al.* (2011), foi confirmada a associação de variantes do gene *PDE4B* com níveis reduzidos de metotrexato poliglutamato e o aumento do risco da recidiva de LLA (YANG *et al.* 2012). O metotrexato é um fármaco comumente utilizado no tratamento da LLA por exercer atividade anti-proliferativa e de anti-metabólito, sendo ministrado em conjunto com outros fármacos, como a prednisona. Nesse estudo quatro variantes do *PDE4B* foram destacadas: rs641262, rs524770, rs6683977 e rs546784, sendo que as três primeiras apresentaram associação com baixos níveis de metotrexato e recidiva da LLA e a última mostrou associação com a recidiva da LLA. Os alelos associados a níveis reduzidos de metotrexato poliglutamato para esses SNPs foram respectivamente os alelos A (fita negativa – alelo ancestral), A (fita positiva – alelo derivado), C (fita negativa – alelo ancestral) e A (fita negativa – alelo ancestral). Destaca-se que o penúltimo SNP (rs6683977) foi o mesmo reportado pelo grupo em 2011, em que o número de cópias do alelo C (fita negativa) desse SNP foi associado à recidiva de LLA

em nativo-americanos. Portanto, há várias evidências da relação dos produtos do gene *PDE4B* com os efeitos dos fármacos empregados no tratamento da LLA.

Ainda há intenso debate na literatura acerca dos mecanismos moleculares exatos que explicam a relação entre a inibição do gene *PDE4B* e o aumento da sensibilidade das células B malignas aos glicocorticoides (GIEMBYCZ & NEWTON 2015, COONEY & AGUIAR 2016) e entre variantes desse gene e a redução do nível de metotrexato poliglutamato nessas células (YANG *et al.* 2012). Tendo em vista que as fosfodiesterases atuam na degradação do AMP cíclico, o qual é um mensageiro secundário que está envolvido em diversos mecanismos, a inibição do gene *PDE4B* causa o acúmulo do AMP cíclico, o que pode levar a diferentes efeitos. Um dos mecanismos mais prováveis se baseia na supressão da atividade das enzimas SYK e PI3Kδ, o que contribui para a redução da cascata de sinalização que estimula eventos de sobrevivência e proliferação celular (COONEY & AGUIAR 2016). Ademais, outros autores também destacam a importância da via AKT/mTOR (KIM *et al.* 2011). Além de exercer um papel no controle da cascata de sinalização intracelular nos linfócitos B, a fosofdiesterase também está envolvida em mecanismos de sinalização nas células de imunidade inata e adquirida (como nas células T), como por exemplo, pelos efeitos do AMP cíclico sobre interleucinas (IL-6 e IL-10) e no microambiente tumoral, onde ocorrem eventos de angiogênese. A Figura 1, adaptada de Cooney & Aguiar (2016), ilustra exemplos das diferentes vias de atuação dos *PDE4*.

Com relação à redução do nível de metotrexato poliglutamato nas células B malignas, há indicações de que pode estar relacionada à fatores como o transporte transmembrana do metotrexato (influxo pela redução do carreador de folato), à glutamilação citosólica do metotrexato em metotrexato poliglutamato pela folilpoliglutamato sintase, e à degradação lisossomal do metotrexato poliglutamato em metotrexato pela y-glutamil hidrolase (KAGER *et al.* 2005). Independente do mecanismo, a redução dos níveis do metotrexato e do metotrexato poliglutamato prejudicam o efeito inibitório desses fatores sobre as enzimas relacionadas à síntese de pirimidina/purinas e a recirculação do folato (dihidrofolato redutase e thimidilato sintase), que promovem a síntese do RNA e DNA no ambiente celular (DE JONGE *et al.* 2005), o que está diretamente relacionado à proliferação das células malignas da LLA. Porém não há registro na literatura sobre a relação do metotrexato com o gene *PDE4B* ou seu produto.

Figura 1. Ilustração das interações moleculares relacionadas às *PDE4* e o AMP cíclico em linfócitos neoplásicos e no microambiente tumoral. Adaptado de Cooney & Aguiar (2016).



AMP: Adenosina monofostato; cAMP: AMP cíclico; TNFa: fator de necrose tumoral alfa; IFNg: interferon gama; IL: interleucina; Treg: células T regulatórias; VEGFA: fator de crescimento vascular endotelial; PKA: proteína quinase A; CSK: C-terminal Src quinase; SFK: família de quinases Src; BTK: tirosina quinase de Bruton; PLCg2: fosfolipase gama 2; SYK: tirosina quinase do baço; AKT: proteína quinase B.

Por sua vez, o gene *MYT1L* é pouco discutido no estudo de YANG *et al.* (2011), porém há evidências de que esse gene exerça um papel em meduloblastomas (ŁASTOWSKA *et al.* 2013). O meduloblastoma é a neoplasia maligna que mais acomete o cérebro de crianças (LOUIS *et al.* 2007) sendo na maioria das vezes originada no cerebelo. O *MYT1L* codifica um fator de transcrição do tipo dedo de zinco cuja expressão foi reportada apenas em tecidos neuronais e cuja função aparentemente está relacionada com o desenvolvimento do sistema nervoso central em mamíferos (KIM & HUDSON 1992, UHLÉN *et al.* 2015). A expressão induzida desse gene leva a maturação de fibroblastos em neurônios ativos e mutações na sua sequência foram relacionadas com doenças de deficiência cognitiva como o autismo e também com a depressão (WANG *et al.* 2010, WANG *et al.* 2016). Já no contexto das enfermidades relacionadas ao câncer, estudos indicaram que esse gene esteve associado ao prognóstico do adenocarcinoma da cárdia (ZHANG *et al.* 2013) e também foi

demonstrado que esse fator de transcrição esteve entre os fatores mais variavelmente expressos em pacientes com LLA (TOMAR *et al.* 2019). Portanto, especialmente por resultados como os apresentados por Tomar *et al.* (2019), há indícios da participação do produto desse gene na etiologia e tratamento da LLA.

Proposta de estudo

Diante dos resultados encontrados por Yang *et al.* (2011) e pela confirmação das evidências por estudos posteriores (YANG *et al.* 2012), há indícios importantes dos papéis dos genes *PDE4B* e *MYT1L* no tratamento da LLA. Ademais, como apresentado, há também resultados que suportam que tais papéis são exercidos de forma diferenciada em genomas com predominância de ancestralidade nativo-americana. Portanto, o estudo detalhado dos padrões de variação dos genes *PDE4B* e *MYT1L* em populações nativo-americanas negligenciadas podem revelar peculiaridades ainda desconhecidas da estrutura genética-populacional desses genes.

Nesse capítulo propõe-se primeiramente analisar a estrutura e a diversidade genética de diferentes populações nativas e miscigenadas distribuídas em países da América Latina com base nos genes *PDE4B* e *MYT1L*. O enfoque é dado nos SNPs que a literatura já revelou como associados à recidiva de LLA (*PDE4B*- rs6683977 e *MYT1L*-rs17039396) e nos SNPs que estejam em desequilíbrio de ligação com esses marcadores. A **hipótese** principal a ser testada é a de que a frequência dos alelos de risco para a recidiva de LLA acompanhará de forma direta a proporção de ancestralidade nativo-americana das populações. Posteriormente, busca-se aprofundar a análise nas regiões adjacentes a esses polimorfismos pontuais a fim de revelar variantes ainda não reportadas e identificar aquelas cujas frequências alélicas se destacam em populações com predominância da ancestralidade nativo-americana. Nesse caso, são também incluídos SNPs do gene *PDE4B* anteriormente reportados como de importância para a recidiva da LLA (-rs546784, -rs641262, -rs524770). Evidências de sinais de regulação gênica em porções dos genes *PDE4B* e *MYT1L* são também analisadas a fim de indicar no âmbito molecular possíveis implicações das variantes encontradas. Por fim, a análise dos resultados é feita em conjunto considerando aspectos relacionados ao tratamento de LLA, em especial considerando o contexto dos países da América Latina.

Procura-se explorar a estrutura genética populacional e as relações interpopulacionais alélicas e haplotípicas a fim de se refinar o estudo sobre esses genes e especialmente o gene *PDE4B* cuja importância é destacada no tratamento da LLA.

Portanto, espera-se reportar dados genético-populacionais que contribuam para o avanço da compreensão acerca da LLA em indivíduos com alta ancestralidade nativo-americana e assim enriquecer o conhecimento sobre a adequação do tratamento dessa doença, especialmente em países de menor desenvolvimento econômico-social, cujas populações sejam predominantemente de ancestralidade nativo-americana.

OBJETIVOS

Objetivo Geral

Revelar os padrões de diversidade genética e estrutura haplotípica em latinos americanos de genes de interesse para o tratamento da LLA e discutir as implicações considerando evidências de regulação gênica e o contexto do tratamento da LLA.

Objetivos Específicos

- Estimar os índices de ancestralidade individual e populacional das amostras estudadas.

- Caracterizar a diversidade genética e estrutura haplotípica de regiões e marcadores dos genes *PDE4B* e *MYT1L* em populações nativo-americanas e avaliar a influência da miscigenação na segregação dos alelos de interesse.

- Avaliar evidências de função regulatória em regiões específicas dos genes *PDE4B* e *MYT1L*.

- Analisar os protocolos de tratamento de LLA para crianças mais utilizados em países da América Latina quanto à aplicação da fase de intensificação tardia.

METODOLOGIA

Amostras e Bancos de Dados

Amostras

As amostras utilizadas para a caracterização genética e de estrutura haplotípica são tanto de indivíduos de origem nativo-americana quanto de indivíduos de populações com histórico de miscigenação. Indivíduos de populações nativas foram amostrados no Peru, Brasil e México (Figura 2). Dessa forma, espera-se que essas populações apresentem prevalência de ancestralidade nativo-americana. As populações peruanas compreendem indivíduos das etnias *Ashaninka* (ASH), *Machiguenga* (MAC), *Quechua* (QUE) e *Aymara* (AYM). As populações brasileiras compreendem indivíduos amostrados em regiões historicamente habitadas por duas etnias de tribos nativas. No estado do Espírito Santo, foram amostrados indivíduos *Tupiniquim* (TUP) e *Guarani*

(GUA). Por sua vez, as amostras do México foram obtidas em localidades habitadas pelos *Huicholes* (HUI) e *Tarahumaras* (TAH). Já populações com histórico de ocorrência de miscigenação foram amostradas no Brasil (Figura 2), em localidades do estado de Minas Gerais: Martinho Campos (MCP), Carmésia (CAR), Resplendor (RES) e São João das Missões (SJM). São localidades próximas a populações historicamente reconhecidas como de origem nativo-americana (*Kaxixó*, *Pataxó*, *Krenak*, e *Xacriabá*) (CEDEFS – Centro de Documentação Eloy Ferreira da Silva). Em razão da história demográfica dessas populações espera-se que esses indivíduos apresentem certo grau de miscigenação com indivíduos de ancestralidade europeia e africana. Em algumas análises essas populações de Minas Gerais são consideradas em conjunto e identificadas pela sigla AMG.

No total foram analisadas 849 amostras (Tabela 1), as quais foram utilizadas para a construção de quatro diferentes grupos de dados baseados em tecnologias de sequenciamento ou genotipagem (**BeadXpress, TaqMan-rs6683977, TaqMan-rs6683977+2.5M e TargetSeq**) e que são descritos com mais detalhes a seguir. Como em muitas circunstâncias os bancos de dados compartilham amostras, a Figura 3 ilustra as interseções e exclusividades de amostras de cada banco de dados utilizado no estudo.

Figura 2. Mapa das localidades e populações amostradas

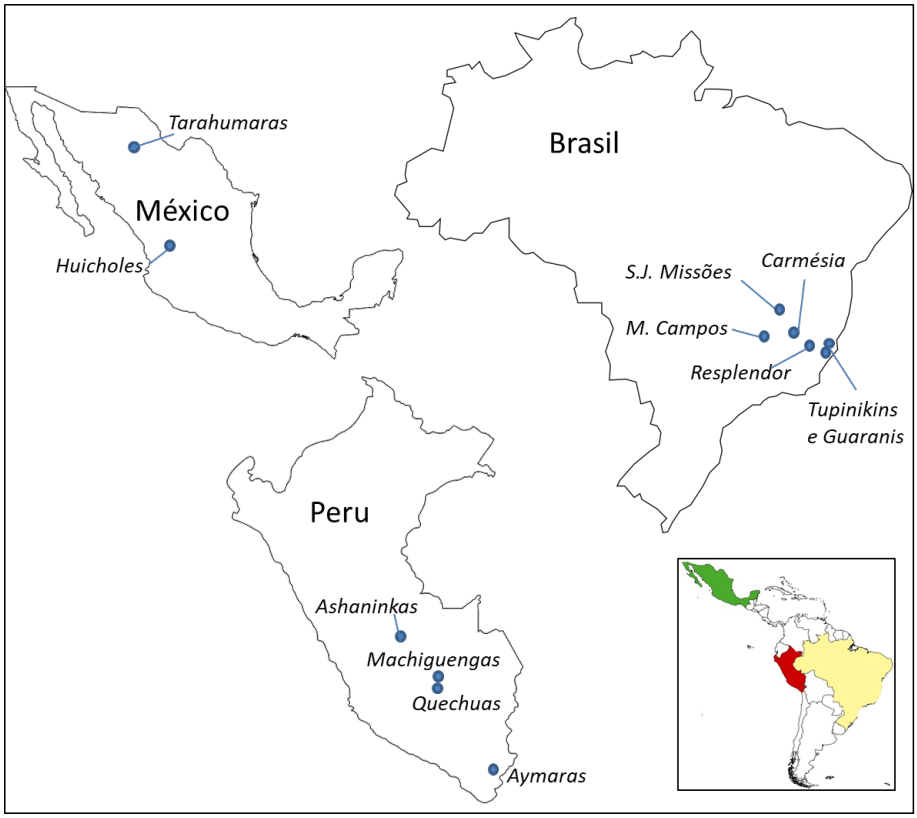
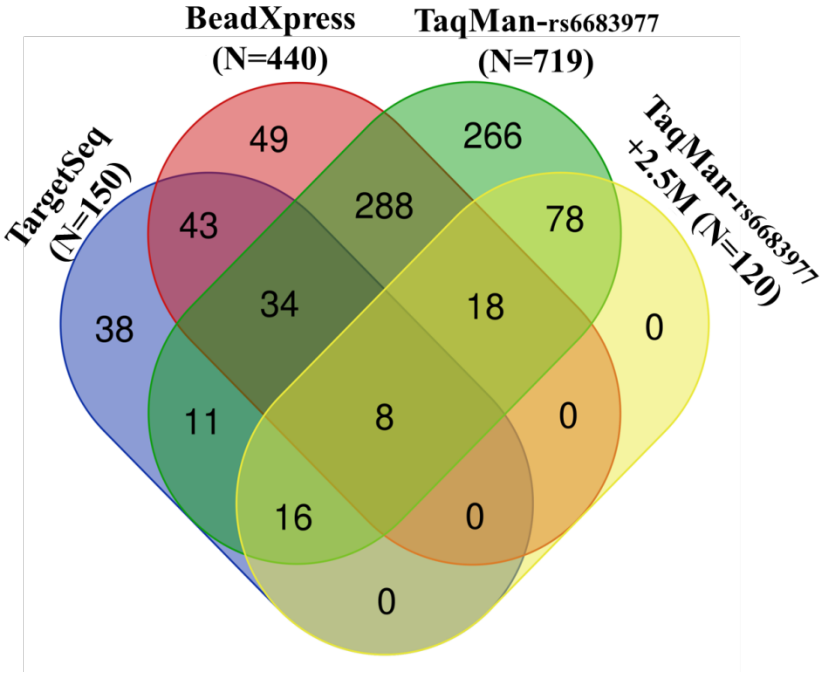


Tabela 1. Países, populações, ancestralidades, bancos de dados e números de amostras.

País	Grupo/ Municipalidade	Grupo etnolinguístico	Ancestralidade	BeadXpress	TargetSeq	TaqMan-rs6683977 +2.5M	TaqMan-rs6683977	Total
Peru	*Quechua*	Quechua	Nativos Andinos		20	21	100	111
	Aymara	Quechuamaran	Nativos Andinos	91	24	16	118	118
	Machiguenga	Arawak	Nativos Amazônicos	72	22	41	177	178
	Ashaninka	Arawak	Nativos Amazônicos	91	16	42	226	231
Brasil	*Tupiniquim*	Tupi	Nativos	45	22			45
	Guarani	Guarani	Nativos	43	24			46
	Carmésia		Miscigenados	19			19	19
	Martinho Campos		Miscigenados	30			30	30
	Resplendor		Miscigenados	24			24	24
	São João das Missões		Miscigenados	25			25	25
México	*Huichol*	Uto-Aztecan	Nativos		12			12
	Tarahumara	Uto-Aztecan	Nativos		10			10
Total				440	150	120	719	849

Figura 3. Número de amostras exclusivas e compartilhadas entre os bancos de dados TargetSeq, BeadXpress, TaqMan-rs6683977 e TaqMan-rs6683977+2.5M.



*Banco de dados **BeadXpress** - Genotipagem*

Um total de 440 amostras foi utilizado para a construção do banco de dados denominado **BeadXpress** (AYM=91, MAC=72, ASH=91, TUP=45, GUA=43, CAR= 19, MCP=30, RES=24, SJM=25) (Tabela 1 e Figura 3). A genotipagem dos marcadores escolhidos foi realizada pela plataforma BeadXpress, Illumina, disponibilizada no Laboratório de Genômica do Centro de Laboratórios Multiusários (CELAM) da UFMG. As amostras foram preparadas de acordo com o protocolo do fabricante para o kit "Golden Gate VeraCode Essay" que permite a genotipagem de 96 variantes escolhidas previamente pelo usuário em cada amostra, em um total de 96 amostras por cada processo de genotipagem. A tecnologia "VeraCode" consiste na análise em solução de "beads" de vidro revestidos com oligonucleotídeos marcados para as variantes escolhidas. Cada "bead" possui um código de barras holográfico interno que permite a especificidade quanto a variante e, além disso, são revestidas por oligonucleotídeos que permitem a complementaridade com as regiões flanqueadoras de cada variante. Oligonucleotídeos marcados com fluorescência de acordo com cada genótipo são amplificados para cada indivíduo no ensaio e posteriormente interagem com os oligonucleotídeos que revestem as "beads". De acordo com o genótipo do indivíduo, as "beads" emitem uma fluorescência específica, a qual permitirá ao leitor a laser dentro do equipamento identificar os genótipos de cada uma das 96 variantes para cada

indivíduo. Os dados são enviados ao computador associado ao equipamento e são analisados por um algoritmo de ajuste de intensidade de sinal para a geração dos genótipos prováveis.

Em razão da geração de dados para outros projetos que não o tratado aqui, a capacidade de genotipagem do kit VeraCode GoldenGate do BeadXpress para o presente projeto foi de 47 SNPs. Dessa forma, buscou-se aumentar a abrangência dos dados obtidos pela escolha da genotipagem de tag-SNPs. Essas variantes são representativas de um conjunto de variantes dos mesmos genes que estejam em desequilíbrio de ligação (DL). Assim, a genotipagem do tag-SNP fornece informação em relação aos genótipos das outras variantes em DL que não precisam ser genotipadas. Para tanto, antes do processo de genotipagem propriamente dito, a então integrante do Laboratorio de Diversidade Genetica Humana (LDGH), Dra. Fernanda Rodrigues Soares, consultou os genótipos das variantes presentes nos genes de interesse nas amostras do projeto HapMap (INTERNATIONAL HAPMAP CONSORTIUM 2003). Dessa forma, especificamente essa análise não foi realizada pelo autor da tese. Na ocasião, esses dados foram analisados no programa GLU (https://code.google.com/archive/p/glu-genetics/) com o uso do módulo Tagzilla para que fossem identificados os grupos de variantes em desequilíbrio de ligação que fossem mais abundantes. Foram selecionadas 47 variantes que cobrissem o maior número de variantes dos genes *PDE4B* e *MYT1L* mas que estivessem a até 500kb de distância em ambas direções dos SNPs rs6683977 e rs17039396, os quais foram relacionados à recidiva de LLA (YANG *et al.*2011) (Tabela 2). Assim, foi possível genotipar no BeadXpress um menor número de variantes as quais foram informativas sobre os genótipos das demais variantes dos grupos de desequilíbrio de ligação.

Tabela 2. Relação das variantes selecionadas com a indicação dos genes onde estão localizadas e o número de variantes cobertas em razão do desequilíbrio de ligação.

Gene	variantes	Locos cobertos
PDE4B	rs6683977	19
	rs12119734	24
	rs6661245	39
	rs10454453	32
	rs1500951	12
	rs12081185	10
	rs6683604	19
	rs11589566	9
	rs1354060	18
	rs12027416	23
	rs2503205	23
	rs6588177	50
	rs10889591	14
	rs12033425	28
	rs12045945	19
	rs1937453	32
	rs1937456	14
	rs522037	26
	rs11208774	29
	rs2186123	27
	rs783036	1
MYT1L	rs17039396	1
	rs12468168	27
	rs11683072	31
	rs13388663	5
MYT1L	rs10195351	19
	rs17338491	27
	rs17039463	5
	rs13026825	10
	rs6705656	31
	rs13034457	5
	rs6548050	14
	rs12986730	9
	rs17039334	11
	rs1862114	12
	rs6729968	13
	rs3748988	10
	rs2241685	6
	rs3748989	6
	rs17039474	10
	rs3924017	22
	rs12991351	24
	rs6728613	20
	rs12468174	9
	rs9284792	14
	rs10178240	8
	rs11127305	14
	TOTAL	831

*Banco de dados **TaqMan-rs6683977** - Genotipagem TaqMan ThermoFisher™*

Para aumentar o tamanho amostral no cálculo da frequência alélica do SNP de interesse *PDE4B*-rs6683977, 719 amostras (QUE=100, AYM=118, ASH=226, MAC=177, CAR=19, MCP=30, RES=24, SJM=25 – Tabela 1 e Figura 3) foram genotipadas para esse SNP pelo método TaqMan (*Assay ID* C___1270960_10 Thermo Fisher™). Várias das amostras foram também utilizadas na construção do Banco de dados **BeadXpress** apresentado anteriormente (Figura 3). Como os resultados iniciais obtidos para esse SNP com o método BeadXpress indicaram indícios da existência de alelo nulo (resultados discutidos adiante), esses dados foram também úteis para a validação dos genótipos do rs6683977 em todas as amostras de nativos peruanos e de brasileiros miscigenados submetidas ao método do BeadXpress. Porém, não foi possível

realizar a validação dos genótipos dos nativos brasileiros por esse método pois as amostras já não estavam mais disponíveis para a genotipagem por TaqMan. Outra estratégia foi utilizada para a validação nessas populações, como será apresentado a seguir.

*Banco de dados **TaqMan-rs6683977+2.5M** - Genotipagem Illumina™* array *HumanOmni2.5-8v1.1*

Alguns bancos de dados já estabelecidos no LDGH incluem genótipos de vários indivíduos nativo-americanos, dentre os quais podem ser encontradas amostras das populações estudadas aqui. Um dos bancos de dados (**2.5M**) inclui genótipos de 2.391.739 de SNPs obtidos pelo uso do *array* HumanOmni2.5-8v1.1 fabricado e comercializado pela Illumina, Inc.. Dentre os SNPs reportados como associados à recidiva da LLA (rs6683977, rs546784, rs641262 e rs524770 do gene *PDE4B* e rs17039396 do gene *MYT1L*), apenas o SNP rs524770 no gene *PDE4B* está presente no *array* HumanOmni2.5-8v1.1. Tendo em vista que algumas amostras presentes no banco **2.5M** são comuns ao banco de dados ***TaqMan-rs6683977*** (Figura 3), apresentado anteriormente, as informações de ambos bancos de dados foram agrupadas. Dessa forma, todas as demais variantes encontradas no gene *PDE4B* disponíveis no *array* 2.5M foram recuperadas para análise dos padrões de desequilíbrio de ligação com o SNP de interesse rs6683977. Portanto, para os estudos propostos aqui, apenas as variantes presentes no gene *PDE4B* foram filtradas do total de variantes do *array* e incluídas nesse banco de dados.

No total, 120 amostras de populações peruanas de nativos (ASH=42, QUE=21, MAC=41, AYM=16) (Tabela 1 e Figura 3) foram incluídas para a formação desse banco de dados, daqui em diante nomeado *Banco de dados **TaqMan-rs6683977+2.5M***. Após a genotipagem de todos os indivíduos, os dados foram manipulados pelo Dr. Victor Borda, integrante do LDGH à época, para a geração de um banco final de genótipos que contemplasse o maior número de SNPs em comum entre todas as amostras.

*Banco de dados **TargetSeq** - Sequenciamento de regiões alvo*

Como apresentado no capítulo 1 dessa tese, regiões de 20Kb intrônicas dos genes *PDE4B* (chr1:66759100-66779100, GRCh37) e *MYT1L* (chr2:2215144-2235144, GRCh37) foram sequenciadas utilizando o método de construção de bibliotecas HaloPlex (Agilent Inc.) e o sequenciamento massivo em paralelo por síntese Illumina.

As regiões sequenciadas incluíram as variantes rs641262, rs524770, rs546784 e rs6683977, do gene *PDE4B*, citadas em estudos de associação com a recidiva de LLA (YANG *et al.* 2011 e 2012), assim como a variante rs17039396, no gene *MYT1L* (YANG *et al.* 2011). Detalhes do processo de geração desses dados e a discussão completa sobre as chamadas de variantes estão apresentados no capítulo 1 dessa tese. Como os reagentes para a construção das bibliotecas foram suficientes para a utilização de apenas 150 amostras, apenas as populações nativo-americanas (TAH=10, HUI=12, QUE=20, AYM=24, MAC=22, ASH=16, TUP=22, GUA=24 - excluindo-se assim as populações de brasileiros miscigenados) (Tabela 1 e Figura 3) foram escolhidas para o sequenciamento das regiões alvo por terem maior proporção de ancestralidade nativa.

Controle de qualidade dos dados

Os dados gerados na plataforma BeadXpress foram analisados no programa GenomeStudio™ Genotyping (Illumina, San Diego, USA) fornecido pela fabricante. As leituras dos genótipos realizadas pelo equipamento foram avaliadas com relação à qualidade dos sinais de cada variante pelos índices de *GenTrain* e *GenCall* e com relação à qualidade de cada amostra medida pelo índice de *Call Rate*. O primeiro índice é uma medida de qualidade do agrupamento dos sinais de cada variante no plano cartesiano considerando todas as amostras em conjunto, já o segundo índice é a métrica de confiança de cada genótipo inferido (variando de 0 a 1). Por sua vez, o índice de *Call Rate* indica a proporção de amostras em que foi possível inferir um genótipo considerando cada um dos locos. Segundo a literatura especializada, índices de *GeneTrain* e *GenCall* (< 30% e <40%) e índices de *Call Rate* (<90%) baixos são indicadores de eliminação de variantes ou amostras (HUGHES-STAMM *et al.* 2013). Já a qualidade dos dados gerados com o uso do ensaio TaqMan (Thermo Fisher™) foi avaliada no software SDS 2.4, disponibilizado pelo fabricante, em que apenas amostras com valores de qualidade superiores a 97 foram mantidas.

Por fim, os dados gerados pelo *array* Illumina HumanOmni2.5-8v1.1 e por sequenciamento também foram controlados considerando parâmetros de qualidade. Os tratamentos e a descrição do banco de dados final estão apresentados no capítulo 1 dessa tese.

Bancos de dados públicos

Bancos de dados públicos foram acessados para a obtenção de dados genômicos de indivíduos de populações de diferentes ancestralidades. Dados dos bancos HapMap,

Projeto 1000 Genomas (1000GP) (1000 GENOMES PROJECT CONSORTIUM) e de outros bancos de relevância foram obtidos e armazenados para análises posteriores. Amostras das populações europeias, africanas, asiáticas, nativo-americanas ou que apresentam alta prevalência de uma dessas ancestralidades (como ocorre, por exemplo, com os indivíduos de origem europeia da população CEU que habitam o estado de Utah nos Estados Unidos) foram obtidas desses bancos públicos e retidas para as análises seguintes.

Os dados do Projeto 1000 Genomas foram obtidos pelo acesso ao banco do Ensembl (FLICEK, 2013), utilizando a ferramenta *VCF to PED converter*. As populações do 1000GP analisadas compreenderam cinco populações européias (EUR): CEU (residentes do estado de Utah com ancestralidade do norte e oeste da Europa), TSI (Toscanos da Italia), FIN (finlandeses da Finlandia), GBR (britânicos da Inglaterra e Escócia), IBS (ibéricos da Espanha); cinco populações africanas (AFR): YRI (*Yoruba* de Ibadan, Nigeria), LWK (*Luhya* de Webuye, Quênia), GWD (gambianos da divisão oeste de Gambia), MSL (*Mende* de Serra Leoa), ESN (*Esan* na Nigeria); quatro populações do continente Americano (AMR): PUR (porto-riquenhos de Porto Rico), CLM (colombianos de Medellin, Colombia), MXL (habitantes de Los Angeles/ EUA com ancestralidade mexicana) e PEL (peruanos de Lima, Peru); cinco populações sul-asiáticas (SAS): GIH (indianos Gujarati de Houston, Texas, EUA), PJL (Punjabi de Lahore, Paquistão), BEB (Bengali de Bangladesh), STU (srilankeses Tamil da Grã-Bretanha), e ITU (Indianos Telugu da Grã-Bretanha); e cinco populações do leste asiático (EAS): CHB (Han Chinese de Pequim, China), JPT (japoneses de Toquio, Japão), CHS (chineses Han do sul, China), CDX (chineses Dai de Xishuangbanna, China), and KHV (Kinh da cidade de Ho Chi Minh, Vietnã). Os dados foram manipulados para que ficassem disponíveis nos formatos utilizados nas análises genéticas, como por exemplo, os formatos *.ped* e *.map* necessários para o uso do software PLINK (PURCELL *et al.* 2007). Foram realizados controles de qualidade para a remoção de variantes com baixos índices de confiança.

Análises de ancestralidade

As análises de ancestralidade foram realizadas primeiramente para os indivíduos genotipados pela plataforma Illumina BeadXpress. Os indivíduos mexicanos, *Huichol* e *Tarahumara*, assim como os peruanos *Quechua*, não foram incluídos nessas análises, mas estudos para análise da ancestralidade já realizados com essas mesmas amostras

(RODRIGUES-SOARES *et al.* 2019) mostraram que esses indivíduos são predominantemente de ancestralidade nativo-americana (*Huichol*: 96%, *Tarahumara*: 94%, *Quechua*: 87%). Por isso, para as demais populações desse estudo, parte-se da hipótese de que as populações com origem nativo-americana terão proporções bastante superiores dessa ancestralidade frente às ancestralidades europeia e africana.

A genotipagem de 96 SNPs informativos de ancestralidade escolhidos entre os 107 reportados por Yaeger e colaboradores (2008) foi realizada com o kit GoldenGate Veracode (Illumina, Inc.) para as 440 amostras utilizadas nessas análises (Tabela 1). O método de geração de dados e contole de qualidade foram os mesmos utilizados para o banco de dados **BeadXpress**. Bancos de dados públicos foram acessados para a obtenção de genótipos de amostras de populações de diferentes continentes para fins de referência nas análises de ancestralidade e nas análises de diversidade genética e estrutura haplotípica. Indivíduos das populações CEU – Utah/EUA (n=174) e YRI-Ibadan/Nigeria (n=176), de origem europeia e africana, respectivamente, foram obtidas do projeto HapMap e incorporados aos dados gerados com o BeadXpress.

O programa ADMIXTURE (ALEXANDER *et al.* 2011) foi utilizado para as análises do perfil de miscigenação das populações e dos indivíduos após o controle de qualidade dos dados. O programa ADMIXTURE indica a ancestralidade individual por meio de um algoritmo que se baseia no modelo de equilíbrio de Hardy-Weinberg para K clusters (i.e. populações parentais) para inferir para cada genoma da amostra a contribuição de cada população parental. O programa infere as porcentagens individuais de cada componente de acordo com o número K de grupos indicado pelo usuário. O método busca a combinação de miscigenações individuais que se adapta melhor ao modelo de Hardy-Weinberg para cada uma das populações parentais consideradas. Após a inferência das estimativas de ancestralidade, o arquivo gerado foi editado para a inserção dos identificadores dos indivíduos e das populações para que um gráfico de barras fosse gerado com o auxílio da plataforma R (TEAM, 2013).

Análises de diversidade genética

As análises básicas de genética de populações realizadas para a caracterização das regiões e variantes de interesse biomédico foram promovidas pela comparação dos dados obtidos nas populações amostradas com os dados das populações europeias (EUR, n=503), africanas (AFR, n=661), leste-asiáticas (EAS, n=504) e sul-asiáticas (SAS, n=489) disponíveis no Projeto 1000 Genomas (1000 GENOMES PROJECT

CONSORTIUM). O cálculo de frequências alélicas e teste de equilíbrio de Hardy-Weinberg foram realizados utilizando o programa PLINK (PURCELL *et al.* 2007). Os cálculos dos índices de F_{ST} par-a-par, F_{CT} e heterozigosidade foram realizados no programa Arlequin (EXCOFFIER & LISCHER 2010) considerando o índice de dados faltantes de até 10%. Com o programa PLINK foi realizada a análise de componentes principais, a qual incluiu amostras das populações com ancestralidades europeia (EUR) e africana (AFR) obtidas do projeto 1000 Genomas, além das amostras das populações latino-americanas genotipadas nesse estudo. Para essa análise foram realizados controles de desequilíbrio de ligação entre os marcadores considerando todas as populações avaliadas.

Diversidade haplotípica e desequilíbrio de ligação

A fase cromossômica dos dados do banco **BeadXpress** foi inferida pelo programa *fastPHASE* (SCHEET & STEPHENS 2006). Esse programa se baseia no modelo oculto de markov para descrever a distribuição espacial dos grupos de haplótipos ao longo dos cromossomos. Considera-se a premissa de que em regiões curtas dos cromossomos as variantes tendem a fazer parte dos mesmos grupos de haplótipos. O *fastPHASE* utiliza o algoritmo de maximização de expectativas para estimar os parâmetros genéticos e as frequências haplotípicas. É um programa que trabalha com grande quantidade de dados sendo eficiente computacionalmente. Os arquivos de saída contendo informações acerca da fase das variantes foram analisados no programa *Haploscope* (SAN LUCAS *et al.* 2012). Esse programa faz uma representação gráfica dos haplótipos encontrados apresentando-os em cores diferentes e indicando as frequências desses haplótipos nas amostras e os possíveis pontos de recombinação entre eles. Por fim, o grau de desequilíbrio de ligação das variantes obtidas nos bancos de dados **TaqMan-rs6683977+2.5M** e **TargetSeq** foi medido com a estatística r^2 (HARTL & CLARK, 2010) utilizando o programa *Haploview* (BARRETT *et al.* 2004) e parâmetros de filtragem em que taxas maiores que 50% de dados faltantes e variantes com frequência do alelo menor (MAF) <0,05 foram excluídas.

Evidências de regiões com função regulatória

Tendo em vista que as regiões sequenciadas nos genes *PDE4B* e *MYT1L* são intrônicas é importante avaliar a existência de motivos de regulação nessas regiões que possam exercer algum papel no fenótipo molecular desses genes. Inicialmente, o posicionamento das variantes com maior diferenciação foi visualizado no UCSC

Genome Browser com relação a dados públicos sobre funções regulatórias, tais como: padrões de modificação de histonas por metilação e acetilação, bem como de predições *in silico* de modificações da cromatina, sítios de clivagem de DNAse e de ligação de fatores de transcrição. Para o detalhamento dos resultados obtidos com o UCSC *Genome* Browser, a base de dados ENCODE (*Encyclopedia of DNA Elements*) (ENCODE 2012) também foi consultada diretamente para indicar os possíveis elementos funcionais existentes nas regiões sequenciadas. O projeto ENCODE permite que apenas a região cromossômica de interesse seja avaliada e possibilita a filtragem dos resultados de acordo com o tecido específico alvo do estudo. Portanto, as regiões sequenciadas nos genes *PDE4B* e *MYT1L* foram consultadas e os resultados foram filtrados para as células dos tecidos medula óssea e hematopoiético, os quais têm relação direta com a LLA. Por fim, os SNPs com maiores índices de F_{CT} de diferenciação (>0,12) foram analisados de acordo com o banco de dados RegulomeDB (BOYLE *et al.* 2012) para a identificação mais precisa da caracterização dos sítios de cada SNP.

Avaliação dos protocolos de tratamento da LLA em países latino-americanos

Para enriquecer as discussões desse estudo foram realizadas pesquisas em sítios eletrônicos da rede mundial de computadores para consultar se os protocolos de tratamento da LLA aplicados nos países latino-americanos incluem a fase de intensificação tardia (recomendada por Yang e colaboradores (2011)). Foram consultados sítios eletrônicos de órgãos nacionais de saúde e de centros de tratamento oncológico nacionais de países latino americanos. Além disso, para ampliar as buscas, foram consultados buscadores da rede mundial de computadores utilizando as seguintes palavras-chave: *protocolo*, *tratamiento*, *leucemia*, *pediatrica*, *America Latina*.

RESULTADOS

Antes de iniciar a apresentação dos resultados cabe esclarecer um problema recorrentemente identificado na literatura ao se reportar alelos de SNPs cujas variantes alélicas compreendem as bases C/G ou A/T. Essas alterações são de certa forma consideradas ambíguas, pois essas bases são complementares quando se considera as duas fitas do DNA. Essa falta de padronização ocorre am vários dos SNPs alvo desse estudo. Os alelos de risco (os quais correspondem aos alelos ancestrais) dos SNPs *PDE4B*-rs6683977, -rs546784 e -rs641262, citados em diversos estudos de associação com a LLA (YANG *et al.* 2011, 2012), correspondem às bases presentes na fita reversa

(3’ → 5’), geralmente chamada fita negativa (-) do DNA. Tendo em vista a necessidade de padronização para evitar desentendimentos, os alelos desses SNPs e de quaisquer outros SNPs tratados aqui correspondem às bases da fita positiva, em concordância com o padrão seguido em bancos de dados públicos, como dbSNP e o 1000GP, e com a literatura científica (NELSON *et al.* 2012). Portanto, na presente Tese os alelos de risco rs6683977.C, rs546784.A e rs641262.A são reportados como rs6683977.G, rs546784.T e rs641262.T.

Ancestralidade nas populações estudadas

Como apresentado anteriormente, as populações alvo dos estudos sobre a diversidade genética e estrutura haplotípica das regiões relacionadas a recidiva de LLA são populações do continente americano com históricos diferentes, sendo que algumas são resultado de eventos de miscigenação entre nativos, europeus e africanos e outras não. Para determinar geneticamente com maior precisão as ancestralidades de cada um dos indivíduos, e consequentemente inferir a origem da população a que pertencem, foram genotipados 96 de 107 SNPs informativos de ancestralidade, não-ligados e que estão distribuídos por todo o genoma, seguindo estudo anterior (YAEGER *et al.* 2008). O processo foi realizado na plataforma Illumina BeadXpress, sendo que 88 SNPs foram mantidos após a realização de controle de qualidade.

As análises dos dados com o programa ADMIXTURE revelaram padrões que corroboraram a hipótese de que as populações mexicanas e peruanas de nativos são menos miscigenadas que as demais populações do estudo (Figura 4) (RODRIGUES-SOARES 2019). Ademais, as populações das localidades do estado de Minas Gerais apresentaram grandes percentuais de miscigenação com indivíduos de origens européia e africana (Figura 4 e Tabela 3). As populações de nativos brasileiros, por sua vez, apesar de apresentarem alta proporção de ancestralidade nativo-americana, também mostraram certa miscigenação com europeus e africanos (Tabela 3).

Figura 4. Representação gráfica das proporções relativas de miscigenação inferidas pelo programa ADMIXTURE a partir de 88 maracadores informativos de ancestralidade para as populações do estudo: CEU (descendentes de europeus/EUA); YRI (*Yoruba*/Nigéria); ASH (*Ashaninka*/Peru); MAC (*Machiguenga*/Peru); AYM (*Aymara*/Peru); QUE (*Quechua*/Peru); MEX (*Taharumara* e *Huichol*/México); GUA (*Guarani*/Brasil); TUP (*Tupiniquim*/Brasil); AMG (Miscigenados/MG - Brasil). Cada barra vertical representa um indivíduo e indica a proporção de ancestralidade de acordo com as cores das populações parentais.

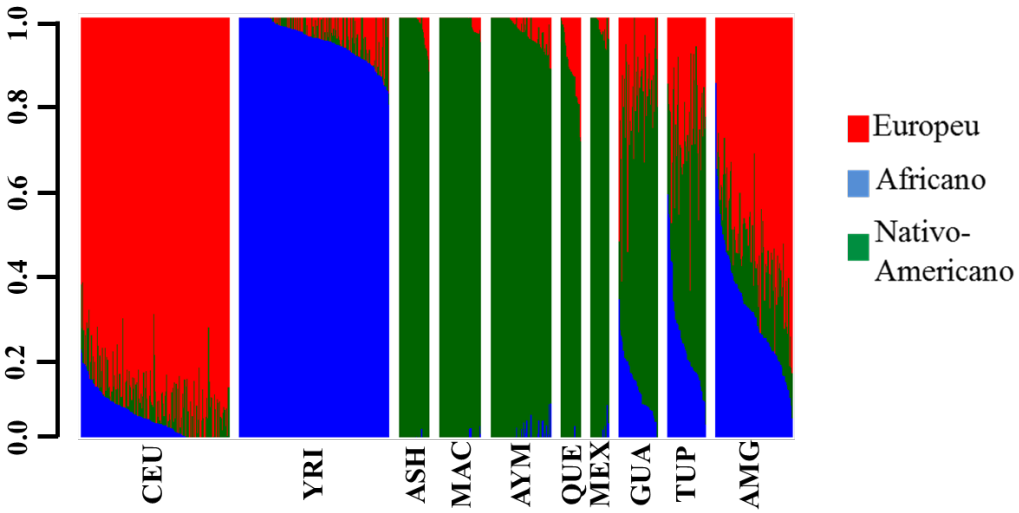


Tabela 3. Proporções de ancestralidade por população obtidas a partir da média das ancestralidades individuais. CEU (descendentes de europeus/EUA); YRI (*Yoruba*/Nigéria); ASH (*Ashaninka*/Peru); MAC (*Machiguenga*/Peru); AYM (*Aymara*/Peru); QUE (*Quechua*/Peru); TAH (*Taharumara*/México); HUI (*Huichol*/México); GUA (*Guarani*/Brasil); TUP (*Tupiniquim*/Brasil); AMG (Miscigenados/MG - Brasil).

Pop\Ancestralidade	**Africana**	**Nativa**	**Europeia**
CEU	0,05	0,06	0,89
YRI	0,95	0,02	0,03
ASH	0,00	0,98	0,02
MAC	0,00	0,99	0,01
AYM	0,01	0,96	0,03
QUE	0,00	0,87	0,13
MEX	0,00	0,97	0,03
GUA	0,16	0,61	0,23
TUP	0,23	0,50	0,27
AMG	0,30	0,15	0,55

***Diversidade e estrutura genética dos genes* PDE4B *e* MYT1L**

Frequências alélicas

Nessa seção é dado enfoque às variantes de interesse dos genes *PDE4B* (rs546784, rs524770, rs6683977, rs641262) e *MYT1L* (rs17039396) (Tabela 4). Porém, as frequências alélicas de outros SNPs presentes nos bancos de dados **BeadXpress** e **TargetSeq** podem ser consultadas nas tabelas do Apêndice desse capítulo.

No caso do banco de dados **BeadXpress**, após a aplicação dos controles de qualidade, dos 47 SNPs selecionados para as análises dos genes *PDE4B* e *MYT1L* três foram eliminados por baixos valores do parâmetro de qualidade *GenTrain*. Dentre os SNPs analisados, alguns apresentaram valores de frequências alélicas que indicam diferenciação entre as populações analisadas (Tabela Apendice 1). Já com relação às variantes identificadas no banco de dados **TargetSeq**, no total (sem considerar qualquer filtro para dados faltantes ou frequência alélica do alelo menor) foram listadas para as populações de nativos 104 variantes na região de 20Kb do gene *PDE4B* e 83 variantes na região de 20Kb do gene *MYT1L*. As frequências alélicas de todas elas, nas diferentes populações, podem ser consultadas nas Tabelas Apêndice 2 e 3.

Para dar destaque às frequências dos alelos das variantes previamente reportadas como associadas à recidiva, elas estão apresentadas na Tabela 4. As frequências alélicas reportadas nessa tabela foram calculadas sobre o maior número de amostras possível, considerando todos os bancos de dados disponíveis (Figura 3). Por fim, as frequências das variantes *PDE4B*- rs6683977.G (alelo ancestral) e *MYT1L*-rs17039396.A (alelo derivado), reportadas como associadas à recidiva de LLA e à ancestralidade Ameríndia (YANG *et al*. 2011), são também apresentadas em função das proporções de ancestralidade nativo-americana das diferentes populações (Figura 5). As análises desses resultados permitem confirmar a hipótese de que a frequência dos alelos de risco para a recidiva de LLA tende a ser maior em populações com maiores índices de ancestralidade nativo-americana.

Um resultado a ser comentado foi o encontrado para o SNP *PDE4B*-rs6683977. Os genótipos obtidos com o banco de dados **TargetSeq** apresentaram baixa concordância no número de heterozigotos identificados no banco de dados **BeadXpress**. Dentre as 150 amostras sequenciadas no **TargetSeq**, 43 também foram previamente genotipadas pela plataforma **BeadXpress** (Tabela 3). Segundo os dados gerados no **BeadXpress**, dentre as 43 amostras, apenas 3 amostras (todas da população de

Aymaras) seriam heterozigotas para esse SNP. Porém, os dados encontrados com o sequenciamento da região indicaram que há 26 amostras heterozigotas, em sua maioria na população *Tupiniquim* (n=13). Outros dois SNPs também estiveram presentes concomitantemente nos bancos de dados **TargetSeq** e **BeadXpress**.

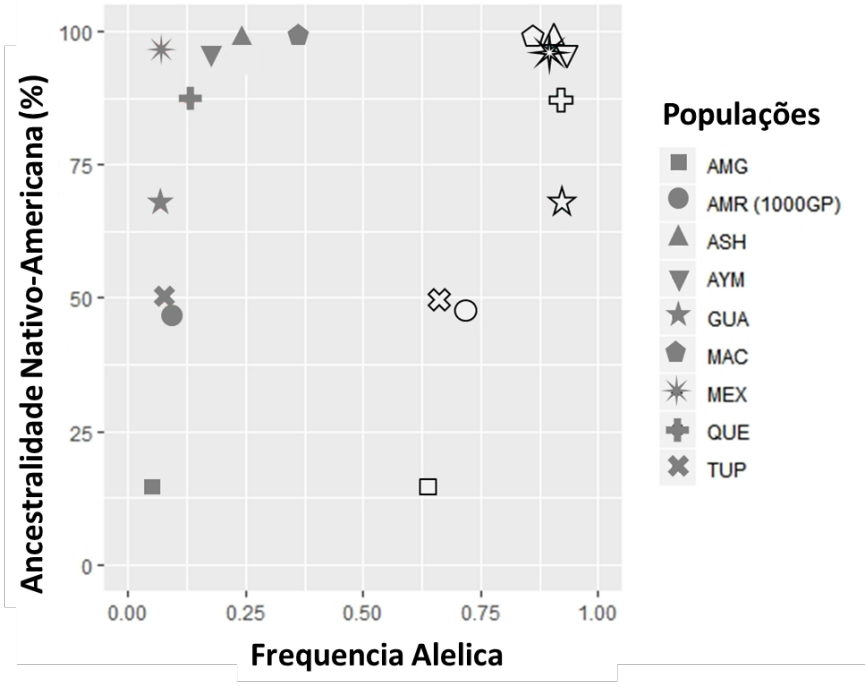
Para avaliar se o problema era exclusivo do SNP *PDE4B*-rs6683977, foram analisadas outras variantes que estavam incluídas nos bancos de dados **TargetSeq** e **BeadXpress**. O SNP rs6683604, no gene *PDE4B*, teve 100% de concordância entre os métodos em todas as populações, exceto nos dados das populações nativas brasileiras, em que a concordância foi de cerca de 75%. Já o SNP rs17039396, localizado no gene *MYT1L*, teve taxa de concordância dos genótipos superior a 95% caso não se considere a população *Aymara*, na qual 50% das amostras tiveram discordância entre os genótipos obtidos.

Em razão das discordâncias observadas entre os métodos para genotipagem utilizados, os dados obtidos no banco **TaqMan-rs6683977** foram utilizados para a validação dos genótipos do SNP rs6683977. Os resultados obtidos com o ensaio TaqMan para as 43 amostras foram 100% concordantes com os genótipos do banco de dados **TargetSeq**. Dessa forma, conclui-se que o ensaio empregado no **BeadXpress** não teve eficiência satisfatória para a genotipagem desse SNP. Por isso, as amostras genotipadas com o BeadXpress que não tiveram algum outro método de validação foram eliminadas das análises, especialmente para o cálculo de frequência alélica do *PDE4B*-rs6683977.

Tabela 4. Frequências dos alelos ancestrais (*PDE4B*- rs546784, rs6683977 e rs641262) e alelos derivados (*PDE4B*- rs524770 e *MYT1L*-rs17039396) reportados como associadas à recidiva de LLA (YANG *et al.* 2011, 2012) em diferentes populações. ASH: *Ashaninka*, MAC: *Machiguenga*, AYM: *Aymara*, QUE: *Quechua*, MEX (*Huicholes* e *Tarahumara*), GUA: *Guarani*, TUP: *Tupiniquim*, AMR: Hispânicos miscigenados/EUA e Latinos Americanos (1000G), EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). n: indica o número total de amostras usado para o cálculo da frequência alélica.Valores em negrito indicam desvio do equilíbrio de Hardy-Weinberg (p<0,05). (-) indica que os SNPs não estavam incluídos nos bancos de dados analisados na população em questão.

Gene	***PDE4B***				***MYT1L***
ID / pop	**rs546784**	**rs524770**	**rs6683977**	**rs641262**	**rs17039396**
ASH	T:0,94 n=16	A:0,82 n=59	G:0,90 n=231	T:0,94 n=16	A:0,24 n=95
MAC	T:0,82 n=22	A:0,78 n=50	G:0,86 n=178	T:0,82 n=22	**A:0,37** n=82
AYM	T:0,92 n=24	A:0,75 n=36	G:0,93 n=118	T:0,92 n=24	**A:0,16** n=85
QUE	T:0,85 n=20	A:0,66 n=35	G:0,92 n=111	T:0,85 n=20	A:0,13 n=20
MEX	T:0,90 n=22	A:0,67 n=22	G:0,90 n=22	T:0,90 n=22	A:0,13 n=22
GUA	T:0,82 n=24	**A:0,44** n=24	G:0,92 n=16	T:0,85 n=24	A:0,07 n=45
TUP	**T:0,58** n=22	**A:0,48** n=22	**G:0,66** n=16	**T:0,62** n=22	A:0,08 n=42
AMG-BRA	-	-	G:0,36 n=98	-	A:0,05 n=98
Populações 1000GP					
AMR	T:0,71 n=347	A:0,56 n=347	G:0,72 n=347	T:0,72 n=347	A:0,10 n=347
EUR	T:0,44 n=503	A:0,37 n=503	G:0,44 n=503	T:0,44 n=503	A:0,11 n=503
AFR	T:0,85 n=661	A:0,27 n=661	G:0,96 n=661	T:0,86 n=661	A:0,01 n=661
EAS	T:0,84 n=504	A:0,47 n=504	G:0,84 n=504	T:0,84 n=504	A:0,24 n=504
SAS	T:0,64 n=489	A:0,51 n=489	G:0,65 n=489	T:0,70 n=489	**A:0,09** n=489

Figura 5. Frequências alélicas dos SNPs de risco para recidiva de LLA *PDE4B*-rs6683977.G (alelo ancestral) (vazado) e *MYT1L*-rs17039396.A (alelo derivado) (sólido) em função da proporção de ancestralidade das populações: AMG: Miscigenados Brasileiros, AMR: Hispânicos Miscigenados dos EUA/Latinos Americanos do 1000GP (populações PUR, CLM, MXL e PEL), ASH: *Ashaninka*, AYM: *Aymara*, GUA: *Guarani*, MAC: *Machiguenga*, MEX: (*Huichol* e *Tarahumara*), QUE: *Quechua*, TUP: *Tupiniquim*.



Índices de heterozigosidade e estatísticas F

Essas análises foram realizadas em dois contextos diferentes, os quais compreendem os bancos de dados **BeadXpress** e **TargetSeq**. O primeiro banco nos permite analisar os genes de forma ampla, já que compreende marcadores situados em regiões de até 1Mb, o que engloba variantes distribuídas por todas as porções de ambos genes. Já os dados obtidos com o **TargetSeq**, por terem sido obtidos pelo sequenciamento de uma parte dos genes, permitem análises mais detalhadas das regiões que flanqueiam os SNPs de interesse *PDE4B*-rs6683977 e *MYT1L*-rs17039396.

*Banco de dados **BeadXpress***

A análise de heterozigosidade é uma medida clássica de variabilidade que permite identificar o grau de diversidade genética da população e outras questões relacionadas à dinâmica da população. As medidas de heterozigosidade utilizando os dados do banco **BeadXpress** foram tomadas em cada uma das populações amostradas considerando todos os marcadores (Tabela 5).

Tabela 5. Índices de heterozigosidade observada intrapopulacional. ASH: *Ashaninka*, AYM: *Aymara*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, MCP: Martinho Campos, CAR: Carmésia, RES: Resplendor, SJM: São João das Missões, Projeto 1000 Genomas: EUR: Europeus, AFR: Africanos, EAS: Leste Asiáticos, SAS: Sul-Asiáticos.

	Peru			**Nativos Brasil**		**Miscigenados Brasil**				**1000G**			
	ASH	**AYM**	**MAC**	**GUA**	**TUP**	**CAR**	**MCP**	**RES**	**SJM**	**EUR**	**AFR**	**EAS**	**SAS**
PDE4B	0,25	0,33	0,31	0,32	0,35	0,35	0,37	0,38	0,32	0,35	0,30	0,38	0,38
MYT1L	0,21	0,21	0,16	0,24	0,37	0,28	0,23	0,26	0,23	0,27	0,30	0,32	0,28

Com o objetivo de medir o percentual da variabilidade que se deve às diferenças interpopulacionais foi estimado o índice F_{ST} par-a-par (WRIGHT 1949). Como esperado, observa-se na Tabela 6 que os maiores valores de diferenciação genética interpopulacional ocorrem quando há a comparação com amostras africanas (AFR). Ademais, observam-se baixos valores de diferenciação quando grupos de nativos são comparados entre si e quando grupos de miscigenados também são comparados entre si.

Análises de componentes principais foram também realizadas com os SNPs genotipados para avaliar a diferenciação observada entre os indivíduos. Pelas análises de componentes principais (Figura 6) observa-se que há certo grau de diferenciação no plano cartesiano dos indivíduos com ancestralidades diferentes, e como esperado, os indivíduos miscigenados estão localizados em posição intermediária. É interessante ressaltar que o segundo eixo principal possibilitou diferenciar também as amostras ameríndias de acordo com o país de origem, muito embora haja sobreposição considerável.

A diferenciação genética entre grupos também foi avaliada com base nos dados de cada SNP utilizando o índice de F_{CT}. No caso específico dessa análise foi investigado qual a proporção da variabilidade observada se deve à diferenciação entre os grupos de nativos (peruanos e brasileiros), africanos e europeus obtidos do 1000GP. Na Tabela 7 são apresentados os índices de F_{CT} para cada um dos SNPs genotipados em cada gene, sendo que 25 comparações merecem destaque por apresentaram valores acima de 0,12, o que indica um relevante grau de diferenciação segundo Elhaik (2012).

Tabela 6 Índices de F_{ST} par-a-par entre as populações. ASH: *Ashaninka*, AYM: *Aymara*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, MCP: Martinho Campos, CAR: Carmésia, RES: Resplendor, SJM: São João das Missões; Projeto 1000 Genomas: EUR: Europeus, AFR: Africanos, EAS: Leste Asiáticos, SAS: Sul-Asiáticos. (* p<0,005)

PDE4B	**ASH**	**AYM**	**MAC**	**CAR**	**MCP**	**RES**	**SJM**	**GUA**	**TUP**	**AFR**	**EUR**	**EAS**	**SAS**		***MYT1L***
ASH	-	0,037*	0,057*	0,226*	0,187*	0,216*	0,285*	0,097*	0,172*	0,303*	0,201*	0,246*	0,221*	**ASH**	
AYM	0,027*	-	0,080*	0,113*	0,082*	0,098*	0,175*	0,043*	0,099*	0,249*	0,143*	0,217*	0,173*	**AYM**	
MAC	0,014*	0,009*	-	0,385*	0,311*	0,350*	0,446*	0,182*	0,270*	0,362*	0,250*	0,304*	0,272*	**MAC**	
CAR	0,096*	0,108*	0,066*	-	0,024	-0,013	0,001	0,028*	-0,016	0,070*	0,063*	0,124*	0,072*	**CAR**	
MCP	0,092*	0,124*	0,082*	0,019	-	0,029*	0,067*	0,001	-0,003	0,154*	0,068*	0,180*	0,097*	**MCP**	
RES	0,045*	0,073*	0,025*	0,005	0,006	-	0,003	0,035*	-0,014	0,062*	0,043*	0,103*	0,058*	**RES**	
SJM	0,055*	0,067*	0,039*	0,023	0,032*	0,011	-	0,070*	-0,002	0,053*	0,110*	0,157*	0,114*	**SJM**	
GUA	0,075*	0,021*	0,054*	0,090*	0,051*	0,026*	0,025*	-	0,017*	0,208*	0,123*	0,203*	0,138*	**GUA**	
TUP	0,085*	0,039*	0,058*	0,075*	0,027*	0,017*	0,016	0,001	-	0,150*	0,113*	0,179*	0,120*	**TUP**	
AFR	0,276*	0,197*	0,272*	0,155*	0,191*	0,168*	0,129*	0,274*	0,280*	-	0,176*	0,115*	0,163*	**AFR**	
EUR	0,165*	0,143*	0,163*	0,056*	0,033*	0,038*	0,074*	0,129*	0,111*	0,175*	-	0,119*	0,013*	**EUR**	
EAS	0,169*	0,063*	0,163*	0,126*	0,117*	0,090*	0,070*	0,131*	0,142*	0,127*	0,134*	-	0,091*	**EAS**	
SAS	0,148*	0,081*	0,138*	0,071*	0,067*	0,052*	0,061*	0,121*	0,111*	0,113*	0,047*	0,056*	-	**SAS**	

Figura 6. Análise de componentes principais realizada com os SNPs genotipados nos genes *PDE4B* e *MYT1L*. PC1 = 12%, PC2 = 10%. NAT_BRA: Nativos Brasileiros; NAT_PERU: Nativos Peruanos; AMG_BRA: Miscigenados Brasileiros, EUR: Europeus (1000GP), AFR: Africanos (1000GP).

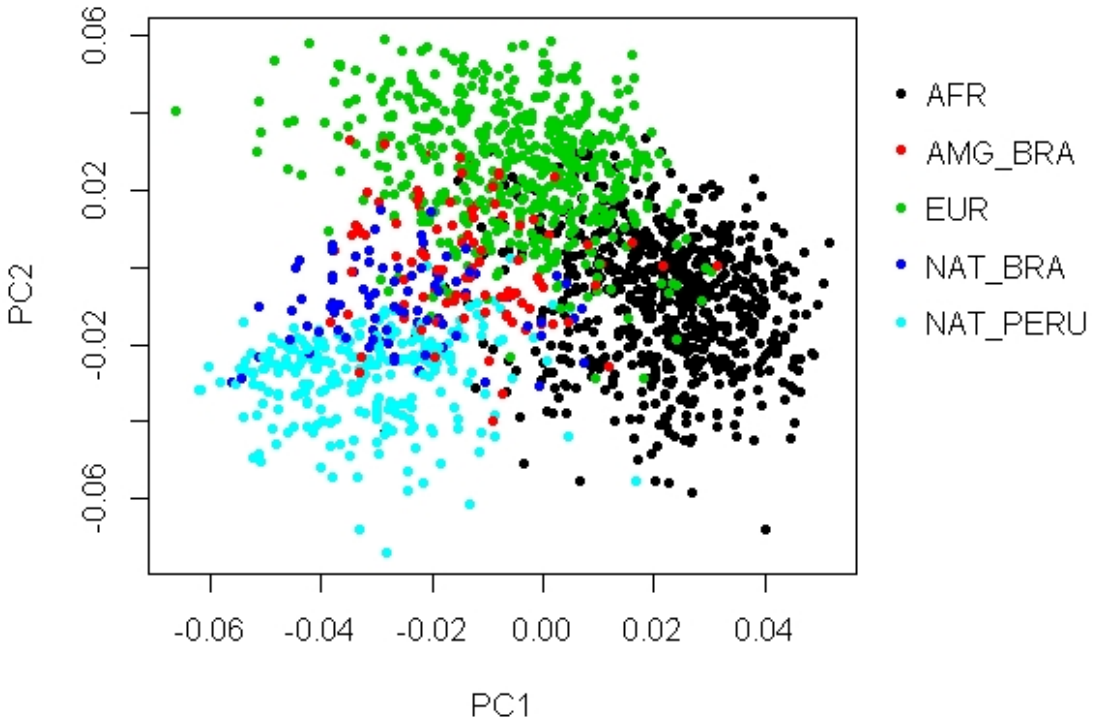


Tabela 7. Índices de F_{CT} par-a-par entre as populações de nativos (brasileiros e peruanos) e as populações de africanos (AFR) e europeus (EUR) do Projeto 1000 Genomas para cada SNP do banco **BeadXpress**. Em negrito valores significativos ($p<0.05$), enquanto que em cinza valores de $F_{CT}>0,12$. NAN: valores não obtidos pela aplicação de filtros de qualidade.

F_{CT} - PDE4B			***F_{CT} - MYT1L***		
SNP	**NAT x EUR**	**NAT x AFR**	**SNP**	**NAT x EUR**	**NAT x AFR**
rs12081185	**0,079**	**0,028**	rs3924017	**0,346**	0,012
rs11589566	**0,025**	**0,192**	rs6728613	**0,081**	**0,593**
rs11208774	NAN	NAN	rs11127305	**0,065**	**0,148**
rs1937456	0,001	**0,150**	rs2241685	-0,006	0,039
rs2186123	-0,007	0,016	rs3748988	**0,054**	**0,393**
rs2503205	NAN	**0,050**	rs3748989	-0,002	0,000
rs10889591	NAN	NAN	rs12986730	**0,045**	**0,277**
rs1937453	0,004	0,007	rs12991351	0,014	**0,033**
rs6588177	**0,510**	**0,692**	rs6729968	0,019	**0,536**
rs12033425	**0,187**	-0,010	rs6548050	**0,079**	**0,171**
rs1354060	-0,004	**0,028**	rs10178240	-0,008	**0,262**
rs1500951	**0,107**	**0,015**	rs9284792	**0,089**	**0,602**
rs12119734	**0,211**	**0,030**	rs12468174	**0,033**	**0,034**
rs12027416	0,022	**0,056**	rs17039334	**0,034**	**0,012**
rs10454453	**0,075**	**0,352**	rs11683072	0,003	**0,015**
rs12045945	-0,007	**0,014**	rs6705656	**0,756**	**0,405**
rs6661245	**0,065**	**0,316**	rs12468168	**0,039**	**0,143**
rs522037	**0,029**	**0,313**	rs17338491	**0,047**	0,002
rs6683604	**0,340**	-0,001	rs17039396	0,029	**0,248**
rs6683977	**0,340**	0,025	rs17039463	-0,001	**0,071**
rs783036	**0,053**	**0,372**	rs17039474	0,002	**0,030**
			rs13034457	**0,039**	0,003
			rs13388663	NAN	NAN
			rs10195351	0,021	-0,018

Banco de dados **TargetSeq**

Análises moleculares de variância (AMOVA) foram realizadas considerando populações do 1000GP e as amostras de nativos americanos desse estudo. Os resultados para os valores de F_{CT} para as variantes previamente reportadas como associadas à recidiva de LLA, calculados par-a-par entre a população continental nativo-americana (a qual inclui as populações brasileiras, mexicanas e peruanas) e as populações do 1000GP, estão apresentados na Tabela 8. Da mesma forma como ocorrido no cálculo das frequências alélicas, como os SNPs rs524770 e rs6683977 foram genotipados por outros métodos (*array* e TaqMan respectivamente) o cálculo do F_{CT} para essas variantes incluiu um número maior de amostras (Tabela 1). Além desses, na seção Apêndice desse capítulo são também apresentados os valores de F_{CT} par-a-par entre as amostras de nativos e as demais populações do 1000GP para as demais variantes do banco de dados **TargetSeq** do gene *PDE4B* (Tabela Apêndice 4) e *MYT1L* (Tabela Apêndice 5), dentre as quais se destacam algumas com alto valor de diferenciação.

Tabela 8. Valores dos índices de F_{CT} par-a-par entre as populações para as variantes previamente reportadas como associadas à recidiva de LLA genotipadas por sequenciamento direcionado. As significâncias (p-valor) são também apresentados. NAT: populações nativas brasileiras, peruanas e mexicanas, Projeto 1000 Genomas: EUR: Europeus, AFR: Africanos, EAS: Leste-Asiáticos, SAS: Sul-Asiáticos.

Gene	SNP	Pops	EUR	AFR	SAS	EAS
PDE4B	rs6683977	NAT	0,36*	0,02***	0,14**	0,00
	rs524770		0,19**	0,32*	0,07**	0,10**
	rs641262		0,25**	0,00	0,04	0,00
	rs546784		0,25***	0,00	0,07***	0,00
MYT1L	rs17039396		0,02***	0,35*	0,05***	0,00

*p<0.001, **p<0.01, ***p<0.05

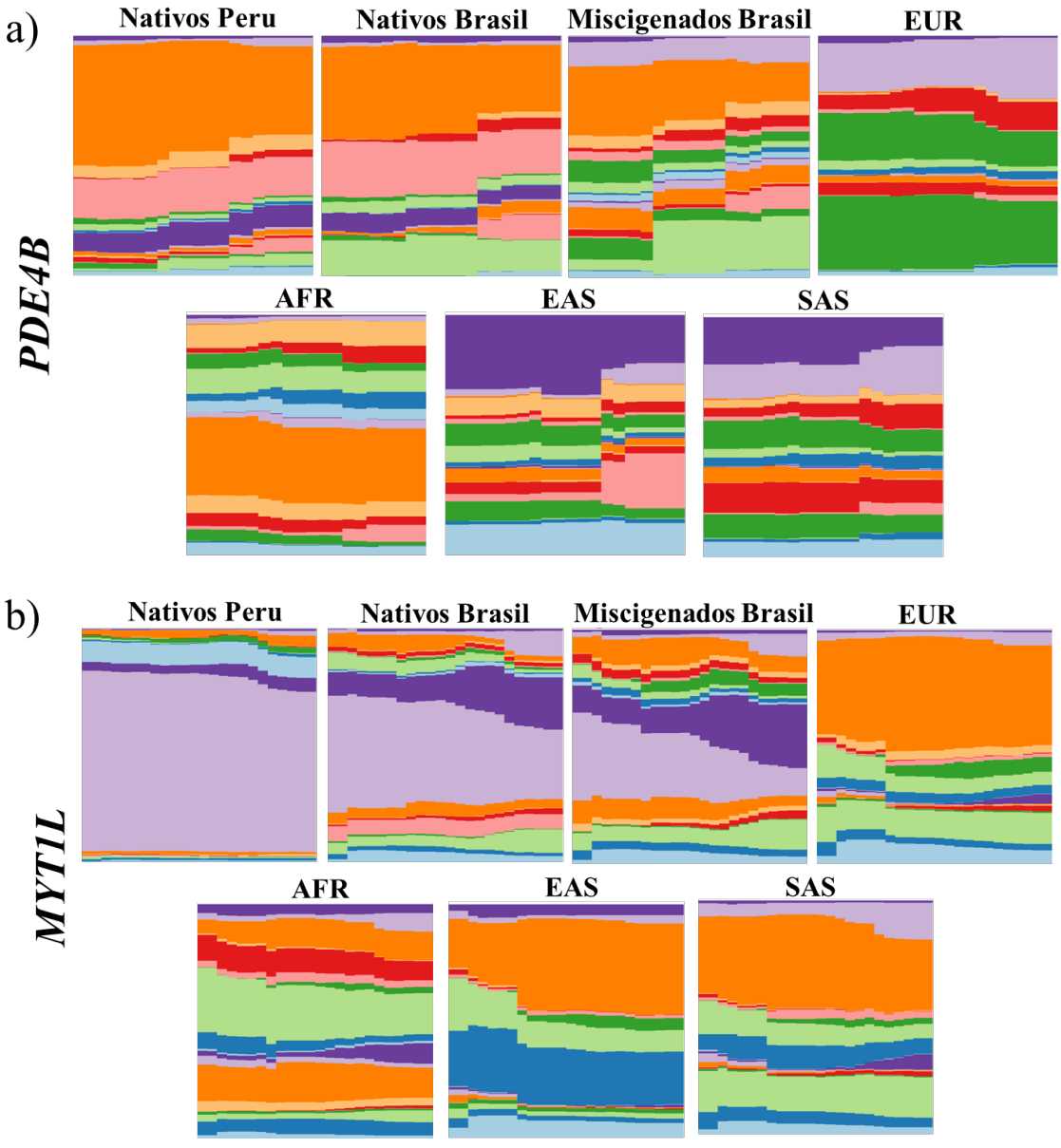
Diversidade haplotípica e desequilíbrio de ligação

Para as análises de diversidade haplotípica e desequilíbrio de ligação foram utilizados os bancos de dados **TargetSeq**, **BeadXpress** e **TaqMan-rs6683977+2.5M**, considerando que cada um deles possui diferentes marcadores, distribuídos por regiões diversas dos genes *PDE4B* e *MYT1L*. Assim, cada análise pode revelar padrões de diversidade haplotípica distintos e também indicar diferentes marcadores em DL com os SNPs de interesse.

Banco de dados ***BeadXpress***

A seguir são apresentados os gráficos com o padrão de distribuição haplotípica de cada gene em cada população gerados com o programa *Haploscope* (Figura 7). Interessante notar que para ambos os genes há maior diversidade de haplótipos nas populações de miscigenados brasileiros quando comparados aos nativos do Peru e Brasil, o que seria resultado da miscigenação com europeus e africanos. Também notam-se frequências haplotípicas similares entre os nativos, e a presença de alguns haplótipos que são encontrados em maior frequência nas populações asiáticas (EAS e SAS).

Figura 7. Distribuição dos haplótipos (representados pelas cores) dos genes e suas frequências relativas (tamanho dos blocos haplotípicos) de cada gene relacionado à recidiva da LLA em cada grupo estudado. Cada linha do gráfico representa um haplótipo de um indivíduo. a) *PDE4B*; b) *MYT1L*. EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP).



As análises de desequilíbrio de ligação realizadas com o programa *Haploview* indicaram que nenhum dos 27 SNPs do gene *MYT1L* estão em desequilíbrio de ligação (r^2>0.8) com o SNP rs17039396 nas populações estudadas. Com relação aos 21 SNPs do gene *PDE4B*, apenas o SNP rs6683604 está em desequilíbrio de ligação (r^2>0.8) com o rs6683977 nas populações de nativos peruanos *Ashaninkas* e *Machiguengas* e na população de miscigenados brasileiros (AMG). A mesma associação entre esses dois SNPs foi encontrada para as populações EUR, AMR, EAS e SAS do 1000GP (Tabela 9).

Banco de dados **TaqMan-rs6683977+2.5M**

Tendo em vista que o banco de dados **TaqMan-rs6683977+2.5M** foi construído unicamente com marcadores do gene *PDE4B*, apenas esse gene foi analisado. Em razão da disponibilidade de amostras, os dados desse banco só foram gerados para as populações de nativos peruanos (Tabela 1).

As variantes do *PDE4B* das amostras dos nativos peruanos foram filtradas do painel de aproximadamente 2,5 milhões de SNPs. No total, foram mantidas 408 variantes ao longo de toda a extensão do gene (Chr1: 66.258.197-66.840.259 - GRCh37/hg19). O programa *Haploview* utiliza o método padrão indicado por Gabriel e colaboradores (2002) para avaliar os índices de ligação entre os marcadores de uma mesma região para delimitar blocos de ligação. As análises indicaram 23 blocos de ligação variando de <1Kb a até 100 Kbs englobando de dois a 29 SNPs. Destaca-se o bloco com 37 Kb em que estão inseridos os SNPs de interesse rs6683977 e rs524770. Considerando todas as populações peruanas em conjunto, a soma da frequência dos haplótipos que possuem o alelo G (alelo ancestral), reportado como de risco para recidiva de LLA no SNP rs6683977 (YANG *et al*. 2011, 2012), totaliza cerca de 91%. Já a soma da frequência dos haplótipos que possuem o alelo A do SNP rs524770 (alelo derivado), reportado como de risco para a recidiva de LLA (YANG *et al.* 2012), alcança cerca de 79%. Analisando esse bloco, nota-se que o alelo de risco do SNP rs524770 sempre ocorre em haplótipos em que o alelo de risco do SNP rs6683977 está presente, indicando certo grau de desequilíbrio de ligação entre eles. Porém, a análise dos índices de DL (r^2) indicaram poucos SNPs com valores acima de 0,8 quando a comparação é realizada com o SNP de interesse rs6683977. Nas populações de nativos peruanos a associação de maior destaque ocorre com o SNP rs519044, o qual está em desequilíbrio

de ligação (r^2>0,8) com rs6683977 em todas as populações de nativos peruanos. O mesmo padrão de desequilíbrio de ligação entre esses dois SNPs foi observado para a maioria das populações do 1000GP, com exceção dos Africanos (Tabela 9).

Banco de dados **TargetSeq**

Em razão da eliminação de variantes com MAF<0,05 e cuja taxa de dados faltantes fosse maior que 25% nas populações alguns polimorfismos foram desconsiderados das análises. As populações de nativos brasileiros apresentaram valores expressivos de dados faltantes e por isso o número de comparações entre polimorfismos foi menor nessas populações.

As análises de desequilíbrio de ligação para as regiões sequenciadas dos genes *PDE4B* e *MYT1L* são apresentadas na Tabela 9 e Figuras 8 e 9. É importante notar que as análises do gene *PDE4B* indicaram um número muito superior de pares de DL que incluíssem a variante de interesse (Tabela 9), do que as análises com o SNP do gene *MYT1L*. No caso do *MYT1L*, apesar de serem observados blocos de ligação relativamente robustos (Figura 8), as comparações realizadas indicaram que apenas a variante rs60515206 está em DL (r^2>0,8) com o SNP de interesse rs17039396; e esse resultado é apenas observado em duas populações de nativos: *Ashaninkas* e *Machiguengas*.

A Tabela 9 mostra que para o gene *PDE4B* há muitas variantes em DL com o SNP de interesse rs6683977 em muitas populações, com exceção da população africana. Os blocos de ligação inferidos pelo *Haploview* para o gene *PDE4B* nas populações estudadas incluíram, além do SNP de interesse rs6683977, outros SNPs relacionados à recidiva de LLA (Figura 9). As frequências dos haplótipos que incluem esses SNPs podem ser encontradas na Tabela 10. De fato, nota-se que as frequências dos haplótipos que incluem os alelos de risco rs6683977.G (alelo ancestral), rs546784.T (alelo ancestral) e rs641262.T (alelo ancestral) são maiores dos que as encontradas nas populações do 1000GP, em especial quando a comparação é realizada com os europeus.

Tabela 9. SNPs do banco de dados TargetSeq, BeadXpress e Taqman-rs6683977+2.5M em desequilíbrio de ligação LD (r^2>0.80) com *PDE4B*-rs6683977 e *MYT1L*-rs17039396 em diferentes populações. EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP), MEX (*Tarahumara* e *Huichol*), QUE: *Quechua*, AYM: *Aymara*, ASH: *Ashaninka*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, AMG-BRA: Miscigenados Brasileiros. 0: nenhuma associação encontrada; (-) população sem dados para o banco em questão.

Pop.	SNPs em DL (r^2>0.80) com *PDE4B*-rs6683977			SNPs em DL (r^2>0.80) com *MYT1L*-rs17039396	
	TargetSeq (83 SNPs)	BEADXPRESS (21 SNPS)	TAQMAN-rs6683977+2.5M (408 SNPS)	TargetSeq (66 SNPs)	BEADXPRESS (27 SNPS)
EUR	rs638111, rs641262, rs494735, rs495477, rs12137115, rs12137080, rs6683604, rs546784, rs6668516, rs782967	rs522037, rs6683604	rs486136, rs519044	0	0
AFR	0		rs6664618, rs12731764 rs12141125	0	0
EAS	rs494735, rs12137115, rs6683604, rs6668516, rs546784, rs495477, rs641262, rs782967, rs12137080	rs6683604	rs519044, rs6664618, rs486136	0	0
SAS	rs12137080, rs12137115, rs495477, rs546784, rs6683604, rs494735, rs6668516, rs782967	rs6683604	rs519044, rs486136	0	0
AMR	rs6668516, rs6683604, rs494735, rs12137115, rs495477, rs641262, rs546784, rs12137080, rs782967	rs6683604	rs519044	0	0
MEX	rs641262, rs495477, rs12137115, rs6683604, rs546784, rs6668516	-	-	0	-
QUE	rs641262, rs495477, rs12137115, rs12137080, rs6683604, rs546784, rs6668516	-	rs519044	0	-
AYM	rs641262, rs495477, rs12137115, rs6683604, rs546784, rs6668516	0	rs519044	0	0
ASH	rs641262, rs495477, rs12137115, rs6683604, rs546784, rs6668516	rs6683604	rs519044	rs60515206	0

Pop.	SNPs em DL (r^2>0.80) com *PDE4B*-rs6683977			SNPs em DL (r^2>0.80) com *MYT1L*-rs17039396	
	TargetSeq (83 SNPs)	BEADXPRESS (21 SNPS)	TAQMAN-rs6683977+2.5M (408 SNPS)	TargetSeq (66 SNPs)	BEADXPRESS (27 SNPS)
MAC	rs641262, rs495477, rs12137115, rs12137080, rs6683604, rs546784, rs6668516	rs6683604	rs486136, rs519044	rs60515206	0
TUP	0	0	-	0	0
GUA	0	0	-	0	0
AMG-BRA	-	rs6683604	-	-	0

Figura 8. Desequilíbrio de ligação par-a-par entre os polimorfismos do gene *MYT1L* nas populações peruanas (PERU), mexicanas (MEX), africanas (AFR) e europeias (EUR). Quadros sólidos indicam desequilíbrio (r^2>0,8), quadros em escala de cinza indicam escala gradual de intensidade de desequilíbrio, enquanto que quadros claros indicam alta probabilidade de recombinação. O retângulo em destaque indica a posição do SNP de interesse rs17039396.

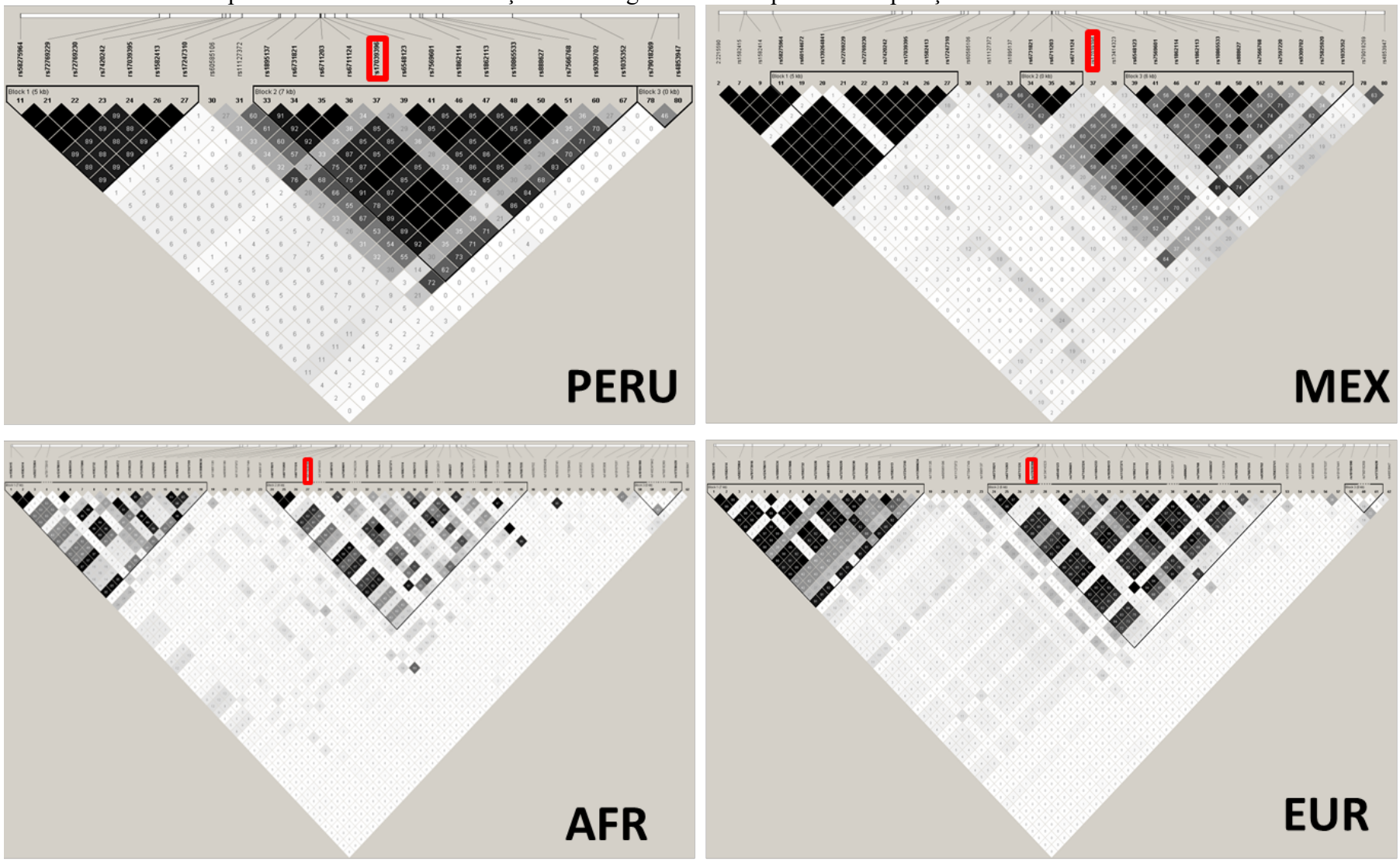


Figura 9. Desequilíbrio de ligação par-a-par entre os polimorfismos do gene *PDE4B* nas populações peruanas (PERU), mexicanas (MEX), africanas (AFR) e europeias (EUR). Quadros sólidos indicam desequilíbrio (r^2>0,8), quadros em escala de cinza indicam escala gradual de intensidade de desequilíbrio, enquanto que quadros claros indicam alta probabilidade de recombinação. O triângulo, o círculo, o retângulo e o triângulo invertido em destaques indicam a posição dos SNPs de interesse rs546784, rs524770, rs6683977 e rs641262, respectivamente.

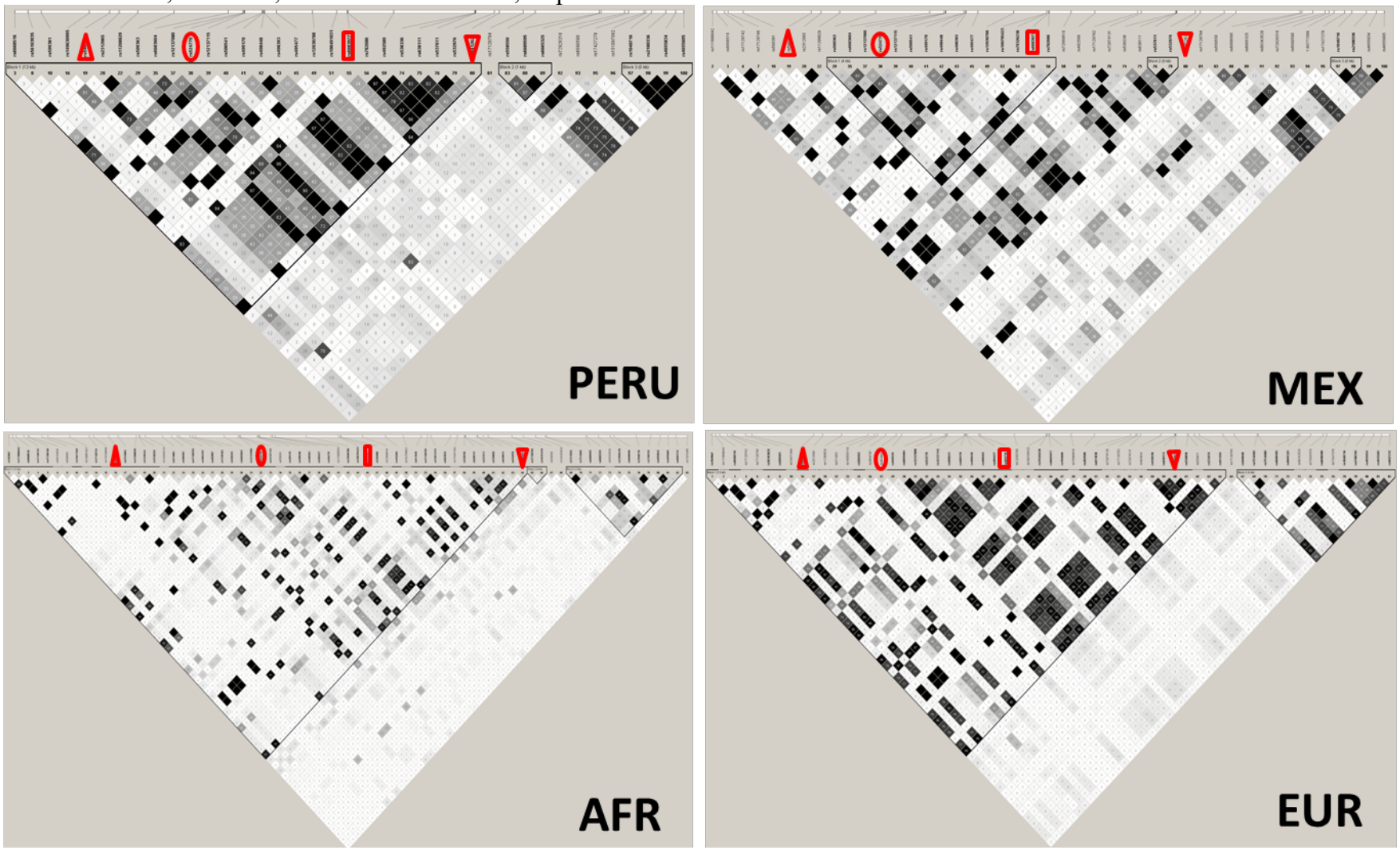


Tabela 10. Haplótipos formados pelas variantes em desequilíbrio de ligação (r^2>0.8) com o SNP *PDE4B*-rs6683977 (destacado em negrito) no banco de dados TargetSeq (em ordem: rs6668516, rs546784, rs6683604, rs12137080, rs12137115, rs495477, rs494735, rs6683977, rs638111, rs641262). Os números de amostras estão apresentados entre parênteses. Populações do 1000GP: AMR= Hispânicos miscigenados dos EUA e Latinos Americanos, SAS= Sul-asiáticos, EAS= Leste-asiáticos, EUR= Europeus. MEX = nativo-americanos do Mexico (*Tarahumaras* e *Huicholes*); populações peruanas de nativos: QUE=*Quechuas*, AYM=*Aymaras*, ASH=*Ashaninkas*, MAC=*Machiguengas*. (-) indica ausência ou haplótipos em baixa frequência.

Populações/ Haplótipos	AMR (347)	SAS (489)	EAS (504)	EUR (503)	MEX (22)	QUE (20)	AYM (24)	ASH (16)	MAC (22)
GTTCCAT**G**CT	0,62	0,51	0,48	0,41	0,80	0,70	0,63	0,84	0,75
ACCTGGC**C**TC	0,27	0,29	0,16	0,56	0,09	0,13	0,08	0,06	0,18
GTTCCAT**G**TT	0,05	0,10	0,35	0,02	0,07	0,15	0,21	0,03	0,07
GTTTCAC**G**CT	0,03	0,02	-	-	0,04	-	0,08	0,06	-

Evidências de regiões com função regulatória

O *Genome Browser* do UCSC permitiu visualizar as variantes mais diferenciadas (F_{CT}>0,12) com relação apenas à população EUR e também com relação tanto à EUR quanto à AFR em um cenário em que estão também dispostas informações de bancos de dados públicos relativas a evidências de função regulatória. As Figuras 10, 11 e 12 trazem os SNPs dos genes *PDE4B* e *MYT1L* que estão incluídos no banco de dados TargetSeq com maior diferenciação (F_{CT}>0,12). As variantes estão dispostas em conjunto com os dados recuperados de diversos bancos relacionados às funções regulatórias disponibilizados pelo Projeto ENCODE.

As seções H3K4Me1e H3K27Ac mostram as posições em que a modificação das proteínas histonas são sugestivas da existência de acentuadores (*enhancers*) (ou em menor escala, de outras atividades regulatórias). Já a seção H3K4Me3 indica modificações em histonas que estão relacionadas a promotores. Esses dados estão disponíveis para 7 linhagens celulares diferentes. Porém, para melhor visualização, bem como para restringir segundo os propósitos deste estudo, são apresentados dados de apenas 2 linhagens celulares do ENCODE: (i) K-562, a qual foi estabelecida a partir da efusão pleural de uma mulher de 53 anos portadora de leucemia mieloide crônica; e (ii) GM12878, a qual é uma linhagem linfoblastóide B obtida do sangue de uma doadora mulher com ancestralidade do continente Europeu. Em seguida a essas evidências

experimentais, são mostradas os sinais preditos *in silico* pelo programa ChromHMM (*Chromatin state discovery and characterization*) (ERNST & KELLIS, 2012) para as linhagens celulares K652 e GM12878.

Na seção seguinte são ilustradas as posições que estão susceptíveis à clivagem por DNAse em 125 linhagens celulares diferentes, tendo em vista que regiões *cis*-regulatórias em geral (em especial promotores) são particularmente sensíveis à clivagem por essa enzima. A seção posterior (*Txn Factor ChIP*) indica evidências de regiões *trans*-regulatorias, correspondentes aos locais onde fatores de transcrição se ligam. O ensaio *ChipSeq* consiste no sequenciamento dos fragmentos DNA imunoprecipitados com anticorpos específicos para fatores de transcrição. Por limitações do UCSC *Genome Browser* não foi possível filtrar por linhagens de células de interesse nas seções de DNAse e *Txn Factor ChIP*, tais como para a visualização apenas dos sinais para as células GM12878 e K652. Por fim, a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151.

Na Figura 10, relativa ao gene *PDE4B*, notam-se diversos SNPs diferenciados em regiões de função regulatória, de ligação de fatores de transcrição e susceptíveis à ação da DNAse. Destacam-se os SNPs de interesse rs524770 (altamente diferenciado em nativos com relação a europeus, F_{CT}=0,14, e africanos, F_{CT}=0,27) e rs641262 (altamente diferenciado em nativos com relação aos europeus, F_{CT}=0,25), os quais estão em regiões onde há fortes evidências de metilação de histonas em linfócitos B (GM12878). Para melhorar a visualização da região com evidências mais fortes de regulação, a Figura 11 mostra com mais detalhes a região de aproximadamente 1Kb compreendida entre os SNPs rs524770 e rs6683977 e indica os SNPs que estão localizados nessa região. Por outro lado, a Figura 12 relativa ao *MYT1L*, não apresenta muitos SNPs diferenciados e poucas evidências de elementos regulatórios na região. Porém, chama atenção o SNP rs79018269, o qual é altamente diferenciado em nativos com relação a africanos (F_{CT}=0,24) e europeus (F_{CT}=0,23) e encontra-se em uma região com evidências de muitas ligações de fatores de transcrição e sequências expostas à DNAse.

As consultas ao banco de dados ENCODE retornaram os resultados apresentados na Tabela 11 e permitiram refinar a análise apresentada no UCSC *Genome Browser.* Ao todo, três elementos com função regulatória foram encontrados na região de 20Kb sequenciada do gene *PDE4B* e outros dois elementos na região de 20Kb

sequenciada do gene *MYT1L*. Os resultados foram filtrados para mostrarem apenas os elementos com evidência nas células dos tecidos hematopoiético e da medula óssea, apesar do resultado positivo em diversas outras células e tecidos. Os resultados apresentados foram considerados apenas quando o índice *Z-score* fosse superior a 1,64, o que significa acima do limiar de 95% da distribuição. Nota-se que há vários SNPs dentro ou próximos às regiões com função regulatória. Por fim, as análises no banco de dados do RegulomeDB (BOYLE *et al.* 2012) (Tabela 12) indicaram que para o gene *PDE4B*, dos 30 SNPs com F_{CT}>0,12, 21 apresentaram alguma indicação de estarem em regiões com importância regulatória (RegulomeDB *score*), enquanto que para o gene *MYT1L* foi inferida função regulatória para seis dos sete SNPs incluídos nessa análise.

Figura 10. Ilustração da região de 20Kb do gene *PDE4B* em formato UCSC *Genome Browser*. Posicionamento dos SNPs com F_{CT} > 0,12 (para com EUR – azul e EUR e AFR – vermelho) em relação às regiões de modificação de histonas (**H3KMe1, H3K4Me3**, **H3K27Ac**) e aos resultados de modelagem para a presença ou ausência de marcas de cromatina (**ChromHMM** - Laranja: acentuador forte, Amarelo: acentuador fraco, Verde claro: transcrição fraca) em células linfoblastóide (**GM12878** - rosa) e de leucemia mielóide (**K562** - roxo); às regiões de sítios de ligação de elementos regulatórios em *cis*- (clivagem por DNAse) e em *trans*- (ligação de fatores de transcrição - **Txn Factor ChIP**). a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151. As posições dos SNPs de interesse inicial (rs546784, rs524770, rs6683977, rs641262) estão em linhas verticais em azul claro.

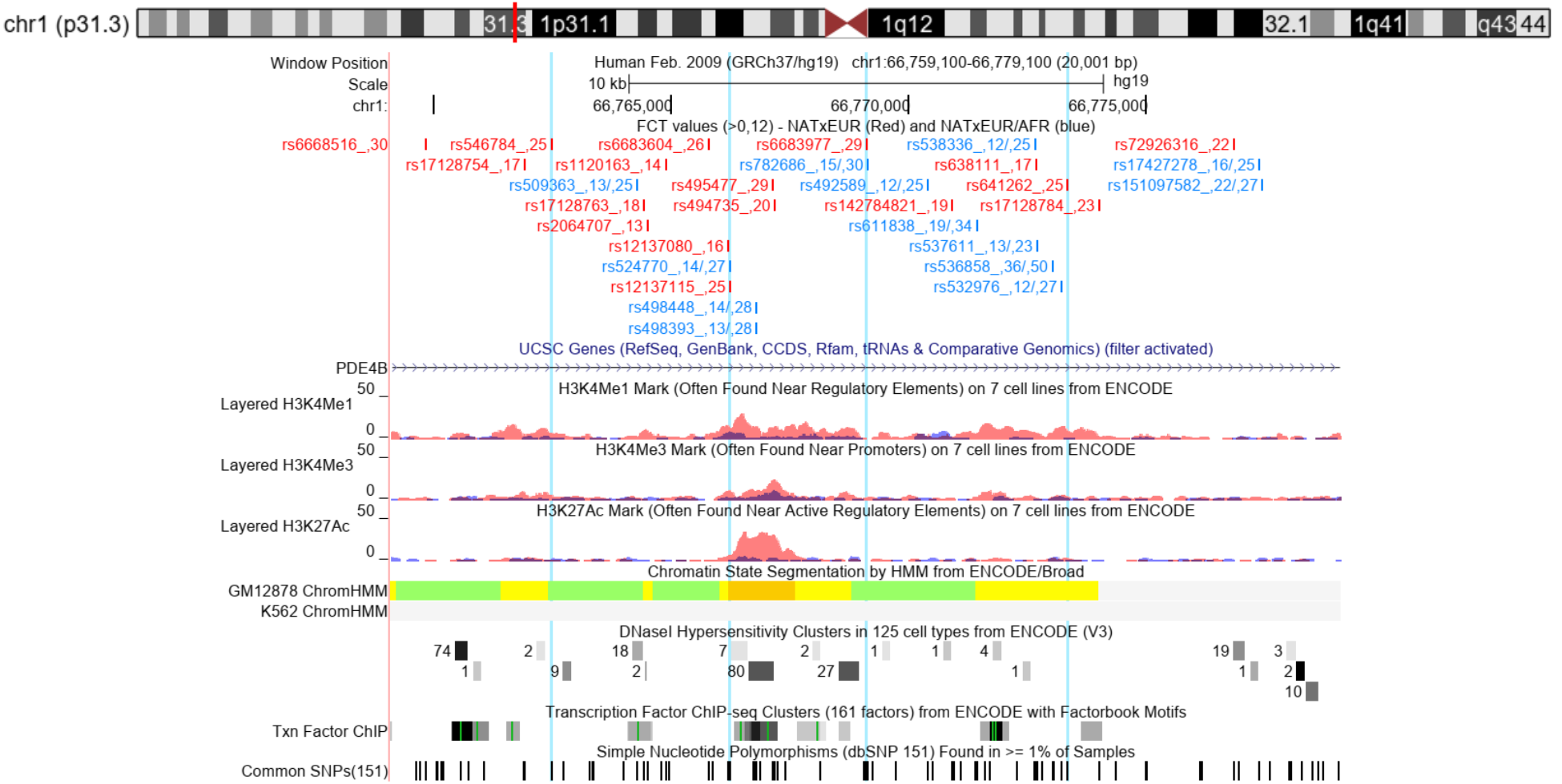


Figura 11. Enfoque na região do gene *PDE4B* (GRCh37, chr1:66.766.000 – 66.769.000) com evidências mais fortes de funçao reulatória em formato UCSC Genome Browser. Posicionamento dos SNPs com F_{CT} > 0,12 (para com EUR – azul e EUR e AFR – vermelho) em relação às regiões de modificação de histonas (H3KMe1, H3K4Me3, H3K27Ac) e aos resultados de modelagem para a presença ou ausência de marcas de cromatina (ChromHMM - Laranja: acentuador forte, Amarelo: acentuador fraco, Verde claro: transcrição fraca) em células linfoblastóide (GM12878 - rosa) e de leucemia mielóide (K562 - roxo); às regiões de sítios de ligação de elementos regulatórios em cis- (clivagem por DNAse) e em trans- (ligação de fatores de transcrição - Txn Factor ChIP). a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151. As posições dos SNPs de interesse inicial (rs524770, rs6683977) estão em linhas verticais em azul claro.

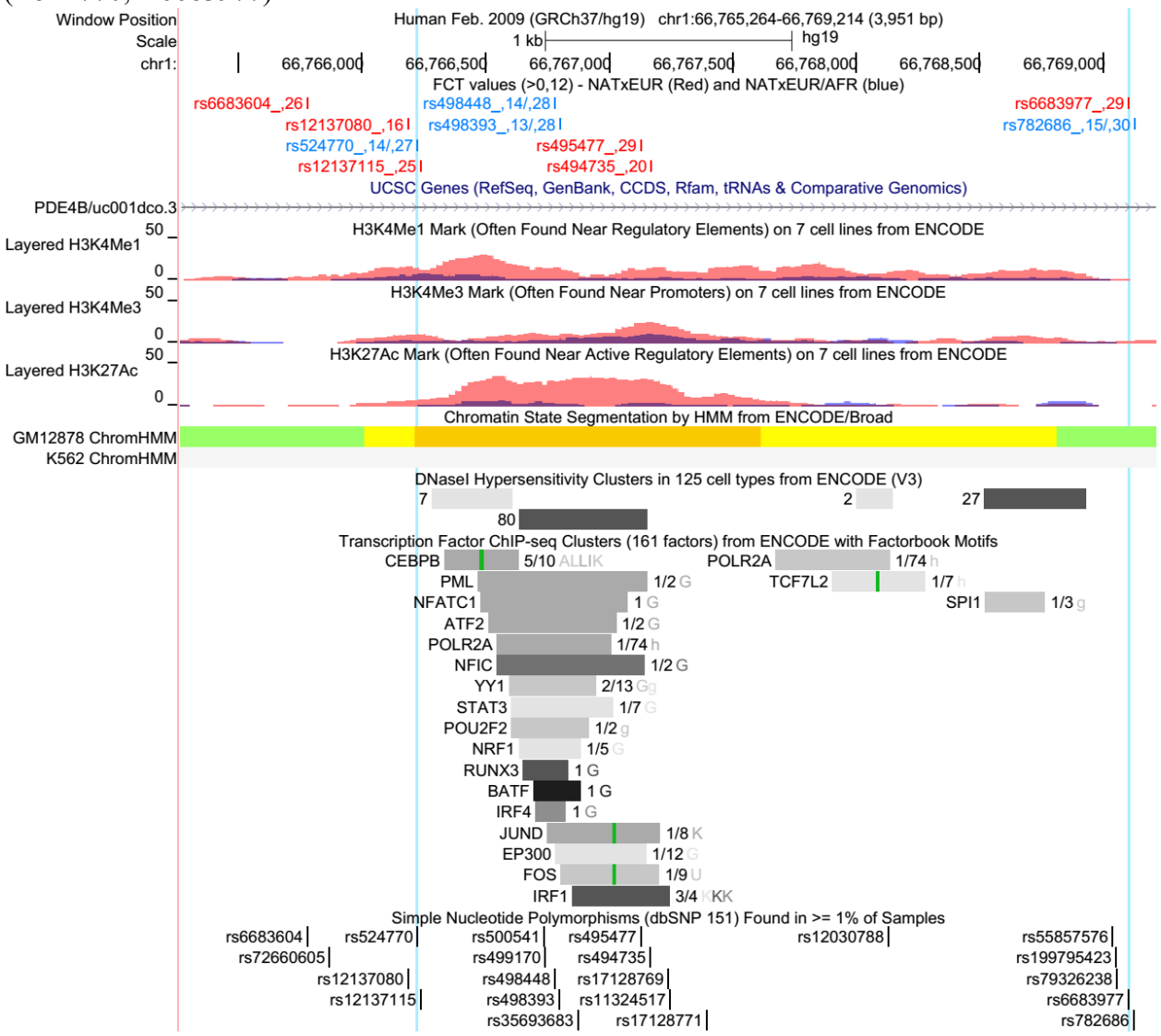


Figura 12. Ilustração da região de 20Kb do gene *MYT1L* em formato UCSC *Genome Browser*. Posicionamento dos SNPs com F_{CT} > 0,12 (para com EUR – azul e EUR e AFR – vermelho) em relação às regiões de modificação de histonas (**H3KMe1, H3K4Me3**, **H3K27Ac**) e aos resultados de modelagem para a presença ou ausência de marcas de cromatina (**ChromHMM** – *não há sinais em evidência*) em células linfoblastóide (**GM12878** - rosa) e de leucemia mielóide (**K562** - roxo); às regiões de sítios de ligação de elementos regulatórios em *cis*- (clivagem por DNAse) e em *trans*- (ligação de fatores de transcrição - **Txn Factor ChIP**). a última seção indica a densidade de SNPs localizados na região disponíveis no banco dbSNP 151. A posição do SNP de interesse inicial rs17039396 está em destaque por uma linha vertical em azul claro.

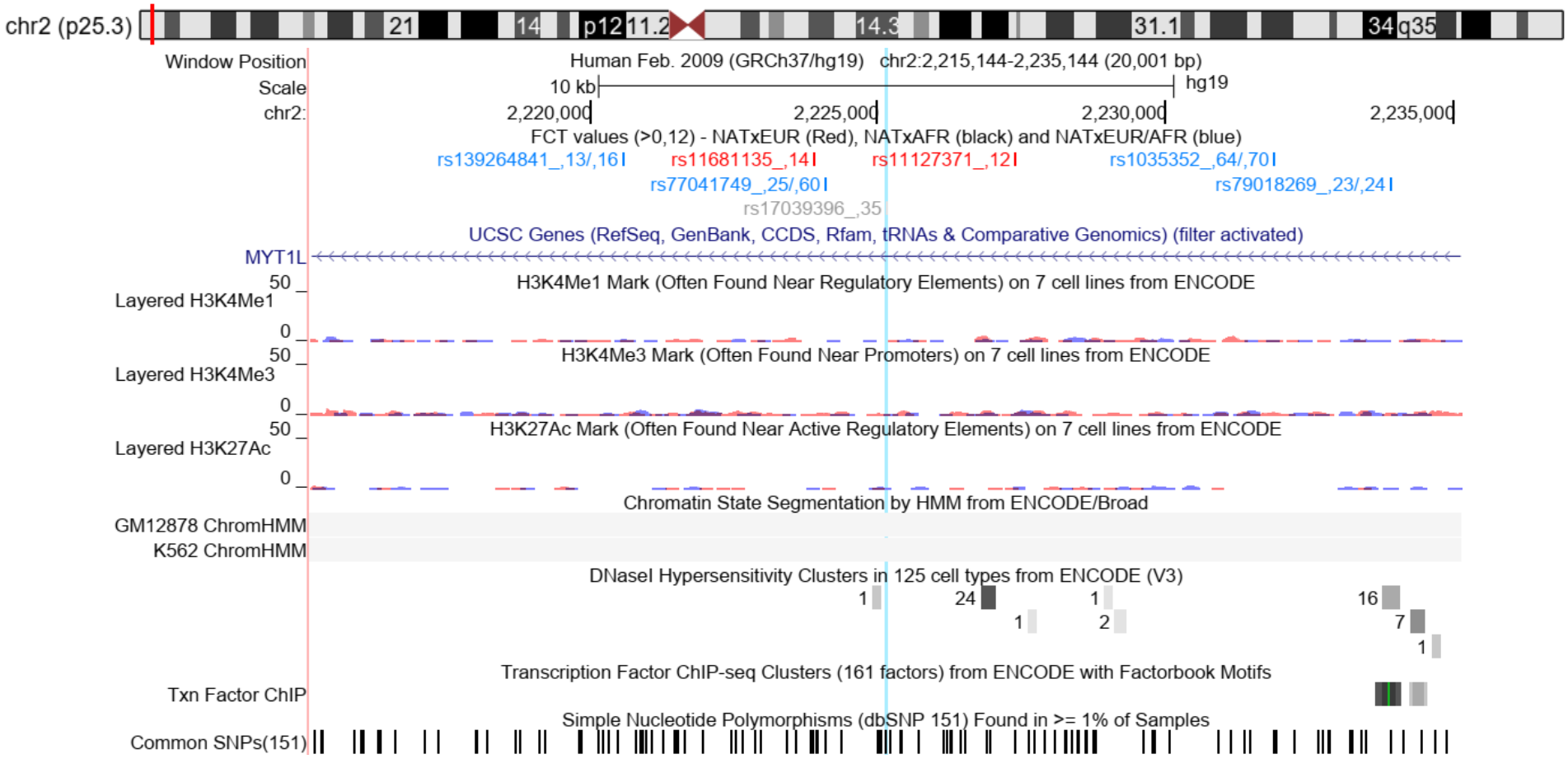


Tabela 11. Evidências da atuação de elementos funcionais nas regiões sequenciadas dos genes *PDE4B* e *MYT1L* obtidos pelo acesso ao banco de dados ENCODE. Em negrito os SNPs em que o índice de diferenciação F_{CT} é maior que 0,12 para comparações com os grupos de africanos e/ou europeus.

Gene	Número de Acesso	Tecido	Tipo Celular	DNAse (Z-score)	H3K4me3 (Z-score)	H3K27ac (Z-score)	Posição Inicial (GRCh37)	Extensão (pb)	SNPs incluídos/próximos (posição)
PDE4B	EH37E0095230	Medula óssea	HS-27A	1,97			66.760.457	556	rs191581125 (66.760.461) rs599381 (66.760.560) rs143612526 (66.760.638) 1:66760675 (66.760.675) rs72677260 (66.760.725)
PDE4B	EH37E0095230	Medula óssea	Tronco Mesenquimal			2,23	66.760.457	556	
PDE4B	EH37E0095230	Medula óssea	HS-5	2,10			66.760.457	556	
PDE4B	EH37E0095235	hematopoietico	Célula T	1,87	1,78		66.766.392	177	**rs12137080** (66.766.185) **rs524770** (66.766.222) **rs12137115** (66.766.236)
PDE4B	EH37E0095235	hematopoietico	Célula multipotente	1,94			66.766.392	177	
PDE4B	EH37E0095235	Medula óssea	Tronco Mesenquimal			2,44	66.766.392	177	
PDE4B	EH37E0095236	hematopoietico	Célula B	2,08		2,03	66.766.652	471	rs500541 (66.766.735) rs499170 (66.766.741) **rs498448** (66.766.780) **rs498393** (66.766.798) rs1040717 (66.766.908) **rs495477** (66.767.128) 1:66767159 (66.767.159) **rs494735** (66.767.167)
PDE4B	EH37E0095236	hematopoietico	Célula T	2,76	2,04		66.766.652	471	
PDE4B	EH37E0095236	hematopoietico	Célula multipotente	2,45			66.766.652	471	
PDE4B	EH37E0095236	Medula óssea	Tronco Mesenquimal			2,45	66.766.652	471	
PDE4B	EH37E0095236	Medula óssea	HS-5	2,17		2,45	66.766.652	471	
MYT1L	EH37E0498148	Medula óssea	Tronco Mesenquimal			1,78	2.226.701	400	rs10204613 (2.226.513)
MYT1L	EH37E0498149	Medula óssea	Tronco Mesenquimal			1,93	2.227.657	475	rs1862113 (2.227.635) rs10865533 (2.227.707) rs12052817 (2.227.894) rs888627 (2.228.072)

Tabela 12. Indicação dos *scores* obtidos pela consulta ao banco de dados RegulomeDB para os 30 SNPs com F_{CT}>0,12. 2b: ligação de fator de transcrição + qualquer motivo + DNase *Footprint* + pico de DNAse; 4: fator de transcrição + pico de DNAse; 5: fator de transcrição ou pico de DNAse; 6: outro.

PDE4B						***MYT1L***	
SNP	***Regulome score***	***SNP***	***Regulome score***	***SNP***	***Regulome score***	***SNP***	***Regulome score***
rs495477	2b	rs498393	4	rs72926316	5	rs79018269	4
rs509363	4	rs494735	4	rs6668516	6	rs139264841	6
rs17128763	4	rs17128754	5	rs546784	6	rs11681135	6
rs2064707	4	rs12137080	5	rs782686	6	rs77041749	6
rs524770	4	rs142784821	5	rs536858	6	rs17039396	6
rs12137115	4	rs538336	5	rs532976	6	rs11127371	6
rs498448	4	rs17128784	5	rs641262	6		

Protocolos utilizados para o tratamento da LLA em países latino americanos

A Tabela 13 mostra os resultados encontrados na busca por protocolos para tratamento da LLA em crianças utilizados em países da América Latina. Do total de 17 países da América Latina pesquisados, as buscas possibilitam identificar protocolos dos seguintes países: Peru, Costa Rica, Colômbia, Argentina, Chile, Equador, México, Guatemala e Brasil (Tabela 13). Infelizmente, documentos oficiais não estão disponíveis para todos os países, portanto alguns protocolos foram identificados por estudos acadêmicos (como dissertações e teses). Os documentos encontrados permitiram identificar que nos nove países os protocolos reportados foram baseados no protocolo BFM, o qual é o mais difundido e utilizado globalmente e que possui indicação de aplicação da fase de intensificação ou reindução. Porém, em alguns documentos, ainda que oficiais, não foi possível identificar quais seriam os fármacos, as dosagens e o tempo total da fase de intensificação. Além dos países apresentados na Tabela 13, as buscas não retornaram resultados relevantes para os seguintes países: Venezuela, Uruguai, Bolívia, Paraguai, Honduras, Nicarágua, Panamá e El Salvador.

Tabela 13. Documentos e referências encontrados sobre os protocolos de tratamento da LLA utilizados em países latino-americanos

Países	Instituição	Título	Sítio na rede ou fonte
Peru	Ministerio de Salud	Plan nacional para la atención integral de la leucemia linfática aguda em pacientes de1 a 21 años (Plan Salvador: 2017- 2021)	▪ http://bvs.minsa.gob.pe/local/MINSA/4233.pdf
	Ministerio de Salud	Guías de práctica clínica de leucemia linfática aguda	▪ https://portal.inen.sld.pe/guias-tecnicas/
Costa Rica	Ministerio de Salud	Manual de Normas para el Tratamiento de Cáncer em Costa Rica	▪ https://www.imprentanacional.go.cr/editorialdigital/libros/textos%20juridicos/manual_contra_el_cancer_edincr.pdf
Colômbia	Ministerio de Salud y Protección Social - Colciencias	Guía de Práctica Clínica para la detección oportuna, diagnóstico y seguimiento de leucemia linfoide aguda y leucemia mieloide aguda em niños, niñas y adolescentes	▪ http://www.iets.org.co/reportes-iets/Paginas/leucemia-linfoide-y-mieloide-aguda.aspx
	Revista Colombiana de Cancerología	Guía de atención integral para la detección oportuna, diagnóstico, tratamiento y seguimiento de leucemia linfoide aguda en ninos, ninas y adolescentes	▪ Rev Colomb Cancerol. 2016;20(1):17-27
Argentina	Sociedad Argentina de Hematología	Guías de diagnóstico y tratamiento - Leucemias Agudas	▪ http://sah.org.ar/docs/2017/006-Leucemias%20Agudas.pdf
Chile	Ministerio de Salud	Guía Clínica Leucemia en Personas Menores de 15 Años	▪ https://www.minsal.cl/sites/default/files/GPC_Leucemia_infantil.pdf
	Revista Medica de Chile	Leucemia linfoblástica aguda estirpe B Philadelphia negativa en adolescentes y adultos jóvenes. Resultados del Protocolo Terapéutico LLA 15-30, Programa Nacional de Cáncer del Adulto (PANDA), Ministerio de Salud, Chile	▪ Rev Med Chile 2014; 142: 707-715
Equador	Universidad Central del Ecuador	Estado nutricional y evolución de leucemia linfoblástica en niños de Solca Quito periodo enero 2009 a diciembre 2014	▪ http://www.dspace.uce.edu.ec/bitstream/25000/11140/1/T-UCE-0006-008-2017.pdf
Mexico	Archives of Medical Research	Results of Treating Childhood Acute Lymphoblastic Leukemia in a Low-middle Income Country: 10 Year Experience in Northeast Mexico	▪ Archives of Medical Research 47 (2016) 668e676
	Instituto Nacional De Salud Pública	Protocolo de la Atención Para Leucemia Linfoblástica. Guía Clínica Y Esquema de Tratamiento	▪ http://www.salud.gob.mx/unidades/dgpfs/micsitio/ptcia/recursos/LEUCEMIA.pdf
Guatemala	Universidad de San Carlos de Guatemala	Toxicidad de la Quimioterapia En Pacientes Con Diagnóstico de Leucemia Linfocítica Aguda, Leucemia Mielocítica Aguda Y Linfomas	▪ http://biblioteca.usac.edu.gt/tesis/05/05_8750.pdf
Brasil	Ministério da Saúde - Secretaria de Atenção à Saúde	Portaria Nº 115, de 10 de Fevereiro de 2012 - Diretrizes Diagnósticas e Terapêuticas - Tratamento Da Leucemia Linfoblástica Aguda Cromossoma Philadelphia Positivo De Criança E Adolescente Com Mesilato De Imatinibe	▪ http://conitec.gov.br/images/Protocolos/DDT/Tratamento_LeucemiaLinfoblasticaAguda-CriancaeAdolescente.pdf
	Hospital Albert Einstein	Guia de Protocolos e Medicamentos para Tratamento em Oncologia e Hematologia 2013	▪ https://medicalsuite.einstein.br/pratica-medica/guias-eprotocolos/Documents/Guia_Oncologia_Einstein_2013.pdf

DISCUSSÃO

Tendo em vista a reconhecida diversidade genética humana, resultado da atuação diferenciada de forças evolutivas nas diversas populações ao longo da história, os padrões genético-populacionais de genes de interesse biomédico e suas variantes possibilitam compreender os aspectos que envolvem a incidência e o tratamento de doenças em diferentes populações. O caso aqui tratado da associação da ancestralidade nativo-americana com a recidiva da LLA revela características próprias das variantes de genes que apresentam fortes evidências de atuação nos mecanismos moleculares envolvidos com a resposta ao tratamento. Os resultados encontrados corroboram estudos anteriores ao revelarem a relação da proporção da ancestralidade nativo-americana das populações com as variantes dos genes *PDE4B* e *MYT1L*. Por isso, reforçam o potencial da utilização da informação dessas regiões gênicas e suas variantes para a adequação de protocolos de tratamento para a LLA em populações ainda pouco estudadas. Os resultados mostram que para esses genes as populações nativas apresentam perfis genético-populacionais próprios que as diferenciam das demais (vide também KAROL *et al*. 2017 para mais genes relacionados), em especial das de origem europeia, que são as mais estudadas e usadas como base para o desenvolvimento de tratamentos.

Controle de qualidade dos dados - Problemas metodológicos

As regiões sequenciadas dos genes relacionados à recidiva de LLA (banco de dados **TargetSeq**) incluíram algumas variantes que foram também genotipadas pelo método GoldenGate VeraCode Illumina (banco de dados **BeadXpress**), o que possibilitou a avaliação da concordância entre os métodos para três SNPs (rs6683977, rs17039396 e rs6683604). A ineficácia da identificação do alelo C do SNP rs6683977 pelo método BeadXpress da Illumina (comparação das Tabelas Apêndice 1 e 3) pode ser explicado pela baixa qualidade do DNA utilizado (considerando que esse método seria mais sensível à qualidade do DNA) ou pela presença de alelos nulos. Como o DNA utilizado para o processo estava de acordo com os parâmetros de qualidade necessários, a hipótese de ocorrência de alelos nulos foi considerada. O programa GenomeStudio, utilizado para a chamada de variantes geradas no BeadXpress, se baseia especialmente em testes de equilíbrio de *Hardy-Weinberg* para a determinação dos genótipos. Tendo em vista que o teste do equilíbrio de *Hardy-Weinberg* não é sensível o suficiente para evitar que alelos nulos sejam identificados (CARLSON *et al.* 2006),

levantou-se a hipótese de que o SNP rs6683977 apresentou problemas de amplificação pela existência desse tipo de alelo, o que gerou a discordância observada entre os métodos.

Para confirmar a possibilidade de existência de variantes que levassem à ocorrência de alelos nulos nesse SNP, os dados de sequenciamento direcionado foram reanalisados para a identificação de possíveis mutações nos sítios de anelamento dos *primers* utilizados no ensaio do BeadXpress. A análise revelou que a sequência de anelamento dos *primers* incluem na quinta posição de sua extensão de 27 bases um sítio polimórfico (rs782686). O sítio possui alelos A e G, porém a análise do arquivo de sequencias dos *primers* fornecido pela fabricante do kit VeraCode indicou que os *primers* complementares dessa região traziam apenas a base T nessa posição, o que permitia o pareamento apenas com o alelo A. A análise cuidadosa das sequências geradas pelo método **TargetSeq**, revelou que os haplótipos que possuem o alelo C na posição polimórfica do rs6683977, o qual não foi satisfatoriamente detectado pelo ensaio do BeadXpress, são justamente os haplótipos que possuem o alelo G na posição do rs782686. Uma forte evidência desse problema foi a genotipagem nula pelo BeadXpress do indivíduo SH247 (da etnia *Machiguenga*, nativo peruano), o único que é homozigoto para o alelo C na posição do rs6683977 (dado confirmado pelo sequenciamento direcionado e pelo ensaio TaqMan) e que portanto é também homozigoto para o alelo G na posição do rs782686.

Apesar do problema de alelo nulo encontrado no SNP rs6683997, alguns outros marcadores foram validados, indicando que os dados do BeadXpress para as demais variantes são confiáveis. Porém, o sistema BeadXpress se mostrou bastante inconsistente em vários aspectos, especialmente em razão de constantes falhas mecânicas, revelando a fragilidade dessa metodologia. Não é de se surpreender que a plataforma tenha sido descontinuada e já não é mais oferecida pela fabricante.

A influência da ancestralidade na diversidade genética das populações

Os resultados encontrados para os marcadores informativos de ancestralidade revelaram proporções de ancestralidade esperadas, bem como alguns padrões de miscigenação não evidentes. A hipótese testada de que as populações nativas dos países latino-americanos seriam essencialmente de ancestralidade nativo-americana não foi corroborada para todas as populações nativas estudadas (Figura 4). De forma mais específica, as populações brasileiras de origem indígena apresentaram proporções

expressivas de ancestralidade europeia (TUP=27%, GUA=23%) e africana (TUP=23%, GUA=16%) (Tabela 3). Esse resultado não seria esperado caso as populações fossem isoladas e ocorresse pouco fluxo gênico com indivíduos de outras origens. Porém, esse isolamento não é completo e a crescente chegada de europeus e africanos às regiões hoje habitadas pelos índios *Tupiniquins* e *Guaranis*, no estado do Espírito Santo, pode justificar o resultado encontrado. De toda forma, apesar da influência europeia e africana, passada ou atual, e da possível ocorrência de miscigenação, a ancestralidade predominente observada foi a nativo-americana o que corrobora com a identidade cultural indígena desses grupos e sua autodeterminação como povos nativos do Brasil.

Por outro lado, a redução ou pouca participação da ancestralidade nativo-americana nas populações miscigenadas brasileiras fica evidente pela análise das populações do estado de MG, em que a ancestralidade nativa representa apenas 15% em média (Tabela 3). Esses resultados são concordantes com o que se conhece da estrutura genetico-populacional da população brasileira, em que a ancestralidade nativa tem pouca representatividade em certos centros, em especial pela redução da contribuição dessa ancestralidade ao longo dos séculos no processo de formação das populações brasileiras atuais (KEHDY *et al.* 2015).

Os resultados das análises de heterozigosidade realizadas com os marcadores dos genes *PDE4B* e *MYT1L* do banco de dados **BeadXpress** podem ser interpretados em consonância com as proporções de ancestralidades inferidas nas populações estudadas (Tabela 5). As populações nativo-americanas, em razão de sua história evolutiva essencialmente marcada por eventos de redução populacional (*bottleneck* e efeito fundador) e isolamento, possuem menores níveis de diversidade genética em ambos genes, especialmente no gene *MYT1L*, quando comparadas às populações europeia e africana. Nesse sentido, os resultados das inferências de ancestralidade contribuem para explicar os resultados de heterozigosidade observados. A influência das maiores proporções de ancestralidade europeia e africana nas populações brasileiras indicam que o aumento da diversidade genética observada nos genes *PDE4B* e *MYT1L* nas populações brasileiras foi devido ao processo de miscigenação ocorrido nos últimos séculos (Figura 4). Da mesma forma, os valores de F_{ST} par-a-par (Tabela 6) refletem índices de diferenciação populacional que podem ser explicados pelos processos de miscigenação ocorridos nas populações brasileiras. Nota-se que os índices de diferenciação das populações brasileiras para com as populações europeias e africanas

são menores do que os índices de diferenciação das populações peruanas para com as populações europeias e africanas. Esse resultado é também evidente na análise de PCA representada na Figura 6, em que as amostras de populações com maior proporção de miscigenação estão dispostas entre as populações de nativos, europeus e africanos. Esses resultados corroboram o que já vinha sendo descrito para as populações brasileiras em outros estudos realizados (KEHDY *et al.* 2015, ADHIKARI *et al.* 2017).

Com relação aos padrões de diversidade haplotípica dos genes *PDE4B* e *MYT1L*, os resultados revelados pelo banco de dados **BeadXpress** também podem ser discutidos de acordo com o grau de miscigenação das populações estudadas. Observa-se no gráfico de distribuição dos haplótipos do gene *MYT1L* (Figura 7b) certa gradação de distribuição dos haplótipos e suas frequências considerando as populações nativas peruanas, nativas brasileiras, miscigenados brasileiros e por fim o padrão visto na população de origem europeia. Há também haplótipos das populações africanas encontrados entre os haplótipos das populações brasileiras, o que é esperado em razão da contribuição dos africanos para formação das populações do Brasil (KEHDY *et al.* 2015, ADHIKARI *et al.* 2017). Os haplótipos do gene *PDE4B* e as frequências com que são observados nas populações estudadas indicam que há estruturação continental, pois as populações peruanas e brasileiras, incluindo as que possuem alto grau de miscigenação, se diferenciam das populações europeia, africana e asiáticas, sendo essas também bastante diferenciadas entre si (Figura 7a). A diversidade haplotípica observada indica que não apenas uma variante ou mutação é responsável pela diferenciação das populações predominantemente nativas, mas que há grande diversidade na frequência dos haplotipos como um todo. Assim, há grupos de variantes inter-relacionadas com frequências diferenciadas quando são comparadas as populações nativas e as demais populações.

A relação entre ancestralidade nativo-americana e a recidiva de LLA

Frequências alélicas e estruturação

O cálculo das frequências dos alelos das variantes de interesse para a recidiva de LLA indicou que os alelos de risco de todos os quatro SNPs do gene *PDE4B* apresentaram altas frequências nos nativo-americanos (Tabela 4), contrastando muitas vezes com as frequências reportadas em outras populações do globo. Esses resultados sugerem que, de fato, essa região do cromossomo 1 pode estar relacionada a alguma resposta diferenciada de indivíduos nativo-americanos submetidos a tratamentos que

utilizam inibidores do *PDE4B*, ou fármacos que atuam nas vias metabólicas em que esse gene atua, como por exemplo fármacos antiinflamatórios (como a prednisona) e antimetabólitos (como o metotrexato). Por sua vez, para o alelo de risco do gene *MYT1L* (rs17039396.A) as frequências alélicas nas populações nativo-americanas são maiores quando são analisadas as populações peruanas (com exceção da população *Quechua*). Já as demais populações de nativos mexicanas e brasileiras apresentam valores menores. É interessante notar que a população do leste-asiático (EAS) do 1000GP apresenta valores de frequência um pouco maiores do alelo de risco, próximo aos valores das populações de nativos peruanos. Esse resultado é bastante interessante, pois considerando que há um estudo associando o SNP *MYT1L* (rs17039396) ao prognóstico favorável de câncer gástrico em chineses (ZHANG *et al.* 2013), há indícios da existência de possíveis mecanismos moleculares que associam esse gene e SNP com fenótipos relacionados ao câncer. Portanto, a análise conjunta das proporções de ancestralidade nativo-americana populacional e das frequências dos alelos de risco dos SNPs de interesse dos genes *PDE4B* e *MYT1L* permitem corroborar a hipótese de que as frequências desses alelos têm relação direta com a proporção de ancestralidade nativa (Figura 5).

Cabe ressaltar que os dados gerados pelo método de sequenciamento permitiram a identificação de muitas variantes novas, ainda não reportadas em bancos de dados públicos (como dbSNP e Projeto 1000 Genomas). Das 104 variantes encontradas no gene *PDE4B*, 17 não têm registro nos bancos de dados mais conhecidos (tais como dbSNP) (Tabela Apêndice 3). Já no gene *MYT1L*, das 81 variantes, 15 não estão presentes nos bancos públicos (Tabela Apêndice 2). Como esperado, a maioria dessas variantes são exclusivas e quando não são *singletons* ainda assim ocorrem em baixa frequência nas populações estudadas. O crescimento populacional recente e acelerado tem um efeito conhecido sobre o espectro de frequência das variantes neutras de uma população, causando o aumento do número observado de variantes raras (KEINAN & CLARK 2012). Dessa forma, as variantes raras encontradas podem ser resultado do histórico demográfico dos ancestrais dessas populações, que passaram por vários eventos de crescimento populacional acelerado (GRAVEL *et al.* 2011). Há evidências de expansões ao longo da história, em especial para os nativos-americanos cujos ancestrais ocuparam áreas ao longo da Cordilheira dos Andes (BOLNICK *et al.* 2016). Os resultados encontrados corroboram a expectativa, já que as populações peruanas andinas (*Quechua* e *Aymara*) possuem grande parte das variantes novas e de baixa frequência encontradas. Para o gene *PDE4B* oito das 17 mutações novas encontradas

foram identificadas nessas populações, enquanto que para o gene *MYT1L* são cinco mutações observadas dentre as 16 mutações ainda não reportadas.

Os índices de diferenciação entre os grupos para cada variante do banco de dados **BeadXpress**, indicados pelos valores de F_{CT} (Tabela 7), permitem que sejam destacadas as variantes com maior diferenciação entre os grandes grupos; no caso específico desse estudo, entre os grupos de nativos (peruanos e brasileiros), africanos e europeus (do 1000GP). As variantes com maiores diferenciações analisadas são todas intrônicas e não apresentam evidência imediata de impacto sobre o fenótipo molecular e nem há estudos que reportam importância dessas variantes. Porém deve-se ressaltar que esses locos podem ter funções ainda desconhecidas e/ou estarem em desequilíbrio de ligação com variantes de importância.

Com relação ao banco de dados **TargetSeq**, a análise da estatística F_{CT} por variante para cada par de populações (Tabela 8) mostra que os SNPs já reportados como associados à recidiva de LLA apresentam valores elevados de diferenciação nas populações nativo-americanas, especialmente considerando os SNPs do gene *PDE4B* e quando as comparações são feitas com a população europeia (>0,15). Outras variantes no gene *PDE4B* também apresentaram altos valores de F_{CT} (Tabela Apêndice 4), com níveis de significância abaixo do esperado ao acaso, o que indica que a região em análise tende a ser diferenciada em nativo-americanos. Já o índice de F_{CT} do SNP de interesse do gene *MYT1L* mostrou valor de diferenciação alto quando a comparação é realizada com a população africana (Tabela 9), o que pode ser explicado pelo fato do alelo de risco rs17039396.A ser muito raro nessa população (A:0,01). De fato, em geral, valores menores de F_{CT} dos SNPs desse gene (quando comparados aos valores dos SNPs de *PDE4B*) foram observados (Tabela Apêndice 5). Dentre as variantes do gene *MYT1L* analisadas, sete tiveram valores altos de diferenciação (>0,12) quando a comparação é feita com a população europeia.. Portanto, os resultados obtidos apresentam variantes que indicam diferenciação da população nativa com relação às demais, especialmente com a população europeia, a qual é a base para a realização de testes clínicos e estudos diversos, incluindo os voltados ao desenvolvimento de fármacos (POPEJOY & FULLERTON 2016, SIRUGO *et al.* 2019).

Desequilíbrio de ligação (DL) e diversidade haplotípica

Os resultados das análises de DL para os bancos de dados **BeadXpress** e **TaqMan+2.5M** mostraram poucas variantes ligadas aos SNPs de interesse (*PDE4B*-

rs6683977 e *MYT1L*-rs17039396), especialmente nas populações nativo-americanas (Tabela 9). Não há registros na literatura de que os SNPs que apresentaram valores de correlação (r^2) maiores que 0,8 nos testes de DL para algumas populações de nativo-americanos - *PDE4B*- rs519044 e -rs6683604 - tenham qualquer consequência biomédica relevante. Esse resultado indica que a associação observada entre os SNPs de interesse e a condição de recidiva de LLA em nativo-americanos provavelmente não está diretamente relacionada aos SNPs testados nesses bancos de dados.

No caso do banco de dados **TargetSeq**, as análises de DL não indicaram haver SNPs em DL (r^2>0,8) com o *MYT1L*-rs17039396 em nenhuma das populações. Por outro lado, muitas variantes foram encontradas em DL com o *PDE4B*-rs6683977 em várias populações, incluindo as populações do 1000GP. Porém, nota-se que não há DL na população de africanos do 1000GP. Isso permite compreender, em parte, a alta frequência do alelo de risco *PDE4B*-rs6683977.G (alelo ancestral) em africanos e a falta de sinal de associação com a recidiva da LLA nessa população, o que sugere que o SNP *PDE4B*-rs6683977 não é o responsável direto pelo fenótipo observado. Uma forte evidência dos resultados apresentados aqui é de que esse SNP pode estar em DL em populações de nativo-americanos com alguma variante (ou mesmo algum grupo de variantes) que leva à associação com a recidiva de LLA. Deve-se destacar dois SNPs anteriormente relacionados à recidiva da LLA (rs546784 e rs641262), os quais estão em DL com o SNP *PDE4B*-rs6683977 e também entre si quando analisadas as populações de nativos peruanos, mexicanos e europeus (Figura 9 e Tabela 9). A frequência dos haplótipos que possuem os alelos de risco para esses SNPs alcança valores altos nas populações nativo-americanas em comparação às demais populações do 1000GP (Tabela 10).

Aprofundando as análises, a avaliação dos padrões de DL com os dados do **TargetSeq** mostram blocos de ligação com maior suporte nas populações com menores índices de diversidade como, por exemplo, nas populações peruanas (Figuras 8 e 9). Para o gene *PDE4B* há evidência de três blocos formados, com destaque para o maior deles, com 27 variantes na população peruana e o qual engloba as 14 variantes do bloco identificado na população mexicana. Esse bloco também apresenta importância por incluir na população mexicana duas e na população peruana as quatro variantes reportadas em Yang e colaboradores (2012) como relacionadas à recidiva de LLA. Cabe aqui discutir que há visível diferença entre as populações peruanas e mexicanas no que

diz respeito a esses blocos de ligação. Essa é mais uma evidência de que há certa heterogeneidade genética entre populações nativas, o que muitas vezes indica que extrapolar conclusões obtidas para uma população de certa origem nativa para populações nativo-americanas de outras origens não seria adequado. Em estudo recente, Naranjo e colaboradores (2018) notaram que há diferenças significativas na variabilidade de genes do complexo *CYP* quando populações de origem nativo-americana do norte e do sul do continente são comparadas. Portanto, no âmbito dos estudos farmacogenéticos, como o discutido no presente capítulo, deve-se levar em conta a diversidade genética entre diferentes populações, mesmo entre as que possuem a mesma ancestralidade predominante.

Os blocos de ligação do gene *MYT1L* apresentam maior suporte e, diferentemente do observado para o gene *PDE4B*, praticamente são concordantes entre as populações peruana e mexicana (Figura 8). O SNP de interesse rs17039396 não está em DL com os SNPs da região em quase todas as populações, exceto nos peruanos, onde se notam valores de r^2 um pouco maiores do que nas demais populações. Porém, esses valores de correlação são ainda pouco expressivos quando são comparados aos valores de r^2 dos demais SNPs da região (Figura 8). Dessa forma, o sinal de associação entre o *MYT1L*-rs17039396 e a recidiva de LLA observado por Yang e colaboradores (2011) pode estar relacionado a alguma outra variante em DL que não foi identificada entre os SNPs analisados nos bancos de dados desse estudo. Para confirmar será preciso aumentar a área sequenciada do gene *MYT1L*, ou mesmo buscar em outras regiões genômicas possíveis explicações para a associação observada.

Evidências de função regulatória

Em particular para o gene *PDE4B,* a visualização dos dados genômicos disponibilizadas pelo UCSC *Genome Browser* permitiu identificar que há diversos indícios de que essa região desempenhe um papel regulatório (Figura 10). Especialmente ao longo de 1Kb entre as variantes rs524770 e rs6683977 (Figura 11). Por outro lado, esse padrão já não é tão evidente para o gene *MYT1L* (Figura 12). É notável observar que para o gene *PDE4B* os sinais são robustos tanto para os padrões de metilação (H3K4Me1) quanto para acetilação de histonas (H3K27Ac). Também há sinais de sítios de ligação de DNAse, de ligação de diversos fatores de transcrição e de acentuadores, preditos por análises *in silico* (ChromHMM). É importante ressaltar que os sinais são em linhagens celulares diretamente relacionadas à LLA, já que se tratam

de linfócitos B sadios (GM12878) e células em condição de leucemia mieloide crônica (K562). Portanto, em uma análise inicial, são várias as evidências independentes de que essa região está envolvida com funções regulatórias.

Em razão dos sinais revelados pelo UCSC *Genome Browser*, os bancos de dados ENCODE e RegulomeDB foram consultados diretamente para o detalhamento dessas possíveis atividades regulatórias considerando as posições dos SNPs mais diferenciados. A análise dos resultados obtidos pela consulta ao ENCODE reforçaram que há fortes indícios de que elementos funcionais atuam nas regiões sequenciadas em tipos celulares relacionados ao sistema de desenvolvimento da LLA (Tabela 11). Além disso, os resultados da consulta ao banco de dados Regulome mostram que algumas variantes estão em posições onde há múltiplas evidências de funções regulatórias, como é o caso por exemplo da variante rs495477, a qual tem *score* 2b - ligação de fator de transcrição + qualquer motivo + DNase *Footprint* + pico de DNAse - (Tabela 12) e que está em desequilíbrio de ligação com o rs6683977 em todas as populações de nativos peruanos e mexicanos (Tabela 10). Dentre os resultados, observam-se sítios de ligação de DNAse, que indicam a disponibilidade dessas regiões à transcrição. Como há fortes evidências da exposição dessas regiões à maquinaria de transcrição, os perfis de expressão desses genes, os quais são influenciados por acentuadores (*enhancers)* ou silenciadores, podem ser afetados de forma diferenciada tendo em vista os diversos haplótipos e variantes dessas regiões. Não se pode descartar também a possibilidade de haver desequilíbrio de ligação com demais variantes fora da região sequenciada, tais como regiões anteriores ou posteriores, em outros íntrons ou éxons, e que possam ter efeito direto sobre o fenótipo.

Esses resultados corroboram o que foi encontrado por Schmitt e colaboradores (2016) em um estudo amplo que buscou identificar em âmbito genômico as regiões dos cromossomos humanos que apresentassem níveis particularmente altos de interações da cromatina e as possíveis especificidades de acordo com diferentes tecidos. Os autores deram destaque a algumas regiões e entre elas está a região do *PDE4B*, que inclui os SNPs rs6683977 e rs546784*,* a qual foi identificada como de alta atividade regulatória em linhagens celulares de linfoblastóide B (GM12878). Os resultados mostraram que essa região é particularmente composta por diversas unidades de *Fire (frequently interacting regions)*, que seriam regiões enriquecidas por acentuadores (*enhancers)*. Por

ser uma região que apresenta grupos de *Fires,* foi classificada como *super-FIRE*, o que ressalta a importância dessa porção para a expressão do *PDE4B*.

Ademais, vale destacar e discutir recente estudo conduzido por Liu e colaboradores (2017) em que foram analisados os efeitos regulatórios de variantes relacionadas a diversos tipos de câncer reportadas em uma ampla gama de estudos de associação. Dentre as milhares variantes analisadas, diversas estão localizadas na mesma porção intrônica do gene *PDE4B* que os SNPs anteriormente reportados como associados à recidiva de LLA (rs546784, rs524770, rs6683977 e rs641262). Apesar dos resultados para o SNP rs6683977 terem sido inconclusivos, três variantes foram classificadas como 'elementos regulatórios positivos' (rs494735, rs502958, and rs12142375), já que os resultados mostraram que levavam à sobreexpressão gênica em ensaios *in vitro*. Foi também constatado que esses SNPs estão em DL com o *PDE4B*-rs546784 em populações chinesas do 1000GP. Assim, usando o rs546784 para inferir os genótipos, o mesmo estudo aprofundou as análises de regulação da expressão gênica pelo SNP rs12142375 e constatou que há diferenças na expressão gênica de acordo com o alelo (referência ou alternativo) testado. A conclusão foi de que há evidências de que a regulação da expressão gênica do *PDE4B* está associada ao risco de LLA.

Apesar de não abordar diretamente o efeito sobre o tratamento da LLA, mas sim o risco de desenvolvimento da doença, os resultados encontrados nesse estudo podem ser extrapolados para indicar possíveis relações com a resposta ao tratamento já que foi mostrado o papel de uma variante na região intrônica do *PDE4B* sobre a expressão desse gene. Tendo em vista as evidências do alto grau de DL entre a variante analisada (rs12142375) e a variante *PDE4B*-rs546784, a qual está incluída no banco de dados **TargetSeq**, deve-se destacar as diferenças de frequência dos haplótipos que incluem o alelo *PDE4B*-rs546784.T nas populações aqui estudadas. Segundo Liu e colaboradores (2017) o alelo *PDE4B*-rs546784.T em homozigose representa o fenótipo molecular de sobreexpressão gênica do *PDE4B*. Interessante notar que, pelas análises dos dados de **TargetSeq,** os haplótipos que incluem esse alelo são muito mais frequentes nas populações de nativos peruanos do que na população europeia (Tabela 10), e além disso, em razão do DL, também incluem o alelo ancestral G do SNP rs6683977 e demais alelos de SNPs associados à ancestralidade nativa e à recidiva da LLA (Tabela 10).

Em especial para o gene *PDE4B* os resultados de diversidade genética em regiões intrônicas reforçam estudos anteriores que indicam que a ancestralidade pode estar associada aos padrões de expressão desse gene (YANG *et al.* 2011, 2012, TSUNODA *et al.* 2012). Vale destacar dois estudos que relacionaram variantes dessa região com a ocorrência de esquizofrenia em indivíduos da população japonesa (FATEMI *et al.* 2008; NUMATA *et al.* 2008). Um desses estudos encontrou associação entre a variante rs498448 (a qual apresentou alta diferenciação das populações nativo-americanas em relação às populações europeia (F_{CT}=0,14) e africana (F_{CT}=0,24) em nossos resultados – Figura 10, 11 e Tabela Apêndice 4) com a esquizofrenia nessa população. Em ambos estudos os autores ressaltam que variações no *PDE4B* podem ter um papel importante na etiologia da esquizofrenia.

Com relação aos resultados obtidos para o gene *MYT1L*, as evidências de que os SNPs com alta diferenciação estão em regiões que podem exercer algum papel regulatório são menores. A análise dos sinais revelados pelo UCSC *Genome Browser* não indica evidências robustas de função regulatória na região em questão (Figura 12). Dentre os SNPs identificados, apenas um deles (rs79018269) apresentou *score* 4 no RegulomeDB, o que indica sinais de ligação de fatores de transcrição e pico de exposição à DNAse. Enfim, há menos evidências de que essa região do gene *MYT1L* tenha relevante importância regulatória e seja mais diferenciada em nativos. De toda forma, além do detalhamento da diversidade genética e das possíveis funções regulatórias de outras regiões desse gene, também é necessário que o papel molecular do *MYT1L* seja melhor caracterizado no contexto da LLA. Especialmente porque há evidências de que esse gene possa exercer funções de relevância em neoplasias, tais como no prognóstico do adenocarcinoma da cárdia (ZHANG *et al.* 2013) e na patogênese da LLA, como indicado por estudo recente (TOMAR *et al.* 2019).

Portanto, estudos mais detalhados sobre a caracterização funcional de variantes e haplótipos ainda não reportados, tanto para o gene *PDE4B* quanto para o gene *MYT1L*, podem indicar diferenças no fenótipo molecular desses genes. Esses resultados certamente contribuirão para o avanço no conhecimento acerca dos mecanismos moleculares que levam à diferenciação da resposta ao tratamento em pacientes de LLA e a influência da ancestralidade nesse processo (KAROL *et al.* 2017, LEE & YANG 2017).

Implicações da diversidade e estrutura genética dos genes PDE4B e MYT1L para os aspectos clínicos da LLA

Demais estudos, assim como os resultados apresentados nesse trabalho, ressaltam a importância de se considerar a diversidade genética humana global para que sejam alcançados avanços nos tratamentos de doenças (POPEJOY & FULLERTON 2016, SIRUGO *et al.* 2019). Vários esforços vêm sendo realizados para melhor compreender os mecanismos que levam à susceptibilidade ao câncer pediátrico, mas estudos focados na análise dos fatores que levam à metástase e os relacionados à recidiva ainda são relativamente pouco explorados (SWEET-CORDERO & BIEGEL 2019). Como apresentado, a diversidade genética dos genes associados à recidiva de LLA varia de acordo com a ancestralidade populacional, que se mostra um fator importante a ser observado nas questões médicas que envolvem o tratamento dessa doença (LEE & YANG 2017). As evidências de que a diversidade genética e o perfil de variantes relacionadas à ancestralidade influenciam a recidiva da doença são significativas e implicam atenção (YANG *et al.* 2011, BHATIA *et al.* 2012, KAROL *et al.* 2017, LEE & YANG 2017). Essa cautela deve ocorrer especialmente nos centros de tratamento onde a estratificação do risco não ocorre de forma adequada ou que, mesmo ocorrendo, não implica em ajustes práticos ao tratamento. Como é o caso de países da América Latina (NAVARRETE *et al.* 2014), onde as populações possuem grande proporção de ancestralidade nativo-americana.

Os resultados de diferenciação genética apresentados nesse estudo revelam que o conhecimento detalhado sobre as implicações dessas variações no tratamento da LLA podem contribuir para o ajuste dos protocolos aplicados em populações com maior proporção de ancestralidade nativo-americana (KAROL *et al*. 2017, LEE & YANG 2017). Estudos funcionais podem revelar como essa diversidade influencia na resposta aos fármacos e aos protocolos aplicados (BHATLA *et al.* 2014). A necessidade desse tipo de validação é crescente e cada vez mais se faz necessária tendo em vista a grande quantidade de dados gerados em razão dos avanços nas metodologias de sequenciamento e genotipagem (SWEET-CORDERO & BIEGEL 2019). Especificamente para o caso do gene *PDE4B*, evidências já foram reportadas de que a ancestralidade nativa pode influenciar na resistência aos glicocorticóides e outros fármacos (YANG *et al.* 2011, 2012). Portanto, os padrões de variação apresentados nessa tese podem direcionar as análises e validações que venham a ser realizadas.

Ademais, especial atenção deve ser dada aos resultados que mostram haver diferenças na estrutura e diversidade genética entre as populações nativo-americanas (peruanas, mexicanas e brasileiras). Assim como já visto para outros alvos farmacogenéticos (genes do complexo CYP - NARANJO *et al.* 2018), a heterogeneidade genética existente entre as populações nativas pode também revelar diferentes fenótipos relacionados ao tratamento da LLA.

Os resultados das buscas pelos tratamentos aplicados na América Latina foram insuficientes para que se pudesse concluir acerca dos protocolos mais utilizados nos países da região e se há inclusão de fases de intensificação tardia após a remissão. Os resultados foram específicos para alguns países e algumas instituições e, por serem apenas diretrizes de órgãos governamentais ou associações médicas, indicam que pode não haver concordância entre os tratamentos adotados nos diversos centros de oncologia dos países em questão. Ademais, o protocolo BFM possui diversas versões e adaptações são feitas constantemente para adequações locais. Sendo assim, a fase de intensificação tardia ministrada pode incluir diversos fármacos, diferentes dosagens e tempo de tratamento variável, de acordo com as disponibilidades dos centros. Portanto, os resultados encontrados não garantem que a fase de intensificação tardia seja aplicada em todos os países da America Latina, bem como, naqueles em que se obteve dados, não se pode afirmar que os protocolos aplicados em todos os centros de tratamento de LLA seguem alguma versão que inclui a fase de intensificação tardia como a indicada em Yang e colaboradores (2011).

De toda forma, é preciso destacar que não é importante apenas acrescentar a fase da intensificação tardia nos tratamentos de LLA. Associado a essa medida, o uso de fármacos que levam à inibição do *PDE4B* pode ser importante para o aumento da sensibilidade das células a glicocorticoides e a outros fármacos (como o metotrexato) usados no tratamento da LLA (MEYERS *et al.* 2007). A relação da ancestralidade com a resposta aos fármacos inibidores do *PDE4B* precisará ser levada em consideração, mas possivelmente revelará importantes peculiaridades de acordo com o perfil de variantes desse gene em cada população. Como há evidências da influência dessas variantes na atividade molecular da fosfodiesterase (YANG *et al.* 2012, LIU *et al.* 2017), isso pode indicar possível influência na eficácia dos inibidores do gene *PDE4B*, os quais possuem atividade antiangiogênica em células tumorais (COONEY & AGUIAR 2016).

Mesmo que estudos mostrem que nos países latino-americanos sejam necessários avanços além da inclusão da fase de intensificação nos tratamentos aplicados (JAIME-PÉREZ *et al.* 2016), a garantia da aplicação dessa fase é ainda mais importante tendo em vista dificuldades socioeconômicas específicas desses locais. Estudos mostram que há alta taxa de abandono ao longo do tratamento que estão relacionadas ao nível de desenvolvimento do país (RODRIGUEZ-GALINDO *et al.* 2013, FRIEDERICH *et al.* 2015). Para o caso específico das populações com alta ancestralidade nativo-americana (>10%) essa realidade é expressiva (ROSSEL *et al.* 2015, FRIEDERICH *et al.* 2016). Além disso, em razão de dificuldades econômicas, não é difícil conceber que alternativas que visem reduzir os custos com o tratamento sejam adotadas, tais como a eliminação de fases do tratamento que impliquem o uso intensivo de fármacos (PUI 2013, VORA *et al.* 2013, SCHRAPPE *et al.* 2016).

A adequação do tratamento dos diferentes tipos de câncer de acordo com o perfil genético-molecular dos pacientes, conhecida como oncologia de precisão, já é realidade em grandes centros de tratamento em países de maior desenvolvimento e é uma tendência cada vez mais abordada em periódicos científicos especializados (LEE & YANG 2017, DUBOIS *et al.* 2019). Infelizmente, as limitações econômicas enfrentadas por países de menor índice de desenvolvimento são obstáculo para a aplicação dos avanços alcançados na área e também fazem com que os alvos das pesquisas e testes clínicos sejam populações com ancestralidade predominantemente europeia. Porém, para o caso específico da recidiva da LLA, pequenos ajustes em protocolos já bastante difundidos (baseados no protocolo BFM) podem fazer com que melhores resultados sejam alcançados, mesmo em países com menor índice de desenvolvimento socioeconômico. Como mostra o resultado da análise dos protocolos de tratamento aplicados à LLA na América Latina (Tabela 13), apesar da utilização do protocolo BFM e suas variações já ocorrer em alguns países, é preciso que os responsáveis pelos tratamentos ministrados tenham consciência dos resultados reportados quando da aplicação e quando da não-aplicação da fase de intensificação tardia para pacientes com maior proporção (>10%) de ancestralidade nativo-americana. Esse conhecimento prévio pode contribuir para a interpretação de resultados, como os de doença residual mínima após a fase de remissão, os quais são tomados como indicadores para a adequação do tratamento.

Portanto, a discrepância das taxas de sobrevivência dos acometidos pela LLA em países de baixo e médio IDH em relação aos países desenvolvidos mostra que ainda há reais possibilidades de melhora nos tratamentos empregados nos países menos desenvolvidos. Além disso, também revela a existência de fatores biológicos e não-biológicos (como questões socioeconômicas) que precisam ser levados em conta para o alcance de melhores resultados. Estudos como o apresentado aqui buscam não apenas progressos para a maior compreensão dos mecanismos moleculares que envolvem a resposta ao tratamento da LLA, mas também destacam a importância do papel dos gestores, que têm poder de decisão em políticas de saúde pública, para a melhora dos indicadores gerais de tratamento da LLA em países negligenciados e/ou com baixo IDH.

REFERÊNCIAS

1000 GENOMES PROJECT CONSORTIUM *et al*. An integrated map of genetic variation from 1,092 human genomes. **Nature**, v. 491, n. 7422, p. 56, 2012.

ADHIKARI, Kaustubh *et al*. The genetic diversity of the Americas. **Annual review of genomics and human genetics**, v. 18, p. 277-296, 2017.

ALEXANDER, David H.; LANGE, Kenneth. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. **BMC bioinformatics**, v. 12, n. 1, p. 246, 2011.

ANTILLÓN, Federico G. *et al*. The treatment of childhood acute lymphoblastic leukemia in Guatemala: Biologic features, treatment hurdles, and results. **Cancer**, v. 123, n. 3, p. 436-448, 2017.

ARIËS, Ingrid M. *et al*. Towards personalized therapy in pediatric acute lymphoblastic leukemia: RAS mutations and prednisolone resistance. **Haematologica**, v. 100, n. 4, p. e132, 2015.

BHATIA, Smita *et al*. Nonadherence to oral mercaptopurine and risk of relapse in Hispanic and non-Hispanic white children with acute lymphoblastic leukemia: a report from the children's oncology group. **Journal of Clinical Oncology**, v. 30, n. 17, p. 2094, 2012.

BHATLA, Teena *et al*. The biology of relapsed acute lymphoblastic leukemia: opportunities for therapeutic interventions. **Journal of pediatric hematology/oncology**, v. 36, n. 6, p. 413, 2014.

BARBIERI, Chiara *et al*. Between A ndes and A mazon: The genetic profile of the A rawak-speaking Y anesha. **American journal of physical anthropology**, v. 155, n. 4, p. 600-609, 2014.

BARRETT, Jeffrey C. *et al*. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, v. 21, n. 2, p. 263-265, 2004.

BATLEVI, Connie Lee *et al*. Novel immunotherapies in lymphoid malignancies. **Nature reviews Clinical oncology**, v. 13, n. 1, p. 25, 2016.

BAYER, Ronald; GALEA, Sandro. Public health in the precision-medicine era. **New England Journal of Medicine**, v. 373, n. 6, p. 499-501, 2015.

BERMEJO, Justo Lorenzo *et al*. Subtypes of Native American ancestry and leading causes of death: Mapuche ancestry-specific associations with gallbladder cancer risk in Chile. **PLoS genetics**, v. 13, n. 5, p. e1006756, 2017.

BHOJWANI, Deepa; PUI, Ching-Hon. Relapsed childhood acute lymphoblastic leukaemia. **The lancet oncology**, v. 14, n. 6, p. e205-e217, 2013.

BIONDI, Andrea; SCRIDELI, Carlos Alberto; CAZZANIGA, Giovanni. Acute Lymphoblastic Leukemia. In: **Molecular Pathology in Clinical Practice**. Springer, Cham, 2016. p. 561-577.

BOLNICK, Deborah A. *et al*. Native American genomics and population histories. **Annual Review of Anthropology**, v. 45, p. 319-340, 2016.

BOYLE, Alan P. *et al*. Annotation of functional variation in personal genomes using RegulomeDB. **Genome research**, v. 22, n. 9, p. 1790-1797, 2012.

BRYC, Katarzyna *et al*. The genetic ancestry of african americans, latinos, and european Americans across the United States. **The American Journal of Human Genetics**, v. 96, n. 1, p. 37-53, 2015.

BURKE, Wylie *et al*. Extending the reach of public health genomics: what should be the agenda for public health in an era of genome-based and "personalized" medicine?. **Genetics in Medicine**, v. 12, n. 12, p. 785, 2010.

CARLSON, Christopher S. *et al*. Direct detection of null alleles in SNP genotyping data. **Human molecular genetics**, v. 15, n. 12, p. 1931-1937, 2006.

CEDEFES, Centro de Documentação Eloy Ferreira da Silva. Povos Indígenas Em Minas Gerais - Quem São? Disponível em: <http://www.cedefes.org.br/povos-indigenas-destaque/>. Acesso em 04 de abr. de 2019.

COONEY, Jeffrey D.; AGUIAR, Ricardo CT. Phosphodiesterase 4 inhibitors have wide-ranging activity in B-cell malignancies. **Blood, The Journal of the American Society of Hematology**, v. 128, n. 25, p. 2886-2890, 2016.

COOPER, Stacy L.; BROWN, Patrick A. Treatment of pediatric acute lymphoblastic leukemia. **Pediatric Clinics**, v. 62, n. 1, p. 61-73, 2015.

CSERVENKA, Anita; YARDLEY, Megan M.; RAY, Lara A. Pharmacogenetics of alcoholism treatment: Implications of ethnic diversity. **The American journal on addictions**, v. 26, n. 5, p. 516-525, 2017.

DING, Lili *et al*. African ancestry is associated with cluster-based childhood asthma subphenotypes. **BMC medical genomics**, v. 11, n. 1, p. 51, 2018.

DRAKE, Katherine A. *et al*. A genome-wide association study of bronchodilator response in Latinos implicates rare variants. **Journal of Allergy and Clinical Immunology**, v. 133, n. 2, p. 370-378. e15, 2014.

DUBOIS, Steven G. *et al*. Ushering in the next generation of precision trials for pediatric cancer. **Science**, v. 363, n. 6432, p. 1175-1181, 2019.

ELHAIK, Eran. Empirical distributions of FST from large-scale human polymorphism data. **PloS one**, v. 7, n. 11, p. e49837, 2012.

ENCODE PROJECT CONSORTIUM *et al*. An integrated encyclopedia of DNA elements in the human genome. **Nature**, v. 489, n. 7414, p. 57, 2012.

ERNST, Jason; KELLIS, Manolis. ChromHMM: automating chromatin-state discovery and characterization. **Nature methods**, v. 9, n. 3, p. 215, 2012.

EXCOFFIER, Laurent; LISCHER, Heidi EL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Molecular ecology resources**, v. 10, n. 3, p. 564-567, 2010.

FATEMI, S. Hossein *et al*. PDE4B polymorphisms and decreased PDE4B expression are associated with schizophrenia. **Schizophrenia research**, v. 101, n. 1-3, p. 36-49, 2008.

FIGUEROA, Jonine D. *et al*. Identification of a novel susceptibility locus at 13q34 and refinement of the 20p12. 2 region as a multi-signal locus associated with bladder cancer risk in individuals of European ancestry. **Human molecular genetics**, v. 25, n. 6, p. 1203-1214, 2016.

FLICEK, Paul *et al*. Ensembl 2014. **Nucleic acids research**, v. 42, n. D1, p. D749-D755, 2013.

FRIEDRICH, Paola *et al*. Magnitude of treatment abandonment in childhood cancer. **PloS one**, v. 10, n. 9, p. e0135230, 2015.

FRIEDRICH, Paola *et al*. Determinants of treatment abandonment in childhood cancer: results from a global survey. **PLoS One**, v. 11, n. 10, p. e0163090, 2016.

GABRIEL, Stacey B. *et al*. The structure of haplotype blocks in the human genome. **Science**, v. 296, n. 5576, p. 2225-2229, 2002.

GALANTER, Joshua M. *et al*. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. **Journal of Allergy and Clinical Immunology**, v. 134, n. 2, p. 295-305, 2014.

GIEMBYCZ, Mark A.; NEWTON, Robert. Potential mechanisms to explain how LABAs and PDE4 inhibitors enhance the clinical efficacy of glucocorticoids in inflammatory lung diseases. **F1000prime reports**, v. 7, 2015.

GONZÁLEZ BURCHARD, Esteban *et al*. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. **American journal of public health**, v. 95, n. 12, p. 2161-2168, 2005.

GRAVEL, Simon *et al*. Demographic history and rare allele sharing among human populations. **Proceedings of the National Academy of Sciences**, v. 108, n. 29, p. 11983-11988, 2011.

HARRISON, Christine J. Cytogenetics of paediatric and adolescent acute lymphoblastic leukaemia. **British journal of haematology**, v. 144, n. 2, p. 147-156, 2009.

HARTL, Daniel L.; CLARK, Andrew G. **Princípios de Genética de Populações-4**. Artmed Editora, 2010.

HERNANDEZ-PACHECO, Natalia *et al*. What ancestry can tell us about the genetic origins of inter-ethnic differences in asthma expression. **Current allergy and asthma reports**, v. 16, n. 8, p. 53, 2016.

HIJIYA, Nobuko; VAN DER SLUIS, Inge M. Asparaginase-associated toxicity in children with acute lymphoblastic leukemia. **Leukemia & lymphoma**, v. 57, n. 4, p. 748-757, 2016.

HUGHES-STAMM, Sheree *et al*. Initial Evaluation of A96-Plex Goldengate® Genotyping SNP Assay with Suboptimal and Whole Genome Amplified Samples. **Journal of Forensic Investigation**, 2013.

HUNGER, Stephen P.; MULLIGHAN, Charles G. Acute lymphoblastic leukemia in children. **New England Journal of Medicine**, v. 373, n. 16, p. 1541-1552, 2015.

INTERNATIONAL HAPMAP CONSORTIUM *et al*. The international HapMap project. **Nature**, v. 426, n. 6968, p. 789, 2003.

IYENGAR, Sudha K.; ELSTON, Robert C. The genetic basis of complex traits. In: **Linkage Disequilibrium and Association Mapping**. Humana Press, 2007. p. 71-84.

JABBOUR, Elias *et al*. New insights into the pathophysiology and therapy of adult acute lymphoblastic leukemia. **Cancer**, v. 121, n. 15, p. 2517-2528, 2015.

JACKSON, Rosanna K.; IRVING, Julie AE; VEAL, Gareth J. Personalization of dexamethasone therapy in childhood acute lymphoblastic leukaemia. **British journal of haematology**, v. 173, n. 1, p. 13-24, 2016.

JAIME-PÉREZ, José C. *et al*. Results of treating childhood acute lymphoblastic leukemia in a low-middle income country: 10 year experience in Northeast Mexico. **Archives of medical research**, v. 47, n. 8, p. 668-676, 2016.

JAIME-PÉREZ, José C. *et al*. Relapse of childhood acute lymphoblastic leukemia and outcomes at a reference center in Latin America: organomegaly at diagnosis is a significant clinical predictor. **Hematology**, v. 23, n. 1, p. 1-9, 2018.

JIMÉNEZ-HERNÁNDEZ, Elva *et al*. Survival of Mexican children with acute lymphoblastic leukaemia under treatment with the protocol from the Dana-Farber Cancer Institute 00-01. **BioMed research international**, v. 2015, 2015.

KAROL, Seth E. *et al*. Genetics of ancestry-specific risk for relapse in acute lymphoblastic leukemia. **Leukemia**, v. 31, n. 6, p. 1325, 2017.

KAUR, Harsimar B. *et al*. Association of tumor-infiltrating T-cell density with molecular subtype, racial ancestry and clinical outcomes in prostate cancer. **Modern Pathology**, v. 31, n. 10, p. 1539, 2018.

KEINAN, Alon; CLARK, Andrew G. Recent explosive human population growth has resulted in an excess of rare genetic variants. **Science**, v. 336, n. 6082, p. 740-743, 2012.

KEHDY, Fernanda SG *et al*. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. **Proceedings of the National Academy of Sciences**, v. 112, n. 28, p. 8696-8701, 2015.

KHOURY, Muin J.; IADEMARCO, Michael F.; RILEY, William T. Precision public health for the era of precision medicine. **American journal of preventive medicine**, v. 50, n. 3, p. 398, 2016.

KIM, J. G.; HUDSON, L. D. Novel member of the zinc finger superfamily: A C2-HC finger that recognizes a glia-specific gene. **Molecular and Cellular Biology**, v. 12, n. 12, p. 5632-5639, 1992.

KIM, Sang-Woo; RAI, Deepak; AGUIAR, Ricardo CT. Gene set enrichment analysis unveils the mechanism for the phosphodiesterase 4B control of glucocorticoid response in B-cell lymphoma. **Clinical cancer research**, v. 17, n. 21, p. 6723-6732, 2011.

KURZROCK, Razelle; GUTTERMAN, Jordan U.; TALPAZ, Moshe. The molecular genetics of Philadelphia chromosome–positive leukemias. **New England Journal of Medicine**, v. 319, n. 15, p. 990-998, 1988.

ŁASTOWSKA, Maria *et al*. Identification of a neuronal transcription factor network involved in medulloblastoma development. **Acta neuropathologica communications**, v. 1, n. 1, p. 35, 2013.

LEE, Shawn HR; YANG, Jun J. Pharmacogenomics in acute lymphoblastic leukemia. **Best Practice & Research Clinical Haematology**, v. 30, n. 3, p. 229-236, 2017.

LEVINE, Selena R.; MCNEER, Jennifer L.; ISAKOFF, Michael S. Challenges faced in the treatment of acute lymphoblastic leukemia in adolescents and young adults. **Clinical Oncology in Adolescents and Young Adults**, v. 6, p. 11, 2016.

LIM, Joshua Yew-Suang *et al*. Genomics of racial and ethnic disparities in childhood acute lymphoblastic leukemia. **Cancer**, v. 120, n. 7, p. 955-962, 2014.

LIU, Song *et al*. Systematic identification of regulatory variants associated with cancer risk. **Genome biology**, v. 18, n. 1, p. 194, 2017.

LOUIS, David N. *et al*. The 2007 WHO classification of tumours of the central nervous system. **Acta neuropathologica**, v. 114, n. 2, p. 97-109, 2007.

LU, Yi-Fan *et al*. Personalized medicine and human genetic diversity. **Cold Spring Harbor perspectives in medicine**, v. 4, n. 9, p. a008581, 2014.

MA, Xiaotu *et al*. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. **Nature communications**, v. 6, p. 6604, 2015.

MAGRATH, Ian *et al*. Paediatric cancer in low-income and middle-income countries. **The lancet oncology**, v. 14, n. 3, p. e104-e116, 2013.

MEYERS, John A. *et al*. Phosphodiesterase 4 inhibitors augment levels of glucocorticoid receptor in B cell chronic lymphocytic leukemia but not in normal circulating hematopoietic cells. **Clinical Cancer Research**, v. 13, n. 16, p. 4920-4927, 2007.

MIRANDA-FILHO, Adalberto *et al*. Epidemiological patterns of leukaemia in 184 countries: a population-based study. **The Lancet Haematology**, v. 5, n. 1, p. e14-e24, 2018.

MITCHELL, Kevin J. What is complex about complex disorders?. **Genome biology**, v. 13, n. 1, p. 237, 2012.

MORENO, Diana J. *et al*. Genetic Ancestry and Susceptibility to Late-Ons*et al*zheimer Disease (LOAD) in the Admixed Colombian Population. **Alzheimer Disease & Associated Disorders**, v. 31, n. 3, p. 225-231, 2017.

NARANJO, Maria-Eugenia G. *et al*. Interethnic Variability in CYP2D6, CYP2C9, and CYP2C19 Genes and Predicted Drug Metabolism Phenotypes Among 6060 Ibero-and Native Americans: RIBEF-CEIBA Consortium Report on Population Pharmacogenomics. **Omics: a journal of integrative biology**, v. 22, n. 9, p. 575-588, 2018.

NAVARRETE, M. *et al*. Treatment of childhood acute lymphoblastic leukemia in central America: A lower-middle income countries experience. **Pediatric blood & cancer**, v. 61, n. 5, p. 803-809, 2014.

NELSON, Sarah C. *et al*. Is 'forward'the same as 'plus'?… and other adventures in SNP allele nomenclature. **Trends in Genetics**, v. 28, n. 8, p. 361-363, 2012.

NGUYEN, Kim *et al*. Factors influencing survival after relapse from acute lymphoblastic leukemia: a Children's Oncology Group study. **Leukemia**, v. 22, n. 12, p. 2142, 2008.

NORDLING, Linda. How the genomics revolution could finally help Africa. **Nature News**, v. 544, n. 7648, p. 20, 2017.

NUMATA, Shusuke *et al*. Positive association of the PDE4B (phosphodiesterase 4B) gene with schizophrenia in the Japanese population. **Journal of psychiatric research**, v. 43, n. 1, p. 7-12, 2008.

OGAWA, Ryosuke *et al*. Inhibition of PDE4 phosphodiesterase activity induces growth suppression, apoptosis, glucocorticoid sensitivity, p53, and p21WAF1/CIP1 proteins in human acute lymphoblastic leukemia cells. **Blood**, v. 99, n. 9, p. 3390-3397, 2002.

PAGE, C. P.; SPINA, D. Phosphodiesterase inhibitors in the treatment of inflammatory diseases. In: **Phosphodiesterases as Drug Targets**. Springer, Berlin, Heidelberg, 2011. p. 391-414.

PEREZ-ANDREU, Virginia *et al*. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. **Nature genetics**, v. 45, n. 12, p. 1494, 2013.

POPEJOY, Alice B.; FULLERTON, Stephanie M. Genomics is failing on diversity. **Nature News**, v. 538, n. 7624, p. 161, 2016.

PRICE, Alkes L. *et al*. A genomewide admixture map for Latino populations. **The American Journal of Human Genetics**, v. 80, n. 6, p. 1024-1036, 2007.

PUI, Ching-Hon; EVANS, William E. Treatment of acute lymphoblastic leukemia. **New England Journal of Medicine**, v. 354, n. 2, p. 166-178, 2006.

PUI, Ching-Hon. Reducing delayed intensification therapy in childhood ALL. **The Lancet Oncology**, v. 14, n. 3, p. 178-179, 2013.

PURCELL, Shaun *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American journal of human genetics**, v. 81, n. 3, p. 559-575, 2007.

QUINTANA-MURCI, Lluis. Understanding rare and common diseases in the context of human evolution. **Genome biology**, v. 17, n. 1, p. 225, 2016.

QUIROZ, Elisa *et al*. The emerging story of acute lymphoblastic leukemia among the Latin American population–biological and clinical implications. **Blood reviews**, v. 33, p. 98-105, 2019.

RABBANI, Bahareh *et al*. Next generation sequencing: implications in personalized medicine and pharmacogenomics. **Molecular BioSystems**, v. 12, n. 6, p. 1818-1830, 2016.

REICH, David *et al*. Reconstructing native American population history. **Nature**, v. 488, n. 7411, p. 370, 2012.

RODRIGUES-SOARES, Fernanda et al. Genomic ancestry, CYP 2D6, CYP 2C9 and CYP 2C19 among Latin-Americans. **Clinical Pharmacology & Therapeutics**, 2019.

RODRIGUEZ-GALINDO, Carlos *et al*. Global challenges in pediatric oncology. **Current opinion in pediatrics**, v. 25, n. 1, p. 3-15, 2013.

RODRIGUEZ-GALINDO, Carlos *et al*. Toward the cure of all children with cancer through collaborative efforts: pediatric oncology as a global challenge. **Journal of Clinical Oncology**, v. 33, n. 27, p. 3065, 2015.

ROSSELL, Nuria; GIGENGACK, Roy; BLUME, Stuart. Childhood cancer in El Salvador: A preliminary exploration of parental concerns in the abandonment of treatment. **European Journal of Oncology Nursing**, v. 19, n. 4, p. 370-375, 2015.

SAN LUCAS, F. Anthony; ROSENBERG, Noah A.; SCHEET, Paul. Haploscope: a tool for the graphical display of haplotype structure in populations. **Genetic epidemiology**, v. 36, n. 1, p. 17-21, 2012.

SCHEET, Paul; STEPHENS, Matthew. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. **The American Journal of Human Genetics**, v. 78, n. 4, p. 629-644, 2006.

SCHMIEGELOW, Kjeld *et al*. Consensus definitions of 14 severe acute toxic effects for childhood lymphoblastic leukaemia treatment: a Delphi consensus. **The Lancet Oncology**, v. 17, n. 6, p. e231-e239, 2016.

SCHMITT, Anthony D. *et al*. A compendium of chromatin contact maps reveals spatially active regions in the human genome. **Cell reports**, v. 17, n. 8, p. 2042-2059, 2016.

SCHORK, Nicholas J. Personalized medicine: time for one-person trials. **Nature News**, v. 520, n. 7549, p. 609, 2015.

SCHRAPPE, Martin *et al*. Reduced intensity delayed intensification in standard-risk patients defined by minimal residual disease in childhood acute lymphoblastic leukemia: results of an international randomized trial in 1164 patients (Trial AIEOP-BFM ALL 2000). 2016.

SELDIN, Michael F.; PASANIUC, Bogdan; PRICE, Alkes L. New approaches to disease mapping in admixed populations. **Nature Reviews Genetics**, v. 12, n. 8, p. 523, 2011.

SIRUGO, Giorgio; WILLIAMS, Scott M.; TISHKOFF, Sarah A. The missing diversity in human genetic studies. **Cell**, v. 177, n. 1, p. 26-31, 2019.

SIVE, Jonathan I. *et al*. Outcomes in older adults with acute lymphoblastic leukaemia (ALL): results from the international MRC UKALL XII/ECOG 2993 trial. **British journal of haematology**, v. 157, n. 4, p. 463-471, 2012.

SKOGLUND, Pontus *et al*. Genetic evidence for two founding populations of the Americas. **Nature**, v. 525, n. 7567, p. 104, 2015.

SOUSA, Daniel Willian Lustosa de *et al*. Acute lymphoblastic leukemia in children and adolescents: prognostic factors and analysis of survival. **Revista brasileira de hematologia e hemoterapia**, v. 37, n. 4, p. 223-229, 2015.

STANISH, Charles. The origin of state societies in South America. **annual review of anthropology**, v. 30, n. 1, p. 41-64, 2001.

SUN, Celi *et al*. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. **Nature genetics**, v. 48, n. 3, p. 323, 2016.

SWEET-CORDERO, E. Alejandro; BIEGEL, Jaclyn A. The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. **Science**, v. 363, n. 6432, p. 1170-1175, 2019.

SWINNEY, Ryan M. *et al*. Polymorphisms in CYP1A1 and ethnic-specific susceptibility to acute lymphoblastic leukemia in children. **Cancer Epidemiology and Prevention Biomarkers**, v. 20, n. 7, p. 1537-1542, 2011.

TARAZONA-SANTOS, Eduardo *et al*. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. **The American Journal of Human Genetics**, v. 68, n. 6, p. 1485-1496, 2001.

TASIAN, Sarah K.; LOH, Mignon L.; HUNGER, Stephen P. Philadelphia chromosome–like acute lymphoblastic leukemia. **Blood, The Journal of the American Society of Hematology**, v. 130, n. 19, p. 2064-2072, 2017.

TEAM, R. Core *et al*. R: A language and environment for statistical computing. 2013. Vienna, Austria.

TERWILLIGER, T.; ABDUL-HAY, MJBCJ. Acute lymphoblastic leukemia: a comprehensive review and 2017 update. **Blood cancer journal**, v. 7, n. 6, p. e577-e577, 2017.

TOMAR, Anil Kumar; AGARWAL, Rahul; KUNDU, Bishwajit. Most Variable Genes and Transcription Factors in Acute Lymphoblastic Leukemia Patients. **Interdisciplinary Sciences: Computational Life Sciences**, p. 1-11, 2019.

TSUNODA, Toshiyuki *et al*. Inhibition of phosphodiesterase-4 (PDE4) activity triggers luminal apoptosis and AKT dephosphorylation in a 3-D colonic-crypt model. **Molecular cancer**, v. 11, n. 1, p. 46, 2012.

UHLÉN, Mathias *et al*. Tissue-based map of the human proteome. **Science**, v. 347, n. 6220, p. 1260419, 2015.

VORA, Ajay *et al*. Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. **The lancet oncology**, v. 14, n. 3, p. 199-209, 2013.

WALSH, K. M. *et al*. Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. **Leukemia**, v. 27, n. 12, p. 2416, 2013.

WANG, Sijia *et al*. Genetic variation and population structure in Native Americans. **PLoS genetics**, v. 3, n. 11, p. e185, 2007.

WANG, Sijia *et al*. Geographic patterns of genome admixture in Latin American Mestizos. **PLoS genetics**, v. 4, n. 3, p. e1000037, 2008.

WANG, Ti *et al*. Common SNPs in myelin transcription factor 1-like (MYT1L): association with major depressive disorder in the Chinese Han population. **PloS one**, v. 5, n. 10, p. e13662, 2010.

WANG, Tianyun *et al*. De novo genic mutations among a Chinese autism spectrum disorder cohort. **Nature communications**, v. 7, p. 13316, 2016.

WRIGHT, Sewall. The genetical structure of populations. **Annals of eugenics**, v. 15, n. 1, p. 323-354, 1949.

XU, Heng *et al*. ARID5B genetic polymorphisms contribute to racial disparities in the incidence and treatment outcome of childhood acute lymphoblastic leukemia. **Journal of clinical oncology**, v. 30, n. 7, p. 751, 2012.

YAEGER, Rona *et al*. Comparing genetic ancestry and self-described race in African Americans born in the United States and in Africa. **Cancer Epidemiology and Prevention Biomarkers**, v. 17, n. 6, p. 1329-1338, 2008.

YANG, Jun J. *et al*. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. **Nature genetics**, v. 43, n. 3, p. 237, 2011.

YANG, Jun J. *et al*. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. **Blood**, v. 120, n. 20, p. 4197-4204, 2012.

ZHANG, Yangmei *et al*. Clinical significance of MYT1L gene polymorphisms in Chinese patients with gastric cancer. **PloS one**, v. 8, n. 8, p. e71979, 2013.

ZHAO, Zhiguo *et al*. Association of genetic susceptibility variants for type 2 diabetes with breast cancer risk in women of European ancestry. **Cancer Causes & Control**, v. 27, n. 5, p. 679-693, 2016.

APÊNDICES

Tabela Apêndice 1. Frequências alélicas dos SNPs do banco BeadXpress localizados nos genes *PDE4B* e *MYT1L* com destaque (em cinza) para os reportados em Yang *et al.* (2011). ASH: *Ashaninka*, AYM: *Aymara*, MAC: *Machiguenga*, GUA: *Guarani*, TUP: *Tupiniquim*, MCP: Martinho Campos, CAR: Carmésia, RES: Resplendor, SJM: São João das Missões, EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). Nan: genótipos não obtidos.

Gene	SNP	Nativos Peru			Nativos Brasil		Miscigenados Brasil				1000G	1000G	1000G	1000G
		ASH	AYM	MAC	GUA	TUP	CAR	MCP	RES	SJM	EUR	AFR	EAS	SAS
PDE4B	rs12081185.A	0,68	0,65	0,55	0,56	0,63	0,29	0,48	0,42	0,66	0,42	0,50	0,68	0,61
	rs11589566.C	0,38	0,31	0,28	0,21	0,23	0,13	0,08	0,21	0,18	0,19	0,07	0,09	0,18
	rs11208774.C	0,19	Nan	0,19	0,15	0,12	0,33	Nan	Nan	Nan	0,47	0,88	0,81	0,76
	rs1937456.C	0,77	0,45	0,65	0,34	0,47	0,79	0,75	0,69	0,68	0,64	0,83	0,55	0,57
	rs2186123.A	0,78	0,56	0,68	0,43	0,30	0,79	0,68	0,65	0,62	0,64	0,70	0,55	0,56
	rs2503205.C	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,02	0,08	0,00	0,07	0,00	0,00
	rs1937453.A	0,77	0,57	0,66	0,67	0,53	0,74	0,70	0,71	0,70	0,71	0,72	0,57	0,61
	rs6588177.A	0,87	0,76	0,99	0,83	0,86	0,76	0,70	0,77	0,80	0,27	0,13	0,37	0,36
	rs12033425.C	0,24	0,45	0,28	0,00	0,01	0,16	0,00	0,15	0,14	0,02	0,20	0,26	0,11
	rs1354060.A	0,39	0,54	0,56	0,35	0,39	0,63	0,40	0,54	0,54	0,45	0,59	0,43	0,53
	rs1500951.A	0,12	0,13	0,19	0,10	0,08	0,00	0,00	0,02	0,00	0,01	0,08	0,35	0,21
	rs12119734.C	1,00	0,99	0,99	0,91	0,90	0,79	0,76	0,73	0,88	0,70	0,90	0,98	0,83
	rs12027416.C	0,93	0,86	0,90	1,00	1,00	0,90	1,00	0,85	0,86	0,97	0,79	0,73	0,89
	rs10454453.A	0,76	0,73	0,73	0,66	0,64	0,58	0,47	0,44	0,36	0,52	0,26	0,56	0,64
	rs12045945.A	0,00	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,14	0,08
	rs6661245.C	0,10	0,25	0,12	0,27	0,21	0,34	0,12	0,13	0,46	0,06	0,62	0,26	0,12
	rs522037.C	0,81	0,66	0,79	0,64	0,59	0,50	0,47	0,65	0,60	0,59	0,28	0,40	0,45
	rs6683604.C	0,12	0,07	0,15	0,06	0,23	0,42	0,45	0,35	0,34	0,56	0,15	0,16	0,36
	rs6683977.C	0,01	0,06	0,00	0,01	0,00	0,03	0,03	0,02	0,02	0,56	0,04	0,16	0,35
	rs783036.A	0,37	0,26	0,43	0,21	0,29	0,47	0,30	0,29	0,32	0,49	0,79	0,15	0,59
MYT1L	rs3924017.C	0,80	0,76	0,94	0,72	0,68	0,55	0,63	0,58	0,44	0,34	0,72	0,67	0,38
	rs6728613.C	0,15	0,24	0,01	0,28	0,28	0,47	0,43	0,44	0,58	0,37	0,83	0,66	0,38
	rs11127305.A	1,00	0,97	1,00	1,00	1,00	1,00	1,00	0,98	0,98	0,90	0,80	0,66	0,90
	rs2241685.C	0,79	0,93	0,98	0,95	0,90	0,87	0,93	0,83	0,86	0,90	0,80	0,65	0,87
	rs3748988.A	0,74	0,83	0,91	0,74	0,66	0,55	0,68	0,62	0,48	0,63	0,29	0,36	0,63
	rs3748989.C	0,79	0,90	0,94	0,97	0,87	0,95	0,95	0,87	0,86	0,91	0,86	0,66	0,88
	rs12986730.A	0,06	0,14	0,01	0,23	0,34	0,37	0,28	0,29	0,46	0,26	0,53	0,30	0,25
	rs12991351.G	0,00	0,00	0,00	0,01	0,02	0,00	0,07	0,06	0,12	0,03	0,05	0,00	0,01
	rs6729968.A	0,93	0,82	0,98	0,73	0,54	0,47	0,78	0,46	0,37	0,73	0,22	0,52	0,64
	rs6548050.C	0,00	0,08	0,00	0,05	0,08	0,21	0,12	0,29	0,16	0,17	0,28	0,34	0,18
	rs10178240.A	0,07	0,10	0,01	0,17	0,28	0,29	0,10	0,25	0,48	0,10	0,48	0,11	0,15
	rs6548048.G	0,96	0,88	1,00	0,83	0,63	0,75	0,92	1,0	1,0	0,70	0,22	0,52	0,65
	rs12468174.A	0,00	0,09	0,00	0,05	0,06	0,08	0,08	0,19	0,08	0,11	0,12	0,34	0,14

Gene	SNP	Nativos Peru			Nativos Brasil		Miscigenados Brasil				1000G	1000G	1000G	1000G
		ASH	AYM	MAC	GUA	TUP	CAR	MCP	RES	SJM	EUR	AFR	EAS	SAS
	rs17039334.C	0,00	0,00	0,00	0,00	0,00	0,05	0,05	0,00	0,02	0,04	0,02	0,08	0,06
	rs11683072.G	0,99	0,92	1,00	0,89	0,83	0,84	0,93	0,83	0,90	0,91	0,88	0,87	0,93
	rs6705656.C	0,00	0,08	0,00	0,09	0,17	0,26	0,23	0,33	0,14	0,84	0,55	0,88	0,93
	rs12468168.C	1,00	0,91	1,00	0,93	0,87	0,79	0,85	0,71	0,92	0,86	0,72	0,88	0,94
	rs17338491.A	0,00	0,09	0,00	0,06	0,09	0,11	0,10	0,23	0,04	0,14	0,07	0,10	0,06
	rs17039396.A	0,23	0,22	0,36	0,07	0,08	0,03	0,05	0,10	0,02	0,11	0,01	0,24	0,08
	rs17039463.A	0,00	0,00	0,00	0,02	0,01	0,03	0,02	0,02	0,04	0,00	0,10	0,17	0,42
	rs17039474.G	0,01	0,02	0,00	0,00	0,00	0,03	0,02	0,06	0,00	0,00	0,05	0,00	0,00
	rs13034457.A	0,40	0,59	0,38	0,47	0,52	0,58	0,78	0,68	0,66	0,62	0,53	0,17	0,42
	rs13388663.A	0,54	Nan	0,36	0,80	0,71	0,61	0,58	Nan	Nan	0,27	0,36	0,08	0,47
	rs10195351.C	1,00	0,99	1,00	0,77	0,63	0,58	0,43	0,58	0,48	0,76	0,84	0,94	0,69

Tabela Apêndice 2. Frequências alélicas (relativas ao número de amostras do banco de dados TargetSeq) em diferentes populações das variantes encontradas entre as posições genômicas 2.215.144 e 2.235.144 do gene *MYT1L* (GRCh37). ASH: *Ashaninka*, MAC: *Machiguenga*, AYM: *Aymara*, QUE: *Quechua*, TUP: *Tupiniquim*, HUI:*Huichol*, EUR: Europeans (1000GP), AFR: Africans (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). Destaque (em cinza) para o SNP rs17039396

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
2:2215590	A:0,97	C:0,03	A:1,0	C:0,0	A:1,0	C:0,0	A:0,87	C:0,13	A:1,0	C:0,0	A:0,97	C:0,03	A:0,96	C:0,04	A:0,79	C:0,21	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0
2:2215876	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:0,95	G:0,05	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0
2:2216012	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:0,87	C:0,13	A:0,80	C:0,20	A:0,95	C:0,05	A:0,78	C:0,22
2:2216078	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:0,96	G:0,04	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0
rs1582415	T:0,96	G:0,04	T:1,0	G:0,0	T:0,97	G:0,03	T:1,0	G:0,0	T:0,92	G:0,08	T:1,0	G:0,0	T:0,95	G:0,05	T:0,78	G:0,22	G:0,14	T:0,86	G:0,12	T:0,88	G:0,1	T:0,9	G:0,06	T:0,94
2:2216359	T:0,83	C:0,17	T:0,95	C:0,05	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0
rs1582414	G:0,94	T:0,06	G:1,0	T:0,0	G:0,94	T:0,06	G:0,95	T:0,05	G:0,95	T:0,05	G:0,97	T:0,03	G:0,96	T:0,04	G:0,79	T:0,21	T:0,14	G:0,86	T:0,12	G:0,88	T:0,1	G:0,9	T:0,06	G:0,94
2:2217073	G:1,0	C:0,0	G:1,0	C:0,0	G:0,98	C:0,02	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0
rs58275964	C:0,94	T:0,06	C:1,0	T:0,0	C:0,94	T:0,06	C:0,9	T:0,1	C:0,95	T:0,05	C:0,87	T:0,13	C:0,96	T:0,04	C:0,75	T:0,25	T:0,14	C:0,86	T:0,12	C:0,88	T:0,1	C:0,9	T:0,06	C:0,94
rs78173916	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,98	C:0,03	T:0,98	C:0,02	T:0,84	C:0,16	T:1,0	C:0,0	T:1,0	C:0,0	C:0,13	T:0,87	C:0,03	T:0,97	C:0,05	T:0,95	C:0,22	T:0,78
rs12478611	G:0,93	A:0,07	G:1,0	A:0,0	G:0,98	A:0,02	G:0,97	A:0,03	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,8	A:0,2	A:0,14	G:0,86	A:0,07	G:0,93	A:0,1	G:0,9	A:0,06	G:0,94
rs10865534	A:0,94	T:0,06	A:1,0	T:0,0	A:0,98	T:0,02	A:0,97	T:0,03	A:1,0	T:0,0	A:1,0	T:0,0	A:0,96	T:0,04	A:0,88	T:0,13	T:0,14	A:0,86	T:0,07	A:0,93	T:0,1	A:0,9	T:0,06	A:0,94
rs73177806	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,96	C:0,04	T:1,0	C:0,0	T:1,0	C:0,0	C:0,13	T:0,87	C:0,19	T:0,81	C:0,05	T:0,95	C:0,22	T:0,78
rs1592732	C:0,97	G:0,03	C:1,0	G:0,0	C:0,93	G:0,07	C:0,91	G:0,09	C:0,97	G:0,03	C:0,97	G:0,03	C:1,0	G:0,0	C:0,91	G:0,09	G:0,14	C:0,86	G:0,24	C:0,76	G:0,1	C:0,9	G:0,06	C:0,94
rs61688379	T:0,92	C:0,08	T:1,0	C:0,0	T:1,0	C:0,0	T:0,9	C:0,1	T:0,94	C:0,06	T:1,0	C:0,0	T:1,0	C:0,0	T:0,73	C:0,27								
rs72769228	T:0,97	G:0,03	T:1,0	G:0,0	T:0,95	G:0,05	T:0,97	G:0,03	T:0,97	G:0,03	T:1,0	G:0,0	T:1,0	G:0,0	T:0,87	G:0,13	G:0,14	T:0,86	G:0,07	T:0,93	G:0,1	T:0,9	G:0,06	T:0,94
rs60144672	T:0,94	C:0,06	T:1,0	C:0,0	T:0,94	C:0,06	T:0,9	C:0,1	T:0,79	C:0,21	T:0,64	C:0,36	T:0,96	C:0,04	T:0,74	C:0,26	C:0,29	T:0,71	C:0,45	T:0,55	C:0,15	T:0,85	C:0,29	T:0,71
rs139264841	A:0,94	G:0,06	A:0,91	G:0,09	A:1,0	G:0,0	A:0,95	G:0,05	A:0,96	G:0,04	A:0,88	G:0,12	A:0,83	G:0,17	A:0,95	G:0,05	G:0,0	A:1,0	G:0,0	A:1,0	G:0,01	A:0,99	G:0,0	A:1,0
rs72769229	A:0,94	T:0,06	A:1,0	T:0,0	A:0,94	T:0,06	A:0,9	T:0,1	A:0,95	T:0,05	A:0,91	T:0,09	A:0,96	T:0,04	A:0,79	T:0,21	T:0,14	A:0,86	T:0,07	A:0,93	T:0,1	A:0,9	T:0,06	A:0,94
rs72769230	G:0,94	A:0,06	G:1,0	A:0,0	G:0,94	A:0,06	G:0,9	A:0,1	G:0,95	A:0,05	G:0,9	A:0,1	G:0,96	A:0,04	G:0,75	A:0,25	A:0,15	G:0,85	A:0,06	G:0,94	A:0,1	G:0,9	A:0,06	G:0,94
rs7420242	G:0,94	A:0,06	G:1,0	A:0,0	G:0,94	A:0,06	G:0,9	A:0,1	G:0,95	A:0,05	G:0,88	A:0,12	G:0,96	A:0,04	G:0,75	A:0,25	A:0,14	G:0,86	A:0,12	G:0,88	A:0,1	G:0,9	A:0,06	G:0,94
rs17039395	G:0,94	A:0,06	G:1,0	A:0,0	G:0,94	A:0,06	G:0,88	A:0,13	G:0,9	A:0,1	G:0,63	A:0,37	G:0,96	A:0,04	G:0,75	A:0,25	A:0,27	G:0,73	A:0,46	G:0,54	A:0,15	G:0,85	A:0,28	G:0,72

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
2:2222593	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,96	T:0,04	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0
rs1582413	A:0,94	T:0,06	A:1,0	T:0,0	A:0,94	T:0,06	A:0,89	T:0,11	A:0,91	T:0,09	A:0,68	T:0,32	A:0,96	T:0,04	A:0,76	T:0,24	T:0,27	A:0,73	T:0,46	A:0,54	T:0,15	A:0,85	T:0,28	A:0,72
rs17247310	T:0,94	C:0,06	T:1,0	C:0,0	T:0,94	C:0,06	T:0,88	C:0,13	T:0,84	C:0,16	T:0,62	C:0,38	T:0,96	C:0,04	T:0,75	C:0,25	C:0,37	T:0,63	C:0,49	T:0,51	C:0,24	T:0,76	C:0,38	T:0,62
rs113009634	C:0,89	T:0,11	C:1,0	T:0,0	C:0,94	T:0,06	C:0,91	T:0,09	C:1,0	T:0,0	C:0,9	T:0,1	C:1,0	T:0,0	C:0,86	T:0,14	T:0,13	C:0,87	T:0,12	C:0,88	T:0,09	C:0,91	T:0,06	C:0,94
rs11681135	C:0,97	T:0,03	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,95	T:0,05	C:1,0	T:0,0	C:1,0	T:0,0	T:0,22	C:0,78	T:0,07	C:0,93	T:0,08	C:0,92	T:0,25	C:0,75
rs60585106	C:0,75	T:0,25	C:0,61	T:0,39	C:0,81	T:0,19	C:0,93	T:0,08	C:0,91	T:0,09	C:0,74	T:0,26	C:0,92	T:0,08	C:0,85	T:0,15	T:0,06	C:0,94	T:0,1	C:0,9	T:0,14	C:0,86	T:0,03	C:0,97
rs11127372	T:0,53	C:0,47	T:0,52	C:0,48	T:0,42	C:0,58	T:0,42	C:0,58	T:0,43	C:0,57	T:0,43	C:0,57	T:0,33	C:0,67	T:0,35	C:0,65	T:0,32	C:0,68	T:0,16	C:0,84	T:0,06	C:0,94	T:0,20	C:0,8
rs77041749	G:0,57	C:0,43	G:0,47	C:0,53	G:0,62	C:0,38	G:0,67	C:0,33	G:0,81	C:0,19	G:0,96	C:0,04	G:0,88	C:0,13	G:0,69	C:0,31	C:0,04	G:0,96	C:0,0	G:1,0	C:0,0	G:1,0	C:0,02	G:0,98
rs1895137	T:0,63	G:0,38	T:0,5	G:0,5	T:0,51	G:0,49	T:0,46	G:0,54	T:0,48	G:0,52	T:0,46	G:0,54	T:0,41	G:0,59	T:0,5	G:0,5	T:0,35	G:0,65	T:0,27	G:0,73	T:0,15	G:0,85	T:0,46	G:0,54
rs6731821	A:0,59	G:0,41	A:0,5	G:0,5	A:0,54	G:0,46	A:0,45	G:0,55	A:0,49	G:0,51	A:0,46	G:0,54	A:0,43	G:0,57	A:0,72	G:0,28	A:0,3	G:0,7	A:0,23	G:0,77	A:0,17	G:0,83	A:0,32	G:0,68
rs6711203	C:0,6	T:0,4	C:0,5	T:0,5	C:0,54	T:0,46	C:0,45	T:0,55	C:0,49	T:0,51	C:0,46	T:0,54	C:0,48	T:0,52	C:0,72	T:0,28	C:0,3	T:0,7	C:0,23	T:0,77	C:0,23	T:0,77	C:0,32	T:0,68
rs6711124	C:0,61	T:0,39	C:0,49	T:0,51	C:0,54	T:0,46	C:0,45	T:0,55	C:0,46	T:0,54	C:0,46	T:0,54	C:0,48	T:0,52	C:0,72	T:0,28	C:0,31	T:0,69	C:0,23	T:0,77	C:0,23	T:0,77	C:0,32	T:0,68
rs17039396	G:0,71	A:0,29	G:0,61	A:0,39	G:0,87	A:0,13	G:0,88	A:0,13	G:0,92	A:0,08	G:0,81	A:0,19	G:0,88	A:0,13	G:1,0	A:0,0	A:0,11	G:0,89	A:0,01	G:0,99	A:0,24	G:0,76	A:0,08	G:0,92
rs13414323	C:0,91	T:0,09	C:1,0	T:0,0	C:0,96	T:0,04	C:0,98	T:0,03	C:1,0	T:0,0	C:1,0	T:0,0	C:0,83	T:0,17	C:0,9	T:0,1	T:0,02	C:0,98	T:0,03	C:0,97	T:0,21	C:0,79	T:0,18	C:0,82
rs6548123	C:0,53	T:0,47	C:0,5	T:0,5	C:0,5	T:0,5	C:0,44	T:0,56	C:0,45	T:0,55	C:0,48	T:0,52	C:0,35	T:0,65	C:0,55	T:0,45	C:0,28	T:0,72	C:0,2	T:0,8	C:0,02	T:0,98	C:0,14	T:0,86
2:2225675	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0
rs7569601	G:0,53	C:0,47	G:0,5	C:0,5	G:0,5	C:0,5	G:0,43	C:0,57	G:0,52	C:0,48	G:0,47	C:0,53	G:0,33	C:0,67	G:0,55	C:0,45	G:0,28	C:0,72	G:0,15	C:0,85	G:0,02	C:0,98	G:0,14	C:0,86
rs71442324	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,0	A:1,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	A:0,13	G:0,87	A:0,01	G:0,99	A:0	G:1,0	A:0,02	G:0,98
rs11694233	A:1,0	T:0,0	NA	NA	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:0,35	T:0,65	A:0,35	T:0,65	A:0,25	T:0,75	A:0,39	T:0,61
rs10204613	T:0,56	G:0,44	T:0,47	G:0,53	T:0,5	G:0,5	T:0,32	G:0,68	T:0,65	G:0,35	T:0,39	G:0,61	T:0,57	G:0,43	T:0,64	G:0,36	T:0,31	G:0,69	T:0,48	G:0,52	T:0,2	G:0,8	T:0,32	G:0,68
rs11127371	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:0,9	G:0,1	A:0,97	G:0,03	A:1,0	G:0,0	A:1,0	G:0,0	G:0,23	A:0,77	G:0,14	A:0,86	G:0,19	A:0,81	G:0,31	A:0,69
rs1862114	C:0,6	T:0,4	C:0,48	T:0,53	C:0,53	T:0,47	C:0,43	T:0,57	C:0,47	T:0,53	C:0,58	T:0,42	C:0,5	T:0,5	C:0,68	T:0,32	C:0,31	T:0,69	C:0,23	T:0,77	C:0,22	T:0,78	C:0,32	T:0,68
rs1862113	T:0,53	G:0,47	T:0,5	G:0,5	T:0,48	G:0,52	T:0,41	G:0,59	T:0,45	G:0,55	T:0,45	G:0,55	T:0,33	G:0,67	T:0,55	G:0,45	T:0,28	G:0,72	T:0,15	G:0,85	T:0,02	G:0,98	T:0,14	G:0,86
rs10865533	T:0,63	C:0,38	T:0,5	C:0,5	T:0,53	C:0,47	T:0,42	C:0,58	T:0,49	C:0,51	T:0,5	C:0,5	T:0,5	C:0,5	T:0,65	C:0,35	T:0,30	C:0,7	T:0,21	C:0,79	T:0,2	C:0,8	T:0,33	C:0,67
rs12052817	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:0,98	A:0,03	C:1,0	A:0,0	C:0,98	A:0,02	C:1,0	A:0,0	C:1,0	A:0,0	A:0,05	C:0,95	A:0,04	C:0,96	A:0,02	C:0,98	A:0,06	C:0,94
rs888627	A:0,63	G:0,38	A:0,5	G:0,5	A:0,54	G:0,46	A:0,45	G:0,55	A:0,52	G:0,48	A:0,5	G:0,5	A:0,5	G:0,5	A:0,65	G:0,35	A:0,30	G:0,7	A:0,27	G:0,73	A:0,22	G:0,78	A:0,33	G:0,67

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
rs7566768	T:0,63	A:0,38	T:0,51	A:0,49	T:0,52	A:0,48	T:0,45	A:0,55	T:0,53	A:0,47	T:0,49	A:0,51	T:0,5	A:0,5	T:0,61	A:0,39	T:0,31	A:0,69	T:0,24	A:0,76	T:0,22	A:0,78	T:0,33	A:0,67
rs114751773	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	A:0,0	G:1,0	A:0,04	G:0,96	A:0,0	G:1,0	A:0,0	G:1,0
rs11900037	A:0,48	G:0,52	A:0,49	G:0,51	A:0,54	G:0,46	A:0,45	G:0,55	A:0,42	G:0,58	A:0,48	G:0,52	A:0,37	G:0,63	A:0,5	G:0,5	A:0,29	G:0,71	A:0,12	G:0,88	A:0,02	G:0,98	A:0,15	G:0,85
2:2228435	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:0,97	A:0,03	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0
rs1808360	T:0,0	C:1,0	T:0,5	C:0,5	T:0,86	C:0,14	T:1,0	C:0,0	T:0,0	C:1,0	NA	NA	NA	NA	NA	NA	T:0,31	C:0,69	T:0,23	C:0,68/ G:0,09	T:0,22	C:0,78	T:0,33	C:0,67
rs13413284	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:0,0	C:1,0	G:1,0	C:0,0	G:1,0	C:0,0	NA	NA	NA	NA	C:0,02	G:0,98	C:0,03	G:0,97	C:0,21	G:0,79	C:0,17	G:0,83
rs116293447	A:0,97	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:0,93	C:0,0	A:1,0	C:0,0	A:0,97	C:0,0	A:0,93	C:0,0	A:0,95	C:0,0	A:0,32	C:0,68	A:0,31	C:0,69	A:0,22	C:0,78		
2:2228993	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,98	C:0,02	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0
rs7597220	C:0,5	G:0,5	C:0,57	G:0,43	C:0,49	G:0,51	C:0,37	G:0,63	C:0,55	G:0,45	C:0,42	G:0,58	C:0,35	G:0,65	C:0,58	G:0,42	C:0,27	G:0,73	C:0,12	G:0,88	C:0,01	G:0,99	C:0,15	G:0,85
rs7607555	G:0,54	A:0,46	G:0,47	A:0,53	G:0,36	A:0,64	G:0,3	A:0,7	G:0,32	A:0,68	G:0,43	A:0,57	G:0,27	A:0,73	G:0,56	A:0,44	G:0,36	A:0,64	A:0,39	G:0,61	G:0,24	A:0,76	G:0,39	A:0,61
rs9309702	G:0,91	A:0,09	G:0,89	A:0,11	G:0,67	A:0,33	G:0,6	A:0,4	G:0,63	A:0,37	G:0,73	A:0,28	G:0,61	A:0,39	G:0,65	A:0,35	A:0,22	G:0,78	A:0,03	G:0,97	A:0,29	G:0,71	A:0,08	G:0,92
rs140595851	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,9	T:0,1	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0
rs75825920	A:0,93	G:0,07	A:1,0	G:0,0	A:0,96	G:0,04	A:0,97	G:0,03	A:1,0	G:0,0	A:1,0	G:0,0	A:0,83	G:0,17	A:0,95	G:0,05	G:0,0	A:1,0	G:0,0	A:1,0	G:0,01	A:0,99	G:0,0	A:1,0
rs142858460	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	G:0,0	C:1,0	G:0,04	C:0,96	G:0,0	C:1,0	G:0,0	C:1,0
rs35633741	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:0,93	T:0,07	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	T:0,18	G:0,82	T:0,01	G:0,99	T:0,08	G:0,92	T:0,15	G:0,85
rs185850219	C:0,96	T:0,04	C:1,0	T:0,0	C:0,96	T:0,04	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,95	T:0,05	C:1,0	T:0,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0
rs17039409	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:0,97	G:0,03	C:1,0	G:0,0	C:1,0	G:0,0	G:0,0	C:1,0	G:0,05	C:0,95	G:0,0	C:1,0	G:0,0	C:1,0
rs1035352	A:0,53	G:0,47	A:0,49	G:0,51	A:0,56	G:0,44	A:0,68	G:0,32	A:0,69	G:0,31	A:0,7	G:0,3	A:0,63	G:0,38	A:0,6	G:0,4	G:0,01	A:0,99	G:0,0	A:1,0	G:0,0	A:1,0	G:0,01	A:0,99
2:2231954	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0
rs1035351	T:1,0	C:0,0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	C:0,34	T:0,66	T:0,43	C:0,57	T:0,41	C:0,59	T:0,44	C:0,56
2:2232416	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,95	C:0,05	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0
rs1465388	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,03	A:0,98	C:0,13	A:0,87	C:0,1	A:0,9	C:0,0	A:1,0	C:0,0	A:1,0	C:0,22	A:0,78	C:0,01	A:0,99	C:0,0	A:1,0	C:0,09	A:0,91
rs10167537	C:0,21	T:0,79	C:0,09	T:0,91	C:0,07	T:0,93	C:0,35	T:0,65	C:0,32	T:0,68	C:0,45	T:0,55	C:0,17	T:0,83	C:0,0	T:1,0	T:0,13	C:0,87	C:0,44	T:0,56	C:0,3	T:0,7	C:0,46	T:0,54
2:2233238	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0
rs10167441	C:0,33	T:0,67	C:0,08	T:0,92	C:0,06	T:0,94	C:0,41	T:0,59	C:0,16	T:0,84	C:0,42	T:0,58	C:0,17	T:0,83	C:0,0	T:1,0	T:0,17	C:0,83	C:0,42	T:0,58	C:0,34	T:0,66	T:0,39	C:0,61

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
rs10184190	A:1,0	C:0,0	A:1,0	C:0,0	A:0,96	C:0,04	A:0,95	C:0,05	A:0,87	C:0,13	A:0,7	C:0,3	A:0,96	C:0,04	A:0,95	C:0,05	C:0,37	A:0,63	C:0,17	A:0,83	C:0,14	A:0,86	C:0,3	A:0,7
rs140247942	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,95	A:0,05	A:0,0	G:1,0	A:0,01	G:0,99	A:0,0	G:1,0	A:0,0	G:1,0
2:2233867	G:0,94	A:0,06	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0
rs79018269	G:0,94	A:0,06	G:0,93	A:0,07	G:0,87	A:0,13	G:0,79	A:0,21	G:0,8	A:0,2	G:0,92	A:0,08	G:0,92	A:0,08	G:0,9	A:0,1	A:0,0	G:1,0	A:0,01	G:0,99	A:0,15	G:0,85	A:0,01	G:0,99
rs17338428	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:0,88	C:0,12	G:1,0	C:0,0	G:1,0	C:0,0	C:0,09	G:0,91	C:0,08	G:0,92	C:0,04	G:0,96	C:0,14	G:0,86
rs4853947	A:0,25	G:0,75	A:0,14	G:0,86	A:0,15	G:0,85	A:0,39	G:0,61	A:0,36	G:0,64	A:0,38	G:0,62	A:0,17	G:0,83	A:0,1	G:0,9	A:0,45	G:0,55	A:0,25	G:0,75	A:0,18	G:0,82	A:0,17	G:0,83
rs148331727	T:0,94	C:0,06	T:1,0	C:0,0	T:0,94	C:0,06	T:0,95	C:0,05	T:0,94	C:0,06	T:0,98	C:0,02	T:1,0	C:0,0	T:1,0	C:0,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,01	T:0,99
2:2235031	T:0,97	C:0,03	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0

Tabela Apêndice 3. Frequências alélicas (relativas ao número de amostras do banco de dados TargetSeq) em diferentes populações das variantes encontradas entre as posições genômicas 66.759.100 e 66.779.100 do gene *PDE4B* (GRCh37). ASH: *Ashaninka*, MAC: *Machiguenga*, AYM: *Aymara*, QUE: *Quechua*, TUP: *Tupiniquim*, HUI:*Huichol*, EUR: Europeans (1000GP), AFR: Africans (1000GP), EAS: Leste-Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP). Destaque (cinza) para os SNPs rs546784, rs524770, rs6683977, rs641262

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
rs782967	C:1,0	T:0,0	C:0,84	T:0,16	C:0,95	T:0,05	C:0,88	T:0,12	C:0,96	T:0,04	C:0,8	T:0,2	C:0,88	T:0,13	C:1,0	T:0,0	C:0,44	T:0,56	T:0,15	C:0,85	T:0,16	C:0,84	T:0,36	C:0,64
rs111886642	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:0,87	G:0,13	A:1,0	G:0,0	G:0,03	A:0,97	G:0,05	A:0,95	G:0	A:1,0	G:0	A:1,0
rs6668516	G:0,94	A:0,06	G:0,86	A:0,14	G:0,92	A:0,08	G:0,85	A:0,15	G:0,92	A:0,08	G:0,77	A:0,23	G:0,88	A:0,13	G:0,95	A:0,05	G:0,44	A:0,56	A:0,14	G:0,86	A:0,16	G:0,84	A:0,36	G:0,64
1:66760010	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:0,98	G:0,03	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0
1:66760022	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:0,98	G:0,03	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0
rs17128742	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,88	C:0,13	T:1,0	C:0,0	C:0,03	T:0,97	C:0,05	T:0,95	C:0	T:1,0	C:0	T:1,0
rs17128748	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:0,88	T:0,13	A:1,0	T:0,0	T:0,03	A:0,97	T:0,05	A:0,95	T:0	A:1,0	T:0	A:1,0
rs56163035	C:0,44	T:0,56	C:0,79	T:0,21	C:0,9	T:0,1	C:0,85	T:0,15	C:0,89	T:0,11	C:0,77	T:0,23	C:0,92	T:0,08	C:1,0	T:0,0	T:0,08	C:0,92	T:0	C:1,0	T:0,01	C:0,99	T:0,16	C:0,84
rs191581125	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	T:0	C:1,0	T:0	C:1,0	T:0	C:1,0	T:0	C:1,0
rs599381	A:0,06	G:0,94	A:0,26	G:0,74	A:0,11	G:0,89	A:0,17	G:0,83	A:0,09	G:0,91	A:0,17	G:0,83	A:0,24	G:0,76	A:0,37	G:0,63	A:0,21	G:0,79	A:0,01	G:0,99	A:0,06	G:0,94	A:0,14	G:0,86
rs143612526	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,92	C:0,08	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,95	C:0,05	C:0	T:1,0	C:0	T:1,0	C:0	T:1,0	C:0	T:1,0

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
1:66760675	T:1,0	A:0,0	T:1,0	A:0,0	T:0,98	A:0,02	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0
rs72677260	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	T:0,03	G:0,97	T:0,05	G:0,95	T:0	G:1,0	T:0	G:1,0
rs138861027	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,98	C:0,02	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	C:0	T:1,0	C:0,03	T:0,97	C:0	T:1,0	C:0	T:1,0
rs184146896	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:1,0	C:0,0	G:0,95	C:0,05	G:1,0	C:0,0	G:1,0	C:0,0	C:0	G:1,0	C:0	G:1,0	C:0	G:1,0	C:0	G:1,0
rs149630995	A:0,97	G:0,03	A:0,93	G:0,07	A:0,92	G:0,08	A:0,93	G:0,08	A:1,0	G:0,0	A:0,98	G:0,02	A:1,0	G:0,0	A:1,0	G:0,0	G:0	A:1,0	G:0	A:1,0	G:0	A:1,0	G:0	A:1,0
rs17128754	T:0,94	C:0,06	T:1,0	C:0,0	T:0,92	C:0,08	T:1,0	C:0,0	T:0,63	C:0,37	T:0,88	C:0,12	T:0,92	C:0,08	T:1,0	C:0,0	C:0	T:1,0	C:0,05	T:0,95	C:0,01	T:0,99	C:0,02	T:0,98
rs116109684	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:0,98	G:0,02	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	G:0	T:1,0	G:0,03	T:0,97	G:0	T:1,0	G:0	T:1,0
rs546784	T:0,94	C:0,06	T:0,82	C:0,18	T:0,92	C:0,08	T:0,85	C:0,15	T:0,82	C:0,18	T:0,58	C:0,42	T:0,88	C:0,13	T:0,95	C:0,05	T:0,44	C:0,56	C:0,15	T:0,85	C:0,16	T:0,84	C:0,36	T:0,64
rs2312065	C:0,97	T:0,03	C:0,93	T:0,07	C:0,79	T:0,21	C:0,85	T:0,15	C:0,96	T:0,04	C:0,95	T:0,05	C:0,88	T:0,13	C:1,0	T:0,0	T:0,02	C:0,98	T:0,19	C:0,81	T:0,34	C:0,66	T:0,11	C:0,89
1:66763253	G:1,0	T:0,0	G:1,0	T:0,0	G:0,96	T:0,04	G:0,98	T:0,03	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0
rs11208829	C:0,97	A:0,03	C:0,93	A:0,07	C:0,79	A:0,21	C:0,85	A:0,15	C:0,98	A:0,02	C:0,95	A:0,05	C:0,88	A:0,13	C:1,0	A:0,0	A:0,02	C:0,98	A:0,09	C:0,91	A:0,23	C:0,77	A:0,06	C:0,94
rs7530444	T:0,97	A:0,03	T:1,0	A:0,0	T:0,93	A:0,07	T:1,0	A:0,0	T:0,88	A:0,12	T:0,97	A:0,03	T:0,96	A:0,04	T:1,0	A:0,0	A:0	T:1,0	A:0,05	T:0,95	A:0,01	T:0,99	A:0,02	T:0,98
rs7518325	C:0,97	T:0,03	C:1,0	T:0,0	C:0,93	T:0,07	C:1,0	T:0,0	C:0,88	T:0,12	C:0,97	T:0,03	C:0,96	T:0,04	C:1,0	T:0,0	T:0	C:1,0	T:0,05	C:0,95	T:0,01	C:0,99	T:0,02	C:0,98
rs149945219	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:0,97	A:0,03	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	A:0	T:1,0	A:0,03	T:0,97	A:0	T:1,0	A:0	T:1,0
rs511983	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,9	T:0,1	C:1,0	T:0,0	C:0,9	T:0,1	T:0	C:1,0	T:0,25	C:0,75	T:0,02	C:0,98	T:0	C:1,0
1:66764030	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,98	T:0,03	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0
rs140585110	T:1,0	C:0,0	T:0,89	C:0,11	T:0,98	C:0,02	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	C:0,01	T:0,99	C:0	T:1,0	C:0	T:1,0	C:0	T:1,0
rs509363	A:0,84	G:0,16	A:0,75	G:0,25	A:0,63	G:0,38	A:0,7	G:0,3	A:0,46	G:0,54	A:0,52	G:0,48	A:0,54	G:0,46	A:0,85	G:0,15	A:0,37	G:0,63	A:0,27	G:0,73	A:0,47	G:0,53	G:0,49	A:0,51
rs17128763	G:0,94	T:0,06	G:1,0	T:0,0	G:0,92	T:0,08	G:1,0	T:0,0	G:0,61	T:0,39	G:0,88	T:0,13	G:0,92	T:0,08	G:1,0	T:0,0	T:0	G:1,0	T:0,05	G:0,95	T:0,01	G:0,99	T:0,02	G:0,98
rs2064707	C:1,0	T:0,0	C:0,94	T:0,06	C:0,81	T:0,19	C:0,79	T:0,21	C:1,0	T:0,0	C:0,97	T:0,03	C:0,83	T:0,17	C:1,0	T:0,0	T:0,02	C:0,98	T:0,09	C:0,91	T:0,23	C:0,77	T:0,06	C:0,94
rs482097	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,93	C:0,07	T:1,0	C:0,0	T:0,89	C:0,11	C:0	T:1,0	C:0,25	T:0,75	C:0,02	T:0,98	C:0	T:1,0
rs1120163	T:1,0	C:0,0	T:1,0	C:0,0	T:0,92	C:0,08	T:1,0	C:0,0	T:0,64	C:0,36	T:0,93	C:0,07	T:0,91	C:0,09	T:1,0	C:0,0	C:0	T:1,0	C:0,05	T:0,95	C:0,01	T:0,99	C:0,02	T:0,98
rs558325	C:0,96	G:0,04	C:0,94	G:0,06	C:0,88	G:0,12	C:0,91	G:0,09	C:1,0	G:0,0	C:1,0	G:0,0	C:0,79	G:0,21	C:1,0	G:0,0	G:0,06	C:0,94	G:0,49	C:0,51	G:0,36	C:0,64	G:0,11	C:0,89
rs6683604	T:0,94	C:0,06	T:0,82	C:0,18	T:0,92	C:0,08	T:0,85	C:0,15	T:0,87	C:0,13	T:0,56	C:0,44	T:0,88	C:0,13	T:0,95	C:0,05	T:0,44	C:0,56	C:0,15	T:0,85	C:0,16	T:0,84	C:0,36	T:0,64
1:66765840	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,98	T:0,03	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0
rs12137080	C:0,88	T:0,13	C:0,82	T:0,18	C:0,83	T:0,17	C:0,85	T:0,15	C:0,47	T:0,53	C:0,55	T:0,45	C:0,79	T:0,21	C:0,95	T:0,05	C:0,44	T:0,56	T:0,21	C:0,79	T:0,17	C:0,83	T:0,38	C:0,62

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
rs524770	A:0,84	G:0,18	A:0,74	G:0,26	A:0,66	G:0,34	A:0,7	G:0,3	A:0,44	G:0,56	A:0,48	G:0,52	A:0,54	G:0,46	A:0,85	G:0,15	A:0,37	G:0,63	A:0,27	G:0,73	A:0,47	G:0,53	G:0,49	A:0,51
rs12137115	C:0,94	G:0,06	C:0,81	G:0,19	C:0,91	G:0,09	C:0,85	G:0,15	C:0,81	G:0,19	C:0,6	G:0,4	C:0,88	G:0,13	C:0,95	G:0,05	C:0,44	G:0,56	G:0,15	C:0,85	G:0,16	C:0,84	G:0,36	C:0,64
rs500541	A:0,06	C:0,94	A:0,25	C:0,75	A:0,1	C:0,9	A:0,15	C:0,85	A:0,14	C:0,86	A:0,32	C:0,68	A:0,21	C:0,79	A:0,35	C:0,65	A:0,21	C:0,79	A:0,01	C:0,99	A:0,05	C:0,95	A:0,14	C:0,86
rs499170	C:0,97	T:0,03	C:0,93	T:0,07	C:0,79	T:0,21	C:0,85	T:0,15	C:0,91	T:0,09	C:0,78	T:0,23	C:0,75	T:0,25	C:0,9	T:0,1	T:0,06	C:0,94	T:0,49	C:0,51	T:0,36	C:0,64	T:0,1	C:0,9
rs498448	C:0,84	T:0,16	C:0,75	T:0,25	C:0,63	T:0,38	C:0,7	T:0,3	C:0,48	T:0,52	C:0,52	T:0,48	C:0,54	T:0,46	C:0,85	T:0,15	C:0,37	T:0,63	C:0,25	T:0,75	C:0,47	T:0,53	T:0,49	C:0,51
rs498393	A:0,84	G:0,16	A:0,75	G:0,25	A:0,63	G:0,38	A:0,7	G:0,3	A:0,45	G:0,55	A:0,51	G:0,49	A:0,54	G:0,46	A:0,85	G:0,15	A:0,37	G:0,63	A:0,25	G:0,75	A:0,47	G:0,53	G:0,49	A:0,51
rs1040717	G:1,0	T:0,0	G:1,0	T:0,0	G:0,98	T:0,02	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	G:1,0	T:0,0	T:0	G:1,0	T:0	G:1,0	T:0	G:1,0	T:0	G:1,0
rs495477	A:0,94	G:0,06	A:0,82	G:0,18	A:0,92	G:0,08	A:0,86	G:0,14	A:0,94	G:0,06	A:0,67	G:0,33	A:0,87	G:0,13	A:0,95	G:0,05	A:0,44	G:0,56	G:0,14	A:0,86	G:0,16	A:0,84	G:0,36	A:0,64
1:66767159	A:0,94	T:0,06	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0
rs494735	T:0,92	C:0,08	T:0,78	C:0,22	T:0,84	C:0,16	T:0,87	C:0,13	T:0,67	C:0,33	T:0,57	C:0,43	T:0,81	C:0,19	T:1,0	C:0,0	T:0,44	C:0,56	C:0,21	T:0,79	C:0,17	T:0,83	C:0,38	T:0,62
rs115006563	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	A:0	G:1,0	A:0,01	G:0,99	A:0	G:1,0	A:0	G:1,0
rs12030788	A:0,97	G:0,03	A:0,93	G:0,07	A:0,79	G:0,21	A:0,85	G:0,15	A:0,95	G:0,05	A:0,95	G:0,05	A:0,88	G:0,13	A:1,0	G:0,0	G:0,02	A:0,98	G:0,07	A:0,93	G:0,23	A:0,77	G:0,06	A:0,94
1:66768353	C:1,0	G:0,0	C:0,93	G:0,07	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0
rs190491031	C:0,88	A:0,13	C:1,0	A:0,0	C:0,92	A:0,08	C:0,95	A:0,05	C:1,0	A:0,0	C:0,95	A:0,05	C:0,96	A:0,04	C:1,0	A:0,0	A:0	C:1,0	A:0	C:1,0	A:0	C:1,0	A:0	C:1,0
1:66768834	C:1,0	T:0,0	C:1,0	T:0,0	C:0,98	T:0,02	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0
rs199795423	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,97	C:0,03	T:1,0	C:0,0	T:0,89	C:0,11	C:0,01	T:0,99	C:0,26	T:0,74	C:0,02	T:0,98	C:0	T:1,0
rs79326238	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,9	C:0,1	T:1,0	C:0,0	T:0,89	C:0,11	C:0,11	T:0,89	C:0,33	T:0,67	C:0,13	T:0,87	C:0,1	T:0,9
rs6683977	G:0,94	C:0,06	G:0,82	C:0,18	G:0,92	C:0,08	G:0,88	C:0,13	G:0,93	C:0,08	G:0,66	C:0,34	G:0,88	C:0,13	G:0,95	C:0,05	G:0,44	C:0,56	C:0,04	G:0,96	C:0,16	G:0,84	C:0,35	G:0,65
rs782686	A:0,84	G:0,16	A:0,75	G:0,25	A:0,63	G:0,38	A:0,73	G:0,28	A:0,44	G:0,56	A:0,53	G:0,48	A:0,54	G:0,46	A:0,85	G:0,15	A:0,37	G:0,63	A:0,25	G:0,75	A:0,47	G:0,53	G:0,49	A:0,51
1:66769338	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:0,96	G:0,04	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0	T:1,0	G:0,0
rs72660610	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:0,88	G:0,13	C:1,0	G:0,0	G:0,03	C:0,97	G:0,01	C:0,99	G:0	C:1,0	G:0	C:1,0
rs492589	T:0,84	C:0,16	T:0,75	C:0,25	T:0,63	C:0,38	T:0,7	C:0,3	T:0,44	C:0,56	T:0,51	C:0,49	T:0,54	C:0,46	T:0,85	C:0,15	T:0,37	C:0,63	T:0,27	C:0,73	T:0,47	C:0,53	C:0,49	T:0,51
1:66770420	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0
1:66770718	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:0,98	A:0,03	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0
rs142784821	T:1,0	C:0,0	T:1,0	C:0,0	T:0,94	C:0,06	T:1,0	C:0,0	T:0,77	C:0,23	T:0,97	C:0,03	T:0,94	C:0,06	T:1,0	C:0,0	C:0	T:1,0	C:0,05	T:0,95	C:0,01	T:0,99	C:0,02	T:0,98
rs116036705	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:0,95	A:0,05	T:1,0	A:0,0	T:0,97	A:0,03	T:0,96	A:0,04	T:1,0	C:0,0	T:1,0	C:0,0

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
1:66771378	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,98	C:0,02	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0
rs611838	T:0,86	C:0,14	T:0,78	C:0,23	T:0,57	C:0,43	T:0,77	C:0,23	T:0,58	C:0,42	T:0,47	C:0,53	T:0,57	C:0,43	T:0,94	C:0,06	T:0,38	C:0,62	T:0,27	C:0,73	T:0,47	C:0,53	C:0,49	T:0,51
1:66771520	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0
rs72674130	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:0,97	G:0,03	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	G:0,01	C:0,99	G:0	C:1,0	G:0	C:1,0	G:0,01	C:0,99
rs17128782	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,87	T:0,13	C:1,0	T:0,0	T:0,03	C:0,97	T:0,05	C:0,95	T:0	C:1,0	T:0	C:1,0
rs72674131	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:1,0	G:0,0	C:0,87	G:0,13	C:1,0	G:0,0	G:0,03	C:0,97	G:0,05	C:0,95	G:0	C:1,0	G:0	C:1,0
rs146585449	T:1,0	C:0,0	T:1,0	C:0,0	T:0,98	C:0,02	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	C:0	T:1,0	C:0	T:1,0	C:0,01	T:0,99	C:0	T:1,0
rs563034	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:1,0	C:0,0	A:0,89	C:0,11	A:1,0	C:0,0	A:0,9	C:0,1	C:0	A:1,0	C:0,25	A:0,75	C:0,02	A:0,98	C:0	A:1,0
rs115437344	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,98	A:0,02	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	A:0	G:1,0	A:0,03	G:0,97	A:0	G:1,0	A:0	G:1,0
rs538336	C:0,84	T:0,16	C:0,75	T:0,25	C:0,63	T:0,38	C:0,7	T:0,3	C:0,45	T:0,55	C:0,5	T:0,5	C:0,54	T:0,46	C:0,85	T:0,15	C:0,38	T:0,62	C:0,27	T:0,73	C:0,46	T:0,54	T:0,49	C:0,51
rs638111	C:0,91	T:0,09	C:0,75	T:0,25	C:0,71	T:0,29	C:0,7	T:0,3	C:0,76	T:0,24	C:0,52	T:0,48	C:0,75	T:0,25	C:0,95	T:0,05	C:0,42	T:0,58	T:0,31	C:0,69	C:0,49	T:0,51	T:0,46	C:0,54
rs537611	C:0,84	T:0,16	C:0,75	T:0,25	C:0,63	T:0,38	C:0,7	T:0,3	C:0,49	T:0,51	C:0,5	T:0,5	C:0,54	T:0,46	C:0,85	T:0,15	C:0,38	T:0,62	C:0,29	T:0,71	C:0,47	T:0,53	T:0,49	C:0,51
rs536858	G:0,96	C:0,04	G:0,8	C:0,2	G:0,65	C:0,35	G:0,78	C:0,22	G:0,5	C:0,5	G:0,58	C:0,42	G:0,63	C:0,38	G:0,82	C:0,18	G:0,38	C:0,62	G:0,27	C:0,73	G:0,46	C:0,54	C:0,49	G:0,51
rs639911	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:0,89	C:0,11	C:0	T:1,0	C:0,25	T:0,75	C:0,02	T:0,98	C:0	T:1,0
rs532976	T:0,84	G:0,16	T:0,77	G:0,23	T:0,63	G:0,38	T:0,7	G:0,3	T:0,42	G:0,58	T:0,5	G:0,5	T:0,54	G:0,46	T:0,85	G:0,15	T:0,38	G:0,62	T:0,25	G:0,75	T:0,46	G:0,54	G:0,49	T:0,51
rs641262	T:0,94	C:0,06	T:0,82	C:0,18	T:0,92	C:0,08	T:0,85	C:0,15	T:0,85	C:0,15	T:0,62	C:0,38	T:0,88	C:0,13	T:0,95	C:0,05	T:0,44	C:0,56	C:0,14	T:0,86	C:0,16	T:0,84	C:0,3	T:0,7
rs17128784	C:0,78	T:0,22	C:0,84	T:0,16	C:0,7	T:0,3	C:0,88	T:0,13	C:0,66	T:0,34	C:0,63	T:0,37	C:0,79	T:0,21	C:0,8	T:0,2	T:0,04	C:0,96	T:0,37	C:0,63	T:0,31	C:0,69	T:0,2	C:0,8
rs76791470	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,95	A:0,05	G:1,0	A:0,0	G:1,0	A:0,0	A:0	G:1,0	A:0,12	G:0,88	A:0	G:1,0	A:0	G:1,0
rs558550	T:0,29	A:0,71	T:0,59	A:0,41	T:0,74	A:0,26	T:0,68	A:0,33	T:0,49	A:0,51	T:0,54	A:0,46	T:0,74	A:0,26	T:0,95	A:0,05	A:0,39	T:0,61	A:0,23	T:0,77	A:0,19	T:0,81	A:0,28	T:0,72
rs146254634	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	A:0	G:1,0	A:0,03	G:0,97	A:0	G:1,0	A:0	G:1,0
rs7514592	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:0,96	T:0,04	A:0,79	T:0,21	A:1,0	T:0,0	A:0,9	T:0,1	T:0,07	A:0,93	T:0,25	A:0,75	T:0,03	A:0,97	T:0,19	A:0,81
rs7514666	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:1,0	G:0,0	A:0,91	G:0,09	A:0,83	G:0,18	A:1,0	G:0,0	A:0,9	G:0,1	G:0,07	A:0,93	G:0,25	A:0,75	G:0,03	A:0,97	G:0,19	A:0,81
1:66775071	T:1,0	C:0,0	T:0,93	C:0,07	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0	T:1,0	C:0,0
rs6689595	C:0,74	T:0,26	C:0,56	T:0,44	C:0,38	T:0,62	C:0,41	T:0,59	C:0,53	T:0,47	C:0,54	T:0,46	C:0,38	T:0,63	C:0,11	T:0,89	C:0,47	T:0,53	T:0,22	C:0,78	C:0,2	T:0,8	C:0,48	T:0,52
rs6665325	A:0,74	G:0,26	A:0,56	G:0,44	A:0,38	G:0,62	A:0,4	G:0,6	A:0,53	G:0,47	A:0,51	G:0,49	A:0,5	G:0,5	A:0,15	G:0,85	G:0,44	A:0,56	G:0,08	A:0,92	A:0,2	G:0,8	A:0,49	G:0,51
1:66776450	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,98	T:0,02	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:1,0	T:0,0	C:0,998	T:0,002	C:1,0	T:0,0	C:1,0	T:0,0

ID / pop	ASH		MAC		AYM		QUE		GUA		TUP		HUI		TAH		EUR		AFR		EAS		SAS	
rs80343528	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:1,0	A:0,0	C:0,88	A:0,13	C:1,0	A:0,0	A:0,01	C:0,99	A:0	C:1,0	A:0	C:1,0	A:0	C:1,0
rs72926316	C:0,91	T:0,09	C:1,0	T:0,0	C:0,9	T:0,1	C:0,83	T:0,18	C:0,53	T:0,47	C:0,82	T:0,18	C:0,88	T:0,13	C:1,0	T:0,0	T:0,01	C:0,99	T:0,1	C:0,9	T:0,08	C:0,92	T:0,04	C:0,96
rs6690590	G:0,75	A:0,25	G:0,57	A:0,43	G:0,38	A:0,63	G:0,4	A:0,6	G:0,53	A:0,47	G:0,52	A:0,48	G:0,38	A:0,63	G:0,15	A:0,85	G:0,46	A:0,54	A:0,29	G:0,71	G:0,19	A:0,81	G:0,48	A:0,52
1:66777089	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:0,88	A:0,13	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0	T:1,0	A:0,0
rs17427278	C:0,97	T:0,03	C:0,75	T:0,25	C:0,9	T:0,1	C:0,88	T:0,13	C:0,87	T:0,13	C:0,97	T:0,03	C:1,0	T:0,0	C:0,85	T:0,15	T:0,01	C:0,99	T:0	C:1,0	T:0,02	C:0,98	T:0	C:1,0
rs151097582	G:0,91	A:0,09	G:1,0	A:0,0	G:0,9	A:0,1	G:0,83	A:0,18	G:0,61	A:0,39	G:0,95	A:0,05	G:0,92	A:0,08	G:1,0	A:0,0	A:0	G:1,0	A:0	G:1,0	A:0	G:1,0	A:0	G:1,0
rs1040716	A:0,75	T:0,25	A:0,82	T:0,18	A:0,38	T:0,63	A:0,4	T:0,6	A:0,52	T:0,48	A:0,52	T:0,48	A:0,5	T:0,5	A:0,15	T:0,85	T:0,44	A:0,56	T:0,08	A:0,92	A:0,2	T:0,8	A:0,49	T:0,51
rs2180336	C:0,75	T:0,25	C:0,82	T:0,18	C:0,39	T:0,61	C:0,39	T:0,61	C:0,56	T:0,44	C:0,52	T:0,48	C:0,5	T:0,5	C:0,15	T:0,85	T:0,44	C:0,56	T:0,08	C:0,92	C:0,2	T:0,8	C:0,49	T:0,51
rs4655834	C:0,75	T:0,25	C:0,82	T:0,18	C:0,38	T:0,63	C:0,4	T:0,6	C:0,52	T:0,48	C:0,5	T:0,5	C:0,38	T:0,63	C:0,05	T:0,95	C:0,39	T:0,61	C:0,37	T:0,63	C:0,17	T:0,83	C:0,3	T:0,7
rs4655605	A:0,75	G:0,25	A:0,81	G:0,19	A:0,37	G:0,63	A:0,4	G:0,6	A:0,52	G:0,48	A:0,5	G:0,5	A:0,35	G:0,65	A:0,05	G:0,95	A:0,39	G:0,61	A:0,37	G:0,63	A:0,17	G:0,83	A:0,3	G:0,7
rs374463121	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:0,9	A:0,1	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0	G:1,0	A:0,0
rs6697215	G:0,78	T:0,22	G:0,87	T:0,13	G:0,44	T:0,56	G:0,32	T:0,68	G:0,65	T:0,35	G:0,78	T:0,22	G:0,39	T:0,61	G:0,13	T:0,88	G:0,47	T:0,53	T:0,22	G:0,78	G:0,2	T:0,8	G:0,48	T:0,52
rs13375308	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:1,0	T:0,0	A:0,93	T:0,07	A:1,0	T:0,0	A:1,0	T:0,0	T:0	A:1,0	T:0,05	A:0,95	T:0	A:1,0	T:0	A:1,0

Tabela Apêndice 4. Valores de F_{CT} para cada uma das variantes do banco de dados TargetSeq do gene *PDE4B* (posições genômicas 66.759.100 a 66.779.100 - GRCh37) calculados par-a-par entre as populações que compõem o banco de dados TargetSeq**.** Destaque (cinza) para os SNPs rs546784, rs524770, rs6683977, rs641262

	EUR x NAT		**AFR x NAT**		**SAS x NAT**		**EAS x NAT**	
	F_{CT}	**p-value**	**F_{CT}**	**p-value**	**F_{CT}**	**p-value**	**F_{CT}**	**p-value**
rs782967	variante com muitos genótipos faltantes							
rs111886642	0,008	0,024	0,019	0,000	-0,004	0,028	-0,004	0,028
rs6668516	0,296	0,005	-0,005	0,655	0,113	0,012	0,002	0,278
1:66760010	0,003	0,022	0,005	0,004	0,003	0,033	0,003	0,029
1:66760022	0,003	0,026	0,005	0,000	0,003	0,021	0,003	0,032
rs17128742	0,008	0,027	0,019	0,000	-0,001	0,032	-0,001	0,018
rs17128748	0,008	0,026	0,019	0,001	-0,002	0,025	-0,001	0,022
rs56163035	0,046	0,021	0,370	0,000	-0,006	0,202	0,246	0,023
rs599381	0,005	0,185	0,252	0,001	-0,006	0,532	0,046	0,018
rs143612526	0,013	0,016	0,012	0,008	0,013	0,018	0,007	0,030
1:66760675	variante com muitos genótipos faltantes							
rs72677260	0,021	0,001	0,031	0,001				
rs138861027	0,003	0,023	0,016	0,000	0,003	0,019	0,003	0,030
rs184146896	0,006	0,024	0,011	0,002	0,006	0,023	0,006	0,029
rs149630995	0,077	0,017	0,096	0,001	0,076	0,018	0,053	0,023
rs17128754	0,173	0,016	0,011	0,012	0,059	0,034	0,130	0,029
rs116109684	0,003	0,028	0,016	0,001	0,003	0,023	0,003	0,028
rs546784	0,249	0,014	-0,004	0,355	0,069	0,015	-0,006	0,551
rs2312065	0,053	0,015	0,029	0,004	-0,001	0,338	0,141	0,002
1:66763253	0,016	0,017	0,022	0,000	0,016	0,021	0,016	0,019
rs11208829	0,057	0,022	-0,005	0,758	0,001	0,133	0,056	0,013
rs7530444	0,023	0,018	0,013	0,013	-0,001	0,300	-0,003	0,140
rs7518325	0,023	0,022	0,013	0,010	-0,001	0,279	-0,003	0,148
rs149945219			0,019	0,000				
rs511983	0,006	0,035	0,142	0,000	0,013	0,025	-0,006	0,489
1:66764030	0,002	0,029	0,005	0,001	0,002	0,035	0,002	0,030
rs140585110	0,002	0,059	0,039	0,001	0,013	0,022	0,027	0,028
rs509363	0,132	0,013	0,254	0,003	0,029	0,036	0,053	0,042
rs17128763	0,178	0,028	0,012	0,009	0,061	0,026	0,135	0,031
rs2064707	0,134	0,021	0,000	0,342	0,022	0,019	0,030	0,025
rs482097	-0,008	0,046	0,149	0,000	-0,006	0,029	-0,006	0,386
rs1120163	0,142	0,031	-0,004	0,152	0,026	0,022	0,092	0,027
rs558325	0,005	0,032	0,317	0,001	0,028	0,015	0,209	0,007
rs6683604	0,257	0,009	-0,005	0,475	0,074	0,019	-0,007	0,544
1:66765840	0,002	0,022	0,005	0,000	0,002	0,033	0,002	0,041
rs12137080	0,160	0,004	0,002	0,108	0,025	0,042	0,015	0,041
rs524770	0,139	0,018	0,265	0,000	0,034	0,020	0,059	0,035
rs12137115	0,245	0,010	-0,004	0,354	0,066	0,012	-0,006	0,522

	EUR x NAT		AFR x NAT		SAS x NAT		EAS x NAT	
	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value
rs500541	-0,004	0,525	0,305	0,002	0,003	0,211	0,109	0,017
rs499170	0,037	0,021	0,208	0,001	0,003	0,069	0,103	0,019
rs498448	0,137	0,016	0,284	0,000	0,032	0,030	0,058	0,030
rs498393	0,134	0,021	0,280	0,000	0,030	0,032	0,056	0,032
rs1040717	-0,001	0,379	0,005	0,002	0,003	0,019	-0,002	0,997
rs495477	0,294	0,012	-0,006	0,632	0,106	0,008	-0,001	0,330
1:66767159	0,002	0,023	0,008	0,000	0,002	0,025	0,002	0,022
rs494735	0,200	0,019	-0,006	1,000	0,052	0,015	-0,001	0,324
rs115006563	0,003	0,022	0,000	0,429	0,003	0,034	0,003	0,028
rs12030788	0,068	0,015	0,003	0,192	0,006	0,057	0,052	0,010
1:66768353	0,009	0,020	0,017	0,002	0,009	0,026	0,009	0,026
rs190491031	0,083	0,017	0,104	0,001	0,081	0,013	0,066	0,017
1:66768834	0,003	0,030	0,005	0,002	0,003	0,035	0,003	0,024
rs199795423	-0,006	0,625	0,155	0,000	0,002	0,044	-0,004	0,506
rs79326238	0,043	0,008	0,202	0,000	0,039	0,013	0,060	0,003
rs6683977	0,287	0,004	0,065	0,008	0,092	0,018	-0,004	0,413
rs782686	0,146	0,014	0,297	0,001	0,038	0,030	0,065	0,029
1:66769338	-0,001	0,028	0,002	0,001	-0,001	0,028	-0,001	0,042
rs72660610	0,007	0,023	-0,006	0,528	-0,003	0,026	-0,005	0,031
rs492589	0,123	0,016	0,248	0,001	0,025	0,054	0,049	0,049
1:66770420	0,003	0,036	0,005	0,001	0,003	0,026	0,003	0,022
1:66770718	0,002	0,026	0,005	0,001	0,002	0,037	0,002	0,026
rs142784821	0,185	0,017	-0,008	0,998	0,000	0,010	0,057	0,017
rs116036705	0,017	0,017	0,022	0,000	-0,002	0,036	-0,002	0,028
1:66771378	0,004	0,031	0,006	0,005	0,003	0,034	0,004	0,028
rs611838	0,191	0,015	0,335	0,001	0,076	0,029	0,108	0,022
1:66771520	variante com muitos genótipos faltantes							
1:66771529	variante com muitos genótipos faltantes							
rs72674130	0,009	0,001			0,005	0,083		
rs17128782	0,007	0,029	0,018	0,000	-0,002	0,021	-0,001	0,028
rs72674131	0,007	0,030	0,018	0,000	-0,003	0,030	-0,002	0,027
rs146585449	-0,001	0,415	0,005	0,000	0,003	0,026	0,001	0,391
rs563034	0,006	0,026	0,141	0,000	0,013	0,026	-0,006	0,364
rs115437344	0,003	0,025	0,015	0,000	0,003	0,031	0,003	0,022
rs538336	0,119	0,013	0,252	0,000	0,028	0,043	0,053	0,049
rs638111	0,166	0,010	-0,002	0,370	0,069	0,015	0,103	0,019
rs537611	0,126	0,015	0,227	0,002	0,031	0,032	0,056	0,028
rs536858	0,356	0,010	0,500	0,000	0,223	0,012	0,270	0,014
rs639911	0,000	0,709	0,164	0,000	-0,001	0,959	0,008	0,161
rs532976	0,114	0,009	0,269	0,001	0,025	0,045	0,052	0,044
rs641262	0,253	0,013	-0,004	0,358	0,035	0,059	-0,007	0,579
rs17128784	0,233	0,013	0,023	0,009	0,004	0,160	0,004	0,047

	EUR x NAT		AFR x NAT		SAS x NAT		EAS x NAT	
	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value
rs76791470	0,001	0,055	0,067	0,000	0,006	0,033	0,007	0,022
rs558550	-0,007	0,426	0,060	0,000	0,022	0,019	0,098	0,011
rs146254634			0,014	0,001				
rs7514592	0,002	0,093	0,115	0,000	0,074	0,013	-0,009	0,401
rs7514666	0,002	0,150	0,114	0,000	0,074	0,012	-0,007	0,365
1:66775071	0,009	0,035	0,017	0,000	0,009	0,024	0,009	0,029
rs6689595	-0,006	0,469	0,211	0,001	-0,005	0,326	0,162	0,016
rs6665325	0,008	0,031	0,461	0,000	-0,004	0,490	0,182	0,010
1:66776450	0,003	0,032	0,005	0,001	0,003	0,026	0,003	0,029
rs80343528	-0,008	0,501	-0,002	0,001	-0,005	0,034	-0,002	0,032
rs72926316	0,220	0,022	0,011	0,085	0,084	0,022	0,031	0,026
rs6690590	-0,005	0,466	0,111	0,000	-0,005	0,430	0,178	0,011
1:66777089	-0,002	0,039	0,007	0,003	-0,003	0,031	-0,002	0,035
rs17427278	0,159	0,013	0,254	0,001	0,183	0,018	0,082	0,020
rs151097582	0,219	0,029	0,268	0,000	0,207	0,018	0,204	0,025
rs1040716	-0,004	0,224	0,414	0,000	-0,005	0,235	0,219	0,017
rs2180336	-0,006	0,259	0,403	0,000	-0,003	0,175	0,232	0,022
rs4655834	0,015	0,042	0,029	0,002	0,071	0,018	0,235	0,016
rs4655605	0,013	0,052	0,028	0,001	0,068	0,021	0,232	0,018
rs374463121	-0,005	0,030	0,002	0,001	-0,005	0,027	-0,005	0,027
rs6697215	-0,005	0,126	0,141	0,000	-0,007	0,075	0,234	0,025
1:66778815	variante com muitos genótipos faltantes							
rs13375308	0,008	0,026	0,017	0,024	0,008	0,034	0,008	0,024

Tabela Apêndice 5. Valores de F_{CT} para cada uma das variantes do banco de dados TargetSeq do gene *MYT1L* (2.215.144 a 2.235.144 - GRCh37) calculados par-a-par entre as populações que compõem o banco de dados TargetSeq. Destaque (cinza) para o SNP rs17039396

	EUR x NAT		AFR x NAT		SAS x NAT		EAS x NAT	
	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value
rs80116641	variante fixada nas populações NAT							
2:2215590	0,096	0,008	0,126	0,004	0,094	0,014	0,097	0,016
2:2215876	0,005	0,034	0,010	0,001	0,005	0,035	0,005	0,031
2:2216078	0,007	0,038	0,011	0,000	0,007	0,033	0,007	0,039
rs1582415	0,025	0,018	0,017	0,000	-0,007	0,862	0,002	0,288
2:2216359	0,032	0,026	0,047	0,002	0,031	0,022	0,032	0,024
rs1582414	0,029	0,011	0,020	0,000	-0,003	0,880	0,006	0,244
2:2217073	0,003	0,025	0,005	0,002	0,003	0,024	0,003	0,028
rs58275964	0,012	0,032	0,005	0,040	0,001	0,138	-0,004	0,513
rs78173916	0,048	0,021	-0,004	0,593	0,109	0,016	0,000	0,299
rs12478611	0,079	0,002	0,032	0,000	0,028	0,005	0,050	0,001

	EUR x NAT		AFR x NAT		SAS x NAT		EAS x NAT	
	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value	F_{CT}	p-value
rs10865534	0,062	0,008	0,017	0,012	0,013	0,018	0,033	0,014
rs73177806	0,076	0,001	0,111	0,001	0,136	0,000	0,021	0,010
rs1592732	0,065	0,001	0,132	0,000	0,013	0,014	0,034	0,014
rs72769228	0,087	0,000	0,038	0,000	0,035	0,000	0,057	0,001
rs60144672	0,057	0,017	0,186	0,000	0,056	0,021	-0,009	0,412
rs139264841	0,125	0,028	0,164	0,002	0,131	0,018	0,082	0,016
rs72769229	0,016	0,020	-0,004	0,849	-0,001	0,291	-0,002	0,462
rs72769230	0,021	0,020	-0,003	0,548	-0,002	0,245	-0,002	0,422
rs7420242	0,015	0,023	0,009	0,014	-0,001	0,211	-0,003	0,456
rs17039395	0,051	0,014	0,192	0,000	0,052	0,015	-0,007	0,430
2:2222593	-0,001	0,031	0,002	0,000	-0,001	0,033	-0,001	0,027
rs1582413	0,066	0,021	0,211	0,002	0,068	0,020	-0,002	0,297
rs17247310	0,113	0,009	0,218	0,001	0,118	0,009	0,022	0,062
rs113009634	0,014	0,015	0,011	0,002	-0,003	0,766	-0,002	0,531
rs11681135	0,136	0,001	0,031	0,000	0,160	0,001	0,036	0,006
rs60585106	0,102	0,021	0,042	0,001	0,188	0,016	0,008	0,096
rs11127372	0,035	0,010	0,207	0,001	0,146	0,001	0,424	0,004
rs77041749	0,251	0,020	0,604	0,001	0,356	0,020	0,608	0,029
rs1895137	0,038	0,012	0,107	0,001	0,001	0,401	0,266	0,000
rs6731821	0,104	0,004	0,192	0,000	0,085	0,009	0,275	0,001
rs6711203	0,104	0,005	0,199	0,001	0,087	0,005	0,194	0,002
rs6711124	0,102	0,002	0,200	0,000	0,087	0,007	0,197	0,004
rs17039396	0,021	0,036	0,348	0,001	0,054	0,022	0,001	0,105
rs13414323	0,000	0,085	-0,002	0,317	0,069	0,007	0,090	0,014
rs6548123	0,081	0,005	0,185	0,000	0,274	0,000	0,614	0,001
2:2225675	0,003	0,032	0,005	0,001	0,003	0,029	0,003	0,027
rs7569601	0,088	0,006	0,262	0,001	0,275	0,001	0,612	0,002
rs10204613	variante com muitos genótipos faltantes							
rs11127371	0,123	0,012	0,059	0,001	0,186	0,002	0,096	0,020
rs1862114	0,086	0,011	0,181	0,002	0,078	0,013	0,186	0,000
rs1862113	0,072	0,012	0,237	0,000	0,251	0,000	0,599	0,002
rs10865533	0,091	0,007	0,208	0,000	0,066	0,013	0,215	0,007
rs12052817	0,019	0,007	0,013	0,003	0,029	0,001	0,002	0,179
rs888627	0,102	0,007	0,134	0,000	0,075	0,003	0,194	0,000
rs7566768	0,089	0,006	0,175	0,000	0,074	0,011	0,188	0,002
rs114751773	0,003	0,030	0,021	0,001	0,003	0,023	0,003	0,027
rs11900037	0,083	0,011	0,370	0,001	0,274	0,013	0,709	0,000
2:2228435	0,009	0,035	0,014	0,002	0,009	0,020	0,009	0,025
rs1808360	variante com muitos genótipos faltantes							
rs13413284	variante com muitos genótipos faltantes							
2:2228993	0,003	0,028	0,005	0,001	0,003	0,023	0,003	0,026
rs7597220	0,065	0,014	0,310	0,001	0,223	0,010	0,645	0,004

	EUR x NAT		AFR x NAT		SAS x NAT		EAS x NAT	
	F_{CT}	**p-value**	**F_{CT}**	**p-value**	**F_{CT}**	**p-value**	**F_{CT}**	**p-value**
rs7607555	0,054	0,016	0,008	0,040	0,030	0,021	0,171	0,024
rs9309702	0,010	0,049	0,362	0,000	0,169	0,009	-0,005	0,577
rs140595851	-0,005	0,027	0,002	0,000	-0,005	0,033	-0,007	0,336
rs75825920	0,053	0,020	0,070	0,002	0,051	0,014	0,001	0,108
rs142858460			0,025	0,000				
rs35633741	0,112	0,001	-0,006	0,565	0,088	0,002	0,035	0,021
rs185850219	0,018	0,019	0,025	0,000	0,018	0,019	0,010	0,029
rs17039409	0,002	0,028	0,024	0,001	0,002	0,024	0,002	0,027
rs1035352	0,636	0,004	0,701	0,000	0,606	0,002	0,665	0,004
2:2231954	0,003	0,029	0,005	0,003	0,003	0,035	0,003	0,019
rs1035351	variante com muitos genótipos faltantes							
rs1465388	0,105	0,004	0,024	0,008	0,012	0,088	0,066	0,023
rs10167537	variante com muitos genótipos faltantes							
rs10167441	variante com muitos genótipos faltantes							
rs10184190	0,168	0,011	0,022	0,009	0,108	0,011	0,007	0,040
rs140247942	-0,004	0,190	0,000	0,382	-0,003	0,030	-0,003	0,025
2:2233867	0,003	0,027	0,008	0,001	0,003	0,028	0,003	0,025
rs79018269	0,233	0,013	0,238	0,000	0,205	0,016	-0,003	0,563
rs17338428	0,036	0,012	0,028	0,002	0,067	0,014	-0,001	0,372
rs4853947	0,071	0,018	-0,005	0,669	0,021	0,024	0,016	0,036
rs148331727	0,064	0,022	0,092	0,001	0,037	0,013	0,065	0,021
2:2235031	0,001	0,030	0,004	0,000	0,001	0,025	0,001	0,031

CAPÍTULO 3. PERSPECTIVAS SOBRE A CONTRIBUIÇÃO DA INFERÊNCIA DA ANCESTRALIDADE LOCAL PARA A IDENTIFICAÇÃO DE SINAIS DE SELEÇÃO POLIGÊNICA

INTRODUÇÃO

A história da dinâmica populacional da espécie humana descreve a existência de diversos eventos de migração e a ocorrência de isolamento populacional durante o povoamento dos continentes (HELLENTHAL *et al.* 2014, NIELSEN *et al.* 2017). Desde o surgimento de populações que se espalharam pela África e que posteriormente alcançaram outras regiões do globo, grupos de indivíduos da espécie humana se isolaram e passaram a habitar diferentes ambientes, caracterizados por distintas condições de sobrevivência (PICKRELL & REICH, 2014; NIELSEN *et al.* 2017). A exposição a pressões seletivas diversas e o isolamento populacional, o qual acarreta na atuação de importantes forças evolutivas, influenciaram de forma distinta as variações genéticas que surgiram ao longo do tempo nesses indivíduos (MARTH *et al*. 2004, JEONG & DI RIENZO, 2014). Dessa forma, as variações do genoma humano foram herdadas diferencialmente nas populações que habitaram distintos ambientes e/ou que permaneceram isoladas umas das outras.

Inúmeros estudos foram e ainda são conduzidos com o objetivo de compreender os mecanismos e as bases genéticas da adaptação em populações humanas. Tradicionalmente os métodos e abordagens utilizados eram (e muitos ainda são) restritos à busca de sinais em locos isolados e independentes do genoma que estão relacionados a fenótipos com características mendelianas clássicas (TISHKOFF *et al.* 2007, QUINTANA-MURCI & BARREIRO, 2010). Além disso, a grande maioria desses estudos ainda foca apenas em sinais de seleção positiva (DING *et al.* 2002, VOIGHT *et al.* 2006, LUISI *et al.* 2015). Porém, a evolução com base na seleção natural não se resume apenas a esses locos e no modo positivo de vantagem adaptativa. Ela se mostra consideravelmente mais complexa, com outros modos de seleção natural, tais como a seleção negativa e seleção balanceadora, os quais podem não apenas atuar em locos únicos, mas em vários locos concomitantemente (BAMSHAD & WOODING 2003, FAN *et al*. 2016). Sabe-se que a seleção natural opera especialmente sobre características fenotípicas que na maioria das vezes são consequências dos produtos de diversos genes, que interagem em redes complexas moleculares que levam ao fenótipo (STRANGER *et al*. 2011, DAUB et *al*. 2013) . Portanto, mecanismos de inter-relação gênica,

como os conhecidos efeitos epistáticos e pleiotrópicos, também compõem o cenário de atuação da seleção natural e seus papéis sobre os fenótipos devem ser considerados quando da análise das consequências das pressões ambientais sobre a diversidade genética existente (SHAO *et al.* 2008, SOHAIL *et al.* 2017).

Classicamente, os estudos focados na busca pelos efeitos da seleção natural ao longo das gerações se restringem basicamente a dois campos de estudos diferentes (NADEAU & JIGGINS 2010): (1) estudos genético-populacionais, que ao negligenciarem os dados fenotípicos buscam na análise de variantes genéticas sinais de seleção natural em genes a fim de encontrarem o efeito de uma variante específica sobre certa característica fenotípica pré-determinada, e (2) estudos de genética quantitativa (QTLs – *quantitative trait loci*) que, ao assumirem que as características de interesse são consequências de inúmeros locos diferentes, se baseiam puramente em diferenças fenotípicas entre aparentados ao longo das gerações para compreenderem a forma de segregação das características observadas.

Nas últimas décadas, com os avanços tecnológicos que permitiram a geração de dados genéticos massivos, novas perspectivas para os estudos focados na busca por sinais de seleção natural passaram a serem traçadas. Os primeiros estudos de associação por todo o genoma (GWAS) demonstraram que as características fenotípicas estão relacionadas a diversas variantes, espalhadas por múltiplas regiões do genoma (FAN *et al.* 2016). Dessa forma, atualmente, com a facilidade de se acessar dados genômicos, as inter-relações entre as variações fenotípicas observadas e suas bases genéticas podem teoricamente ser estabelecidas e estudadas com mais amplitude. Porém, a grande complexidade das redes de interação formada entre os genes, suas variantes, os fenótipos resultantes e as pressões seletivas torna a identificação desses sinais de seleção uma tarefa bastante difícil (BALARESQUE *et al.* 2007, WRAY *et al*. 2013).

Os estudos de GWAS também vêm mostrando que modelos de seleção natural clássicos baseados na busca pelos efeitos pronunciados de um único loco ou gene sobre as características fenotípicas sob seleção são na verdade pequena parcela dos casos (MANOLIO *et al.* 2009). A grande maioria das bases genéticas dos fenótipos adaptativos estão relacionadas a leves mudanças nas frequências de diversas variantes localizadas em múltiplos genes (HANCOCK *et al.* 2010, PRITCHARD & DI RIENZO 2010, STRANGER *et al.* 2011, SCHRIDER *et al.* 2017). As evidências de que os eventos adaptativos ocorrem principalmente pelas pequenas mudanças de frequências em variantes de diversos genes despertaram o interesse pela busca de sinais poligênicos de seleção. Porém, as suaves variações de frequência em vários locos são metodologicamente muito difíceis de serem

diferenciadas das alterações que ocorrem em razão de eventos aleatórios, como os que direcionam as alterações de frequências alélicas em locos neutros do genoma (PRITCHARD *et al.* 2010). Esses obstáculos metodológicos motivaram nos últimos anos o desenvolvimento de alternativas para a busca de sinais poligênicos de seleção (GNECCHI-RUSCONE *et al.* 2018), porém os avanços ainda são insuficientes para o alcance de resultados robustos (CSILLÉRY *et al.* 2018).

Uma das dificuldades encontradas para as análises poligênicas, assim como para análises de variantes distribuídas por todo o genoma, se baseia nos problemas estatísticos gerados pelo enorme número de comparações realizadas em um mesmo estudo. Os ajustes que devem ser aplicados nas análises respeitam princípios estatísticos de correção para testes múltiplos, os quais são empregados para evitar resultados falso-positivos (PE'ER *et al.* 2008, MOSKVINA *et al.* 2008). Porém, esses ajustes são na maioria das vezes extremamente restritivos, o que causa a desconsideração de resultados verdadeiro-positivos e leva à necessidade do uso de grupos com maiores números de amostras (JOHNSON *et al.* 2010, BUSH & MOORE 2012). Outras dificuldades encontradas ocorrem em razão da capacidade computacional requerida para as diversas comparações realizadas. Como o número de variantes e amostras em estudos que lidam com informações obtidas por todo o genoma é muito grande, os processos computacionais empregados são dispendiosos com relação ao tempo gasto para as análises e às capacidades computacionais necessárias (tanto memória de acesso aleatório quanto capacidade de processamento) (WAN *et al.* 2010, THOMPSON & CHARNIGO 2015).

O estudo de populações miscigenadas é historicamente reconhecido como uma forma eficiente de se mapear locos ou genes que exercem efeito sobre características humanas de interesse (BUERKLE & LEXER 2008). Dentre as muitas vantagens que esses tipos de estudo oferecem está a redução do número de comparações realizadas em estudos de mapeamento por mistura. Assim, uma das alternativas para a busca de sinais poligênicos de seleção natural pode se dar pela exploração dos dados de populações miscigenadas, as quais são ainda pouco analisadas apesar de seu potencial em facilitar estudos de evolução poligênica (JEONG *et al.* 2014, RACIMO *et al.* 2018).

Diversidade genética humana e miscigenação

Ao longo do período de povoamento dos continentes, a história descreve a ocorrência de interação entre as populações que se estabeleceram em ambientes diversos e que por isso carregavam diferentes características genéticas. A migração de indivíduos, ou até mesmo o

encontro de populações inteiras, promoveram a interação entre genomas diversificados, contribuindo para a mistura de variantes genéticas únicas e de frequências alélicas distintas (HELLENTHAL *et al.* 2014). Esse processo de mistura é tratado na literatura científica especializada e é reconhecido quando indivíduos de duas populações previamente isoladas se reproduzem e dão origem a descendentes que compartilham material genético de ambas as origens. As populações isoladas dos indivíduos que promovem o processo de mistura são referenciadas como populações ancestrais ou parentais, enquanto que os descendentes, originários desse processo de mistura, compõem as populações ditas misturadas ou miscigenadas (SHRINER, 2013).

As características genéticas das populações miscigenadas trazem grandes vantagens para a realização de estudos que buscam avaliar os padrões de segregação de variantes específicas. Em razão das diferenças na presença e ausência de variantes genéticas, bem como nas frequências alélicas nas diferentes populações parentais, as populações miscigenadas representam oportunidade única de se promover o mapeamento de genes e/ou marcadores relacionados a doenças ou características específicas. Essa abordagem foi inicialmente proposta por Rife, na década de 1950 (RIFE 1954). Na oportunidade, o pesquisador estudou grupos de americanos de descendência africana e de descendência europeia em busca de elucidar padrões de ligação de variantes que contribuíam para as características morfológicas das mãos e da pigmentação da pele. Em seus trabalhos, Rife conclui que há vantagens em estudar populações miscigenadas para identificar padrões de segregação de variantes específicas, mas em razão das limitações metodológicas da época, especialmente pela dificuldade em acessar a informação em nível genômico, outros métodos de mapeamento que não consideravam essa estratégia eram mais aplicados naqueles tempos.

Tendo em vista os recentes avanços nas técnicas de sequenciamento e genotipagem em larga escala do DNA, a estratégia de mapeamento de variantes e genes por meio da análise de populações miscigenadas passou a ser reconsiderada há algumas décadas (SMITH *et al.*, 2004). Vários estudos têm demonstrado que os fenótipos podem variar de acordo com a ancestralidade, como ocorre com as doenças autoimunes, tais como a esclerose múltipla, as quais tendem a ser mais comuns em indivíduos com ancestralidade europeia. Sabe-se também que doenças como hipertensão, câncer de próstata e as relacionadas ao sistema renal são mais prevalentes em indivíduos de ancestralidade africana (SMITH *et al.* 2004, WINKLER *et al.* 2010). O acesso à diversidade em âmbito genômico estimulou o desenvolvimento de métodos mais precisos e estatísticas mais adequadas para a análise dessas variantes de frequências

diferenciadas. Os primeiros métodos se basearam em marcadores específicos, os quais permitem diferenciar as variantes que possuem frequências alélicas divergentes nas populações parentais anteriormente ao evento de formação da população miscigenada. Esse grupo de marcadores é denominado "marcadores informativos de ancestralidade" (MIA) os quais são determinados pela avaliação de diferentes estatísticas obtidas pela análise das populações parentais, dentre as quais está o índice F_{ST} de Wright (WRIGHT 1950, PRITCHARD & ROSENBERG 1999, ENOCH *et al.* 2006).

Muitos estudos têm sido publicados nas últimas décadas utilizando métodos que consideram os padrões de ancestralidade para o mapeamento genético, o que promoveu a consolidação da estratégia proposta por Rife e que passou a ser conhecida como "mapeamento por mistura" (SHRIVER *et al.* 2003, ZHU *et al.* 2005, FREEDMAN *et al.* 2006, PALMER *et al.* 2016) . Ao considerar que o fenótipo buscado ocorre com maior frequência em uma das populações parentais, o mapeamento por mistura permite que as informações sobre a frequência fenotípica nas populações ancestrais sejam relacionadas aos segmentos cromossômicos herdados pelos indivíduos das populações miscigenadas. Assim, é possível que sejam identificadas regiões do genoma que possuem maior probabilidade de estarem relacionadas à característica avaliada (HOGGART *et al.* 2004).

Ancestralidade Local

Na maioria dos casos, os estudos de mapeamento por mistura apontam as regiões candidatas do genoma herdadas de certa população ancestral que estão associadas ao fenótipo observado. Quando a ancestralidade é inferida ao nível cromossômico é denominada ancestralidade local (SANKARARAMAN *et al.* 2008). Nesse processo as análises apontam com certa precisão estatística as ancestralidades das regiões cromossômicas. Em seguida à inferência da ancestralidade local, as análises podem ser posteriormente aprofundadas em um processo de mapeamento-fino, em que a região genômica alvo é explorada com maior detalhamento em busca de variantes que apresentem um sinal mais evidente de associação com o fenótipo (JOHNSON *et al.* 2016). A ancestralidade local inferida para o loco alvo no grupo dos casos pode ser comparada à ancestralidade local inferida para o mesmo loco nos genomas do grupo controle (abordagem caso-controle) ou com a ancestralidade local inferida para outros locos dos genomas dos mesmos indivíduos do grupo caso (abordagem "case-only"). Essa abordagem "case-only" é geralmente mais robusta, já que não há introdução de qualquer ruído estatístico pela análise das amostras do grupo controle (SELDIN 2011).

O método de mapeamento por mistura traz vantagens frente aos demais justamente pela redução do número de marcadores necessários para as análises, o qual é consideravelmente menor do que o utilizado para os estudos de GWAS (SMITH & O'BRIEN, 2005). Os primeiros estudos de mapeamento por mistura geravam resultados mais precisos em casos de miscigenação recente (<20 gerações), mas atualmente novos métodos permitem análises que contemplem eventos de miscigenação de até 100 gerações. Porém, o mapeamento por mistura não possui resolução tão alta quanto à obtida no mapeamento por associação, mas, como já mencionado, análises posteriores podem ser realizadas para detalhar as variantes presentes nessas regiões (SHRINER, 2013).

Como tratado anteriormente, o mapeamento por mistura utiliza a informação acerca da ancestralidade em populações miscigenadas para facilitar as análises de busca por sinais de seleção ao longo do genoma. Porém, esse método ainda se baseia em análises que buscam sinais fortes de seleção localizados. As evidências de que grande parte das características adaptativas decorre de mudanças suaves na frequência de alelos de vários locos de forma não-aleatória (HANCOCK *et al.* 2010, PRITCHARD *et al.* 2010) indica que muitos dos sinais de seleção não estariam sendo detectados por métodos como os empregados no mapeamento por mistura. Porém, as vantagens decorrentes do uso da informação da ancestralidade em populações miscigenadas para o estudo de sinais de seleção, em especial a redução do número de comparações e variantes analisadas, ainda podem ser utilizadas para a identificação de seleção poligênica (JEONG *et al.* 2014, RACIMO *et al.* 2018).

O estudo aqui apresentado propõe uma alternativa para a utilização da informação da ancestralidade local em populações reconhecidamente miscigenadas na busca por sinais de seleção natural em grupos de genes. Parte-se da hipótese de que genes cujos produtos estejam inter-relacionados em redes moleculares complexas estão submetidos a pressões seletivas que atuam sobre suas variantes, levando a suaves alterações da frequência de seus alelos de forma conjunta. Dessa forma, a análise de genes de populações miscigenadas que estão submetidas a pressões seletivas comuns pode revelar alterações não-aleatórias nas frequências das variantes de genes inter-relacionados e para os quais a ancestralidade local tenha sido inferida. Portanto, identificar em um grupo de genes relacionados sinais de excesso ou redução de certa ancestralidade quando comparado ao acaso pode revelar evidências de seleção poligênica. Esse estudo propõe a análise do padrão de distribuição das ancestralidades dos genes em diferentes populações miscigenadas resultantes da interação de grupos de origem europeia, africana e nativo-americana. Espera-se identificar em populações miscigenadas de diferentes localidades sinais comuns e/ou exclusivos de seleção poligênica

que revelem a existência de vantagem evolutiva de grupos de genes com determinadas ancestralidades frente às demais.

Esse estudo foi idealizado e conduzido durante o período de estágio sanduíche, realizado de julho de 2017 a junho de 2018, no laboratório do prof. Dr. Jeffrey Kidd, no Departamento de Genética Humana, Universidade de Michigan/EUA. A experiência durante esse período permitiu que fossem adquiridos conhecimentos teóricos em diversas áreas, incluindo estatística e genômica. Na ocasião, habilidades de programação em linguagem *perl* e *python* foram aprimoradas durante a execução deste e outros projetos sob supervisão do prof. Dr. Jeffrey Kidd.

OBJETIVOS

Objetivo geral:

Identificar sinais de seleção poligênica pela análise da informação de ancestralidade local dos genes de indivíduos de populações miscigenadas

Objetivos específicos:

- Desenvolver um *pipeline* bioinformático para identificar sinais da seleção natural sobre grupos gênicos funcionais

- Aplicar o *pipeline* desenvolvido em diferentes populações miscigenadas em busca de sinais comuns e exclusivos de seleção poligênica

METODOLOGIA

Amostras

Para o desenvolvimento e aplicação do *pipeline* foi necessário obter dados genômicos de indivíduos de populações reconhecidamente miscigenadas. Considerando que populações latino-americanas disponíveis para o uso público pelo Projeto 1000 Genomas (1000GP) (1000 GENOMES PROJECT CONSORTIUM, 2015) são o resultado da miscigenação entre três grandes grupos ancestrais, a saber: europeus, nativos-americanos e africanos, essas populações foram consideradas adequadas para o desenvolvimento do *pipeline* de análises. Além de serem reconhecidamente populações miscigenadas, essas populações foram submetidas à inferência da ancestralidade local das porções cromossômicas dos seus

indivíduos. Para elevar o grau de confiança das inferências foi utilizado o consenso entre quatro métodos de inferência da ancestralidade[6] disponibilizados pelo 1000GP.

Dentre as populações disponíveis para acesso que continham informações de ancestralidade local, foram selecionadas as populações de mexicanos de Los Angeles (MXL), composta por 65 indivíduos, de colombianos de Medellin (CLM), composta por 60 amostras e a de porto-riquenhos (PUR), composta por 55 indivíduos. As proporções de miscigenação dessas amostras estão apresentadas na Tabela 1. Além das populações do 1000GP, também foram utilizados os dados da coorte de Salvador, composta por 1.309 indivíduos e alvo de estudos do Projeto EPIGEN (https://
epigen.grude.ufmg.br) conduzidos por nosso grupo de pesquisa (Tabela 1). A inferência da ancestralidade local dos indivíduos de Salvador foi realizada com o programa RFMix (MAPLES *et al.* 2013) a partir dos dados genéticos obtidos pela genotipagem das amostras com o *array* Illumina HumanOmni2.5–8v1. Apenas porções cromossômicas com valores de confiança de inferência de ancestralidade local do RFMix superiores a 99% foram utilizadas para as análises.

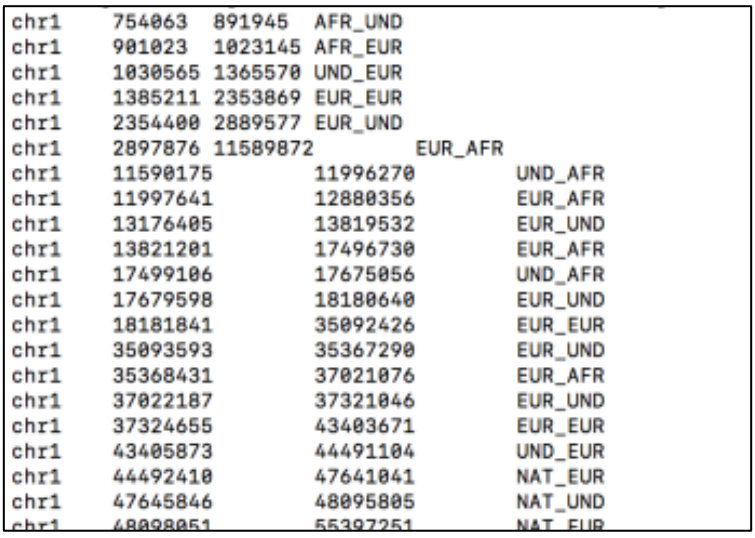
Tanto os dados das populações miscigenadas do 1000GP, quanto os dados da coorte de Salvador foram disponibilizados em arquivos do tipo *.bed* (*browser extensible data*) o qual incluiu informações sobre as posições genômicas de início e fim das porções cromossômicas bem como a identificação da ancestralidade inferida para cada uma dessas porções. Os arquivos obtidos do 1000GP disponibilizam as informações dos cromossomos homólogos de forma concatenada, ou seja, o genótipo de ancestralidade é apresentado para a coordenada cromossômica de ambos os cromossomos em uma mesma linha. A Figura 1 apresenta uma ilustração da estruturação de um arquivo *.bed* obtido do 1000GP após algumas manipulações. Nota-se que as informações de ancestralidade são apresentadas como genótipos, que compreendem as combinações entre as três ancestralidades e também as indicações de quando a ancestralidade não foi determinada.

[6] 1000 Genomes Project phase 1.
<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/ancestry_deconvolution/> Acesso em 09 abr 2019

Tabela 1. Populações do Projeto 1000 Genomas e do Projeto EPIGEN e suas respectivas proporções de ancestralidade. MXL – mexicanos de Los Angeles, CLM – colombianos de Medellin, PUR – porto-riquenhos. EUR – ancestralidade europeia, AFR – ancestralidade africana e NAT – ancestralidade nativo-americana.

Ancestralidades/ Populações	EUR	AFR	NAT
MXL	45%	4%	47%
CLM	65%	7%	24%
PUR	72%	11%	13%
Salvador (EPIGEN)	43%	51%	6%

Figura 1. Ilustração de um dos arquivos *.bed* (*browser extensible data*) contendo informação da ancestralidade local por indivíduo. UND – ancestralidade não determinada; AFR – ancestralidade africana, EUR- ancestralidade europeia, NAT – ancestralidade nativo-americana. As ancestralidades são apresentados em genótipos, considerando os dois cromossomos homólogos. Arquivo texto separado por tabulações com quebra das linhas.



```
chr1    754063  891945  AFR_UND
chr1    901023  1023145 AFR_EUR
chr1    1030565 1365570 UND_EUR
chr1    1385211 2353869 EUR_EUR
chr1    2354400 2889577 EUR_UND
chr1    2897876 11589872        EUR_AFR
chr1    11590175        11996270        UND_AFR
chr1    11997641        12880356        EUR_AFR
chr1    13176405        13819532        EUR_UND
chr1    13821201        17496730        EUR_AFR
chr1    17499106        17675056        UND_AFR
chr1    17679598        18180640        EUR_UND
chr1    18181841        35092426        EUR_EUR
chr1    35093593        35367290        EUR_UND
chr1    35368431        37021076        EUR_AFR
chr1    37022187        37321046        EUR_UND
chr1    37324655        43403671        EUR_EUR
chr1    43405873        44491104        UND_EUR
chr1    44492410        47641041        NAT_EUR
chr1    47645846        48095805        NAT_UND
chr1    48098051        55397251        NAT_EUR
```

Dados de grupos gênicos e coordenadas genômicas

O banco de dados MSig (*Molecular Signature Database v6.1*) (LIBERZON *et al.* 2015), disponível para acesso pelo *Broad Institute*[7] é um dos banco de dados mais amplos disponíveis para uso público. Esse banco possui informações sobre os genes que compõem vários grupos e vias já descritos. A versão 6.1 possui 17.786 grupos de genes, porém dentre esses grupos há aqueles que possuem em sua nomenclatura a identificação de serem grupos de genes ligados, tais como os que ocorrem em certas bandas específicas de alguns cromossomos. Esses grupos foram removidos das análises, já que poderiam levar a resultados

[7] Broad Institute. **Gene Set Enrichment Analysis**. Disponível em: http://software.broadinstitute.org/gsea/index.jsp. Acesso em: 9 de abr. 2019

enviesados. Além desses grupos, também foram eliminadas as vias e grupos gênicos com menos de dez e mais de 500 genes, seguindo sugestões de Mooney e Wilmot (2015). Após essas restrições o banco de dados final totalizou 16.544 grupos de genes. A Figura 2 ilustra a estrutura do arquivo de grupos gênicos obtido do MSig, o qual foi posteriormente adaptado.

Figura 2. Ilustração da estrutura do arquivo com os símbolos dos genes que compõem os grupos gênicos analisados. A primeira coluna traz o nome do grupo e a segunda o sítio eletrônico de origem. Arquivo texto separado por tabulações com quebra das linhas.

qAAANWWTGC_UNKNOWN	http://www.broadinstitute.org/gsea/msigdb/cards/AAANWWTGC_UNKNOWN										MEF2C
X2B	MYH2	SKP2	ZHX3	TGIF1	LUC7L3	PPARGC1A		TSC22D4	SPAG9	MAPK3	PDGFRB	EFNA5
L3	TRAM1	CDC42EP3		OLFM1	CDC42EP5		KLF12	RSPO2	KLF14	NTRK3	EPHA7	SEMA6C
F2	TLE4	CACNG3	INHBA	MPZL3	SGCD	SORT1	PATZ1	STEAP2	PPFIA2	GANAB	NNAT	ITM2C
AP4	PPP1R10	ELAVL4	MID1	FZD7	MRPL24	NRAS	CDH13	DUSP1	BNC2	PHF15	LRRN1	ZNF462
4EBP2	DAB1	ANK2	ANK3	HOXC4	GATA3	LOX	OLIG2	ATOH7	DSCAM	SATB2	SPARCL1	PCTP
3	TFAP4	SOX5	SOX4	CNTFR	SESN2	AFF4	MEIS1	CEPT1	HOXA2	FBXW7	GPR21	HOXA3
EUROG1	NTN1	RBMX	DSEL	FOXP1	FOXP2	SALL3	HOXB2	HOXB6	RRS1	OMG	CACNA1D	MYLK
4F1	IGF2BP1	SHC3	DLG2	ZW10	TRPM3	SSBP3	SOX13	OTX2	ESR1	LEAP2	HEPACAM	MBNL1
L17	CHD2	POU2F1	PIK3R3	ACTB	DCHS2	SNX25	ZNF827	DPYSL5	GRHL3	LIPG	C11orf87	
AAAYRNCTG_UNKNOWN	http://www.broadinstitute.org/gsea/msigdb/cards/AAAYRNCTG_UNKNOWN										LTBP1
C	C14orf1	XRCC1	SEPT7	AQP2	CITED2	EPC1	ZFP91	FAM49A	DCAKD	MAP3K5	GRIN2B	RPLP0
1	CNTLN	STC2	PMCH	STRN3	STRN4	CCNY	EMX2	COLEC10	GLDN	ZNF503	LHFP	NAV3
1	PDGFRA	ATPIF1	PPP2R5E	MYO18A	ADD3	DENND4A	NAGLU	SRSF8	SIRPA	PPP2R3A	FIZ1	ERBB4
X1	PFN2	DMD	EGFLAM	HOXA10	BAI3	DYNC1I1	NUP54	LHX9	IL1RAPL1		PLAGL2	FCER1A
521	HGF	PCDH17	ZNF524	DDX4	PCDH18	FOXG1	EPHA7	KCNQ1DN	ARSG	SEMA6D	ARHGAP44	
5B	HIP1R	RAB5C	ZMAT3	ZNF710	FAM83F	KCNIP2	SORBS2	VWA5A	RTN1	STAC3	PMCHL1	TAS1R2
J	ICAM4	KIAA0182		CMKLR1	PAPD5	LENG9	HN1	PROK2	SLC26A6	ZNF296	ANKRD28	GNAQ

As coordenadas genômicas dos genes que compõem os 16.544 grupos do banco de dados formado a partir do MSig foram obtidas a partir do *UCSC Genome Table Browser* (KENT *et al.* 2002), seguindo a referência do genoma GRCh37/hg19. Os dados foram obtidos em formato *.bed*, o qual foi customizado para apresentar para cada gene apenas o número do cromossomo em que estão localizados, as posições cromossômicas de início e fim e o símbolo dos genes. Em seguida, os dados foram tratados para garantir que os genes tivessem apenas uma única coordenada genômica e que cada gene estivesse representado apenas uma única vez no arquivo (os genes com duas coordenadas foram eliminados para evitar conflitos e contagens duplas).

Cálculo do conteúdo de ancestralidade, estatísticas e automatização

Para o cálculo do conteúdo de ancestralidade foi utilizado o programa BEDOPS (NEPH *et al.* 2012). Foram determinadas nove classes, as quais foram testadas para o excesso ou falta de ancestralidade ou genótipo de ancestralidades. As seis primeiras compreenderam os possíveis genótipos formados pelas combinações das três ancestralidades (EUR_EUR, NAT_NAT, AFR_AFR, EUR_NAT, EUR_AFR, AFR_NAT). As três demais foram EUR (ancestralidade europeia), AFR (ancestralidade africana), NAT (ancestralidade nativo-americana), cujos valores finais corresponderam aos valores das ancestralidades

homozigotas multiplicado por dois e somado aos valores dos genótipos heterozigotos formados pelas respectivas ancestralidades. Para esclarecer, entende-se como ancestralidade homozigota as porções de cromossomos homólogos de mesma ancestralidade, e heterozigota a que representa combinações de ancestralidades diferentes.

Para a geração do modelo nulo foi utilizado o comando *shuffle* do programa BEDTOOLS (QUINLAN & HALL 2010) em que 10.000 permutações foram utilizadas para a distribuição aleatória das posições cromossômicas de cada amostra, considerando apenas os intervalos com informação de ancestralidade. Após o cálculo do p-valor foi aplicada correção para testes múltiplos com o uso do pacote *qvalue* do programa estatístico R (STOREY *et al.* 2019), o qual aplica os princípios da taxa de descoberta de falsos-positivos (FDR). O valor utilizado como limiar foi FDR<0.05. As plotagens foram realizadas com o uso da biblioteca *matplotlib* implementada em linguagem *python.* O *pipeline* para a automatização do processo foi desenvolvido em linguagem *python 3*, sendo que algumas etapas contaram com o interpretador de comandos *bash* em ambiente linux.

RESULTADOS PRELIMINARES

Desenvolvimento do pipeline

O *pipeline* de análises foi construído pela programação de comandos em linguagem *python 3*. Foram gerados *scripts* para a execução de cada uma das etapas do processo, as quais compõem um procedimento único de análise, representado pelo fluxograma da Figura 3. A seguir são destacados os principais passos do processo desenvolvido.

Preparação dos bancos de dados

A filtragem do banco de dados MSig resultou em 16.544 grupos de genes. Para que os grupos gênicos possuíssem apenas genes cujas coordenadas genômicas fossem constantes do banco do UCSC *Genome Table Browser*, uma nova filtragem foi realizada. Ao final, após outras filtragens como, por exemplo, para a retirada de genes no MSig não constantes no banco do UCSC, ou mesmo pela retirada de genes duplicados do banco de dados do UCSC, o número total de 18.152 genes foi mantido no arquivo de coordenadas genômicas, sendo que todos constavam pelo menos uma vez nos grupos gênicos do banco de dados filtrado do MSig.

Cálculo do conteúdo de ancestralidade observada para cada grupo gênico

Uma vez que as informações sobre a ancestralidade local de cada amostra e as coordenadas dos genes dos grupos selecionados para análises foram reunidas, o comando *bedmap* do programa BEDOPS (NEPH *et al.* 2012) foi utilizado para mapear em cada amostra os genes nas porções cromossômicas que apresentaram a informação da ancestralidade. Dessa forma, para cada amostra, em cada intervalo cromossômico com a informação da ancestralidade local, foram alocados os genes cujas coordenadas genômicas estivessem incluídas nesses intervalos. Uma ilustração da estrutura dos dados após o mapeamento está apresentada na Figura 4. Para o caso de genes cujas coordenadas genômicas incluíssem mais de uma ancestralidade, todas elas foram consideradas igualmente para os cálculos realizados. Por sua vez, considerando cada amostra, foram excluídos os genes cujas coordenadas genômicas não estivessem incluídas em nenhuma das porções cromossômicas com informações sobre a ancestralidade. Dessa forma, para cada amostra foi criado um banco de dados em que foram mapeados os genes dos 16.544 grupos anteriormente definidos.

Figura 3. Fluxograma com a indicação das etapas e processos executados no *pipeline* desenvolvido

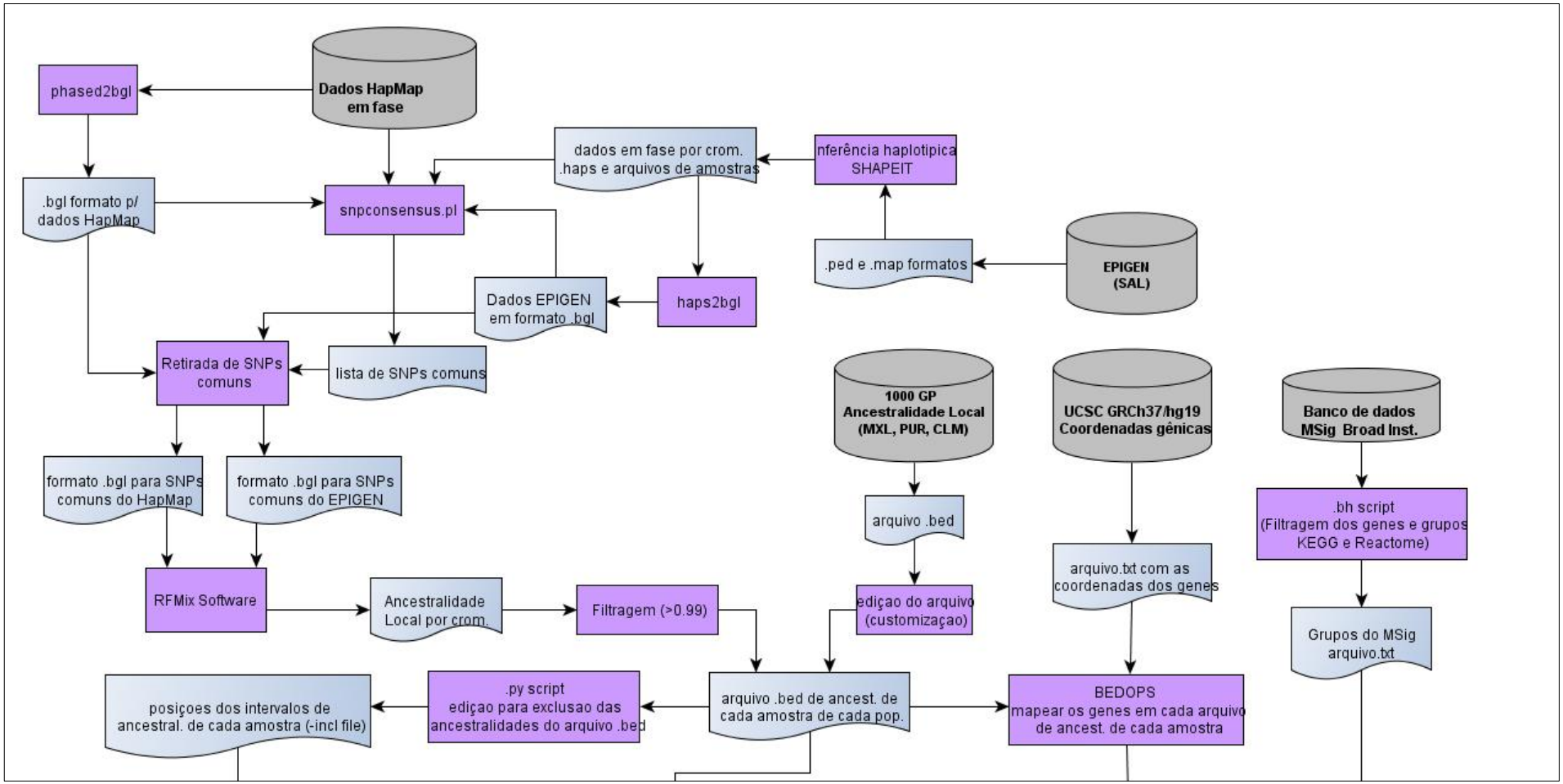


Figura 3. Continuação (quadro menor destacando a etapa de permutação para geração de distribuições nulas)

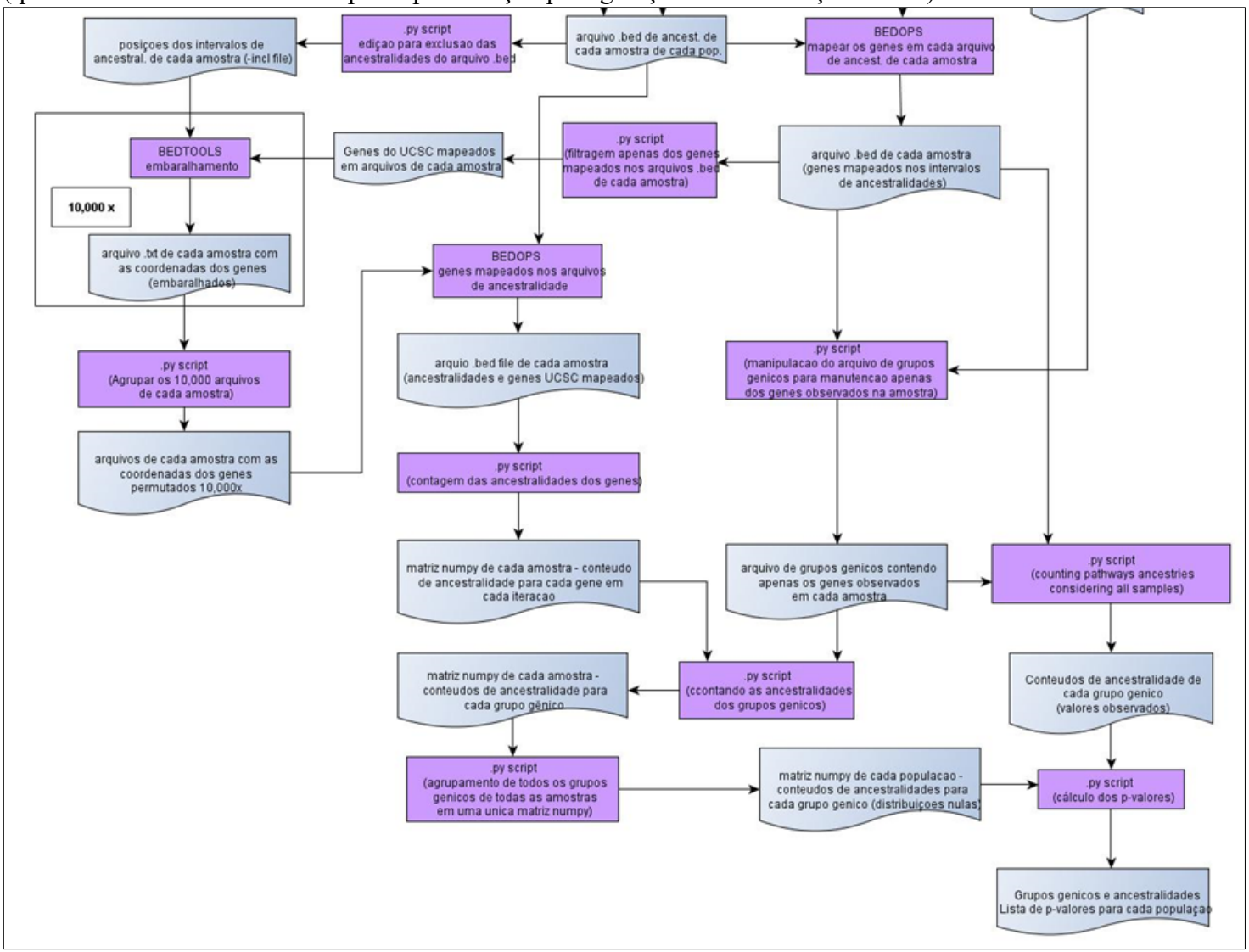


Figura 4. Ilustração da estrutura do arquivo após o mapeamento dos genes de acordo com as ancestralidades e posições genômicas. Primeira coluna indica o cromossomo, segunda coluna a posição inicial, a terceira coluna a posição final, a quarta coluna os genótipos de ancestralidade e da quinta coluna em diante os símbolos gênicos. Arquivo texto separado por tabulações com quebra das linhas.

```
chr1    834830  6334260 AFR_AFR ENTPD5  TXNL4A  INPP5J
M58     HCAR3   SERINC2 ERCC3   PPP1R9B CIDEA   CTH
1D17    ZNF93
chr1    6335503 6814758 UND_AFR ANKRD44 RGS3    TXNDC16
chr1    6816370 16848652        EUR_AFR STK35   MIPEP
AU1     WDR82   PRRC2B  POM121L12       LOC100132356
rf22    SIX2    ALOX15  CUTC    PRDX3   LOC100129935
P1      ARL4D   CEBPA   EIF6    ASPDH   JOSD1   PTMS
chr1    17004807        17874843        UND_AFR LONRF1
chr1    17876051        24495722        AFR_AFR PTPRM
RIB1    PARP10  TNFAIP2 FOXN1   JMJD7-PLA2G4B   SDCCAG8
orf1    LOC283585       PTPRG   KCNH1   DMRT3   MYO7A
chr1    24498696        25184356        UND_AFR LOC90246
chr1    25187557        30996940        EUR_AFR LONP2
RA2B    UBR5    NPY2R   RAB42   C20orf62        PQLC1
```

Tomando cada amostra como uma unidade de análise separada, a lista de grupos gênicos obtida do MSig foi utilizada como referência para a contagem do conteúdo de ancestralidade de cada grupo. Dessa forma, utilizando o arquivo gerado pelo mapeamento do BEDOPS (Figura 4), foi contabilizado o número de genes de cada ancestralidade inferida para cada amostra e para cada grupo gênico, totalizando a soma final de cada ancestralidade para cada grupo gênico. Portanto, foi realizada uma soma simples para cada ancestralidade, em que cada gene contribuiu com o valor de 1 ou mais, dependendo da ancestralidade ou ancestralidades observadas ao longo de sua extensão. Vale ressaltar que os genes que possuíam mais de uma inferência de ancestralidade (e.g. genes mais longos em extensão) contribuíram igualmente para todas as ancestralidades inferidas, independentemente de diferenças de tamanho entre as porções cromossômicas de diferentes ancestralidades de um mesmo gene.

Como a identificação dos sinais poligênicos foi feito no nível populacional, os valores de ancestralidade de cada grupo gênico obtido para cada indivíduo foram somados para o cálculo do conteúdo de ancestralidade para cada grupo gênico em cada população. Por fim, os valores obtidos para cada grupo de genes em cada população foram armazenados em uma estrutura de dicionário *python* considerando cada ancestralidade e cada genótipo de ancestralidade. Ressalta-se que além de avaliar o conteúdo de cada ancestralidade, também foram analisados os conteúdos de cada genótipo de ancestralidade, entendido aqui como as porções de cromossomos homólogos, as quais podem ser homozigotas para cada uma das

ancestralidades, ou heterozigotas com combinações entre as ancestralidades estudadas. Portanto, para cada população, nove valores de conteúdo de ancestralidade para cada grupo de genes foi armazenado, sendo três deles das ancestralidades totais (EUR, AFR e NAT) e outros seis dos possíveis genótipos (EUR_EUR, NAT_NAT, AFR_AFR, EUR_NAT, EUR_AFR, AFR_NAT).

Construção das distribuições nulas de cada população para as classes testadas

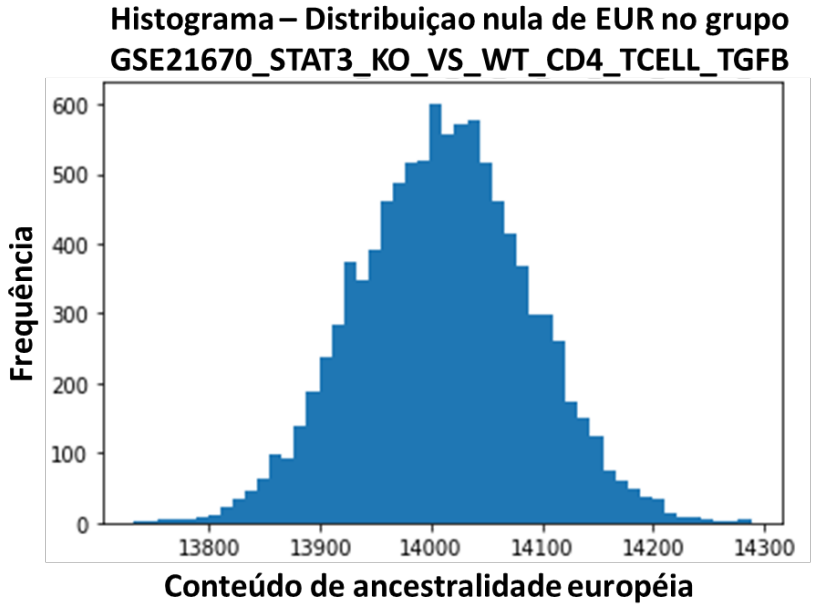
Para gerar a distribuição nula dos valores de cada uma das nove classes testadas foram realizadas 10.000 permutações dos genes presentes no arquivo de coordenadas genômicas obtidos do banco de dados do UCSC. As permutações respeitaram os tamanhos dos genes, apenas alterando-os de posição. Para facilitar a manipulação dos dados, já que foram criados 10.000 arquivos de posições dos genes, os arquivos foram reunidos em apenas um. Porém, antes de reuni-los, a cada par gene-posição, de cada arquivo de permutação, foi acrescentado o numero correspondente à permutação para permitir a identificação de cada permutação após o agrupamento dos dados em um único arquivo.

Contagem dos conteúdos de ancestralidade

Após a geração de um arquivo único reunindo as 10.000 permutações, o mapeamento dos genes nas porções cromossômicas de cada amostra foi realizado por um *script* em *python*, assim como os demais processos descritos a seguir (Figura 3 – quadro em destaque). Em seguida, para cada amostra, os valores de ancestralidade para cada gene em cada permutação foi armazenado em uma matriz do tipo *numpy*. O processo seguinte foi desenvolvido para que a contagem do conteúdo de ancestralidade por amostra e por grupo gênico do banco MSig fosse também armazenada em uma matriz do tipo *numpy*. Para a obtenção dos valores de conteúdo de ancestralidade por grupo gênico foi necessário apenas somar as matrizes *numpy* de cada amostra de cada população. Os valores obtidos para cada uma das nove classes, para cada grupo gênico e em cada população corresponderam à distribuição nula dos conteúdos de ancestralidade. Um exemplo de distribuição nula para um grupo gênico específico está ilustrado na Figura 5. Esses valores foram contrastados aos valores observados de cada classe previamente obtidos, a fim de se analisar a significância estatística. Um *script* em *python* foi desenvolvido para determinar o p-valor dado o valor observado e a distribuição nula. O p-valor foi determinado considerando ambas caudas da distribuição, indicando excesso ou escassez do conteúdo de ancestralidade. Como as correções para testes múltiplos (FDR) aplicada foi realizada no programa R, o *pipeline* desenvolvido não incluiu essa análise.

Portanto, o produto final do *pipeline* é um arquivo final com os nomes das vias e os p-valores obtidos para cada uma das classes a ser utilizado como entrada para a correção FDR..

Figura 5. Exemplo de uma distribuição nula da ancestralidade europeia para um grupo gênico específico construída a partir de 10.000 permutações das posições gênicas



Aplicação do pipeline *para as populações miscigenadas*

Como mencionado anteriormente, três populações do 1000GP e a população brasileira de Salvador foram utilizadas como entrada para o *pipeline* desenvolvido. Após a aplicação do teste de correção múltipla (FDR<0.05) aos arquivos de saída do *pipeline*, alguns grupos gênicos foram identificados como com excesso ou escassez de alguma ancestralidade ou genótipo de ancestralidades. Um exemplo de histograma dos p-valores gerados pela aplicação do índice de FDR está ilustrado na Figura 6. O número de resultados significantes variou entre populações e ancestralidades e está representado na Tabela 2.

Figura 6. Histograma exemplificativo dos p-valores gerado para o cálculo da taxa de falsos-positivos (FDR). População CLM e ancestralidade africana.

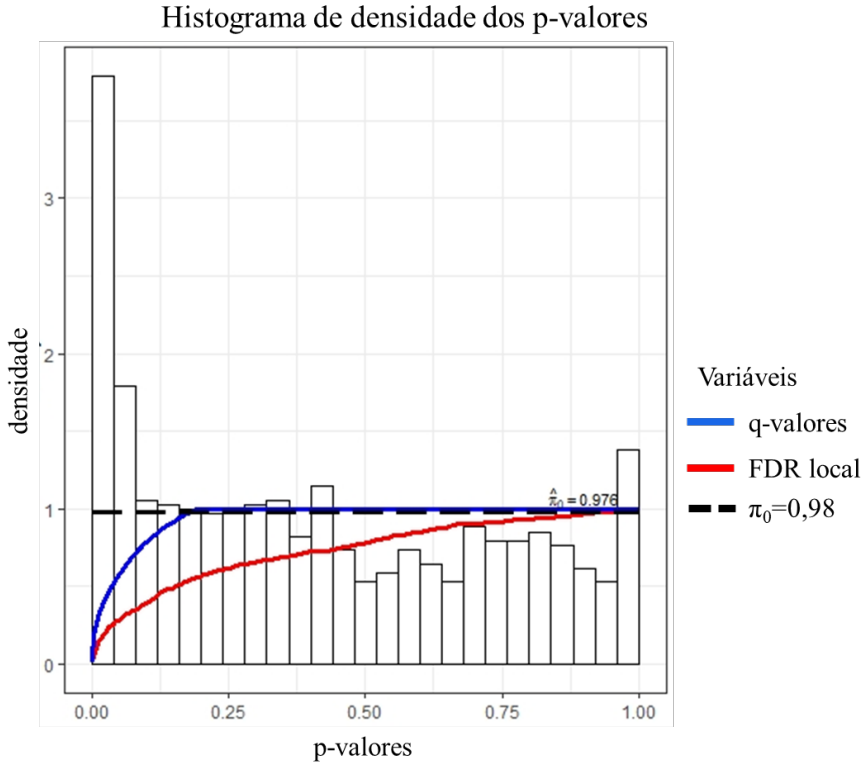


Tabela 2. Número de grupos gênicos significativos em cada população por ancestralidade (EUR – europeia, AFR – africana, NAT- nativo-americana) e combinações de ancestralidade.

		SAL	CLM	MXL	PUR
EUR	Escassez	29	671	59	90
	Excesso	19	14	26	28
AFR	Escassez	10	32	32	10
	Excesso	646	234	129	23
NAT	Escassez	7	23	2811	61
	Excesso	153	59	61	25
EUR_EUR	Escassez	77	305	39	82
	Excesso	12	10	13	43
AFR_AFR	Escassez	8	11	2	9
	Excesso	1250	66	73	18
NAT_NAT	Escassez	0	5	1792	37
	Excesso	0	29	37	30
EUR_AFR	Escassez	0	24	5	30
	Excesso	0	102	79	59
EUR_NAT	Escassez	27	36	45	44
	Excesso	72	16	22	21
AFR_NAT	Escassez	3	23	49	62
	Excesso	62	122	95	12

Os resultados também podem ser analisados com mais detalhes quando se comparam os grupos gênicos significativos para cada população de acordo com a ancestralidade. Os diagramas de Venn indicam as concordâncias e resultados exclusivos entre as populações (Figuras 7 e 8). Os resultados podem ser representados de duas formas distintas. Primeiro analisou-se apenas o nome dos grupos com resultados significativos em cada população, sem considerar se foram significativos pelo excesso ou escassez das ancestralidades (Figura 7). Assim, é possível identificar sinais concomitantes, em que certo grupo gênico tem excesso ou escassez de uma ancestralidade em uma população e também é significativo em outra população, mas não necessariamente pelo mesmo excesso ou escassez. Por outro lado, os diagramas de Venn também podem ser gerados para indicarem quando o grupo gênico e a condição de escassez ou excesso de certa ancestralidade são exatamente os mesmos em diferentes populações (Figura 8). Como ocorre, por exemplo, com o grupo gênico GO_OLFACTORY_RECEPTOR_ACTIVITY, o qual possui genes com escassez da ancestralidade europeia em todas as populações analisadas.

Por fim, a Tabela 3 apresenta os resultados de destaque ao listar os principais grupos gênicos que apresentaram sinais de excesso ou escassez das ancestralidades africana e europeia em todas as populações estudadas.

Figura 7. Diagrama de Venn apresentando os números de grupos gênicos significativos concordantes e exclusivos entre as populações estudadas. Os números representam apenas a quantidade de grupos gênicos, sem considerar a condição de excesso ou escassez de ancestralidade. AFR: ancestralidade africana; EUR: ancestralidade europeia, NAT: ancestralidade nativo-americana. Populações do 1000GP: CLM: colombianos de Medellin, MXL: mexicanos de Los Angeles/EUA; PUR: porto-riquenhos. SAL: população de Salvador – Projeto EPIGEN

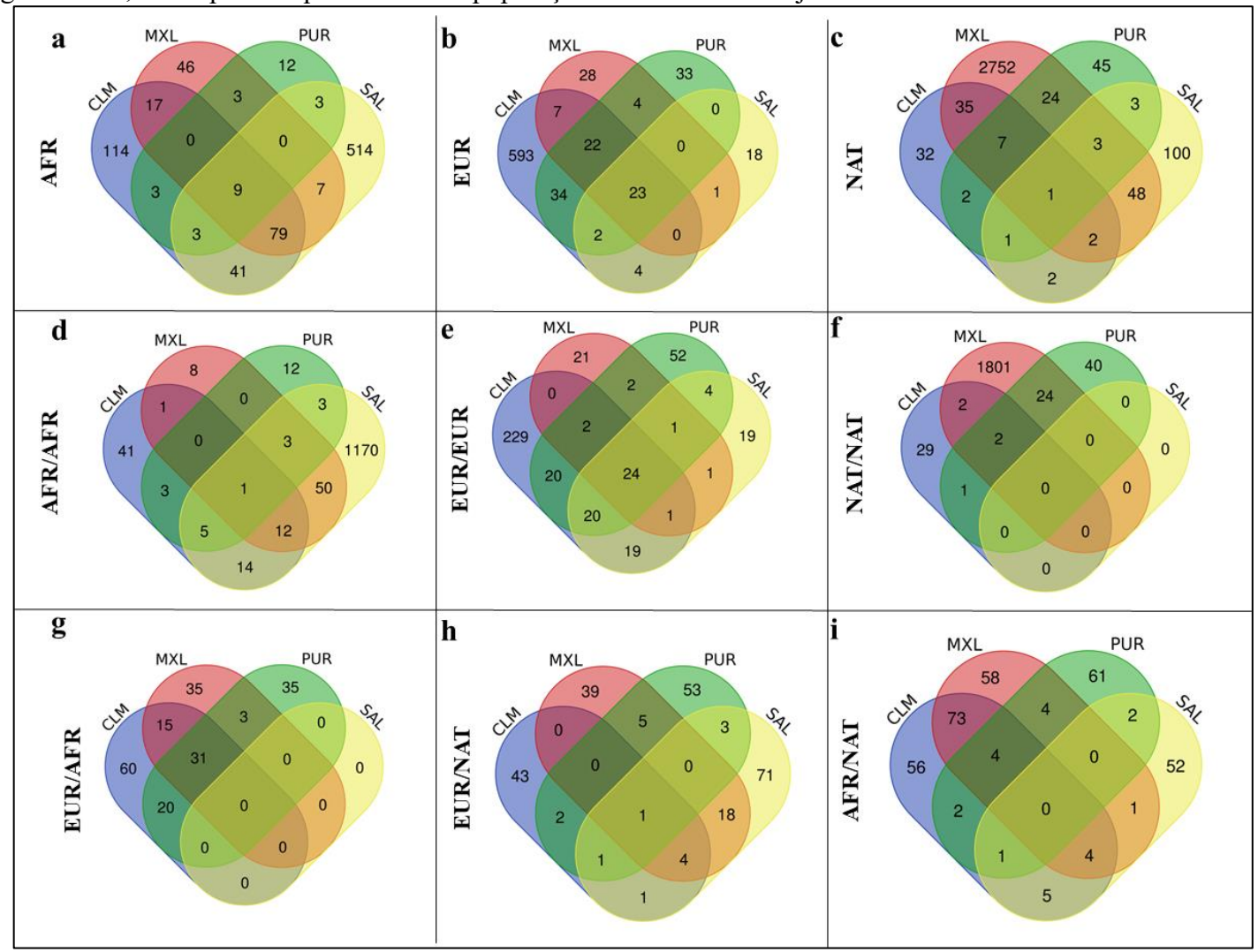


Figura 8. Diagrama de Venn apresentando os números de grupos gênicos significativos concordantes e exclusivos entre as populações estudadas. Os resultados compreendem o número de grupos gênicos considerando a condição de excesso ou escassez de ancestralidade. AFR: ancestralidade africana; EUR: ancestralidade europeia, NAT: ancestralidade nativo-americana. Populações do 1000GP: CLM: colombianos de Medellin, MXL: mexicanos de Los Angeles/EUA; PUR: porto-riquenhos. SAL: população de Salvador – Projeto EPIGEN

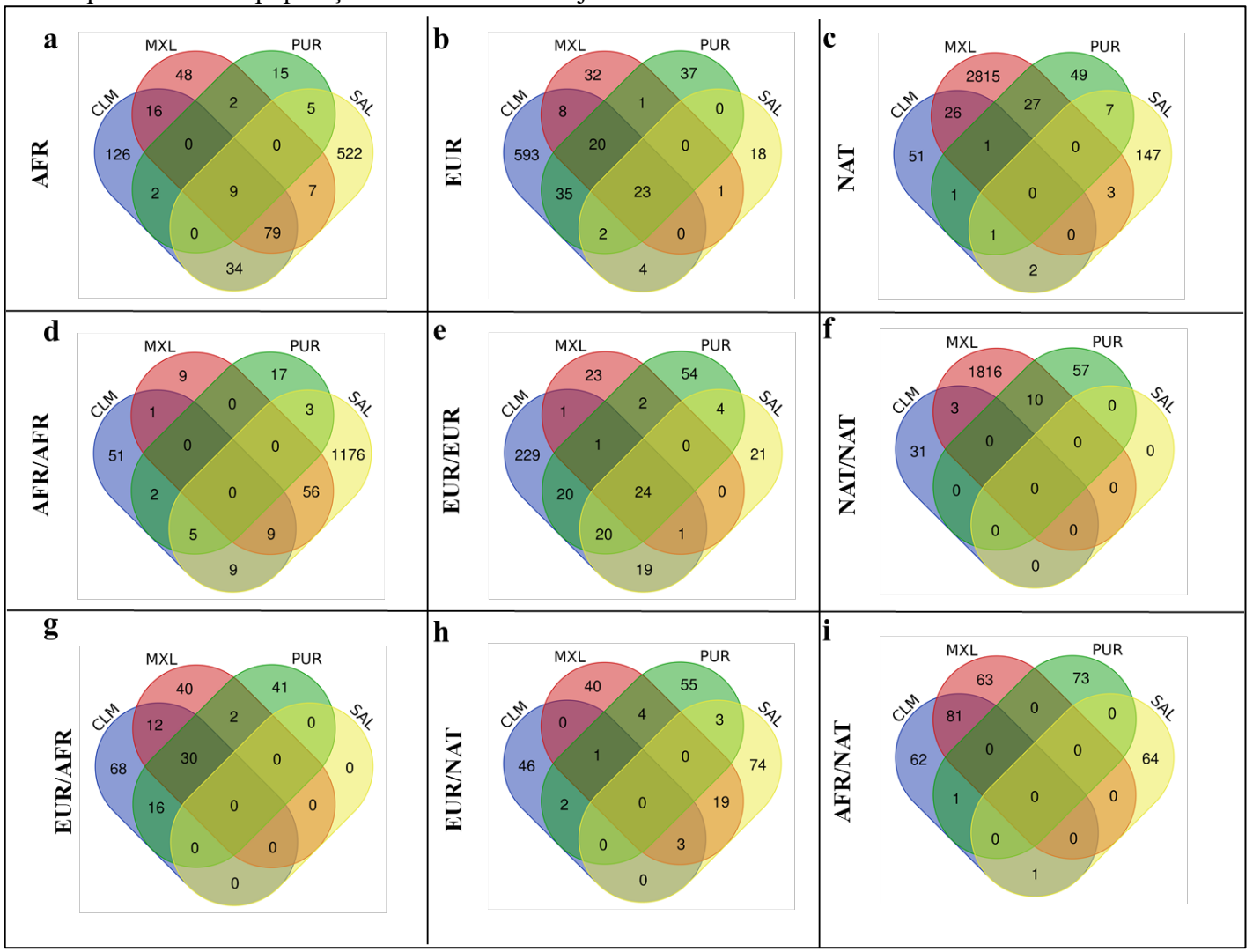


Tabela 3. Nomes dos grupos gênicos, as ancestralidades e suas condições de excesso ou escassez que apresentaram significância em todas as populações analisadas.

Ancestralidade Africana em excesso
REACTOME_RNA_POL_I_TRANSCRIPTION
REACTOME_AMYLOIDS
KRCTCNNNNMANAGC_UNKNOWN
REACTOME_TELOMERE_MAINTENANCE
REACTOME_RNA_POL_I_PROMOTER_OPENING
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
REACTOME_PACKAGING_OF_TELOMERE_ENDS
REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES_AT_THE_CENTROMERE
TTTNNANAGCYR_UNKNOWN
Ancestralidade européia em escassez
REACTOME_MEIOSIS
GO_OLFACTORY_RECEPTOR_ACTIVITY
GO_DNA_REPLICATION_DEPENDENT_NUCLEOSOME_ORGANIZATION
GO_PROTEIN_DNA_COMPLEX_SUBUNIT_ORGANIZATION
KEGG_OLFACTORY_TRANSDUCTION
REACTOME_MEIOTIC_SYNAPSIS
GO_DNA_PACKAGING_COMPLEX
REACTOME_MEIOTIC_RECOMBINATION
REACTOME_AMYLOIDS
REACTOME_TELOMERE_MAINTENANCE
GO_CHROMATIN_SILENCING
ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_DN
REACTOME_PACKAGING_OF_TELOMERE_ENDS
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
REACTOME_RNA_POL_I_TRANSCRIPTION
REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION
GO_CHROMATIN_SILENCING_AT_RDNA
REACTOME_TRANSCRIPTION
KRCTCNNNNMANAGC_UNKNOWN
GO_PROTEIN_DNA_COMPLEX
REACTOME_RNA_POL_I_PROMOTER_OPENING
GO_NUCLEAR_NUCLEOSOME

DISCUSSÃO PRELIMINAR

Desenvolvimento do pipeline

O desenvolvimento do *pipeline* se mostrou um processo com diversos desafios tendo em vista o grande número de grupos gênicos a serem testados, o tamanho das amostras e também pelas características dos arquivos obtidos dos bancos de dados. Os dados dos bancos públicos necessitaram de processamentos prévios, já que foram identificadas inconsistências que poderiam afetar negativamente as análises posteriores. Inicialmente, tratando os dados obtidos do banco MSig, foi necessário eliminar símbolos de genes em duplicata nos grupos gênicos. Além disso, as listas de grupos gênicos incluíam genes que se localizavam agrupados em uma única porção cromossômica, o que viola as premissas desse estudo considerando que necessariamente compartilhariam a mesma ancestralidade. Dessa forma, esses grupos tiveram que ser identificados e removidos das análises. Por fim, ainda com relação às informações desse banco de dados, seguindo orientações de um guia para análises de grupos gênicos (MOONEY & WILMOT 2015), os grupos com tamanhos entre 10 e 200 genes seriam mais adequados para serem analisados. Os autores argumentam que problemas poderiam derivar especialmente da análise de grupos com um numero de genes superior a 200, já que seriam grupos que poderiam envolver diversos processos biológicos. Como o objetivo da análise proposta aqui é exploratório o intervalo do tamanho dos grupos mantido foi um pouco maior, de 10 a 500 genes. Portanto, quaisquer grupos com números diferentes desses foram eliminados das análises. A análise cuidadosa dos resultados finais obtidos indica que esse intervalo aparentemente interferiu negativamente, o que será discutido adiante.

Os dados de coordenadas genômicas obtidos do UCSC também tiveram que ser ajustados em razão de inconsistências. Vários genes com coordenadas duplas e alguns cujos símbolos não estavam representados adequadamente foram eliminados. Esses filtros poderiam influenciar nos resultados finais por interferirem no conteúdo de ancestralidade de certos grupos gênicos em que esses genes são importantes. Porém, o número de genes nessa condição não foi expressivo (<100), o que reduz o impacto sobre o numero final de genes considerado. De toda forma, a necessidade de se editar os arquivos obtidos dos bancos públicos revelou a importância do controle de qualidade dos dados utilizados.

A filtragem do banco de dados de grupos gênicos e coordenadas obtidos do MSig e do UCSC não levou à redução expressiva do número de grupos, totalizando ao final 16.544 grupos gênicos. Dessa forma, em razão do grande número de grupos gênicos, as etapas do

método que consistiam na contagem das ancestralidades dos genes de cada grupo despenderam grande tempo de processamento computacional, o que indicou a necessidade de se estabelecer estratégias de análises em paralelo das amostras. Medidas como o desenvolvimento de *scripts* em formato pbs (*Portable Batch System*) de processamento de amostras em paralelo foram essenciais para permitir a execução das análises em tempo razoável. Outra medida importante para facilitar a manipulação de uma grande quantidade de dados foi a união dos arquivos de permutação de cada individuo em apenas um único arquivo e posteriormente o mapeamento dessas informações nas coordenadas das ancestralidades de cada uma das amostras. A identificação da posição de cada gene em cada permutação permitiu que o processo fosse otimizado para essa etapa sem prejudicar as análises posteriores. Outra medida essencial para a execução computacional em tempo razoável do método desenvolvido foi o uso de matrizes do tipo *numpy* disponíveis em linguagem *python* para a estruturação dos dados. A memória computacional de acesso aleatória bem como o tempo necessário para os cálculos em formato de matriz foi consideravelmente menor caso outra estrutura de dados fosse utilizada, tais como dicionários ou listas. A utilização desse tipo de matriz permitiu que a análise fosse realizada ao nível do indivíduo e posteriormente fossem gerados os dados populacionais por processos de soma de matrizes. De toda forma, ainda são necessários ajustes no método estabelecido para permitir processamentos mais eficientes dos dados e a otimização do tempo e dos recursos computacionais utilizados. Uma das alternativas a serem consideradas é a estruturação das informações em bancos de dados do tipo NoSQL, tais como o MongoDB.

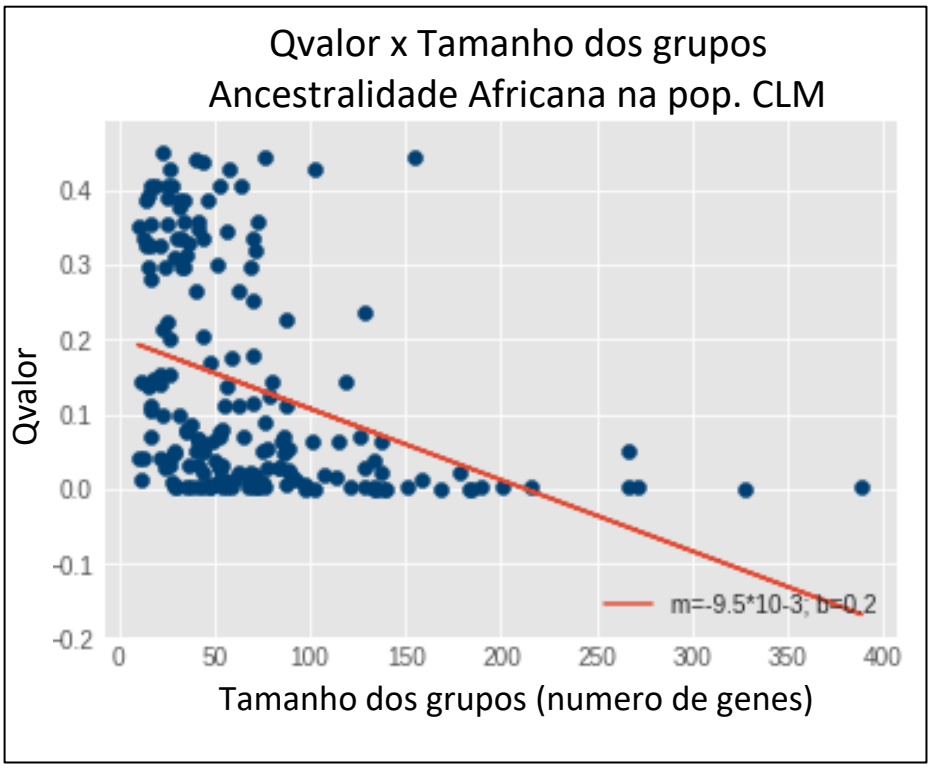
Aplicação do pipeline *às populações miscigenadas*

O *pipeline* desenvolvido está apresentado em formato de fluxograma na Figura 3 e foi executado para as populações miscigenadas do 1000GP e amostras da coorte de Salvador, cujos dados estão vinculados ao Projeto EPIGEN. Como a coorte de Salvador totaliza um número maior de amostras quando comparada às demais populações do 1000GP (1.309 amostras frente ao máximo de 65 amostras da população MXL do 1000GP), a execução de todo o *pipeline* despendeu maior tempo computacional para essa população. O tempo tomado para a análise das amostras de Salvador foi de aproximadamente 72 horas de processamento computacional (considerando processamentos em paralelo de grupos de amostras, como discutido anteriormente). O tempo despendido não é impeditivo das análises realizadas, tendo em vista o grande número amostras dessa coorte, o qual pode ser considerado como parâmetro para a análise de populações menores.

Os resultados obtidos para as populações analisadas trazem padrões interessantes que merecem reflexões sobre suas características. Observa-se, por exemplo, que os resultados estatisticamente significativos foram principalmente obtidos quando a ancestralidade prevalente das populações foi analisada. Como exemplo citam-se os 671 grupos gênicos que excessivamente apresentaram ancestralidade europeia na população CLM, a qual é prevalentemente europeia (65%). Da mesma forma, essa tendência pode ser observada para as populações de Salvador (prevalentemente africana 51%), para o caso de excesso da ancestralidade africana e para a população MXL (prevalentemente nativo-americana – 47%) para a escassez da ancestralidade nativo-americana (Tabela 1). Porém, essa não é uma tendência absoluta, já que a população PUR, a qual possui maior proporção de ancestralidade europeia (72%), ou as populações MXL e Salvador, as quais também possuem alta proporção de ancestralidade europeia (45% e 43% respectivamente), não se destacaram por apresentar altos números de grupos gênicos com estatísticas significativas quando comparado aos demais. É interessante também discutir os resultados quando se comparam as populações pelos diagramas de Venn (Figuras 7 e 8). Nesses casos nota-se que populações diferentes apresentam sinais em comum de excesso ou escassez de ancestralidades para certos grupos gênicos. Destacam-se os sinais de excesso de ancestralidade africana e escassez de ancestralidade europeia em todas as populações analisadas (comparação entre as interseções dos diagramas de Venn das Figuras 7a e 8a, 7b e 8b). Como são sinais completamente independentes para o excesso e escassez dessas ancestralidades, eles podem representar uma evidência de que esses grupos gênicos estão sofrendo processos evolutivos comuns relacionados à dinâmica de miscigenação nos diferentes ambientes.

Por outro lado, análises mais detalhadas revelaram relações interessantes entre os grupos gênicos estudados. Os grupos apresentados na Tabela 3 têm em comum um número grande de genes que estão muito proximamente localizados uns dos outros, o que indica que a região em que estão localizados foi provavelmente inferida como de uma única ancestralidade. Dessa forma, como a permutação dos genes ocorre de forma independente, é baixa a probabilidade de que o mesmo número de genes seja também alocado em coordenadas genômicas que compartilham a mesma ancestralidade. Dessa forma, o conteúdo de ancestralidade observado para esses grupos gênicos tende a ser sempre maior ou menor para as ancestralidades analisadas. Além disso, outra relação interessante observada é a que indica tendência de significância para os grupos gênicos com maior número de genes (Figura 9).

Figura 9. Relação entre q-valor e o tamanho dos grupos gênicos (número de genes) analisados para a ancestralidade africana na população miscigenada de colombianos (CLM)



Duas hipóteses surgem para explicar esse padrão. A primeira se baseia na possibilidade desses grupos terem maior número de genes próximos uns dos outros, o que levaria ao viés entre o valor observado de conteúdo de ancestralidade frente à distribuição nula. A outra hipótese recai sobre o que argumentam Mooney & Wilmot (2015), que indicam que grupos com mais de 200 genes sejam desconsiderados das análises por abarcarem diversos processos biológicos diferentes. Dessa forma, esses grupos tenderiam a apresentar valores de conteúdo de ancestralidade diferentes do acaso por possuírem sub-grupos de genes que na verdade seriam os responsáveis pelo sinal observado.

Essa segunda hipótese é bastante robusta tendo em vista a característica da análise realizada. Como são tomados grupos de genes de bancos de dados públicos, as listas de genes são previamente determinadas e o teste avalia todos os genes de cada grupo gênico da mesma forma. A existência de grupos de genes internos aos grupos testados, ou seja subgrupos, poderia ser uma hipótese a ser testada, já que muito provavelmente esses subgrupos seriam os responsáveis pelos sinais observados. Além disso, a avaliação dos subgrupos seria mais efetiva para a identificação dos sinais que se distinguem do acaso, já que muitas vezes esses sinais podem estar sendo perdidos pela avaliação dos grupos como um todo. Não é surpreendente que essa seja uma das explicações para os resultados obtidos aqui, tendo em

vista inclusive estudo recente (GOUYU *et al.* 2017) que destaca a importância da análise dos subgrupos nos estudos que buscam sinais poligênicos de seleção considerando grupos de genes e vias. Nesse estudo, os autores aplicam o método na busca de sinais de seleção para adaptação a altitude e encontram resultados positivos para vias metabólicas relacionadas à hipóxia.

Conclusão e perspectivas

A literatura especializada mostra que a busca por sinais de seleção em populações humanas ocorre desde que os estudos evolutivos começaram a ser realizados. Mas é possível notar que com os avanços tecnológicos recentes os estudos estabelecidos a nível genômico têm contribuído para o aumento expressivo de resultados que indicam sinais de seleção em porções específicas e isoladas do genoma. Porém, vêm tendo destaque na literatura os argumentos que reforçam a hipótese de que a seleção deve ocorrer na maioria das vezes de forma concomitante sobre locos múltiplos ao longo do genoma. Os sinais conhecidos como *soft-sweeps* aparentemente são prevalentes em razão da atuação da seleção natural alterando de forma suave e não-aleatória as frequências de múltiplos locos e variantes do genoma. Tendo em vista as novas possibilidades de acesso à variação genômica no âmbito populacional, estudos cujos objetivos sejam a análise de sinais poligênicos de seleção têm sido reportados com mais frequência na literatura especializada (HANCOCK *et al.* 2010, DAUB *et al.* 2013, BERG & COOP 2014). Porém, ainda é escassa a literatura que busca estudar a relação dos aspectos que envolvem a miscigenação de populações humanas com os sinais de seleção poligênica. O estudo mais recente que destaca essa relação faz uso de dados da dinâmica histórica da miscigenação entre populações representadas em grafos para indicar possíveis eventos de seleção poligênica (RACIMO *et al.* 2018). Os autores utilizaram estratégias de cadeia de Markov para inferir parâmetros relacionados à atuação da seleção natural e identificaram sinais poligênicos relacionados a fenótipos como altura, formato das sobrancelhas, entre outros. Outra linha de pesquisa que utiliza eventos de miscigenação na busca por sinais de seleção poligênica se estabeleceu após o advento de métodos e técnicas que permitem o estudo de DNA antigo. Por progressos nessa área, atualmente é possível avançar na compreensão das consequências sobre o genoma humano que ocorreram em razão da introgressão de sequências de genoma de outros hominídeos, em especial os Neandertais (DANNEMANN & RACIMO 2018).

Apesar dos exemplos citados anteriormente, ainda é escasso o uso de populações humanas miscigenadas para o estudo de sinais de seleção e uma das possíveis razões seria o

argumento sobre o tempo evolutivo curto decorrido desde os eventos de miscigenação. Tendo em vista a ocorrência de eventos de miscigenação recentes pelo povoamento do Novo Mundo, o tempo evolutivo medido em gerações ainda seria curto para que a seleção atue, o que impossibilitaria a detecção de suas marcas sobre os genomas. Por outro lado, há exemplos de condições ambientais que, por serem severas, levariam a coeficientes de seleção expressivos e portanto gerariam sinais detectáveis em um curto período evolutivo de tempo. Estudos que reportam essas condições podem ser encontrados na literatura especializada. Apesar de não fazerem uso de métodos baseados na miscigenação para a inferência de sinais poligênicos, esses estudos apontam evidências de que os tempos evolutivos curtos podem não ser um entrave para a detecção de sinais recentes de seleção natural, em especial em populações que tenham sofrido pressões evolutivas expressivas. Como exemplos citam-se pressões como a hipóxia por que passam as populações nômades do pacífico quando mergulham em ambientes oceânicos na busca de alimento (ILLARDO *et al.* 2018) ou como as que sofreram os genomas de nativos-americanos frente à diversidade de genomas virais e bacterianos aos quais foram expostos quando da colonização europeia das Américas (LINDO *et al.* 2016). Além disso, outros estudos indicam haver vantagens ao se estudar populações miscigenadas, ou que tenham passado por eventos de mistura em sua história de formação. As populações das cordilheiras do Himalaia são exemplos bastante estudados atualmente, já que há histórico de fluxo gênico entre os grupos que habitam diferentes altitudes. Esses fluxos gênicos fazem parte do contexto em que essas populações evoluíram e influenciaram os processos de seleção natural sofridos por esses grupos (JEONG *et al*. 2014, YANG *et al.* 2017).

Portanto, os genomas das populações miscigenadas são interessantes objetos de estudo tendo em vista as diversas pressões seletivas que sofreram pela alteração brusca dos ambientes onde atualmente habitam quando comparado aos ambientes de suas populações parentais. Porém, os resultados obtidos aqui ainda são preliminares e indicam que o método aplicado ainda carece de aprimoramento. Avanços nos procedimentos de geração da distribuição nula e contagem dos conteúdos de ancestralidade precisam ainda ser alcançados, especialmente para se compreender os desvios notados em favor dos grupos gênicos com maior número de genes. Outras estratégias também podem ser consideradas, como por exemplo, pelo uso de simulações para a geração de genomas hipotéticos, o que pode ser uma alternativa à estratégia de geração de distribuições nulas por permutação.

De toda forma, espera-se avançar no aprimoramento do método apresentado e assim contribuir para o progresso de estudos que buscam sinais adaptativos em populações

humanas, em especial considerando aquelas miscigenadas. Enfim, as hipóteses a serem testadas são variadas e os avanços nas metodologias e nas técnicas genômicas têm permitido que teorias clássicas como as de seleção poligênica possam ser mais testadas em populações de diversas localidades.

REFERÊNCIAS

1000 GENOMES PROJECT CONSORTIUM *et al.* A global reference for human genetic variation. **Nature**, v. 526, n. 7571, p. 68, 2015.

BALARESQUE, Patricia L.; BALLEREAU, Stephane J.; JOBLING, Mark A. Challenges in human genetic diversity: demographic history and adaptation. **Human molecular genetics**, v. 16, n. R2, p. R134-R139, 2007.

BAMSHAD, M.; WOODING, S.P. Signatures of natural selection in the human genome. **Nature Reviews Genetics**, 4(2), p.99, 2003.

BERG, Jeremy J.; COOP, Graham. A population genetic signal of polygenic adaptation. **PLoS genetics**, v. 10, n. 8, p. e1004412, 2014.

BUERKLE, C. Alex; LEXER, Christian. Admixture as the basis for genetic mapping. **Trends in ecology & evolution**, v. 23, n. 12, p. 686-694, 2008.

BUSH, William S.; MOORE, Jason H. Genome-wide association studies. **PLoS computational biology**, v. 8, n. 12, p. e1002822, 2012.

CSILLÉRY, Katalin *et al.* Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution. **Molecular ecology**, v. 27, n. 3, p. 606-612, 2018.

DANNEMANN, Michael; RACIMO, Fernando. Something old, something borrowed: admixture and adaptation in human evolution. **Current opinion in genetics & development**, v. 53, p. 1-8, 2018.

DAUB, Josephine T. *et al.* Evidence for polygenic adaptation to pathogens in the human genome. **Molecular biology and evolution**, v. 30, n. 7, p. 1544-1558, 2013.

DE JONGE, Robert et al. Effect of polymorphisms in folate-related genes on in vitro methotrexate sensitivity in pediatric acute lymphoblastic leukemia. **Blood**, v. 106, n. 2, p. 717-720, 2005.

DING, Yuan-Chun *et al.* Evidence of positive selection acting at the human dopamine receptor D4 gene locus. **Proceedings of the National Academy of Sciences**, v. 99, n. 1, p. 309-314, 2002.

ENOCH, Mary-Anne *et al.* Using ancestry-informative markers to define populations and detect population stratification. **Journal of Psychopharmacology**, v. 20, n. 4_suppl, p. 19-26, 2006.

FAN, Shaohua *et al.* Going global by adapting local: A review of recent human adaptation. **Science**, v. 354, n. 6308, p. 54-59, 2016.

FREEDMAN, Matthew L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. **Proceedings of the National Academy of Sciences**, v. 103, n. 38, p. 14068-14073, 2006.

GNECCHI-RUSCONE, Guido A. *et al.* Evidence of polygenic adaptation to high altitude from Tibetan and Sherpa genomes. **Genome biology and evolution**, v. 10, n. 11, p. 2919-2930, 2018.

GOUY, Alexandre; DAUB, Joséphine T.; EXCOFFIER, Laurent. Detecting gene subnetworks under selection in biological pathways. **Nucleic acids research**, v. 45, n. 16, p. e149-e149, 2017.

HANCOCK, Angela M. *et al.* Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. **Proceedings of the National Academy of Sciences**, v. 107, n. Supplement 2, p. 8924-8930, 2010.

HELLENTHAL, Garrett *et al.* A genetic atlas of human admixture history. **Science**, v. 343, n. 6172, p. 747-751, 2014.

HOGGART, Clive J. *et al.* Design and analysis of admixture mapping studies. **The American Journal of Human Genetics**, v. 74, n. 5, p. 965-978, 2004.

ILARDO, Melissa A. *et al.* Physiological and genetic adaptations to diving in sea nomads. **Cell**, v. 173, n. 3, p. 569-580. e15, 2018.

JEONG, Choongwon; DI RIENZO, Anna. Adaptations to local environments in modern human populations. **Current opinion in genetics & development**, v. 29, p. 1-8, 2014.

JEONG, Choongwon *et al.* Admixture facilitates genetic adaptations to high altitude in Tibet. **Nature communications**, v. 5, p. 3281, 2014.

JOHNSON, Randall C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). **BMC genomics**, v. 11, n. 1, p. 724, 2010.

JOHNSON, Randall C.; WINKLER, Cheryl A.; YEAGER, Meredith. 7 Admixture mapping for disease gene discovery. **Genome-Wide Association Studies: From Polymorphism to Personalized Medicine**, p. 89, 2016.

KAGER, Leo *et al.* Folate pathway gene expression differs in subtypes of acute lymphoblastic leukemia and influences methotrexate pharmacodynamics. **The Journal of clinical investigation**, v. 115, n. 1, p. 110-117, 2005.

KENT, W. James *et al.* The human genome browser at UCSC. **Genome research**, v. 12, n. 6, p. 996-1006, 2002.

LIBERZON, Arthur *et al.* The molecular signatures database hallmark gene set collection. **Cell systems**, v. 1, n. 6, p. 417-425, 2015.

LINDO, John *et al.* A time transect of exomes from a Native American population before and after European contact. **Nature communications**, v. 7, p. 13175, 2016.

LUISI, Pierre *et al.* Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. **Genome Biology and Evolution**, v. 7, n. 4, p. 1141-1154, 2015.

MANOLIO, Teri A. *et al.* Finding the missing heritability of complex diseases. **Nature**, v. 461, n. 7265, p. 747, 2009.

MAPLES, Brian K. *et al.* RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. **The American Journal of Human Genetics**, v. 93, n. 2, p. 278-288, 2013.

MARTH, Gabor T. *et al.* The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. **Genetics**, v. 166, n. 1, p. 351-372, 2004.

MARTIN, Alicia R. *et al.* An unexpectedly complex architecture for skin pigmentation in Africans. **Cell**, v. 171, n. 6, p. 1340-1353. e14, 2017.

MOONEY, Michael A.; WILMOT, Beth. Gene set analysis: A step-by-step guide. **American Journal of Medical Genetics Part B: Neuropsychiatric Genetics**, v. 168, n. 7, p. 517-527, 2015.

MOSKVINA, Valentina; SCHMIDT, Karl Michael. On multiple-testing correction in genome-wide association studies. **Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society**, v. 32, n. 6, p. 567-573, 2008.

NADEAU, Nicola J.; JIGGINS, Chris D. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. **Trends in Genetics**, v. 26, n. 11, p. 484-492, 2010.

NEPH, Shane *et al.* BEDOPS: high-performance genomic feature operations. **Bioinformatics**, v. 28, n. 14, p. 1919-1920, 2012.

NIELSEN, Rasmus *et al.* Tracing the peopling of the world through genomics. **Nature**, v. 541, n. 7637, p. 302-310, 2017.

PALMER, Nicholette D. *et al.* Admixture mapping of serum vitamin D and parathyroid hormone concentrations in the African American—Diabetes Heart Study. **Bone**, v. 87, p. 71-77, 2016.

PE'ER, Itsik *et al.* Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. **Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society**, v. 32, n. 4, p. 381-385, 2008.

PICKRELL, Joseph K.; REICH, David. Toward a new history and geography of human genes informed by ancient DNA. **Trends in Genetics**, v. 30, n. 9, p. 377-389, 2014.

PRITCHARD, Jonathan K.; ROSENBERG, Noah A. Use of unlinked genetic markers to detect population stratification in association studies. **The American Journal of Human Genetics**, v. 65, n. 1, p. 220-228, 1999.

PRITCHARD, Jonathan K.; DI RIENZO, Anna. Adaptation–not by sweeps alone. **Nature Reviews Genetics**, v. 11, n. 10, p. 665, 2010.

PRITCHARD, Jonathan K.; PICKRELL, Joseph K.; COOP, Graham. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. **Current biology**, v. 20, n. 4, p. R208-R215, 2010.

QUINLAN, Aaron R.; HALL, Ira M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, n. 6, p. 841-842, 2010.

QUINTANA-MURCI, Lluis; BARREIRO, Luis B. The role played by natural selection on Mendelian traits in humans. **Annals of the New York Academy of Sciences**, v. 1214, n. 1, p. 1-17, 2010.

RACIMO, Fernando; BERG, Jeremy J.; PICKRELL, Joseph K. Detecting polygenic adaptation in admixture graphs. **Genetics**, v. 208, n. 4, p. 1565-1584, 2018.

RIFE, David C. Populations of hybrid origin as source material for the detection of linkage. **American journal of human genetics**, v. 6, n. 1, p. 26, 1954.

SANKARARAMAN, Sriram *et al.* Estimating local ancestry in admixed populations. **The American Journal of Human Genetics**, v. 82, n. 2, p. 290-303, 2008.

SCHRIDER, Daniel R.; KERN, Andrew D. Soft sweeps are the dominant mode of adaptation in the human genome. **Molecular biology and evolution**, v. 34, n. 8, p. 1863-1877, 2017.

SELDIN, Michael F.; PASANIUC, Bogdan; PRICE, Alkes L. New approaches to disease mapping in admixed populations. **Nature Reviews Genetics**, v. 12, n. 8, p. 523, 2011.

SHAO, Haifeng *et al.* Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. **Proceedings of the National Academy of Sciences**, v. 105, n. 50, p. 19910-19914, 2008.

SHRINER, Daniel. Overview of admixture mapping. **Current protocols in human genetics**, v. 76, n. 1, p. 1.23. 1-1.23. 8, 2013.

SHRIVER, Mark D. *et al.* Skin pigmentation, biogeographical ancestry and admixture mapping. **Human genetics**, v. 112, n. 4, p. 387-399, 2003.

SMITH, Michael W.; O'BRIEN, Stephen J. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. **Nature Reviews Genetics**, v. 6, n. 8, p. 623, 2005.

SMITH, Michael W. *et al.* A high-density admixture map for disease gene discovery in African Americans. **The American Journal of Human Genetics**, v. 74, n. 5, p. 1001-1013, 2004.

SOHAIL, Mashaal *et al.* Negative selection in humans and fruit flies involves synergistic epistasis. Science, v. 356, n. 6337, p. 539-542, 2017. Spencer, C.C., Su, Z., Donnelly, P. and Marchini, J., 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. **PLoS genetics**, 5(5), p.e1000477.

STOREY, J.D., BASS, A.J., DABNEY, A., ROBINSON, D., 2019. qvalue: Q-value estimation for false discovery rate control. R package version 2.14.1, http://github.com/jdstorey/qvalue.

STRANGER, Barbara E.; STAHL, Eli A.; RAJ, Towfique. Progress and promise of genome-wide association studies for human complex trait genetics. **Genetics**, v. 187, n. 2, p. 367-383, 2011.

THOMPSON, Katherine; CHARNIGO, Richard. Parallel Computing in Genome-Wide Association Studies. **Journal of Biometrics & Biostatistics**, v. 6, n. 1, p. 1, 2015.

TISHKOFF, Sarah A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. **Nature genetics**, v. 39, n. 1, p. 31, 2007.

VOIGHT, Benjamin F. *et al.* A map of recent positive selection in the human genome. **PLoS biology**, v. 4, n. 3, p. e72, 2006.

WAN, Xiang *et al.* Detecting two-locus associations allowing for interactions in genome-wide association studies. **Bioinformatics**, v. 26, n. 20, p. 2517-2525, 2010.

WINKLER, Cheryl A.; NELSON, George W.; SMITH, Michael W. Admixture mapping comes of age. **Annual review of genomics and human genetics**, v. 11, p. 65-89, 2010.

WRAY, Naomi R. *et al.* Pitfalls of predicting complex traits from SNPs. **Nature Reviews Genetics**, v. 14, n. 7, p. 507-515, 2013.

Wright, S,, 1950, Genetical structure of populations, Nature, 166, pp,247-49

YANG, Jian *et al.* Genetic signatures of high-altitude adaptation in Tibetans. **Proceedings of the National Academy of Sciences**, v. 114, n. 16, p. 4189-4194, 2017.

ZHU, Xiaofeng *et al.* Admixture mapping for hypertension loci with genome-scan markers. **Nature genetics**, v. 37, n. 2, p. 177, 2005.

CONCLUSÃO GERAL

São enormes os desafios apresentados pela revolução por que a biologia molecular vem passando nas últimas décadas em razão do advento das novas tecnologias de geração de dados genéticos, o que impacta diretamente diversas áreas, desde a pesquisa básica até a aplicada. Inegavelmente há uma imediata necessidade de se aumentar os esforços voltados a analise de dados e processamento da informação genômica que se encontra disponível e em constante expansão. A tese aqui apresentada destacou a importância de se compreender os impactos da Genômica na ciência realizada atualmente como um processo amplo e integrado entre diferentes áreas do conhecimento. É exatamente nessa diversidade de áreas que essa tese se propôs a transitar, contribuindo para avanços, ainda que modestos, nos assuntos tratados.

A diversidade de temas abordados não impede que conclusões gerais sejam apresentadas considerando a tese como um todo. As consequências do aumento da qualidade dos dados biológicos gerados são imediatamente refletidas em diversas áreas, desde as voltadas para a melhora de procedimentos biomédicos até mesmo para as que englobam estudos de base teórica, como os que buscam compreender processos evolutivos. Como apresentado nos resultados do primeiro capítulo deve-se atentar para os métodos de geração de dados genéticos, desde as peculiaridades de cada ensaio realizado até as estratégias empregadas para a identificação das variantes que comporão a base de estudos posteriores. Foi possível concluir que a customização das análises se mostra de grande relevância já que também permite a constante atualização dos métodos de identificação de variantes, os quais, aliados à adequada compreensão da natureza do dado gerado, garantem a qualidade dos estudos conduzidos.

A confiança no dado contribui para a sua aplicação em estudos mais amplos, como o realizado no segundo capítulo da tese. Nesse capítulo, foi possível aplicar diferentes estratégias para a melhor compreensão da estrutura genético-populacional de genes associados à recidiva da leucemia linfoblástica aguda em populações latino americanas. Os resultados permitiram concluir que nas populações com prevalência de ancestralidade nativo-americana, os genes estudados possuem particularidades com relação às suas variantes. A diversidade e estrutura haplotípica desses genes revelaram novas mutações e frequências alélicas que indicam estruturas genéticas próprias dessas populações, as quais podem estar relacionadas às diferentes respostas aos tratamentos da leucemia observados em indivíduos prevalentemente nativo-americanos. Esses resultados indicam novas perspectivas para a

realização de estudos que busquem aprofundar o conhecimento acerca das respostas aos tratamentos da leucemia linfoblástica aguda e suas relações com as diferentes ancestralidades.

Por fim, os resultados preliminares do último capítulo permitiram concluir que a riqueza dos grandes bancos de dados disponíveis atualmente pode ser mais bem aproveitada pelo desenvolvimento de métodos e estratégias que se baseiem nas particularidades apresentadas pelas populações miscigenadas. O uso da informação obtida a partir da inferência da ancestralidade local nessas populações pode ser um dos caminhos para o aprofundamento dos conhecimentos sobre processos adaptativos até então pouco conhecidos. Nesse contexto, a seleção poligênica pode ser tomada como base para o teste de hipóteses formuladas sobre a atuação da seleção natural nas populações humanas, em especial as resultantes de processos históricos de mistura. Porém, o caminho a ser percorrido ainda é longo e as atuais alternativas metodológicas ainda precisam ser aprimoradas e outras ainda desenvolvidas para a realização de estudos mais conclusivos e robustos.

REFERÊNCIAS

BENTLEY, Amy R.; CALLIER, Shawneequa; ROTIMI, Charles N. Diversity and inclusion in genomic research: why the uneven progress?. **Journal of community genetics,** v. 8, n. 4, p. 255-266, 2017.

BHATIA, Smita. Disparities in cancer outcomes: lessons learned from children with cancer. **Pediatric blood & cancer**, v. 56, n. 6, p. 994-1002, 2011.

BOTTLES, Kent; BEGOLI, Edmon; WORLEY, Brian. Understanding the pros and cons of big data analytics. **Physician executive**, v. 40, n. 4, p. 6-12, 2014.

HAGEN, Joel B. The origins of bioinformatics. **Nature Reviews Genetics**, v. 1, n. 3, p. 231, 2000.

MARDIS, Elaine R. The impact of next-generation sequencing technology on genetics. **Trends in genetics,** v. 24, n. 3, p. 133-141, 2008.

MARDIS, Elaine R. A decade's perspective on DNA sequencing technology. **Nature**, v. 470, n. 7333, p. 198, 2011.

MCKEIGUE, Paul M. Prospects for admixture mapping of complex traits. **The American Journal of Human Genetics**, v. 76, n. 1, p. 1-7, 2005.

NEKRUTENKO, Anton; TAYLOR, James. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. **Nature Reviews Genetics**, v. 13, n. 9, p. 667, 2012.

O'CONNOR, Jeremy M. *et al*. Factors Associated With Cancer Disparities Among Low-, Medium-, and High-Income US Counties. **JAMA network open**, v. 1, n. 6, p. e183146-e183146, 2018.

O'RAWE, Jason *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. **Genome medicine**, v. 5, n. 3, p. 28, 2013.

POPEJOY, Alice B.; FULLERTON, Stephanie M. Genomics is failing on diversity. **Nature News**, v. 538, n. 7624, p. 161, 2016.

SELDIN, Michael F.; PASANIUC, Bogdan; PRICE, Alkes L. New approaches to disease mapping in admixed populations. **Nature Reviews Genetics**, v. 12, n. 8, p. 523, 2011.

SIRUGO, Giorgio; WILLIAMS, Scott M.; TISHKOFF, Sarah A. The missing diversity in human genetic studies. **Cell**, v. 177, n. 1, p. 26-31, 2019.

TAUB, Margaret A.; BRAVO, Hector Corrada; IRIZARRY, Rafael A. Overcoming bias and systematic errors in next generation sequencing data. **Genome medicine**, v. 2, n. 12, p. 87, 2010.

WETTERSTRAND K. DNA sequencing costs: data from the NGGRI genome sequencing program (GSP). <http://www.genome.gov/sequencingcosts/>. Acesso em abr 2019

ANEXOS

Outras publicações científicas e principais atividades desenvolvidas pelo aluno durante o período de doutorado (2015-2019)

Principais atividades

2019: Organizador do "*Curso de Next Generation Sequencing para Medicina Genômica. - Teoria e Prática Bioinformática*"; ICB/UFMG - Laboratório de Diversidade Genética Humana: Genômica Translacional (*MOSAICO Translational Genomics*) (01-05/ago/19)

2018: Coordenação do "*Curso teórico-prático de Noções Gerais de Sequenciamento de Nova Geração por Síntese e Construção de Bibliotecas 16S*" 80 hrs (01-30/nov/18). DESA/UFMG

2017 - 2018: Período de 12 meses de estágio sanduíche no laboratório do prof. Dr. Jeffrey Kidd, no Departamento de Genética Humana, Universidade de Michigan/EUA.

2017: Participante do 22º Curso de Verão em Genética Estatística - Universidade de Washington/EUA

2017: Organização e implementação do projeto "Utilização da tecnologia NGS para análises do *HLA* em amostras da população brasileira" – Parceria LGB/UFMG e Símile – Medicina Diagnóstica

2016 – atual: Membro do Projeto de Extensão 402554 - Laboratório de Diversidade Genética Humana: Genômica Translacional (*MOSAICO Translational Genomics*)/SIEX/UFMG

2015 – atual: Membro do Projeto de Extensão 302604 - CELAM (Centro de Laboratórios Multiusuários)/SIEX/UFMG

Manuscritos atualmente submetidos a periódicos científicos

1) **Moreira RG**, Saraiva-Duarte JM, Santolalla ML, Pereira AC, Sosa-Macias M, ... & Rodrigues-Soares F. High frequency of the acute lymphoblastic leukemia relapse biomarker *PDE4B*-rs6683977 gives support for an extra phase of chemotherapy in Amerindians.
2) Gouveia, M. H., Borda, V., Leal, T. P., **Moreira, R. G.**, Bergen, A. W., Aquino, M. M., ... & Machado, M. (**2019**). Origins, admixture dynamics and homogenization of the African gene pool in the Americas. bioRxiv, 652701.

Principais manuscritos publicados ao longo do doutorado

1) Magalhães WCS, Araujo NM, Leal TP, Araujo GS, Viriato PJS, Kehdy FS, Costa GN, Barreto ML, Horta BL, Lima-Costa MF, Pereira AC, Tarazona-Santos E, Rodrigues MR; **Brazilian EPIGEN Consortium***. EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. Genome Res. (**2018)**. Jul;28(7):1090-1095. doi: 10.1101/gr.225458.117. Epub 2018 Jun 14.a. ***Brazilian EPIGEN Consortium**: Isabela O. Alvim, Victor Borda, Mateus H. Gouveia, Moara Machado, **Rennan G. Moreira**, ...
2) França, J. A., de Sousa, S. F., **Moreira, R. G**., Bernardes, V. F., Guimarães, L. M., Santos, J. N., ... & Gomes, C. C. (**2018**). Sporadic granular cell tumours lack recurrent

mutations in PTPN11, PTEN and other cancer-related genes. Journal of clinical pathology, 71(1), 93-94.

3) Lopes, M. R., Batista, T. M., Franco, G. R., Ribeiro, L. R., Santos, A. R., Furtado, C., **Moreira RG**, ... & Lachance, M. A. (**2018**). *Scheffersomyces stambukii* fa, sp. nov., a D-xylose-fermenting species isolated from rotting wood. International journal of systematic and evolutionary microbiology, 68(7), 2306-2312.
4) Pereira, T. D. S. F., Diniz, M. G., França, J. A., **Moreira, R. G.**, de Menezes, G. H. F., de Sousa, S. F., ... & Gomez, R. S. (**2018**). The Wnt/β-catenin pathway is deregulated in cemento-ossifying fibromas. Oral surgery, oral medicine, oral pathology and oral radiology, 125(2), 172-178.
5) de Sousa, S. F., Diniz, M. G., França, J. A., Pereira, T. D. S. F., **Moreira, R. G**., dos Santos, J. N., ... & Gomes, C. C. (**2018**). Cancer genes mutation profiling in calcifying epithelial odontogenic tumour. Journal of clinical pathology, 71(3), 279-283.
6) Batista, T. M., Hilário, H. O., **Moreira, R. G**., Furtado, C., Godinho, V. M., Rosa, L. H., ... & Rosa, C. A. (**2017**). Draft genome sequence of *Metschnikowia australis* strain UFMG-CM-Y6158, an extremophile marine yeast endemic to Antarctica. Genome Announc., 5(20), e00328-17.
7) Batista, T. M., **Moreira, R. G**., Hilário, H. O., Morais, C. G., Franco, G. R., Rosa, L. H., & Rosa, C. A. (**2017**). Draft genome sequence of *Sugiyamaella xylanicola* UFMG-CM-Y1884T, a xylan-degrading yeast species isolated from rotting wood samples in Brazil. Genomics data, 11, 120-121.
8) Santos, J. N., Sousa Neto, E. S., França, J. A., Diniz, M. G., **Moreira, R. G.**, Castro, W. H., ... & Gomes, C. C. (**2017**). Next-generation sequencing of oncogenes and tumor suppressor genes in odontogenic myxomas. Journal of Oral Pathology & Medicine, 46(10), 1036-1039.
9) de Siqueira, E. C., de Sousa, S. F., França, J. A., Diniz, M. G., Pereira, T. D. S. F., **Moreira, R. G.,** ... & Gomes, C. C. (**2017**). Targeted next-generation sequencing of glandular odontogenic cyst: a preliminary study. Oral surgery, oral medicine, oral pathology and oral radiology, 124(5), 490-494.
10) de Sousa, S. F., **Moreira, R. G.**, Gomez, R. S., & Gomes, C. C. (**2016**). Interrogation of cancer hotspot mutations in 50 tumour suppressor genes and oncogenes in calcifying cystic odontogenic tumour. Oral oncology, 100(57), e1-e3.
11) Gomes, C. C., De Sousa, S. F., de Menezes, G. H., Duarte, A. P., Pereira, T. S., **Moreira, R. G.**, ... & Gomez, R. S. (**2016**). Recurrent KRAS G12V pathogenic mutation in adenomatoid odontogenic tumours. Oral oncology, 56, e3.
12) Kehdy, F. S., Gouveia, M. H., Machado, M., Magalhães, W. C., Horimoto, A. R., Horta, B. L., **Moreira R.G**., ... & The Brazilian EPIGEN Project Consortium (**2015**). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proceedings of the National Academy of Sciences, 112(28), 8696-8701.

Artigo submetido ao periódico Leukemia (formato *Letter*) em 27/10/2019

HIGH FREQUENCY OF THE ACUTE LYMPHOBLASTIC LEUKEMIA RELAPSE BIOMARKER *PDE4B*-rs6683977 GIVES SUPPORT FOR AN EXTRA PHASE OF CHEMOTHERAPY IN AMERINDIANS

Running title: Acute lymphoblastic leukemia relapse biomarkers

Rennan G. Moreira[1,2], Julia M. Saraiva-Duarte[1], Meddly L. Santolalla[1,3] , Alexandre C. Pereira[4], Martha Sosa-Macias[5,6], Carlos Galaviz[5,6], Wagner C.S. Magalhães[1,7], Camila Zolini[1], Thiago P. Leal[1], Zsolt Balázs[8,9,10], Adrián Llerena[6,11], Jose Geraldo Mill[12], Robert H. Gilman[3,13], Eduardo Tarazona-Santos[1,3,6,14§], Fernanda Rodrigues-Soares[1,6,15§§]

[1] Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-910, Belo Horizonte, Minas Gerais, Brazil

[2] Centro de Laboratórios Multiusuários, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-910, Belo Horizonte, Minas Gerais, Brazil

[3] Universidad Peruana Cayetano Heredia, 15102, Lima, Peru

[4] Laboratory of Genetics and Molecular Cardiology, Heart Institute, Medical School of University of São Paulo, 05403-900, São Paulo, Brazil

[5] Instituto Politécnico Nacional, CIIDIR Unidad Durango. Durango, Mexico

[6] RIBEF Ibero-American Network of Pharmacogenetics and Pharmacogenomics, Badajoz, Extremadura, Spain

[7] Núcleo de Ensino e Pesquisas do Instituto Mário Penna - NEP-IMP, 30180-490, Belo Horizonte, Minas Gerais, Brazil

[8] Chair of Medical Informatics, Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

[9] Biomedical Informatics, University Hospital of Zurich, Zurich, Switzerland

[10] Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

[11] Instituto de Investigación Biosanitaria de Extremadura, Universidad de Extremadura, SES, Badajoz, Extremadura, Spain

[12] Departamento de Ciências Fisiológicas, Centro de Ciências da Saúde, Universidade Federal do Espírito Santo, 29042-755, Vitória, Espírito Santo, Brazil
[13] Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[14] Instituto de Estudos Avançados Transdisciplinares, Universidade Federal de Minas Gerais, 31270-910, Belo Horizonte, Minas Gerais, Brazil
[15] Departamento de Patologia, Genética e Evolução, Instituto de Ciências Biológicas e Naturais, Universidade Federal do Triângulo Mineiro, 38025-350, Uberaba, Minas Gerais, Brazil

The authors declare no conflict of interest.

Corresponding author

[§]Dr. Eduardo Tarazona-Santos
Departamento de Biologia Geral
Instituto de Ciências Biológicas
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Pampulha
Belo Horizonte, MG, CEP 31270-910, Brazil
Telephone: +55 31 3409-2597
E-mail: edutars@icb.ufmg.br

Ancestry is associated with incidence and treatment outcome of several types of cancer (1). However, our understanding of these associations is limited by the fact that most studies are performed in European/US populations (2, 3). In particular, Native Americans are neglected in studies about the genetic diversity and genetic basis of diseases (3).

Acute lymphoblastic leukemia (ALL) is the most common cancer in <5 years-old children. Its incidence is higher in countries with high Native American ancestry (4). An admixture mapping in US-Hispanics (those who are product of admixture between Native Americans, Europeans and Africans) reported that alleles *PDE4B*-rs6683977.C (minus strand) (intronic) and *MYT1L*-rs17039396.A (intronic) increase the risk of ALL-relapse (5), even after adjusting for known prognostic factors. Since in the studied patients these alleles were preferentially present in genomic regions of Native American origin, these ancestry-based associations account in part for the association between Native American ancestry and ALL-relapse (6). Importantly, this ancestry-associated higher risk of relapse disappeared with an extra phase of chemotherapy, suggesting that at population level an ancestry-adjusted therapy can mitigate the risk of relapse (5). The *PDE4B*-rs6683977.C association with ALL-relapse has been replicated in an ethnically diverse cohort of St. Jude Hospital (TN, U.S.), and other three *PDE4B* SNPs (rs546784.A, rs641262.A and rs524770.A) were associated with ALL-relapse (7).

MYT1L encodes a transcription factor related with neuronal differentiation. Although the role of *MYT1L* in ALL has not been elucidated, *MYT1L* is involved in prognosis of cardia cancer (8) and it is among the most variably expressed transcription factors in ALL patients (9). *PDE4B* encodes for a phosphodiesterase that control cyclic-AMP (cAMP) degradation and is a drug-target in B-cell malignancies, including ALL (10).
Following B-cell receptor activation, cAMP down-modulates signaling pathways responsible for cell proliferation. *PDE4B* overexpression abrogates the cAMP inhibitory effects in B-cell lymphomas and is associated with glucocorticoid resistance in ALL treatment (10). Since glucocorticoids (i.e. dexamethasone and prednisone) are essential in ALL treatment (5), the role of *PDE4B* diversity might be important in ALL treatment outcomes.

The associations between *PDE4B*-rs6683977.C and *MYT1L*-rs17039396.A and ALL-relapse on US-Hispanics suggest that indigenous populations of the Americas have higher frequencies of these ALL-relapse associated variants. However, this suggestion has never

been corroborated in Native American populations (3). Here we test the hypothesis that *PDE4B* rs6683977.C, *MYT1L*-rs17039396.A and their associated genomic regions show higher frequency and are highly differentiated in Native American populations in respect to other continental groups.

PDE4B-rs6683977 has two alleles: C and G. A methodological caveat frequently occurs with SNPs with the pairs of alleles C/G or A/T, requiring caution. These kinds of SNPs are ambiguous because the alternative alleles (that by definition are present in the different units of the pair of homologous chromosomes) are sometimes confused with the complementary bases of the double DNA strands. *PDE4B*-rs6683977.C allele, reported as the risk allele for ALL-relapse, designates the allele present in the minus (-) strand (5, 7). However, to follow the convention (such as the 1000 Genome Project – 1000GP (11)) and be consistent with strand nomenclature (12), alleles reported in this study hereafter will refer to the plus (+) strand. Thus, the *PDE4B* risk alleles will be rs6683977.G, rs546784.T, rs641262.T and rs524770.A.

Populations, samples and assays

We studied populations considered as Native Americans from Mexico (*Tarahumara* and *Huichol* ethnicities), Peru (*Quechua*, *Aymara*, *Machiguenga* and *Ashaninka* ethnicities) and Brazil (*Tupiniquim* and *Guarani* ethnicities) and also Brazilian admixed populations from Minas Gerais State (AMG). We estimated *PDE4B*-rs6683977 and *MYT1L*-rs17039396 allele frequencies after genotyping 790 (Table 1) and 489 (Table S1) individuals, respectively, and estimated continental admixture proportions of each population (see Supplementary Material). Data from 1000GP were considered for comparison. Institutional review board of participant institutions approved this study.

Results

Ancestry and allele frequencies

Native Americans from Peru and Mexico have <5% of non-Native American admixture, except for *Quechuas* (mean 12.6%, Table 1, Figure S1). Native Americans from Brazil have more Old-World admixture, but are mainly Native American (58.8%). Admixed individuals from Minas Gerais are predominantly European (Table 1, Figure S1).

Populations with the highest Native American ancestry proportions showed the highest *PDE4B*-rs6683977.G frequencies (from 0.66 in *Tupiniquim* to 0.93 in *Aymara*, Table 1, Figure

S2), confirming our hypothesis (Figure S3). Comparing to the 1000GP populations, *PDE4B*-rs6683977.G frequencies in Native Americans (except for the *Tupiniquim* population) are similar to Africans (0.96-0.99) and East Asians (0.84), and are higher than in 1000GP admixed US-Hispanics/Latin Americans (0.72) and South Asians (0.65) (Table 1, Figure S2). Differently, the G allele frequency drops in Europeans (0.44). Thus, European admixture influences the *PDE4B*-rs6683977.G frequencies in *Tupiniquim*, AMG, and 1000GP admixed US-Hispanics/Latin Americans populations (Table 1).

The frequencies range of *MYT1L*-rs17039396.A is lower worldwide (5-37%) than for *PDE4B*-rs6683977.G, but they reach the highest frequencies in Native American Peruvian populations [particularly in *Ashaninkas* and *Machiguengas* from Peru (24-37%), Table S1, Figure S2], confirming our hypothesis based on the ALL-relapse admixture mapping on US-Hispanics (Figure S3). However, there is substantial variation among Native American, as observed for other pharmaco-alleles (13). 1000GP East Asians have a *MYT1L*-rs17039396.A frequency similar to the Native Americans populations from Peru (Table S1).

Linkage disequilibrium with admixture mapping hits

To identify SNPs in linkage disequilibrium (LD, $r^2 \geq 0.8$) with *PDE4B*-rs6683977.G and *MYT1L*-rs17039396.A we analyzed three databases including native and admixed Latin American populations and additional SNPs in the genomic regions surrounding these variants (Figure S4): (i) **TargetSeq** dataset, product of next-generation target-sequencing of 20 Kb centered in *PDE4B*-rs6683977 and *MYT1L*-rs17039396 SNPs; (ii) **BeadXpress**, product of genotyping 21 and 27 SNPs in *PDE4B* and *MYT1L,* respectively (including rs6683977 and rs17039396); (iii) **TaqMan-rs6683977+2.5M** dataset, comprised *PDE4B*-rs6683977 genotypes obtained by TaqMan Assay ID: C___1270960_10 (Thermo Fisher^{TM}) and genotypes of 407 *PDE4B* SNPs assessed with the HumanOmni2.5-8v1.1 array (Illumina, Inc.).

The TargetSeq dataset indicated that the hit *MYT1L*-rs17039396 is in LD with the intronic *MYT1L*-rs60585106 in the *Ashaninka* and *Machiguenga* populations, but no other SNP was found in LD ($r^2 \geq 0.8$) with *MYT1L*-rs17039396 in any population or dataset (Table S1).

Conversely, all datasets revealed SNPs in LD with the *PDE4B*-rs6683977, especially the TargetSeq dataset (Table 1). Intronic variants in LD with *PDE4B*-rs6683977 in the BeadXpress and TaqMan-rs6683977+2.5M datasets have not been reported as related with any condition

or disease. However, intronic SNPs *PDE4B*-rs546784 and -rs641262, in LD with rs6683977 in Native Peruvian and Native Mexican of the TargetSeq dataset, and also in 1000GP Europeans, were associated with ALL-relapse in the multiethnic St. Jude cohort (7).

While a functional genomic study did not clarify if rs6683977 directly affects *PDE4B* expression, other three SNPs in the same intron (rs494735, rs502958 and rs12142375) were classified as 'positive regulatory elements' (14). Such SNPs are in high LD in Chinese populations with the *PDE4B*-rs546784 (14), which is included in our TargetSeq dataset and was associated with ALL-relapse in the multiethnic St. Jude study (7). Using the *PDE4B*-rs546784 as a tag-SNP, the same study (14) has shown allelic-specific regulatory effects of *PDE4B*-rs12142375, suggesting that the association of such SNP with ALL risk is related with overexpression of *PDE4B*. This result corroborates robust evidences that the region encompassing the *PDE4B*-rs6683977 is a spatially active chromatin segment harboring active enhancers (15). Indeed, functional databases (see Supplementary Material for details) reveal that the region around the *PDE4B*-rs6683977 has DNase I hypersensitive binding sites, transcription factor ChIP-seq clusters and histone marks of active enhancers (H3K4me1 and H3K27ac) in a lymphoblastoid cell line (Figure 1A). Interestingly, our data revealed that the *PDE4B*-rs546784 SNP is in LD with the *PDE4B*-rs6683977 and rs641262 in the Native Americans populations (Figure 1B). Notably, we show that the sum of frequencies of *PDE4B* haplotypes including the ALL-relapse risk alleles - rs546784.T, rs6683977.G and rs641262.T – is higher in Native Americans from Mexico (0.92) and Peru (0.82-0.93) than in Europeans (0.43) and other 1000GP populations (Table S2).

In conclusion, we clarified a technical issue regarding rs6683977 alleles nomenclature that prevents a better understanding of its effect in ALL. *PDE4B*-rs6683977.G is highly prevalent in Native Americans and Africans, but only in Native Americans it is in LD with other SNPs that have been reported to be associated with ALL-relapse (7). Evidences such as active chromatin likely harboring active enhancers (15) support the regulatory role of the region surrounding the *PDE4B*-rs6683977. Accordingly, the importance of *PDE4B* for the ALL treatment outcome is related with the association of its overexpression with glucocorticoids resistance (5, 10). Our findings are consistent with the admixture mapping hits found by Yang *et al.* (5), expand their results by revealing the very high prevalence of *PDE4B*-rs6683977.G and associated haplotypes in Native Americans, which may be related with *PDE4B* overexpression. Consequently, our results provide support to the therapeutic decision of an

extra phase of chemotherapy (5) in populations with predominant Native American ancestry, such as Mexico, Guatemala and the Andean region of South America.

ACKNOWLEDGEMENTS

The authors declare no conflict of interest. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) funded this study.

REFERENCES

1. Özdemir BC, Dotto GP. Racial differences in cancer susceptibility and survival: more than the color of the skin? Trends in cancer. 2017;**3**:181-197.

2. Soares-Souza G, Borda V, Kehdy F, Tarazona-Santos E. Admixture, Genetics and Complex Diseases in Latin Americans and US-Hispanics. Current Genetic Medicine Reports. 2018;**6**:208-23.

3. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;**177**:26-31.

4. Miranda-Filho A, Piñeros M, Ferlay J, Soerjomataram I, Monnereau A, Bray F. Epidemiological patterns of leukaemia in 184 countries: a population-based study. The Lancet Haematology. 2018;**5**:e14-e24.

5. Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, *et al.*. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nature genetics. 2011;43:237.

6. Yang JJ, Yang W, Cheng C, Devidas M, Cao X, Campana D, *et al.*. Genetically Defined Racial Differences Underlie Risk of Relapse in Childhood Acute Lymphoblastic Leukemia. Blood. 2008; **112**: A-14.

7. Yang JJ, Cheng C, Devidas M, Cao X, Campana D, Yang W, *et al.*. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. Blood. 2012;**120**:4197-204.

8. Zhang Y, Zhu H, Zhang X, Gu D, Zhou X, Wang M, *et al.*. Clinical significance of MYT1L gene polymorphisms in Chinese patients with gastric cancer. PloS one. 2013;**8**:e71979.

9. Tomar AK, Agarwal R, Kundu B. Most Variable Genes and Transcription Factors in Acute Lymphoblastic Leukemia Patients. Interdisciplinary Sciences: Computational Life Sciences. 2019:1-11.

10. Cooney JD, Aguiar RC. Phosphodiesterase 4 inhibitors have wide-ranging activity in B-cell malignancies. Blood. 2016;**128**:2886-90.

11. 1000 Genome Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;**491**:56.
12. Nelson SC, Doheny KF, Laurie CC, Mirel DB. Is 'forward'the same as 'plus'?… and other adventures in SNP allele nomenclature. Trends in Genetics. 2012;**28**:361-3.
13. Moya GE, Penas-Lledo EM, Rodrigues-Soares F, Llerena A, Teran E, Rodeiro I, *et al.*. Genomic ancestry, CYP2D6, CYP2C9 and CYP2C19 among Latin-Americans. Clinical Pharmacology & Therapeutics. 2019. e-pub ahead of print 2 August 2019; doi: 10.1002/cpt.1598
14. Liu S, Liu Y, Zhang Q, Wu J, Liang J, Yu S, *et al.*. Systematic identification of regulatory variants associated with cancer risk. Genome biology. 2017;**18**:194.
15. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, *et al.*. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell reports. 2016;**17**:2042-59.

Figure 1. Functional signals, putative regulatory elements and linkage disequilibrium pattern surrounding the *PDE4B*-rs6683977 in Native Americans and Europeans. (A) *UCSC Genome Browser* visualization of 84,668 bp of *PDE4B* (chr1:66752830-66837497). Browser comprises (see Supplementary Material for details): (i) one custom track (*IMPET*) with predictions of enhancer-promoter interaction for lymphocytes cells; two tracks of histone marks (*H3K4Me1* (ii) and *H3K27Ac* (iii)) for the lymphoblastoid GM12878 (pink) and myelogenous leukemia K562 (purple) cell lines; two tracks with results of modelling the presence of chromatin marks for GM12878 (iv) and K562 (v) cells (*ChromHMM* - Red: Promoter, Orange: strong enhancer, Yellow: weak enhancer, Dark-Green: Transcriptional transition/elongation, Light-Green: weak transcribed); and two tracks of (vi) *cis*- (*DNaseI Hypersensitivity Clusters*) and (vii) *trans*-regulatory elements (*Transcription Factor ChIP-seq Clusters*). (B) Linkage disequilibrium between variants assessed using the TargetSeq dataset surrounding the *PDE4B*-rs6683977 (20Kb - chr1:66759100-66779100); grey scale denotes correlation between markers (black color indicates r^2>0.8). Triangle, ellipse, rectangle and invert triangle represent rs546784, rs524770, rs6683977 and rs641262, respectively. Mexico (MEX), Europeans from the 1000GP (EUR)

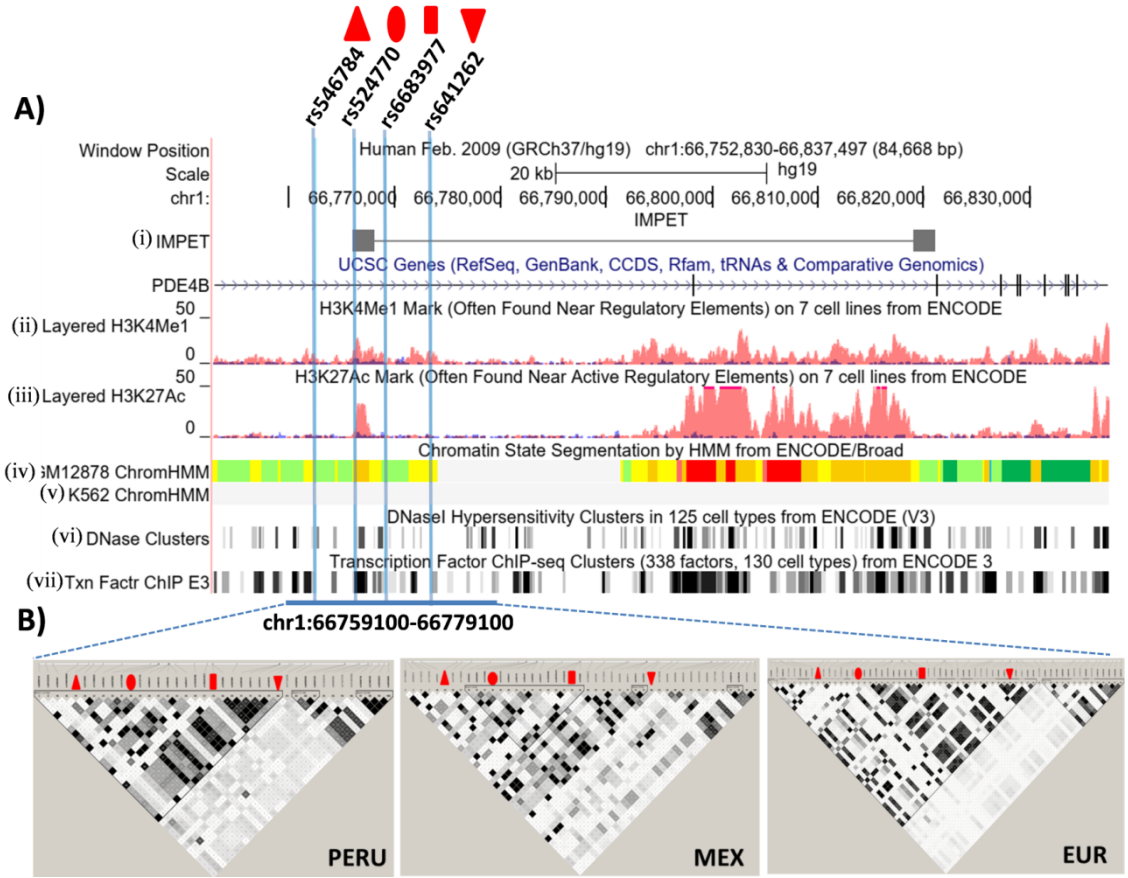


Table 1. Ancestry proportions (Native American-NAT, European-EUR and African-AFR), number of samples (N) used to calculate *PDE4B*-rs6683977 frequencies, p-value for Hardy-Weinberg equilibrium (HWE), and SNPs in linkage disequilibrium (LD) (r^2>0.8) with the *PDE4B*-rs6683977 in different datasets. 1000GP populations: Europeans (CEU, FIN, GBR, ITS, IBS), West Africans (GWD, MSL), West Central Africans (ESN, YRI), East Africans (LWK), East-Asians (CHB, JPT, CHS, CDX, and KHV), South-Asians (GIH, PJL, BEB, STU, and ITU) and Admixed U.S. Hispanics/Latin Americans (PUR, CLM, MXL and PEL). (-) refers to populations without data for a specific dataset.

POPULATION (ETHNO-LINGUISTIC GROUP) - *COUNTRY*	**ANCESTRY PROPORTIONS (%)**			**N**	**rs6683977 FREQUENCIES**				**HWE *p-value***	**SNPS IN LD (r^2>0.80)**		
	NAT	**EUR**	**AFR**		**G**	**CC**	**GC**	**GG**		**TARGETSEQ (83 SNPS)**	**BEADXPRESS (21 SNPS)**	**TAQMAN-rs6683977+2.5M (408 SNPS)**
1000 GENOME PROJECT POPULATIONS												
Europeans				503	0.44	0.34	0.45	0.21	0.046	rs638111, rs641262, rs494735, rs495477, rs12137115, rs12137080, rs6683604, rs546784, rs6668516, rs782967	rs522037, rs6683604	rs486136, rs519044
West Africans				198	0.98	0.00	0.04	0.96	1	0	0	rs6664618, rs12731764
West Central Africans				207	0.99	0.00	0.02	0.98	1	0	0	rs12141125, rs6664618
East Africans				99	0.96	0.00	0.07	0.93	1	0	0	0
East Asians				504	0.84	0.03	0.26	0.71	0.255	rs494735, rs12137115, rs6683604, rs6668516, rs546784, rs495477, rs641262, rs782967, rs12137080	rs6683604	rs519044, rs6664618, rs486136

POPULATION (ETHNO-LINGUISTIC GROUP) - *COUNTRY*	ANCESTRY PROPORTIONS (%)			N	rs6683977 FREQUENCIES				HWE *p-value*	SNPS IN LD (r^2>0.80)		
	NAT	EUR	AFR		G	CC	GC	GG		TARGETSEQ (83 SNPS)	BEADXPRESS (21 SNPS)	TAQMAN-rs6683977+2.5M (408 SNPS)
South Asians				489	0.65	0.12	0.46	0.42	0.766	rs12137080, rs12137115, rs495477, rs546784, rs6683604, rs494735, rs6668516, rs782967	rs6683604	rs519044, rs486136
Admixed U.S. Hispanics/ Latin Americans	47.0	47.0	6.0	347	0.72	0.10	0.37	0.53	0.184	rs6668516, rs6683604, rs494735, rs12137115, rs495477, rs641262, rs546784, rs12137080, rs782967	rs668360	rs519044
AMERICAS												
Tarahumara/Huichol (Uto-Aztecan) *Mexico*	96.7	2.4	0.9	22	0.90	0.0	0.18	0.82	1	rs641262, rs495477, rs12137115, rs6683604, rs546784, rs6668516	-	-
Quechuas (Quechua) *Peru*	87.4	12.5	0.1	111	0.92	0.01	0.14	0.85	0.530	rs641262, rs495477,rs12137115, rs12137080, rs6683604, rs546784, rs6668516	-	rs519044
Aymara (Quechuamaran) *Peru*	96.4	2.9	0.7	118	0.93	0.01	0.12	0.87	0.420	rs641262, rs495477, rs12137115, rs6683604, rs546784, rs6668516	0	rs519044

POPULATION (ETHNO-LINGUISTIC GROUP) - *COUNTRY*	ANCESTRY PROPORTIONS (%)			N	rs6683977 FREQUENCIES				HWE *p-value*	SNPS IN LD (r^2>0.80)		
	NAT	EUR	AFR		G	CC	GC	GG		TARGETSEQ (83 SNPS)	BEADXPRESS (21 SNPS)	TAQMAN-rs6683977+2.5M (408 SNPS)
Ashaninkas (Arawak) *Peru*	98.1	1.9	0.0	231	0.90	0.01	0.18	0.81	1	rs641262, rs495477, rs12137115, rs6683604, rs546784, rs6668516	rs6683604	rs519044
Machiguengas (Arawak) *Peru*	99.1	0.7	0.2	178	0.86	0.02	0.24	0.74	1	rs641262, rs495477, rs12137115, rs12137080, rs6683604, rs546784, rs6668516	rs6683604	rs486136 rs519044
Tupiniquim (Tupi) South East *Brazil*	49.9	27.3	22.8	16	0.66	0	0.81	0.19	0.013	0	0	-
Guarani (Guarani) South East *Brazil*	67.6	20.0	12.4	16	0.92	0	0.13	0.87	1	0	0	-
Brazil Admixed South East Brazil	14.9	54.9	30.2	98	0.64	0.13	0.46	0.41	1	-	rs6683604	-

SUPPLEMENTARY MATERIAL

SAMPLES AND DATASETS

We studied individuals comprising populations consensually considered as Native Americans from Mexico (*Tarahumara* (TAH) and *Huichol* (HUI) ethnicities), Peru (*Quechua* (QUE), *Aymara* (AYM), *Machiguenga* (MAC) and *Ashaninka* (ASH) ethnicities) and Brazil (*Tupiniquim* (TUP) and *Guarani* (GUA) ethnicities) and also Brazilian admixed populations from the Minas Gerais state (AMG). Analyses were performed using all or a combination of different datasets available (described in detail below).

TargetSeq dataset

A total of 150 samples were used to build this dataset (QUE = 20, AYM = 24, MAC = 22, ASH = 16, TUP = 22, GUA = 24, HUI = 12, TAH = 10). Intronic regions sizing 20Kb surrounding the *PDE4B*-rs6683977 (chr1:66759100-66779100, GRCh37) and *MYT1L*-rs17039396 (chr2:2215144-2235144, GRCh37) were sequenced using the HaloPlex library strategy (Agilent Inc., Catalog Number 5190-5436, Lot 6248658). The Agilent SureDesign software was used to design targeting probes according to the Genome Reference Consortium Human Build 37 (GRCh37). A total of 150 samples were grouped into three different sets of approximately 50 samples that were sequenced independently using three Illumina MiSeq v3 600 cycles cartridges. Sequencing runs were performed in the Center of Multiusers Laboratories in the Universidade Federal de Minas Gerais (CELAM/UFMG). Variants were called according to the GRCh37 using a consensus between the GATK v4.0 (Broad Institute) (16) and the Surecall (Agilent Tech.) proprietary software. Data was stored in variant calling format files (.vcf) and comprised only single nucleotide polymorphisms (SNPs). Given that 24 samples were also genotyped with the HumanOmni2.5-8v1.1 array (Illumina, Inc.) we used a customized script developed in *python* to apply the *mpileup* tool of the Samtools (17) software to increase confidence over those SNPs called using the GATK and SureCall. The complete variant calling pipeline is available for public access at https://github.com/ldgh/Targeted_Sequencing_Pipeline.

Beadxpress dataset

This dataset comprised 440 samples (AYM=91, MAC=72, ASH=91, TUP=45, GUA=43, AMG=98). We used the Illumina GoldenGate VeraCode strategy of the BeadXpress system in the CELAM/UFMG to genotype 26 and 21 SNPs along 1 Mb surrounding the *MYT1L*-rs17039396 and the *PDE4B*-rs6683977 respectively. Such 47 SNPs were defined as those showing the highest number of tag-SNPs after linkage disequilibrium analyses performed with the GLU software (https://code.google.com/archive/p/glu-genetics/) using the Tagzilla tool and the YRI and CEU populations from the HapMap Project (18). Overall, those 21 *PDE4B*-SNPs and 27 *MYT1L*-SNPs spanned other 468 and 363 SNPs, respectively. Finally, BeadXpress system genotypes were called using the Illumina GenomeStudio™ Genotyping software after applying the *GenTrain* (<30%), *GenCall* (<40%) and *Call Rate* (<90%) cutoffs (19).

Taqman-rs6683977 dataset

The TaqMan dataset was built by genotyping the *PDE4B*-rs6683977 using the TaqMan Assay ID: C___1270960_10 (Thermo Fisher™) in a total of 719 samples (QUE=100, AYM=118, ASH=226, MAC=177, AMG=98). The SDS 2.4 software was used for genotype calling and only those SNPs with quality index above 97 were kept in the dataset. Considering that among the 719 samples genotyped with the TaqMan method 375 samples were also genotyped using the BeadXpress system (BeadXpress dataset) and the target sequencing strategy (TargetSeq dataset), TaqMan dataset genotypes were used as reference for other datasets quality verification (Figure S4).

Taqman-rs6683977+2.5M dataset

We used the HumanOmni2.5-8v1.1 Illumina array which included 2 391 739 variants to genotype a total of 120 samples (ASH=42, QUE=21, MAC=41, AYM=16). Genotypes were first called using the Illumina GenomeStudio™ Genotyping software and then converted to PLINK (.ped and .map) format (20). Following, several quality control steps were performed, such as the *Call Rate* (<90%) cutoff and also the PLINK *geno* and *mind* filters, keeping only those SNPs with less than 10% of missing data. Considering that the HumanOmni2.5-8v1.1 array did not include the *MYT1L*-rs17039396 and the *PDE4B*-rs6683977 SNPs, this dataset was restricted to the *PDE4B*-rs6683977 and was built by merging the *PDE4B*-rs6683977 genotypes generated in the

other datasets (TargetSeq, BeadXpress and TaqMan) with those genotypes of the 407 *PDE4B* variants assessed with the HumanOmni2.5-8v1.1 (Figure S4).

FREQUENCIES AND LINKAGE DISEQUILIBRIUM ANALYSES

Allele frequencies were calculated using the highest number of samples available including all datasets. Primarily, 849 and 505 samples of the *PDE4B*-rs6683977 and *MYT1L*-rs17039396 SNPs, respectively, were considered for allele frequencies calculation. However, after filtering samples with missing or low quality data, the total numbers of samples were 790 and 489 for *PDE4B*-rs6683977 and *MYT1L*-rs17039396 SNPs respectively. Linkage disequilibrium (LD) analysis of the *PDE4B*-rs6683977 region was performed using four different datasets (TargetSeq, BeadXpress, TaqMan-rs6683977, and TaqMan-rs6683977+2.5M), whereas two of them (TargetSeq, BeadXpress) were used to assess linkage disequilibrium surrounding the *MYT1L*-rs17039396. Haploview (21) was used to LD analyses.

PUBLIC DATABASE

We accessed the 1000 Genome Project (11) phase 3 data in the Ensembl GRCh37 genome reference database. We used the Ensembl *VCF to PED converter* tool to obtain the same *PDE4B* and *MYT1L* variants of the TargetSeq, BeadXpress and Taqman-rs6683977+2.5M datasets from five European populations: CEU (Utah Residents with Northern and Western European Ancestry), TSI (Toscani in Italia), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian Population in Spain); five African populations: YRI (Yoruba in Ibadan, Nigeria), LWK (Luhya in Webuye, Kenya), GWD (Gambian in Western Divisions in the Gambia), MSL (Mende in Sierra Leone), ESN (Esan in Nigeria); four American populations: PUR (Puerto Ricans from Puerto Rico), CLM (Colombians from Medellin, Colombia), MXL (Mexican Ancestry from Los Angeles, USA) and PEL (Peruvians from Lima, Peru); five South-Asian populations: GIH (Gujarati Indian from Houston, Texas, USA), PJL (Punjabi from Lahore, Pakistan), BEB (Bengali from Bangladesh), STU (Sri Lankan Tamil from the UK), and ITU (Indian Telugu from the UK); and five East-Asian populations CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan), CHS (Southern Han Chinese, China), CDX (Chinese Dai in Xishuangbanna, China), and KHV (Kinh in Ho Chi Minh City, Vietnam).

ANCESTRY ANALYSES

We used the Illumina GoldenGate VeraCode strategy of the BeadXpress system, in the CELAM/UFMG, to genotype 96 ancestry informative markers (AIMs) reported by Yaeger *et al.* (22) in all those 440 samples of the BeadXpress dataset described above. The same AIMs of the samples from the CEU – Utah/USA (n=174) and the YRI-Ibadan/Nigeria (n=176) from the HapMap Project (18) were merged with the BeadXpress dataset and considered as reference of European and African ancestries, respectively. The ADMIXTURE (23) software with K set to 3 clusters was used to infer the European, African and Native American ancestry proportions of all the 440 samples (Figure S1). The ancestry proportion of each population was calculated as the average of individuals' proportions (Table 1 and Table S1).

Additionally, we also report ancestries proportions estimation for the admixed U.S. Hispanics/Latin Americans (AMR) populations from the 1000GP (PUR, PEL, MXL and CLM) (Tables 1 and S1). Ancestry proportions of AMR were also inferred using the ADMIXTURE software in unsupervised mode for three ancestry clusters (K=3). Since ADMIXTURE analysis assumes independence among SNPs, we pruned for linkage disequilibrium (LD) removing highly linked SNPs with the PLINK 1.7 with the option indep-pairwise 200 25 0.4.

FUNCTIONAL INFERENCES

Functional importance of the *PDE4B* 20Kb region sequenced in the TargetSeq dataset was assessed and visualized in the UCSC *Genome Browser* (Figure 1). The *Encyclopedia of DNA Elements* (ENCODE) (24) databases "Integrated Regulation" and "Histone Modification" were selected to be shown. We restricted the former database to only depict the *Transcription Factor ChIP-seq Clusters*, *H3K4Me1 Marks* (often found near regulatory elements), *H3K27Ac Marks* (often found near active regulatory elements), and *DNaseI Hypersensitivity Clusters* (indicating DNA exposure due to less condensed chromatin) tracks because these were the most informative traits. The latter database contained the ENCODE/Broad *Chromatin State Segmentation by HMM*, which is an extensive dataset comprising results of the ChromHMM software (25). Such tool integrates multiple chromatin datasets and uses multivariate Hidden Markov Model to explicitly model the presence or absence of chromatin marks. Importantly, we restricted results of the ENCODE tracks to only show signals in GM12878 and K562 cells, which are lymphoblastoid and myelogenous leukemia cell lines, respectively. We also added a

custom track built with the most relevant regions retrieved from the 4D Genome (26). The 4D Genome is a public and curated database of chromatin interaction data obtained from computational predictions and also experimental studies. Results most relevant were those obtained with the 4D-Genome/IM-PET (Integrated Method for Predicting Enhancer Targets) software, which is a statistical predictor for enhancer-promoter interactions. Likewise, we also restricted the custom track to only show results of lymphocytes cells (CD4 naive and CD4 T).

REFERENCES

16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;**20**:1297-303.

17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;**25**:2078-9.

18. International HapMap Consortium .The international HapMap project. Nature. 2003;**426**:789.

19. Hughes-Stamm S, Barash M, Grisedale K, van Daal A. Initial Evaluation of A96-Plex Goldengate® Genotyping SNP Assay with Suboptimal and Whole Genome Amplified Samples. Journal of Forensic Investigation. 2013; **1**:8

20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;**81**:559-75

21. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2004;**21**:263-5

22. Yaeger R, Avila-Bront A, Abdul K, Nolan PC, Grann VR, Birchette MG, et al. Comparing genetic ancestry and self-described race in African Americans born in the United States and in Africa. Cancer Epidemiology and Prevention Biomarkers. 2008;**17**:1329-38.

23. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics. 2011;**12**:246.

24. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;**306**:636-40.

25. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;**9**:215.

26. Teng L, He B, Wang J, Tan K. 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics. 2015;**31**:2560-4.

Figure S1. Relative proportions of the European (red), African (blue) and Native American (green) ancestries in the European descent and African populations from the HapMap database (CEU and YRI) and Native Americans from Peru (QUE:*Quechuas*, AYM:*Aymaras*, ASH:*Ashaninkas*, MAC:*Machiguengas*) and Mexico (MEX:*Tarahumaras* and *Huicholes*), Brazilian Native Americans (TUP:*Tupiniquins* and GUA:*Guaranis*) and Brazilian admixed (AMG) populations.

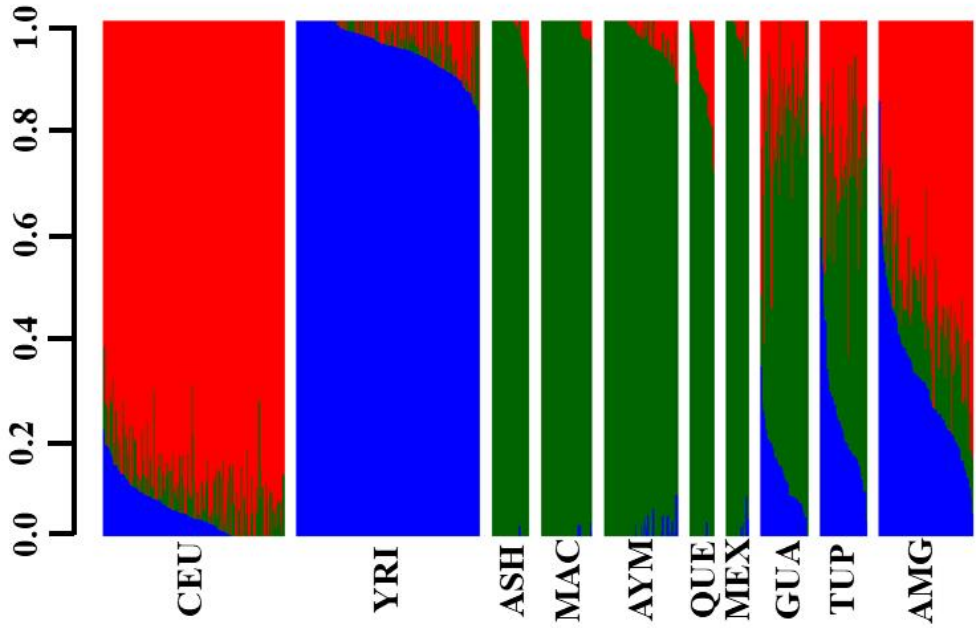


Figure S2. *PDE4B*-rs6683977.**G** and *MYT1L*-rs17039396.**A** worldwide frequencies distributions. Native Americans from Peru (QUE: *Quechuas*, AYM: *Aymaras*, ASH: *Ashaninkas*, MAC: *Machiguengas*) and Mexico (MEX: *Tarahumaras* and *Huicholes*), Brazilian Native Americans (TUP: *Tupiniquins* and GUA: *Guaranis*) and Brazilian admixed (AMG) populations. 1000 Genomes Project populations – AMR: Admixed U.S Hispanics/Latin Americans (PUR, CLM, MXL and PEL), EUR: Europeans (CEU, FIN, GBR, ITS, IBS), WAFR: West Africans (GWD and MSL), WCAFR: West Central Africans (ESN and YRI), EAFR: East African (LWK), SAS: South-Asians (GIH, PJL, BEB, STU, and ITU), EAS: East-Asians (CHB, JPT, CHS, CDX, and KHV).

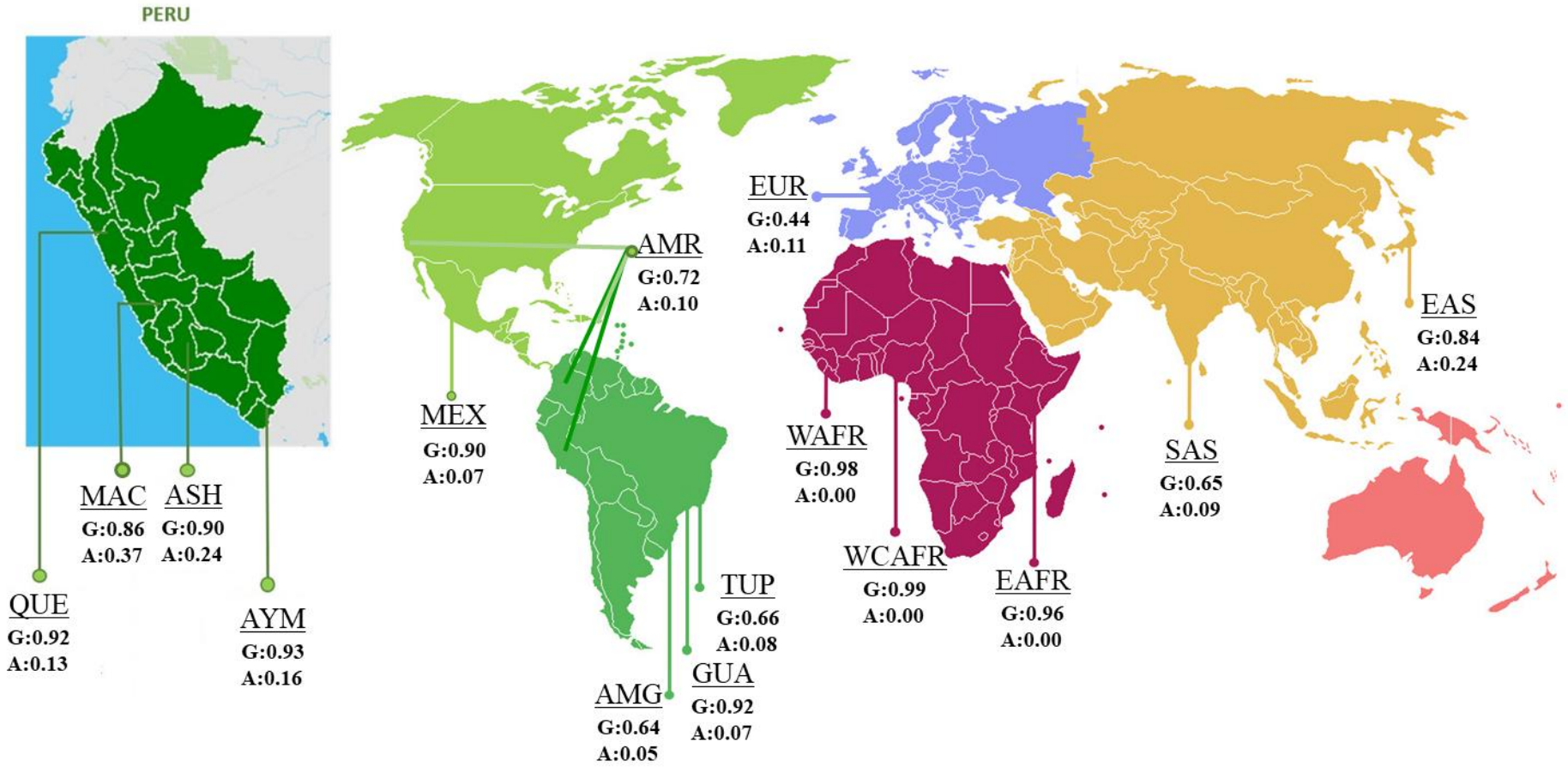


Figure S3. *PDE4B*-rs6683977.**G** (void) and *MYT1L*-rs17039396.**A** (solid) allele frequencies as a function of Native American ancestry proportions in each Native or admixed population. Native Americans from Peru (QUE: *Quechuas*, AYM: *Aymaras*, ASH: *Ashaninkas*, MAC: *Machiguengas*) and Mexico (MEX: *Tarahumaras* and *Huicholes*), Brazilian Native Americans (TUP: *Tupiniquins* and GUA: *Guaranis*) and Brazilian admixed (AMG) populations. AMR: Admixed U.S. Hispanics/Latin Americans from 1000GP (PUR, CLM, MXL and PEL).

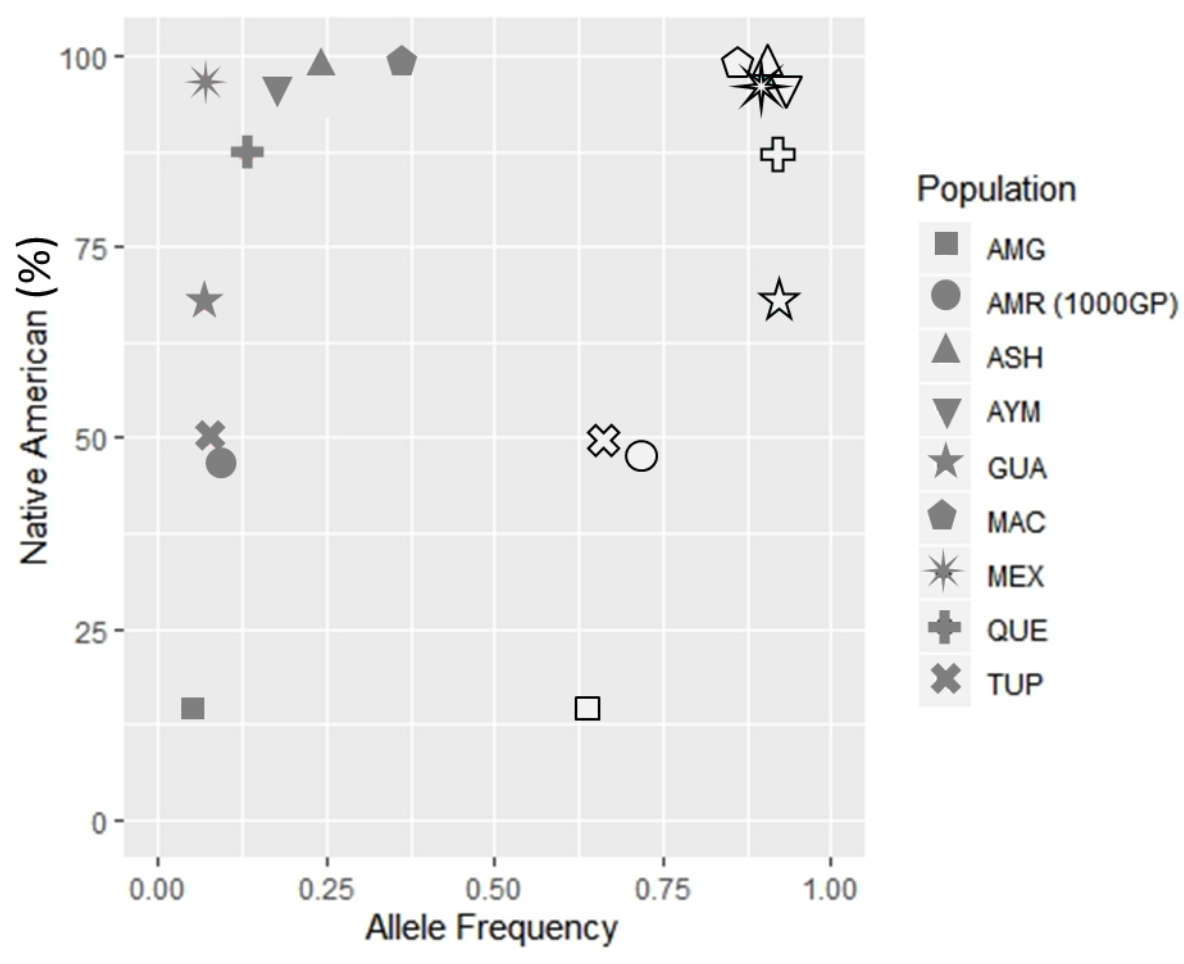


Figure S4. Venn diagram depicting number of samples shared by and exclusive of each dataset: TargetSeq, BeadXpress, TaqMan-rs6683977, and TaqMan-rs6683977+2.5M.

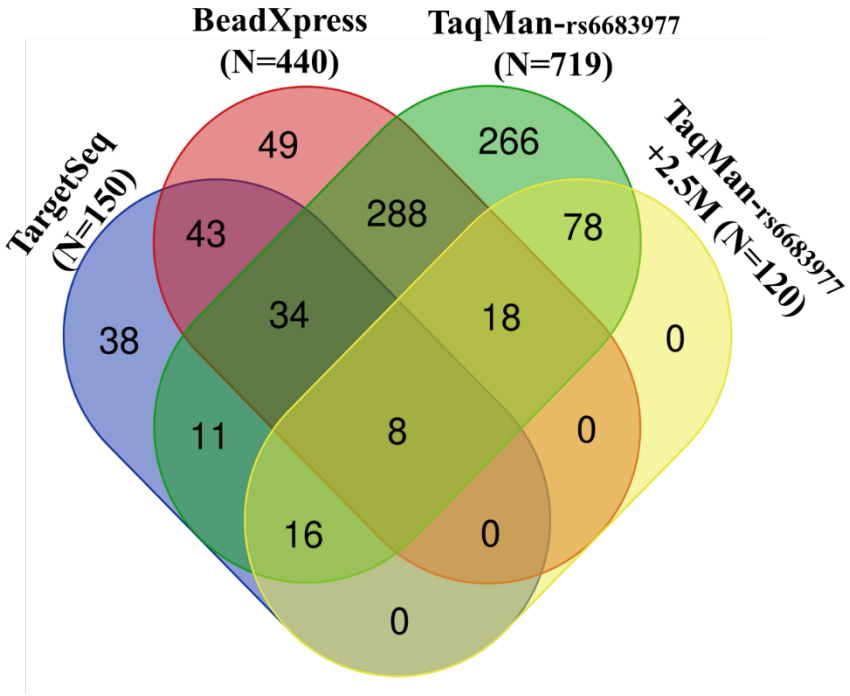


Table S1. Ancestry proportions (Native American-NAT, European-EUR and African-AFR), overall number of samples (N) used to calculate *MYT1L*-rs17039396 frequencies, p-value for Hardy-Weinberg equilibrium (HWE), SNPs in linkage disequilibrium (LD) (r^2>0.8) with the *MYT1L*-rs17039396 in different datasets. 1000GP populations: Europeans (CEU, FIN, GBR, ITS, IBS), West Africans (GWD, MSL), West Central Africans (ESN, YRI), East Africans (LWK), East-Asians (CHB, JPT, CHS, CDX, and KHV), South-Asians (GIH, PJL, BEB, STU, and ITU) and Admixed U.S. Hispanics/Latin Americans (PUR, CLM, MXL and PEL). (-) refers to populations without data for a specific dataset.

POPULATION (ETHNO-LINGUISTIC GROUP) - *COUNTRY*	**ANCESTRIES PROPORTIONS (%)**			**N**	**rs17039396 FREQUENCIES**				**HWE**	**SNPS IN LD (r^2>0.80)**	
	NAT	**EUR**	**AFR**		**A**	**GG**	**GA**	**AA**	***p-value***	**TARGETSEQ (66 SNPS)**	**BEADXPRESS (27 SNPS)**
1000 GENOMES PROJECT POPULATIONS											
Europeans				503	0.11	0.78	0.21	0.01	0.659	0	0
West Africans				198	<0.1	1.00	0.00	0.00	1.000	0	0
West Central Africans				207	<0.1	1.00	0.00	0.00	1.000	0	0
East Africans				99	<0.1	1.00	0.00	0.00	1.000	0	0
East Asians				504	0.24	0.58	0.37	0.05	0.910	0	0
South Asians				489	0.09	0.85	0.13	0.02	0.009	0	0
Admixed U.S. Hispanics/Latin Americans				347	0.10	0.81	0.19	0.00	0.066	0	0

POPULATION (ETHNO-LINGUISTIC GROUP) - *COUNTRY*	ANCESTRIES PROPORTIONS (%)			N	rs17039396 FREQUENCIES				HWE	SNPS IN LD (r^2>0.80)	
	NAT	EUR	AFR		A	GG	GA	AA	*p-value*	TARGETSEQ (66 SNPS)	BEADXPRESS (27 SNPS)
AMERICAS											
Tarahumara/Huichol (Uto-Aztecan) *Mexico*	96.7	2.4	0.9	22	0.07	0.90	0.05	0.05	0.070	0	-
Quechuas (Quechua) *Peru*	87.4	12.5	0.1	20	0.13	0.75	0.25	0.00	1.000	0	-
Aymara (Quechuamaran) *Peru*	96.4	2.9	0.7	85	0.16	0.75	0.18	0.07	0.005	0	0
Ashaninkas (Arawak) *Peru*	98.0	1.9	0.0	95	0.24	0.58	0.36	0.06	0.783	rs60585106	0
Machiguengas (Arawak) *Peru*	99.0	0.7	0.2	82	0.37	0.32	0.62	0.06	0.004	rs60585106	0
Tupiniquim (Tupi) South East *Brazil*	49.9	27.2	22.8	42	0.08	0.86	0.12	0.02	0.238	0	0
Guarani (Guarani) South East *Brazil*	67.6	20.0	12.4	45	0.07	0.84	0.16	0.00	1.000	0	0
Brazil Admixed South East *Brazil*	14.9	54.9	30.2	98	0.05	0.91	0.08	0.01	0.215	-	0

Table S2. Haplotypes of variants in high LD (r^2>0.8) with the *PDE4B*-rs6683977 (in bold) in at least one population (ordered: rs6668516, rs546784, rs6683604, rs12137080, rs12137115, rs495477, rs494735, rs6683977, rs638111, rs641262), number of samples (N) and haplotypes frequencies in different populations assessed using the TargetSeq dataset (Brazilian Native Americans were not considered due to low quality data). 1000GP populations: AMR= Admixed U.S. Hispanics/Latin Americans (PUR, CLM, MXL and PEL), SAS= South-Asians (GIH, PJL, BEB, STU, and ITU), EAS= East-Asians (CHB, JPT, CHS, CDX, and KHV), EUR= Europeans (CEU, FIN, GBR, ITS, IBS); and Native Americans from Mexico (MEX = *Tarahumaras* and *Huicholes*) and Peru (QUE=*Quechuas*, AYM=*Aymaras*, ASH=*Ashaninkas*, MAC=*Machiguengas*). (-) represents absence or very low frequency haplotypes.

Populations / **Haplotypes**	**AMR (N=347)**	**SAS (N=489)**	**EAS (N=504)**	**EUR (N=503)**	**MEX (N=22)**	**QUE (N=20)**	**AYM (N=24)**	**ASH (N=16)**	**MAC (N=22)**
GTTCCAT**G**CT	0.62	0.51	0.48	0.41	0.80	0.70	0.63	0.84	0.75
ACCTGGC**C**TC	0.27	0.29	0.16	0.56	0.09	0.13	0.08	0.06	0.18
GTTCCAT**G**TT	0.05	0.10	0.35	0.02	0.07	0.15	0.21	0.03	0.07
GTTTCAC**G**CT	0.03	0.02	-	-	0.04	-	0.08	0.06	-