

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 Martín 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



ATA DA DEFESA DE TESE


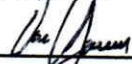
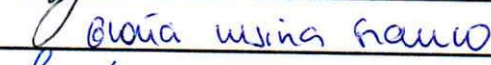


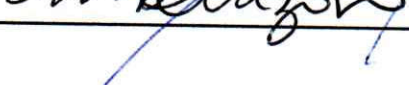
Rennan Garcias Moreira

114/2019
entrada
1º/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	01249405602	APROVADO
Dr. Dani Gamerman	UFMG	43124917715	APROVADO
Dra. Glória Regina Franco	UFMG	623 387. 496-34	APROVADO
Dr. Leandro Machado Colli	USP	306.691.69872	APROVADO
Dra. Maria Cátira Bortolini	UFRGS	38704455253	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-americana 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 discute 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 ABREVIACES

RNA - *Ribonucleic Acid*

DNA - *Deoxyribonucleic Acid*

PCR – *Polymerase Chain Reaction*

GATK – *Genome Analysis Toolkit*

LDGH – *Laboratrio de Diversidade Gentica 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 Linfoblstica Aguda*

IDH – *Índice de Desenvolvimento Humano*

BFM - *Berlin-Frankfurt-Mnster*

DRM - *doena residual mnima*

AMP – *adenosine monophosphate*

CEDEFS – *Centro de Documentao 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 25°-75°; 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 25°-75°; 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 marcadores 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 *MYTIL*-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 *MYTIL*. 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) *MYTIL*. 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 *MYTIL* 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 linfoblásticoide (**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 função regulatória 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 linfoblásticoide (**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 linfoblásticoide (**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 *MYTIL*-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 heterozigidade 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.----- --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 ($r^2 > 0,80$) com *PDE4B*-rs6683977 e *MYT1L*-rs17039396 em diferentes populações. EUR: Europeus (1000GP), AFR: Africanos (1000GP), EAS: Leste Asiáticos (1000GP), SAS: Sul-Asiáticos (1000GP), 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=*Aymarás*, 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 *MYTIL* 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 *MYTIL* (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: Africanos (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 *MYTIL* (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 <i>PIPELINE</i> 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 nucleicos, 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 nucleicos 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-americana 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 sequências de ácidos nucleicos 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 softwares existentes.

A manipulação da amostra, prévia ao sequenciamento do ácido nucleico, 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 nucleicos 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 (SOLOMONENKO *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 nucleicos 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 construção de 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 sequências 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 sequenciamento, 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 softwares, 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 kits 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 sequências 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 Technologies).

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	<i>Quechua</i>	20
	<i>Aymara</i>	24
	<i>Machiguenga</i>	22
	<i>Ashaninka</i>	16
Brasil	<i>Tupiniquim</i>	22
	<i>Guarani</i>	24
México	<i>Huichol</i>	12
	<i>Tarahumara</i>	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.

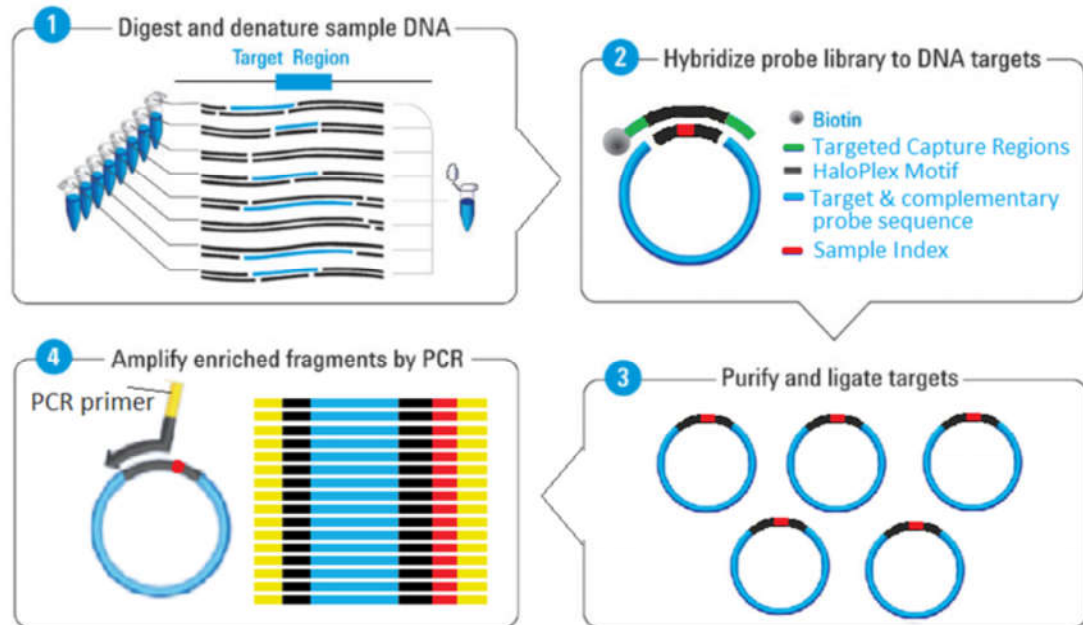
O DNA genômico utilizado como material inicial da construção das bibliotecas foi obtido de amostras de sangue total periférico, por meio de extração com kit próprio (*Gentra*® *Puregene*® *Blood Kit* - Qiagen). A quantificação do DNA após a extração foi realizada no equipamento Nanodrop (ThermoFisher). 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 sequências 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 amplificação 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.

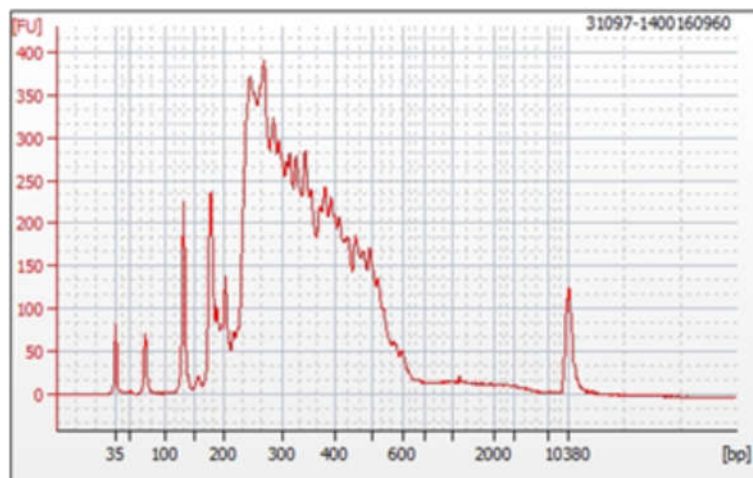
Figura 1. Principais etapas e procedimentos realizados na construção das bibliotecas com o kit Agilent HaloPlex.



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 contribuiu 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 sequências 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 SNPPEP, desenvolvido pela própria Agilent Technologies, 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 SNPPEP 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 SNPPEP, 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
<i>ACAN</i>	15	89346664 - 89418595	<i>GLI3</i>	7	42000538 - 42277479	<i>PDE4B</i>	1	66759100 - 66779100
<i>ADCY1</i>	7	45613729 - 45762725	<i>GSK3B</i>	3	119540160 - 119813274	<i>NT_006576.16</i>	5	4308910 - 4310969
<i>ADCY2</i>	5	7396311 - 7830204	<i>IIPR2</i>	12	26488275 - 26986141	<i>NT_006576.16</i>	5	9965864 - 9968303
<i>ADCY3</i>	2	25042028 - 25142718	<i>MUC4</i>	3	195473626 - 195539158	<i>NT_007592.15</i>	6	14701312 - 14704126
<i>APOB</i>	2	21224291 - 21266955	<i>NRG1</i>	8	31496892 - 32622568	<i>NT_007819.17</i>	7	8287819 - 8291201
<i>APP</i>	21	27251851 - 27252851	<i>PARK2</i>	6	161768442 - 163148844	<i>NT_009952.14</i>	13	21786357 - 21788159
<i>APP</i>	21	27252851 - 27543456	<i>PDE4D</i>	5	58264855 - 59817957	<i>NT_009952.14</i>	13	23286884 - 23288808
<i>BRIP1</i>	17	59756537 - 59940930	<i>PLCBI</i>	20	8112814 - 8949013	<i>NT_010393.16</i>	16	17909619 - 17912456
<i>CALR</i>	19	13048404 - 13049404	<i>PRKCA</i>	17	64298916 - 64806872	<i>NT_010859.14</i>	18	7450207 - 7451781
<i>CALR</i>	19	13049404 - 13055314	<i>PRKCE</i>	2	45878474 - 46415139	<i>NT_011109.16</i>	19	3549641 - 3552069
<i>CASR</i>	3	121901520 - 121902520	<i>PTRPD</i>	9	8314236 - 10612733	<i>NT_011387.8</i>	20	7604013 - 7606701
<i>CASR</i>	3	121902520 - 122005354	<i>RAFI</i>	3	12625090 - 12705735	<i>NT_016354.19</i>	4	29319625 - 29322099
<i>CDH13</i>	16	82660389 - 83830225	<i>RBFOX1</i>	16	6069122 - 7763350	<i>NT_016354.19</i>	4	105450162 - 105451914
<i>CDH9</i>	5	26880699 - 27121267	<i>RELA</i>	11	65421057 - 65430575	<i>NT_021937.19</i>	1	1021368 - 1022843
<i>CNTNAP2</i>	7	145813443 - 148118100	<i>SMAD3</i>	15	67358173 - 67487543	<i>NT_023736.17</i>	8	5094941 - 5096742
<i>CSMD1</i>	8	2792865 - 4852504	<i>SRC</i>	20	35973078 - 36034463	<i>NT_025028.14</i>	18	23004514 - 23006360
<i>CTNNB1</i>	3	41236318 - 41301597	<i>STAT3</i>	17	40465332 - 40540523	<i>NT_025741.15</i>	6	68545632 - 68547631
<i>EGFR</i>	7	55086704 - 55324323	<i>TP63</i>	3	189349195 - 189615078	<i>NT_029419.12</i>	12	9923356 - 9926079
<i>ERBB4</i>	2	212240432 - 213403575	<i>ZNF280A</i>	22	22868050 - 22874623	<i>NT_030059.13</i>	10	70198161 - 70200298
<i>FHIT</i>	3	59735026 - 61237143	<i>ESR1</i>	6	151938366 - 151958366	<i>NT_030059.13</i>	10	79271368 - 79272941
<i>FHIT</i>	3	61237143 - 61238343	<i>ESR1</i>	6	152337857 - 152357858	<i>NT_034772.6</i>	5	36518682 - 36521560
<i>FYN</i>	6	111981525 - 112194665	<i>MYT1L</i>	2	2215144 - 2235144			

Genome Analysis Toolkit

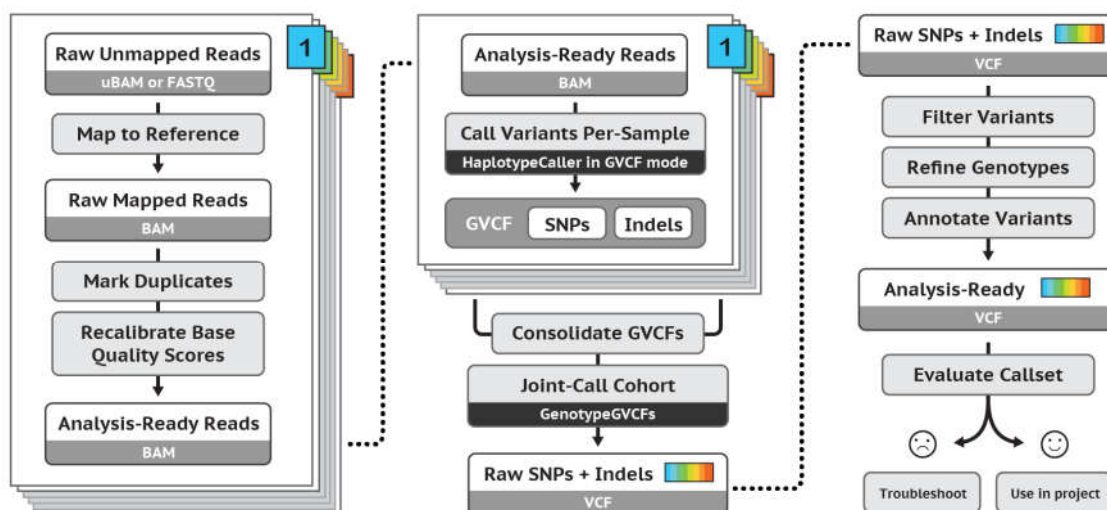
A plataforma de análises GATK (“Genome Analysis Toolkit”) (MCKENNA *et al.* 2010), desenvolvida pelo Broad Institute, Cambridge, MA, EUA, é hoje reconhecidamente a que disponibiliza os melhores algoritmos e estratégias de análise de sequenciamento em larga escala (PIROOZANIA *et al.* 2014, LAURIE *et 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 “HaplotypeCaller”, 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 trabalho³. 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 GATK⁴.

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 sequências 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 sequências 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

genotipagem 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 “kqp” 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 ThermoFisher™) 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

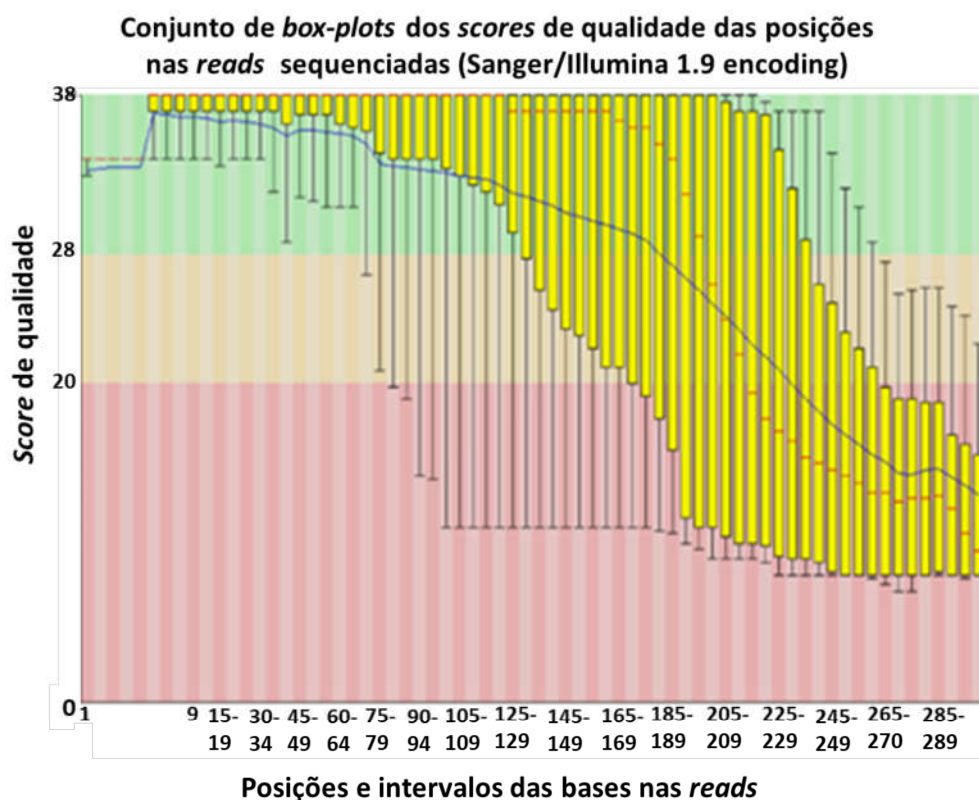
Construção de bibliotecas customizadas

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 oligonucleotídeos. 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 25°-75°; 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

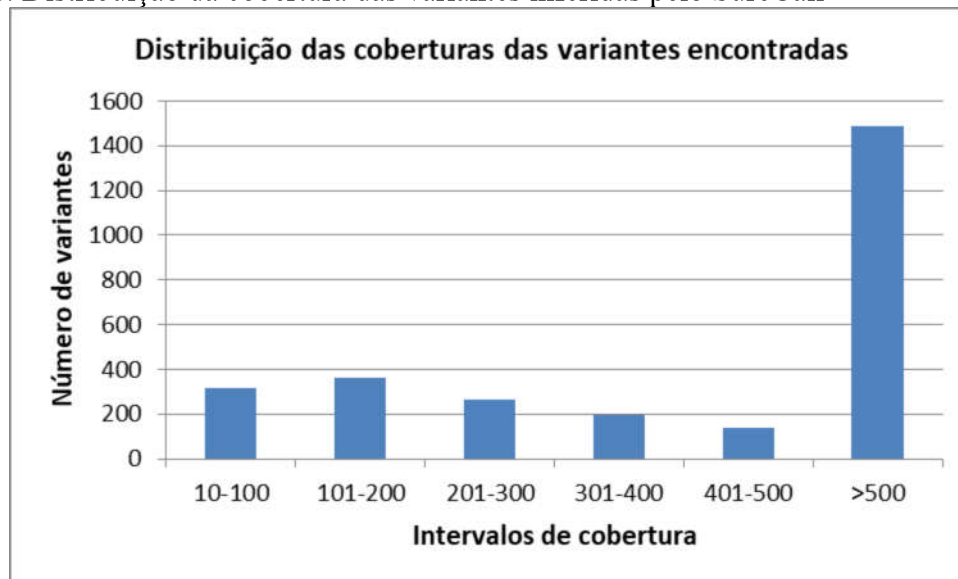
As análises no *SureCall* foram realizadas de forma individual, já que o programa trata cada amostra por vez. Porém, os resultados apresentados a seguir mostram métricas calculadas ao nível global das amostras (n=150) (Tabela 3). Como o próprio programa possui funcionalidades que permitem o pré-processamento 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)*

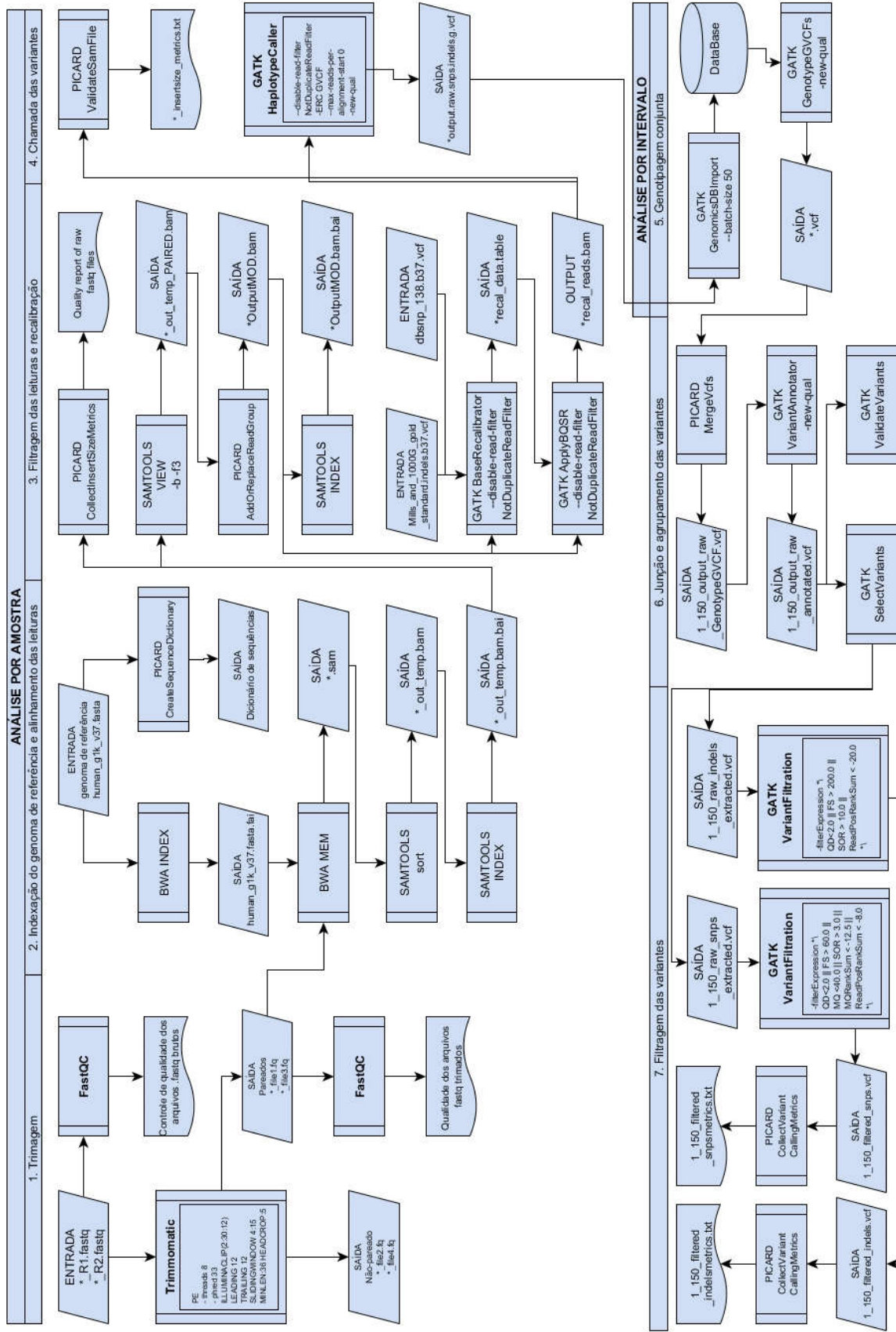


Figura 7. Ilustração adaptada da distribuição da qualidade das bases ordenadas pela posição em todas as *reads* de uma amostra após tratamento pelo programa Trimmomatic. Em verde: >Q28, em amarelo: >Q20, em vermelho: <Q20. Retângulos amarelos: percentis 25°-75°; linhas inferiores e superiores em cada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

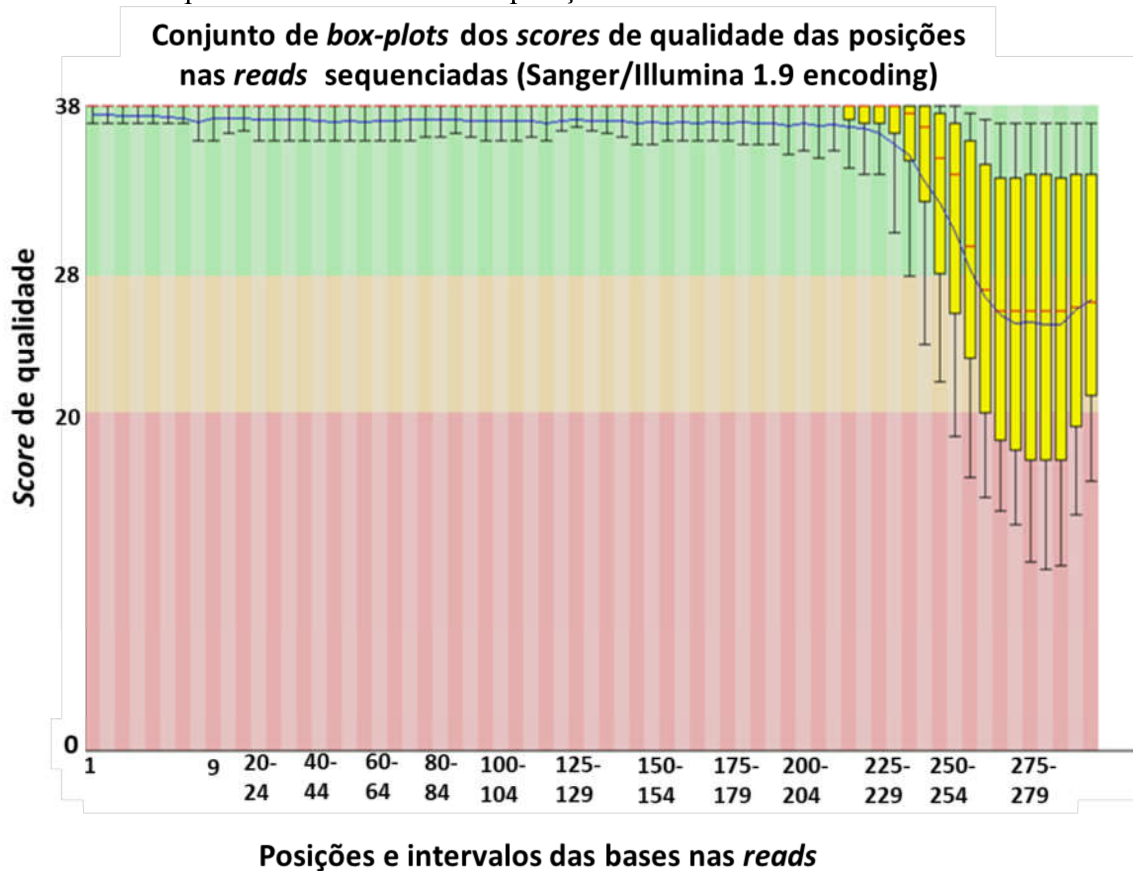


Figura 8. Distribuições dos parâmetros utilizados como indicadores para o processo de filtragem de Indels. SOR: *StrandOddsRatio*, FS: *FisherStrand*, QD: *QualByDepth*, QUAL: *Quality*, ReadPosRankSum: *ReadPosRankSumTest*. Em vermelho, a distribuição obtida com os SNPs do dbSNP, já em azul a distribuição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

Indels

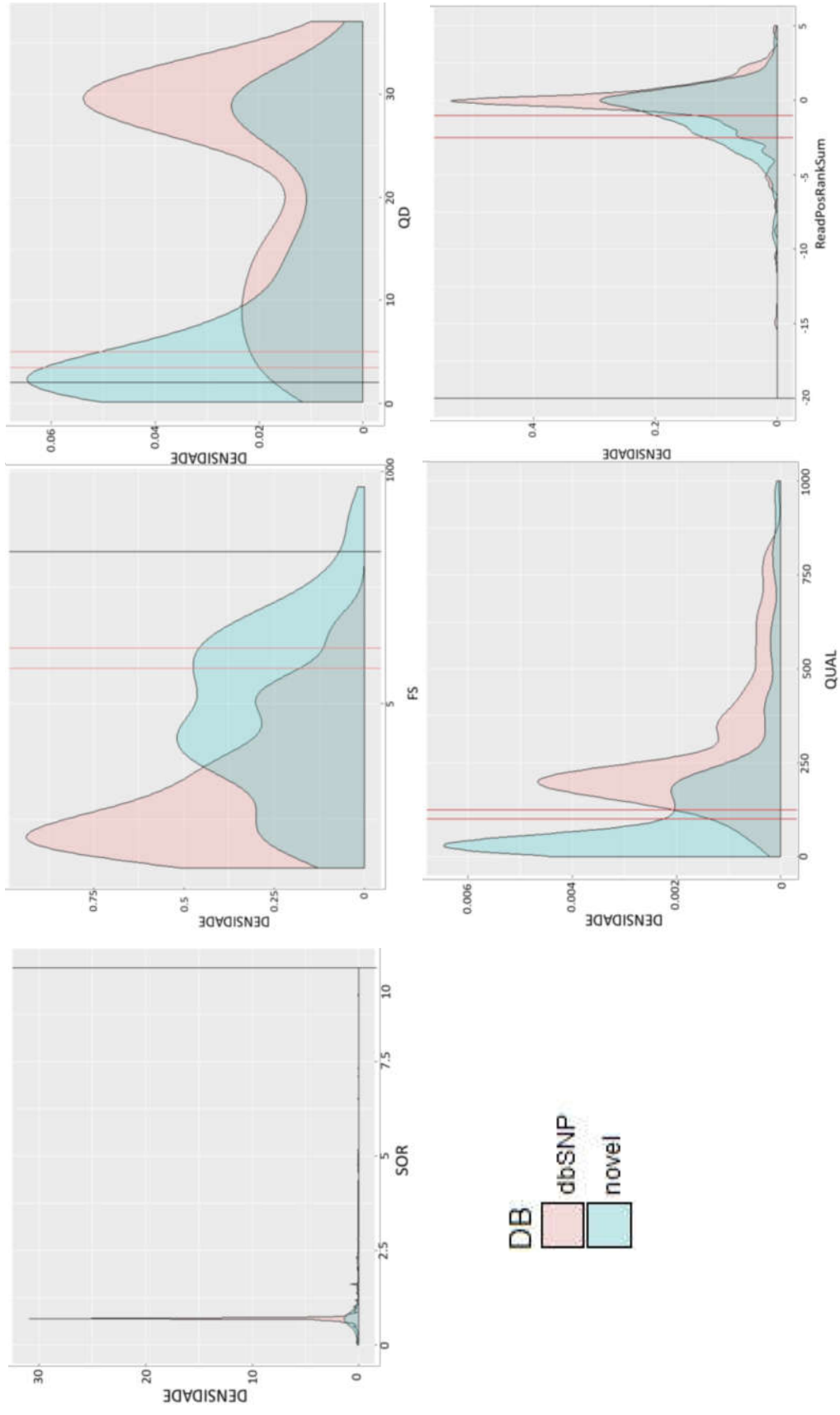


Figura 9. Distribuições dos parâmetros utilizados como indicadores para o processo de filtragem de SNPs. SOR: *StrandOddsRatio*, FS: *FisherStrand*, QD: *QualByDepth*, QUAL: *Quality*, ReadPosRankSum: *ReadPosRankSumTest*. Em vermelho, a distribuição obtida com os SNPs do dbSNP, já em azul a distribuição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

SNPs

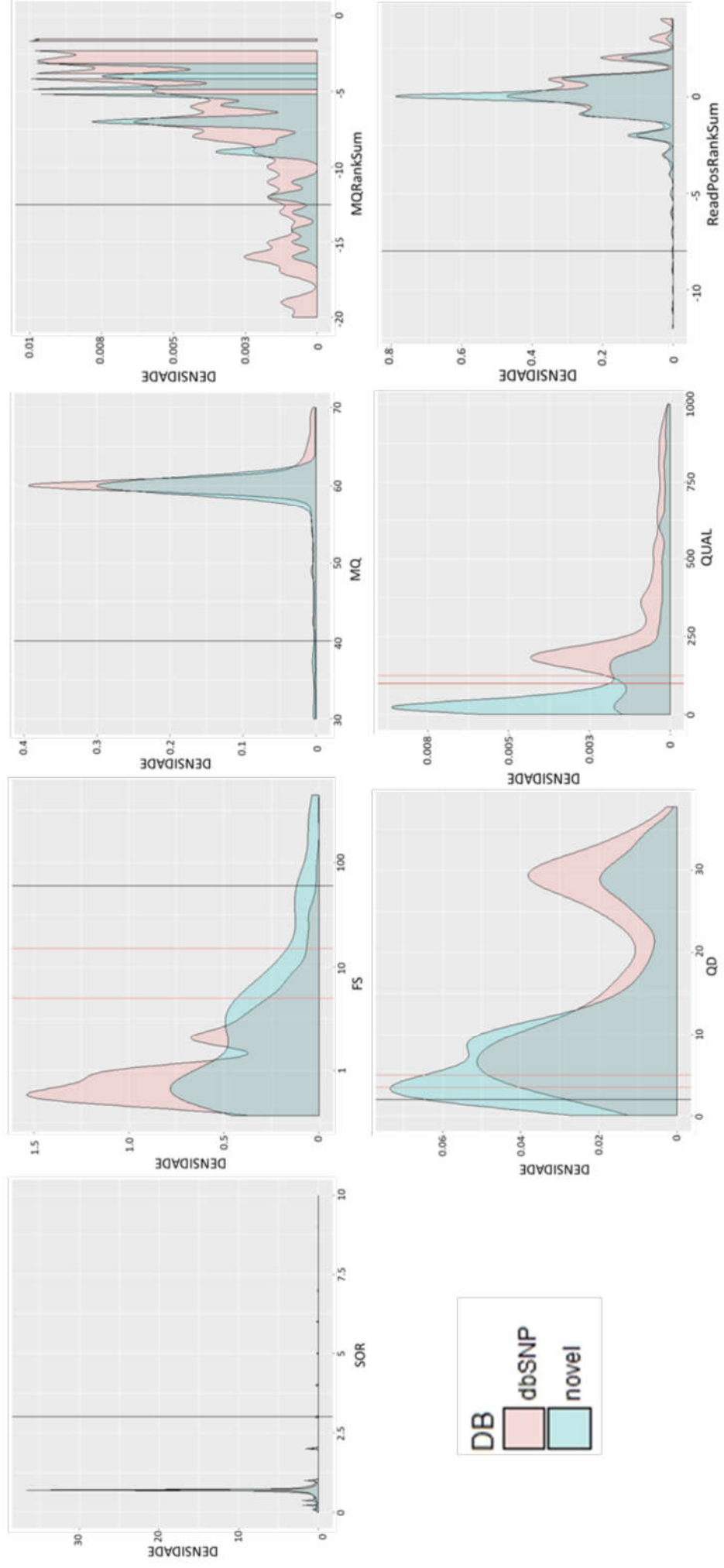
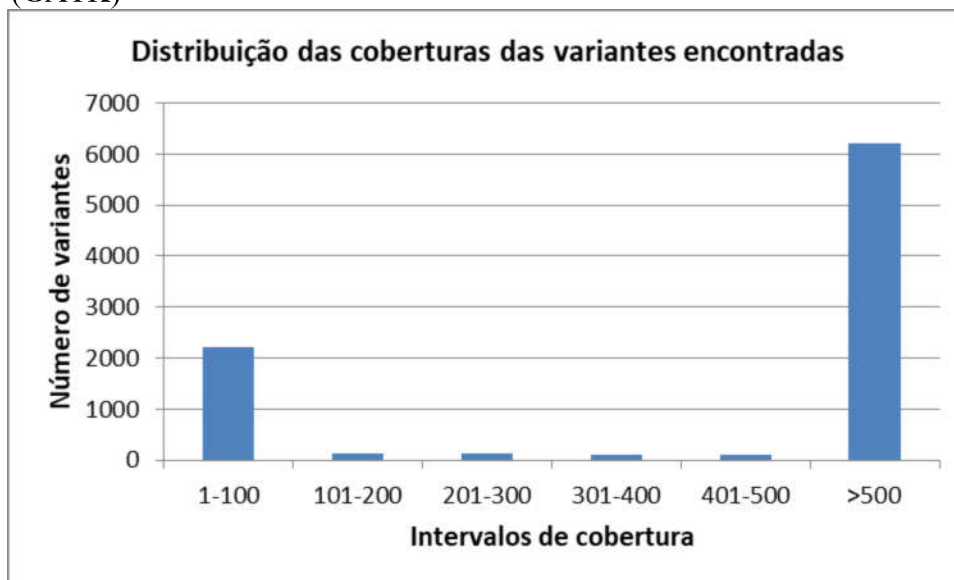


Figura 10. Distribuição da cobertura das variantes inferidas pelo *Genome Analysis Toolkit* (GATK)



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 homocigotos. 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.

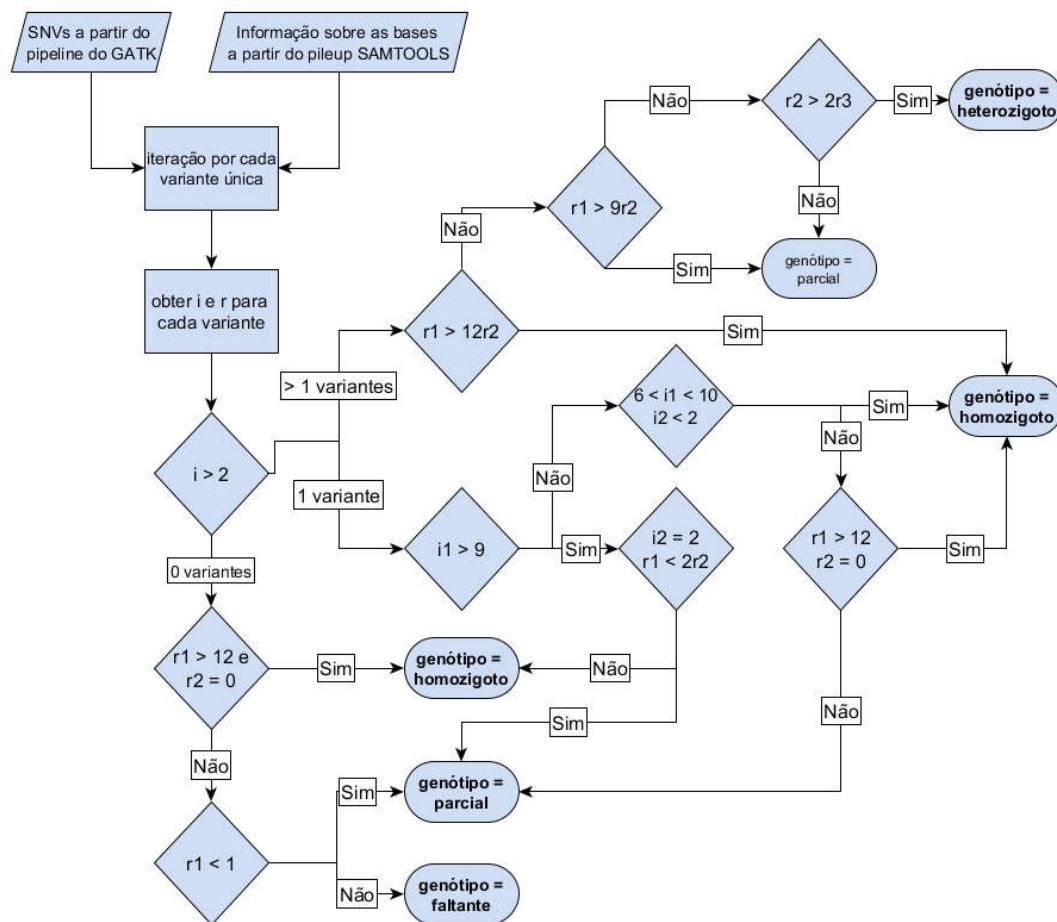


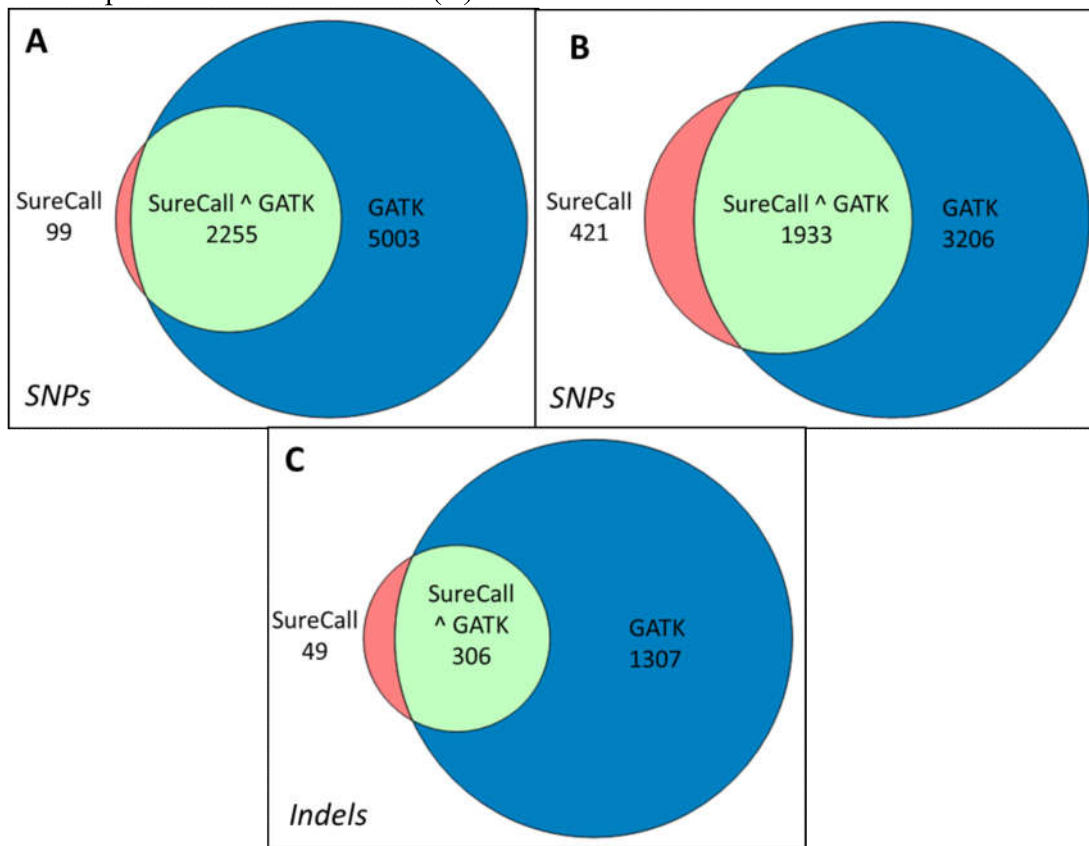
Tabela 5. Números de variantes e suas características obtidos com o *pipeline* customizado do *Genome Analysis Toolkit* (GATK) após correções

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 sequências de ácidos nucleicos 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 sequências 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 teórico⁵. 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 homocigoto para o alelo C encontrado. Nesse caso, o SureCall indicou que essa amostra seria heterocigota, 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 número 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 Sciences**, 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 neurology 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.

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

SOUZA, G. Identificação 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.

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 análise 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