PROD. TEC. ITV DS – N036/2022
DOI 10.29223/PROD.TEC.ITV.DS.2022.36.Dalapicolla

RELATÓRIO TÉCNICO ITV DS

GUIA PARA O USO DE GENOMAS COMPLETOS DE BAIXA COBERTURA PARA ANÁLISES POPULACIONAIS E DE GENÉTICA DA CONSERVAÇÃO

RELATÓRIO PARCIAL

Amazoomics - Eixo Genômica da Conservação

Jeronymo Dalapicolla
Tânia Fontes Quaresma
Lucas Eduardo Costa Canesin
Alexandre Aleixo

Belém / Pará
Dezembro / 2022

INSTITUTO TECNOLÓGICO VALE

Título: Guia para o uso de genomas completos de baixa cobertura para análises populacionais e de genética da conservação	
PROD. TEC. ITV DS N036/2022	**Revisão 01**
Classificação: ( ) Confidencial ( ) Restrita ( ) Uso Interno (X) Pública	

Informações Confidenciais - Informações estratégicas para o Instituto e sua Mantenedora. Seu manuseio é restrito a usuários previamente autorizados pelo Gestor da Informação.

Informações Restritas - Informação cujo conhecimento, manuseio e controle de acesso devem estar limitados a um grupo restrito de empregados que necessitam utilizá-la para exercer suas atividades profissionais.

Informações de Uso Interno - São informações destinadas à utilização interna por empregados e prestadores de serviço

Informações Públicas - Informações que podem ser distribuídas ao público externo, o que, usualmente, é feito através dos canais corporativos apropriados

Citar como
DALAPICOLLA, Jeronymo; QUARESMA, Tânia Fontes; CANESIN, Lucas Eduardo Costa; ALEIXO, Alexandre. **Guia para o uso de genomas completos de baixa cobertura para análises populacionais e de genética da conservação**. Belém: 2022. (Relatório Técnico N036/2022) DOI 10.29223/PROD.TEC.ITV.DS.2022.36.Dalapicolla Acesso em:

Dados Internacionais de Catalogação na Publicação (CIP)

D136g Dalapicolla, Jeronymo
Guia para o uso de genomas completos de baixa cobertura para análises populacionais e de genética da conservação. / Jeronymo Dalapicolla, Tânia Fontes Quaresma, Lucas Eduardo Costa Canesin, Alexandre Aleixo - Belém: ITV, 2022.

54 p. : il.

Relatório Técnico (Instituto Tecnológico Vale) – 2022
PROD.TEC.ITV.DS – N036/2022
DOI 10.29223/PROD.TEC.ITV.DS.2022.36.Dalapicolla

1 Genômica. 2.Genoma completo. 3. Mapeamento sequencial. 4.Gênomica – análise e conservação. I. Quaresma, Tânia Fontes. II. Canesin, Lucas Eduardo Costa. III. Aleixo, Alexandre. IV. Título

CDD 23. ed. 622.752098115

Bibliotecário responsável: Eddie Saraiva / CRB 2 – 058P

RESUMO EXECUTIVO

Umas das vertentes do projeto Amazoomics é o uso de ferramentas de Genômica da Conservação para quantificar os riscos de extinção em espécies de aves e mamíferos amazônicos. O ponto de partida para isso é a delimitação da forma correta de gerar e filtrar marcadores genômicos, chamados SNPs, para permitir uma maior eficiência e rapidez nas entregas e para permitir a comparação entre as diferentes espécies analisadas. Neste relatório, foi criado um guia para as três etapas principais da identificação de SNPs: (*i*) avaliação da qualidade do sequenciamento; (*ii*) mapeamento de genomas de baixa cobertura em um genoma de referência e (*iii*) seleção de SNPs para as análises populacionais e de conservação. Futuros estudos com genomas dentro e fora do projeto Amazoomics poderão utilizar esse guia, bem como pesquisadores de fora do Instituto Tecnológico Vale (ITV).

RESUMO

Os SNPs são marcadores genômicos populares em estudos populacionais e de conservação devido ao seu alto número, baixo custo e facilidade de descoberta no genoma. Contudo, as interpretações dos resultados oriundos dos SNPs dependem da quantidade e da qualidade dos dados gerados. Ainda há discussões em relação à melhor maneira de detectar e selecionar os SNPs. O projeto Amazoomics tem como um dos objetivos realizar estudos de Genômica da Conservação em espécies ameaçadas de extinção de aves e mamíferos da Amazônia. Nesses estudos, SNPs serão identificados utilizando genomas completos de baixa cobertura, uma técnica chamada de *low-coverage Whole Genome Sequencing* (*lcWGS),* para o cálculo de diversidade genética e para a estimativa dos riscos e das ameaças que pode estar afetando a sobrevivência das espécies-alvo. Dessa forma, uma padronização para escolhas dos marcadores genômicos se faz necessária para permitir uma comparação entre as espécies. Este relatório apresenta e discute as três etapas principais da identificação de SNPs: (*i*) avaliação da qualidade do sequenciamento; (*ii*) mapeamento do genoma de baixa cobertura com um genoma de referência e (*iii*) seleção dos SNPs para as análises populacionais e de conservação. São fornecidos neste relatório *scripts* para a automatização de várias etapas e explicações para as escolhas de diferentes programas e filtros de seleção de SNPs. A *pipeline* desenvolvida aqui está disponível no *cluster* do Instituto Tecnológico Vale (ITV) no endereço: *</bio_temp/share_bio/projects/Amazoomics/pipeline/>*.

Palavras-chave: Genomas completos. Filtragem de SNPs. lcWGS. Mapeamento. Sequenciamento.

ABSTRACT

SNPs are popular genomic markers in population genetics and conservation studies due to their high number, low cost, and ease of discovery in the genome. However, interpretations of results from SNPs depend on the quantity and quality of data generated. There are still discussions regarding the best way to detect and select SNPs. One of the objectives of the Amazoomics project is to carry out Conservation Genomics studies on endangered species of birds and mammals in the Amazon. In these studies, SNPs will be identified using low coverage whole genome sequencing (lcWGS), for calculating genetic diversity and estimating risks and threats that may be affecting the survival of the target species. Thus, a standardization for genomic marker choices is necessary to allow comparison between different species. This report presents and discusses the three main steps of SNP and genotype calling: (*i*) sequencing quality assessment, (*ii*) mapping reads from low coverage sequencing using a reference genome, and (iii) filtering SNPs for population and conservation analyses. Scripts for automating steps and explanations for the different softwares and filters are provided in this report as well. The pipeline developed here is available at the Instituto Tecnológico Vale (ITV) cluster at: </bio_temp/share_bio/projects/Amazomics/pipeline/>.

Keywords: Genotype calling. LcWGS. Mapping. Sequencing. SNP calling.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 GENOMAS DE BAIXA COBERTURA

Diferenças no DNA (polimorfismos) entre indivíduos, populações ou espécies tem papel central nos estudos de genética e são identificadas utilizando marcadores moleculares (VIGNAL *et al.*, 2002). O desenvolvimento e aplicação de marcadores moleculares começou na década de 1980, com o uso de marcadores que envolviam sondas e eletroforese em gel, que na década de 1990 foram substituídas por técnicas envolvendo a reação em cadeia da polimerase (PCR) e o sequenciamento do DNA, como os microssatélites (AMITEYE, 2021). Na década de 2000, um novo tipo de marcador codominante denominado SNP (Polimorfismo de Nucleotídeo Único) ganhou popularidade após a introdução de técnicas de sequenciamento de nova geração (NGS) (MORIN; MARTIEN; TAYLOR, 2009; HELYAR *et al.*, 2011).

Os SNPs são cada vez mais populares em estudos de genética populacional devido ao seu alto número, baixo custo e facilidade de descoberta. A aplicação de SNPs como marcador molecular oferece uma oportunidade para identificar a história demográfica, evolutiva e aspectos ecológicos em organismos modelos e não-modelos (LO *et al.*, 2018). Dados de SNPs, antes considerados reservados para estudos de genética populacional, agora estão sendo aplicados em pesquisas ecológicas, filogenéticas, com em escalas de tempo evolutivas profundas e em estudo sobre adaptação e conservação, mostrando-se um marcador molecular versátil (LEACHÉ; OAKS, 2017).

Nos últimos anos, diferentes técnicas e plataformas foram desenvolvidas para detecção e aplicação de SNPs, sendo uma das mais promissoras aquela usando genomas completos de vários indivíduos, uma técnica chamada resequencimento (STRATTON, 2008). Contudo, os custos para a geração de genomas completos e para analisar esses dados ainda são elevados, fazendo com que essa técnica seja ainda pouco utilizada (FUENTES-PARDO; RUZZANTE, 2017). Uma alternativa que surgiu nos últimos anos é utilizar genomas de baixa cobertura, abaixo de 5× de cobertura média, para reduzir os custos computacionais e de sequenciamento, uma técnica chamada em inglês de *low-coverage Whole Genome Sequencing* (lcWGS) (LOU *et al.*, 2021). Atualmente, os custos de produção de bibliotecas e de sequenciamento para *lcWGS*, são equivalentes aos custos de técnicas de representação de genoma reduzido como *Restriction-site Associated DNA* (RAD) ou

Genotyping-by-Sequencing (GBS) (FUENTES-PARDO; RUZZANTE, 2017). O sequenciamento de baixa cobertura é suficiente para estudos populacionais, com estimativas de parâmetros altamente precisas e baseadas em mais locais no genoma (BUERKLE; GOMPERT, 2013; LOU *et al.*, 2021).

As interpretações dos resultados oriundos dos SNPs dependem da quantidade e qualidade dos dados gerados. Ainda há sérios desafios em relação à filtragens e seleção dos SNPs para as análises de dados (LEACHÉ; OAKS, 2017; LO *et al.*, 2018). A qualidade dos dados depende diretamente da cobertura do genoma e por isso, na técnica de *lcWGS,* o maior obstáculo é selecionar de maneira correta os SNPs, identificando corretamente o polimorfismo já que a cobertura dos genomas é baixa (BUERKLE; GOMPERT, 2013).

Uma forma de resolver esse problema é considerar a incerteza na identificação da base nitrogenada dos SNPs (*SNP calling*) e também nos cálculos das frequências alélicas e demais análises populacionais. Essa incerteza é calculada usando a qualidade das bases do sequenciamento com o uso de métodos probabilísticos (e.g., modelos Bayesianos) e algoritmos baseados em aprendizagem de máquina (MARTIN *et al.*, 2010). As probabilidades de um genótipo são chamadas de *Genotype Likelihood* (GL) ou *Phred-scaled Likelihood* (PL), se os valores forem normalizados para escala *Phred* (NIELSEN *et al.*, 2011).

1.2 POR QUE ESSE GUIA É IMPORTANTE?

O projeto Amazoomics tem como um dos objetivos o sequenciamento, montagem e a anotação de genoma de referências para espécies amazônicas de aves e mamíferos ameaçadas de extinção. É objetivo do projeto também realizar estudos de Genômica da Conservação de aves e mamíferos da Amazônia, onde serão identificados SNPs utilizando a técnica de *lcWGS* para o cálculo de diversidade genética e para a estimativa dos riscos de extinção dessas espécies.

As plataformas de sequenciamento de próxima geração (NGS) podem gerar grandes quantidades de dados de sequenciamento, mas geralmente com altas taxas de erros. Para dados de baixa a média profundidade são necessárias ferramentas que implementem análises em um contexto probabilístico, trabalhando com os dados brutos e as incertezas na identificação das bases, na forma de probabilidades de genótipos (GL) (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014).

Para que haja a possibilidade de comparação dos resultados entre todas as espécies-alvo do projeto é necessário a padronização dos métodos de identificação e filtragem dos SNPs utilizando o GL a partir de genomas de baixa cobertura. Dessa forma, este guia oferece um roteiro de identificação e filtragem de SNPs a partir da técnica de *lcWGS* baseados em alguns estudos: McKenna *et al.* (2010), Fuentes-Pardo e Ruzzante (2017), Poplin *et al.* (2017), O'Leary *et al.* (2018) e Lou *et al.* (2021).

Esse relatório vai abranger três etapas principais na identificação dos SNPs: (*i*) avaliação da qualidade do sequenciamento; (*ii*) mapeamento do genoma de baixa cobertura com suporte de genoma de referência; e (*iii*) filtragem de SNPs para as análises populacionais e de conservação.

1.3 MATERIAIS NECESSÁRIOS PARA O GUIA

Para realizar as etapas deste guia, é necessário ter (*i*) acesso a um supercomputador, ou *High-Performance Computing* (HPC – *cluster*) com programas específicos instalados; (*ii*) arquivos **.fastq** dos indivíduos da espécie ou grupo-alvo já sequenciados; e (*iii*) um genoma de referência de alta cobertura para mapeamento, em formato **.fasta**.

Sobre o supercomputador, HPC ou *cluster*, esses equipamentos são necessários, pois as análises deste guia são demoradas e demandam muito esforço computacional. Dentro do Instituto Tecnológico Vale (ITV) há *clusters* disponíveis, mas é preciso solicitar acesso ao administrativo. Caso seu acesso seja remoto, também é necessário solicitar acesso a uma *Virtual Private Network* (VPN). Dentro do *cluster,* é importante que para cada análise e/ou etapa seja criada uma pasta específica e que se evite duplicação de dados e *inputs* para otimizar o espaço do *cluster*. Se as pastas precisam ser públicas, é necessário alterar a permissão delas com o comando **chmod 777** logo após a criação de cada pasta. Uma sugestão é que cada projeto/espécie tenha sua pasta e dentro dela cada etapa/programa seja uma subpasta com seu respectivo arquivo de execução (**.pbs**) dentro dela, por exemplo, para a espécie *Anodorhynchus hyacinthinus*, a organização de pastas poderia ser:

```
/bio_temp/share_bio/projects/Amazoomics/Anodorhynchus
                                                /fastq
                                                /reference
                                                /quality_raw_fastqc
                                                /trimmed_prinseq
                                                /quality_trimmed_fastqc
                                                /mapped_bwa
                                                /sorted_samtools
                                                /vcfs_ind_gatk
```

Figura 01: Esquema de organização de pastas e arquivos para otimização o espaço do *cluster* ou computadores.

No *cluster* ou na máquina que for utilizada para análises de lcWGS é necessário ter alguns programas instalados (Tabela 01). Não é necessário usar esses mesmos programas, podem ser similares. Por exemplo, para a remoção de adaptadores das *reads* foi usado neste relatório o programa Trimmomatic (BOLGER; LOHSE; USADEL, 2014), mas existem outras opções como cutadapt (MARTIN, 2011) ou Prinseq (SCHMIEDER; EDWARDS, 2011). Utilize o programa de sua preferência.

Para gerar os arquivos **.fastq** de indivíduos para a espécie-alvo é necessário criar uma biblioteca genômica por indivíduo e posteriormente sequenciar as amostras combinadas. Este guia não abordará a preparação das bibliotecas e nem o seu sequenciamento. O guia utiliza amostras já sequenciadas em *paired-end (Forward* e *Reverse)*, com arquivos R1 e R2 em formato **.fastq** para cada indivíduo.

Como estamos utilizando *lcWGS* para reduzir a incerteza de identificação dos SNPs precisamos usar como referência um genoma com alta cobertura. Essa referência deve ser da mesma espécie que será estuda. Contudo, na ausência de uma referência para a espécie, pode-se utilizar um genoma de referência de uma espécie próxima, do mesmo gênero ou em último caso, da mesma família da espécie-alvo (GALLA *et al.*, 2018). Esse guia não abordará a montagem de um genoma *de novo.*

Tabela 01: Lista de programas ou *softwares* utilizados neste guia com o *link* para instalação dos mesmos.

Programas	Instalação
ANGSD	http://www.popgen.dk/angsd/index.php/Installation
BaseSpace	https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html)
BCFtools	http://www.htslib.org/download/
Burrows-Wheeler Aligner (BWA)	https://github.com/lh3/bwa
CSVKit	https://csvkit.readthedocs.io/en/0.9.1/install.html
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
GATK	https://github.com/broadinstitute/gatk/releases
NCBI's Sequence Read Archive (SRA)	https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
ngsLD	https://github.com/fgvieira/ngsLD
Picard	https://broadinstitute.github.io/picard/
R	https://cran.r-project.org/
PopLDdecay	https://github.com/hewm2008/PopLDdecay
Seqtk	https://github.com/lh3/seqtk
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
VCFtools	https://vcftools.github.io/index.html
SAMtools	http://www.htslib.org/download/

Há dois tipos de dados necessários para a obtenção de genomas de referência que são encontrados em repositórios como o NCBI (https://www.ncbi.nlm.nih.gov/) ou similares: os arquivos **.fastq** e a referência já montada em .**fasta**. Para os .**fastq** é necessário ter o programa **SRA Tools Kit**. Além disso, é necessário o código do genoma no NCBI. No *site* (https://www.ncbi.nlm.nih.gov/) procure a aba **ALL DATABASES** e selecione nela a opção **SRA** e busque pelo nome da espécie-alvo ou uma espécie próxima. Verifique o tamanho do genoma, se ele é *paired-end* e busque o código SRR ou SRA. Para *Anodorhynchus hyacinthinus* a melhor opção é o SRR7535835 gerado pelo artigo de Hains *et al.* (2020) com cobertura de 120x. Veja o tamanho do genoma, no caso dessa espécie é 1119.46 Mb. Com essas informações é possível fazer o *download* dos .**fastq** R1 e R2 desse genoma, com o seguinte comando:

```
fastq-dump --split-files --gzip [NCBI_CODE]
fastq-dump --split-files --gzip SRR7535835
```

Para a referência em **.fasta** montada, precisa-se do *download* do *Assembly*, para isso, no *site* do **NCBI** (https://www.ncbi.nlm.nih.gov/) >> procure a aba **ALL DATABASES** a opção **ASSEMBLY** >> busque pelo nome da espécie-alvo >> selecione um genoma e abra a página dele. No canto direito da página tem a opção **DOWNLOAD ASSEMBLY** e abaixo em uma tabela com os parâmetros de qualidade. No exemplo do *Anodorhynchus* o *link* desse *Assembly* é <https://www.ncbi.nlm.nih.gov/assembly/GCA_009936445.1>

Existem várias opções para *download* de *assembly*, as duas mais importante são a **Genomic FASTA** e a **CDS Genomic Fasta**, mas normalmente só a primeira fica disponível para a maioria dos genomas. O arquivo ***_genomic.fna.gz (Genomic FASTA)** vem no formato .**fasta** com a(s) sequência(s) genômica(s) da montagem. Sequências repetitivas em eucariotos são mascaradas com letras minúsculas. O arquivo **.genomic.fna.gz** inclui todas as sequências de nível superior na montagem (cromossomos, plasmídeos, organelas, *unlocalized scaffolds*, *unplaced scaffolds* e quaisquer loci alternativos ou *patches* de *scaffolds*). Os *scaffolds* que fazem parte dos cromossomos não são incluídos porque são redundantes com as sequências cromossômicas. O **.fna** é o formato **.fasta** para ácido nucleico (*fasta nucleic acid*) em oposição a **.faa** (fasta para aminoácidos) para sequências de proteínas. Pode-se renomear o arquivo, substituindo o **.fna** por **.fasta** sem problemas. O arquivo ***_cds_from_genomic.fna.gz (CDS do genomic FASTA)** é também um formato **.fasta**, mas com as sequências de nucleotídeos correspondentes a todas as CDS anotadas na montagem. Uma **CoDing Sequence (CDS)** é uma região de DNA ou RNA cuja sequência determina a sequência de aminoácidos em uma proteína. É a lista de genes codificantes que pode ser usada, por exemplo, para mapear apenas áreas codificantes para análises filogenéticas.

1.4 PRINCIPAIS FORMATOS DE ARQUIVOS

É necessário saber alguns detalhes sobre os diferentes tipos de aquivos que serão usados neste guia:

SRA (Sequence Read Archives): O SRA é um arquivo de dados brutos e com informação sobre a qualidade das bases. O SRA aceita arquivos binários como os formatos **.bam**, **.sff** e **.hdf5** e formatos de texto como **.fastq**. É usado para submissão no NCBI e foi criado pela equipe do NCBI.

FASTA: é um formato baseado em texto para representar tanto sequências de nucleotídeos quanto sequências de peptídeos, no qual os nucleotídeos ou aminoácidos são representados usando códigos de uma única letra. Uma sequência em formato **.fasta** começa com uma descrição de uma única linha, seguida por linhas de dados em sequência. A linha de descrição se distingue da sequência dos dados por um símbolo maior-que (">") no início. A palavra que segue o símbolo ">" é o identificador da sequência e o resto da linha é a descrição (ambos são opcionais). Não deve haver nenhum espaço entre o ">" e a primeira letra do identificador. Recomenda-se que os descritores sejam menores que 80 caracteres. A sequência termina se uma outra linha de partida com outro símbolo ">", que indica o início de outra sequência. Um exemplo simples de uma sequência em formato **.fasta**:

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
        LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
        EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
        LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
        GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
        IENY
```

FASTQ: é um formato baseado em texto para armazenar uma sequência biológica (geralmente sequência de nucleotídeos) e suas informações sobre a qualidade dessas bases. Tanto a letra de sequência quanto o índice de qualidade são codificados com um único caractere **.ascii**. Um arquivo FASTQ normalmente usa quatro linhas por sequência. Esse arquivo foi desenvolvido pelo Instituto Sanger.

Linha 1: começa com um caractere '@' e é seguida por um identificador de sequência e uma descrição opcional (como uma linha de título de um **.fasta**);

Linha 2: são as letras que representam a sequência bruta;

Linha 3: começa com um caractere '+' e é opcional, e contém qualquer descrição sobre a amostra;

Linha 4: codifica os valores de qualidade para a sequência na linha 2 e deve conter o mesmo número de símbolos que as letras na sequência. Cada símbolo representa um valor de qualidade, por exemplo:

```
        @SEQ_ID
        GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
        +
        !''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCC65
```

SAM (Sequence Alignment and Map): Os arquivos **.sam** são arquivos de texto que contém as informações de alinhamento de várias sequências mapeadas em relação às sequências de referência. Esses arquivos também podem conter sequências não-mapeadas. São arquivos gerados por alinhadores como o BWA (LI; DURBIN, 2009) ou Bowtie (LANGMEAD; SALZBERG, 2012) a partir de dados brutos de sequência na forma de **.fastq** juntamente com um genoma de referência (geralmente na forma de um arquivo **.fasta**). Os arquivos **.sam** são arquivos contendo as leituras junto com a localização genômica. Como os arquivos **.sam** são arquivos de texto, eles são mais legíveis. O cabeçalho começa com o símbolo '@', que os distingue da seção de alinhamento. As seções de alinhamento têm 11 campos obrigatórios, bem como um número variável de campos opcionais. São arquivos grandes e carregam muita informação.

BAM (Binary Alignment and Map): Os arquivos **.bam** contêm as mesmas informações que os arquivos **.sam**, exceto que estão em formato binário que não é facilmente legível por humanos. Por outro lado, os arquivos **.bam** são menores e mais eficientes para as análises por serem binários, economizando tempo e reduzindo custos de computação e armazenamento. Os dados de alinhamento quase sempre são armazenados em arquivos **.bam** e a maioria dos programas que analisam *reads* alinhadas espera os *inputs* no formato **.bam**. Duas etapas adicionais são necessárias após a geração de um arquivo **.bam**: sua classificação (**sort**) e sua indexação (**index**). Como as leituras usadas para gerar um arquivo **.bam** são (ou pelo menos deveriam ser) aleatórias em relação às suas posições dentro do genoma, os arquivos **.bam** precisam ser classificados pelo identificador de *read*. Assim **.bam** de indivíduos diferentes são classificados de uma mesma maneira, na mesma ordem, permitindo a comparação entre arquivos individuais. A classificação de um arquivo **.bam** pode ser feita por alguns programas como SAMTools (DANECEK *et al.*, 2021) e Picard (BROAD INSTITUTE, 2019). Há duas opções para métodos de classificação: por identificador de sequência ou por coordenadas genômicas (geralmente chamadas respectivamente de localização/*location* ou posição/*position*). A escolha do método dependerá da aplicação dos dados, mas geralmente a classificação por coordenadas é a mais usada para dados genômicos. Os arquivos **.bam** geralmente são acompanhados por um arquivo de índice também conhecido como arquivo **.bai**. Este arquivo sempre será muito menor que o arquivo **.bam** e atua como um índice, indicando onde no arquivo **.bam** uma *read* específica ou um conjunto de *reads* podem

ser encontrados. Como o local das leituras no arquivo provavelmente será alterado com a classificação, é importante gerar ou refazer o arquivo de índice depois que o arquivo **.bam** for classificado. A criação do arquivo de índice pode ser feita novamente usando SAMTools ou Picard. A maioria dos programas que esperam um arquivo **.bam** como *input* também esperam que um arquivo **.bai** complementar, com o mesmo nome, esteja na mesma pasta.

FAI, DICT e outros índices: No decorret deste guia será necessário criar vários arquivos de índices com informações das coordenadas das sequências. O alinhador BWA cria seu próprio *index*, o *index* **.fai** é usado para listar o cromossomo e buscar rapidamente uma sequência dentro do arquivo de referência **.fasta**, e o **.dict** lista os cromossomos, mas também fornece informações sobre as sequências **.fasta** (para ter certeza de que você está usando a mesma referência), o nome do(s) organismo(s), a onde podemos recuperar as sequências, entre outras informações. Este arquivo **.dict** será inserido/comparado com os cabeçalhos dos arquivos **.bam** e **.vcf**.

VCF (Variant Calling Format): é um formato de arquivo de texto (armazenado de forma compactada em **.gz**). Ele contém linhas de meta-informação, um cabeçalho-linha e, em seguida, linhas de dados contendo informações sobre uma posição no genoma. Ele possui a informação do polimorfismo no genoma, bem como o genótipo e pode conter as informações sobre o GL, PL, qualidade de bases, dentre outros dados importantes.

GVCF (Genomic Variant Calling Format): é similar ao **.vcf**. A principal diferença é que o **.gvcf** possui registros para todos os sítios/bases, haja ou não variação. O objetivo é ter todos os sites representados no arquivo para fazer a análise conjunta de uma coorte nas etapas subsequentes. Os registros em um **.gvcf** incluem uma estimativa precisa de quão confiantes é a determinação de que os sítios são homozigotos ou não.

BCF (Binary Variant Calling Format): é a versão binária do arquivo **.vcf**, que mantém as mesmas informações do **.vcf**, mas organizadas de forma diferente que permitem a eficiência maior no processamento e leitura, especialmente em conjuntos de dados muito grandes. A relação entre o **.vcf** e o **.bcf** é similar à relação entre os arquivos **.sam** e **.bam**.

BEAGLE (.beagle): é um arquivo com informações dos genótipos faseados (*phasing*), com a separação dos haplótipos. É uma forma eficiente para analisar grandes conjuntos de dados.

2 AVALIAÇÃO DA QUALIDADE DO SEQUENCIAMENTO

Para facilitar o trabalho foram criados *shell scripts* (**.sh**) e *scripts* em R (**.R**) para usar dentro do *cluster* e automatizar a maioria das análises deste guia. Os *scripts* estão disponíveis na pasta **SCRIPTS** disponibilizada junto com este relatório e também no *cluster* do ITV, no endereço:

```
/bio_temp/share_bio/projects/Amazoomics/pipeline/
```

A preparação das bibliotecas genômicas não será abordada por este guia. Como foi explicado anteriormente, para este guia é necessário ter em mãos os arquivos **.fastq** individuais. Contudo, é necessário confirmar que estamos lidando com dados brutos oriundos de um bom sequenciamento. Para isso, tem que se verificar a qualidade dos dados, obtendo os *outputs* do sequenciador, seja pelo *BaseSpace* da *Illumina* (SanDiego, CA, USA) ou solicitar esses arquivos para quem programou o sequenciador e verificar se o padrão do sequenciamento se aproxima do recomendado (Tabela 02).

Tabela 02: Padrão recomendado pelo fabricante para sequenciamento *Illumina* com relação a densidade de *clusters*, tamanhos dos *outputs* e o número de *reads* de acordo com o equipamento e kit utilizados.

	MiSeq		**MiniSeq**		**NextSeq**	
	V2	**V3**	**Mid**	**High**	**Mid**	**High**
Densidade de *Clusters*	1000-1400 K/mm²	1200-1400 K/mm²	170-220 K/mm²			
Tamanho do *Output* (Gb)	-	**150:** 3.3-3.8 **600:** 13.2-15.0	**300:** 2.1-2.4	**150:** 3.3-3.75 **300:** 6.6-7.5	**150:** 19.5 **300:** 39	**150:** 60 **300:** 120
Número de *Reads* (Milhões)	-	44-50	14-16	44-50	400	800

Outras informações importantes sobre o sequenciamento são: qual o kit de preparação utilizado, kit de sequenciamento, lote e validade dos kits, qual sequenciador, data do sequenciamento, número de amostras, % de bases com qualidade acima de 30 (+ que 80% é recomendado), % de bases que passaram pela filtragem (+ que 80% é recomendado).

Se a qualidade do sequenciamento estiver dentro do esperado, a próxima etapa é o *Demultiplex* das amostras de acordo com os *barcodes*. Este guia não abordará esse tema, mas essa etapa é feita automaticamente dentro do ITV e no final são gerados os arquivos **.fastq** por indivíduo.

2.1 INFORMAÇÕES BÁSICAS SOBRE A QUALIDADE

Para verificar a qualidade dos **.fastq** gerados é necessário recuperar algumas informações por amostra, como número de *reads*, comprimento das *reads*, cobertura, %GC, presença de adaptadores, entre outros. Para isso utiliza-se o programa FastQC (ANDREWS, 2010) ou Prinseq (SCHMIEDER; EDWARDS, 2011)**.** Um exemplo de comando para o uso do FastQC é o seguinte:

```
fastqc -t 64 INPUT.fastq.gz -o caminho/da/pasta/para/salvar/os/resultados
```

Onde **-t** é o número de *threads* ou *cores* utilizados e os *outputs* saem com o mesmo nome do arquivo de início (*input*). Os termos em negrito indicam campos que podem ser editados. Os resultados são arquivos **.html** que devem ser abertos e avaliados individualmente no seu computador pessoal. Em resumo, deve-se atentar às informações abaixo sobre a qualidade e verificar que tipos de limpezas extras são necessários:

1) Basic Statistics: informações sobre o número de *reads*, comprimento das reads (duplicados se for *paired-end*) e %GC;

2) Sequence Length Distribution: se a maioria das sequências tem o mesmo comprimento. Preocupante se for vários tamanhos diferentes;

3) Per base sequence content, Per sequence GC content e **Per base N content** mostram o teor por base da % A/C/G/T; %GC; %N em cada posição da sequência e na sequência inteira. Preocupante se variar bastante por toda a sequência, normalmente variações no início e fim da sequência são esperadas;

4) Per base sequence quality, **Per tile sequence quality** e **Per sequence quality scores**: Índice de qualidade por base, *tile* e sobre as leituras. Preocupante se for menor que 20;

5) K-mers Content: um *k-mer* é um padrão de comprimento k observado mais de uma vez em uma mesma sequência, por exemplo, a repetição de GAGG, mesmo que espaçadas: tccGAGGaaggGAGGaag. Qualquer *K-mer* deve ser representado uniformemente ao longo do comprimento da leitura. Uma lista de *k-mers* que

aparecem em posições específicas com frequência maior que a esperada é relatada no gráfico. Os *K-mers* tendenciosos perto do início da leitura provavelmente são devidos ao processo de cisalhamento, tagmentação ou fragmentação do DNA durante a preparação da biblioteca. *K-mers* no meio da sequência são mais preocupantes. O arquivo R2 é comum ter *K-mers* no meio da sequência, melhor conferir pelo arquivo R1.

Para calcular a cobertura do sequenciamento por amostra é necessário multiplicar o número de *reads* pelo comprimento das *reads* em pares de bases (R1 + R2 se for *paired-end*) e dividir pelo tamanho do genoma usado como referência. Genomas de baixa cobertura tem valores < 5×, genomas com média cobertura entre 5–15× e com boa cobertura acima de 15× (MEISNER; ALBRECHTSEN, 2018).

```
COBERTURA = (n° de reads * (Comprimento R1 + Comprimento R2))/Tamanho do Genoma
```

2.2 REMOÇÃO DE DADOS BAIXA QUALIDADE

Para remover os adaptadores e as bases com bases de baixa qualidade dos arquivos **.fastq** é necessário utilizar algum programa como o Trimmomatic (BOLGER; LOHSE; USADEL, 2014), cutadapt (MARTIN, 2011) ou Prinseq (SCHMIEDER; EDWARDS, 2011). Neste guia vamos usar o Trimmomatic:

```
java -jar /caminho/para/Trimmomatic.jar PE -threads 64 INPUT.R1.fastq.gz INPUT.R2.fastq.gz OUTPUT.R1.Ptrim.fq OUPUT.R1.Utrim.fq OUTPUT.R2.Ptrim.fq OUTPUT.R2.Utrim.fq ILLUMINACLIP:/caminho/para/lista/adaptadores/TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:10:20 LEADING:20 TRAILING:20 MINLEN:35 AVGQUAL:20
```

Onde precisa ser indicado o caminho do programa em Java. *PE* indica que foram usados arquivos *paired-end* e *-threads* o número de processadores. *ILLUMINACLIP* informa o tipo de adaptador utilizado no sequenciamento e é necessário indicar o tamanho e o movimento da janela deslizante para a análise da qualidade das bases (*SLIDINGWINDOW*). Neste caso, o limiar para remoção de dados de baixa qualidade é 20. O *MINLEN* indica o tamanho mínimo da *read* para ser mantida nos arquivos **.fastq** após a remoção de bases nitrogenadas de baixa qualidade. A ordem dos arquivos para *paired-end* é sempre: **R1.Ptrim** (*reads* da R1 que parearam), **R1.Utrim** (*reads* da R1 que não parearam), **R2.Ptrim** (*reads* da R2 que parearam), **R2.Utrim** (*reads* da R2 que não parearam). Lembrando que uma

qualidade de 20 (Q20) corresponde a chance de erro de 1% e que em Q30 a chance de erro é de 0.01%. Para *lcWGS*, um Q20 já é o suficiente e não elimina muitos dados já que a cobertura é baixa (LOU *et al.*, 2021).

Após a remoção das *reads* e bases ruins é necessário reavaliar mais uma vez a qualidades dos arquivos .**fastq** e verificar se os filtros foram aplicados corretamente.

```
fastqc -t 64 INPUT.Ptrim -o caminho/da/pasta/para/salvar/os/resultados
```

Com os arquivos **.fastq** filtrados e com boa qualidade, a próxima etapa é gerar os arquivos **.bam** com o mapeamento na referência. Com esses dados filtrados em **.fastq** também é possível montar os mitogenomas e gerar análises de AAF (Figura 02). Contudo, essas análises não serão abordadas neste guia.

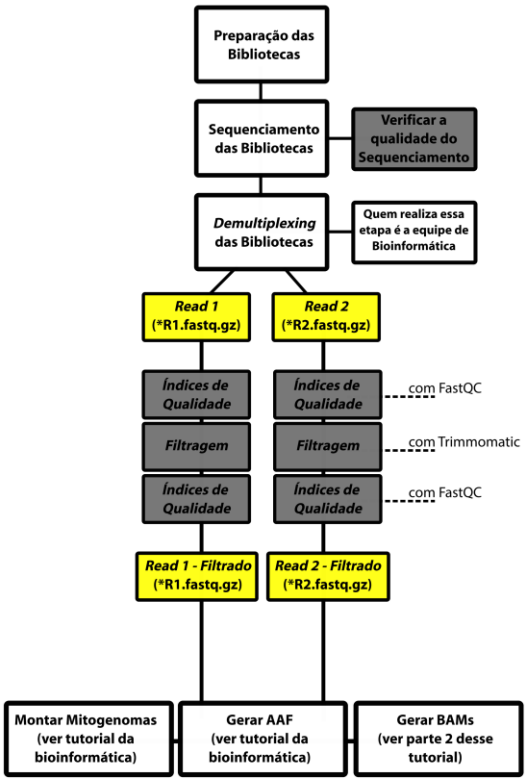

Figura 02: Resumo da etapa de avaliação da qualidade do sequenciamento.

3 MAPEAMENTO DO SEQUENCIAMENTO DE BAIXA COBERTURA

A conversão dos **.fastq** filtrados para arquivos **.bam** é feito em cada indivíduo separadamente e são necessários dois passos: (*i*) mapear as *reads* filtradas no genoma de referência, e (*ii*) remover as *reads* duplicadas e possíveis contaminações. Uma terceira etapa de subamostragem pode ser incluída caso você queira reduzir a cobertura de um arquivo **.bam** (Figura 03). Com esses arquivos **.bam** é possível criar os *inputs* para análises demográficas e para a identificação de SNPs que podem ser usadas nas análises populacionais ou filogenéticas (Figura 03).

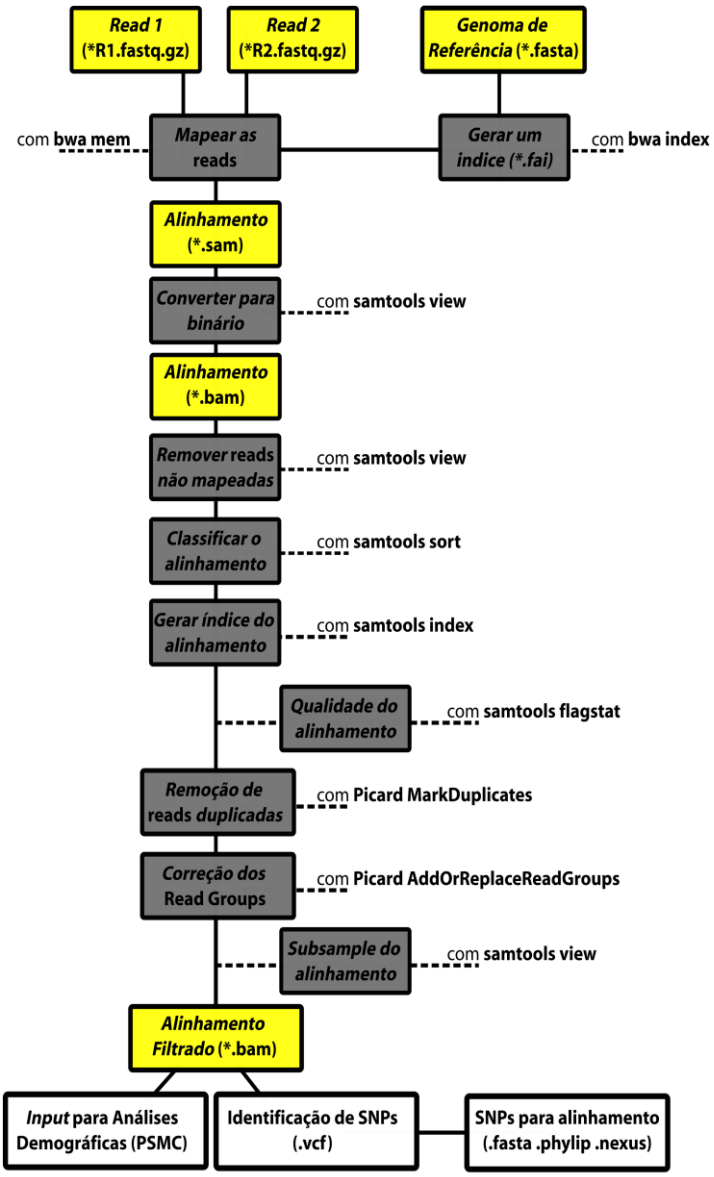

Figura 03: Resumo da etapa de mapeamento de *reads* no genoma de referência.

3.1. MAPEAMENTO DE *READS* NO GENOMA DE REFERÊNCIA

O primeiro passo é criar um índice para o arquivo de referência, indicando onde no arquivo **.fasta** de referência uma *read* específica ou um conjunto de *reads* podem ser encontrados. É realizada uma única vez e faz com que a etapa de alinhamento seja mais rápida, usando o mesmo arquivo em todos os indivíduos para padronizar o alinhamento. Os índices devem ser mantidos na mesma pasta da referência. Para o mapeamento neste guia, utilizou-se o programa BWA (LI; DURBIN, 2009).

```
bwa index genoma_de_referência _em_.fasta
```

Segundo passo é mapear as *reads* na referência, criando um arquivo em formato **.sam** que marca quais *reads* estava presentes tanto na referência quanto na amostra. Deve ser feito para cada indivíduo e pode-se usar multiprocessadores com o argumento (**-t**):

```
bwa mem -t 64 REFERÊNCIA.fasta INPUT_R1_filtrado.fastq.gz
INPUT_R2_filtrado.fastq.gz > OUTPUT.sam
```

Após o mapeamento é necessário converter o arquivo **.sam** em **.bam**. Essa conversão diminui o tamanho dos arquivos e o tempo de processamento. Deve ser feito para cada indivíduo. Nesta etapa usamos o SAMTools (DANECEK *et al.,* 2021) e para permitir multi-processadores deve-se usar o argumento (**-@**)

```
samtools view -h -b -S -@ 64 INPUT.sam > OUTPUT.bam
```

Outro passo importante é a remoção de *reads* não-mapeadas, que reduz o tamanho dos arquivos. Também deve ser feito para cada indivíduo:

```
samtools view -b -F 4 -@ 64 INPUT.bam > OUTPUT_UNFLAGS.bam
```

O passo seguinte é classificar o arquivo **.bam** resultante e indexá-lo para que todos os indivíduos tenham a mesmas posições de *reads:*

```
samtools sort -@ 64 INPUT_UNFLAGS.bam -o OUTPUT_SORTED.bam
samtools index OUTPUT_SORTED.bam
```

Após o alinhamento e remoção das *reads* não-alinhadas é necessário ver a qualidade e número de alinhamento. Essa etapa não permite o uso de múltiplos processadores. O *output* dessa etapa virá no arquivo **.log** na mesma ordem dos indivíduos listado, com 13 linhas para cada indivíduo. Nessa etapa as *reads* alinhadas devem ser de 100% e as pareadas (R1+R2) deveriam ser maior de 80%.

```
samtools flagstat INPUT_SORTED.bam


20082880 + 0 in total (QC-passed reads + QC-failed reads)  #quantas reads passaram + não passaram no controle de qualidade
0 + 0 secondary
191547 + 0 supplementary
0 + 0 duplicates
20082880 + 0 mapped (100.00% : N/A)  #quantas reads foram mapeadas com a referência
19891333 + 0 paired in sequencing
9962487 + 0 read1
9928846 + 0 read2
19243568 + 0 properly paired (96.74% : N/A)  #quantas reads foram pareadas (R1 + R2)
19805454 + 0 with itself and mate mapped
85879 + 0 singletons (0.43% : N/A)
486128 + 0 with mate mapped to a different chr
291302 + 0 with mate mapped to a different chr (mapQ>=5)
```

3.2. REMOÇÃO DE *READS* DUPLICADAS

Essa etapa usa o programa Picard (BROAD INSTITUTE, 2019) não permite a paralelização. Assim é necessário aumentar a memória (*-Xmx)*, usando um único processador.

O primeiro passo é remover *reads* duplicados devido à PCR. Isso reduz o tamanho dos arquivos. Há uma discussão se essa etapa é necessária ou não para o *variant calling.* Ebbert *et al.* (2016) diz que as duplicatas de PCR tem efeito mínimo na acurácia na identificação das variantes, mas é um passo feito rotineiramente, para cada indivíduo, usando o arquivo **.bam** após a etapa de classificação e indexação:

```
java -Xmx100g -jar /caminho/para/PICARD.jar MarkDuplicates REMOVE_DUPLICATES=true I=INPUT_SORTED.bam O=OUTPUT_WITHOUT_DUPLICATES.bam M=marked_dup_metrics.txt
```

Às vezes é necessário fazer a correção dos *Read Groups* do arquivo **.bam,** que são os nomes individuais das amostras**.** Os nomes das amostras não pode ter menos que três caracteres, ter acento gráfico e cada amostra tem que ter um nome único. Sem nomes únicos por indivíduo, o programa GATK (POPLIN et. al., 2017) não funcionará.

```
java -Xmx100g -jar /caminho/para/PICARD.jar AddOrReplaceReadGroups I=INPUT_WITHOUT_DUPLICATES.bam O=OUTPUT_RG.bam RGID=4 RGLB=WGS RGPL=illumina RGPU=barcode CREATE_INDEX=True RGSM=[nome_da_amostra]
```

3.3. SUBAMOSTRAGEM DE *READS*

Esse passo final é facultativo e só deve ser usado para fazer um subamostragem do arquivo **.bam**, reduzindo o seu tamanho ou a cobertura dele. Essa redução facilita as análises e em seguida é necessário indexar o arquivo novamente.

```
samtools view -s 0.3166 -b INPUT_RG.bam > OUTPUT_reduced.bam

samtools index INPUT_reduced.bam

#para confirmar que houve redução do arquivo para a cobertura desejada:
samtools depth INPUT_reduced.bam | awk '{sum+=$3} END { print "Average = ",sum/NR}'
```

O valor 0.3166 indica que serão selecionadas aleatoriamente 31,66% das *reads* originais para a subamostragem.

4 IDENTIFICAÇÃO DOS SNPS COM GATK

Para a estimativa dos GL ou PL podem ser utilizadas diferentes programas, como SAMTools (DANECEK *et al.,* 2021) ou GATK (POPLIN *et al.*, 2017). Alguns estudos sugeriram que o GATK é melhor para dados de baixa cobertura (LIU *et al.*, 2013; POPLIN *et al.*, 2017; LIU; SHEN; BAO*,* 2022). Há a opção de usar o programa ANGSD (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014) que permite a identificação dos SNPs, calculando o GL com diferentes algoritmos, inclusive com o GATK. Neste guia, será explicado a abordagem utilizando GATK que é mais lenta e também utilizando o ANGSD que faz a identificação e a filtragem de SNPs ao mesmo tempo.

4.1 IDENTIFICAÇÃO DE SNPS

Inicialmente é necessário criar um arquivo **.gvcf** para cada indivíduo separadamente usando os **.bam** filtrados, classificados e com índices. Além disso, também é preciso do índice **.dict** para a referência em formato **.fasta** (Figura 04). O programa não permite vários *threads* ou processadores, então é melhor usar um processador com muita memória para fazer essas análises. Essa é a etapa mais demorada do processo. Em média, cada amostra com 5 milhões de *reads* demora cerca de 70 horas, com 100GB de memória RAM.

```
samtools faidx REFERÊNCIA.fasta

java -Xmx100g -jar /caminho/para/PICARD.jar CreateSequenceDictionary R=
REFERÊNCIA.fasta O= REFERÊNCIA.dict

gatk --java-options "-Xmx100g" HaplotypeCaller --reference
REFERÊNCIA.fasta -ERC GVCF --input INPUT.bam --output OUTPUT.gvcf
```

Após todos os **.gvcf** individuais serem concluídos, o próximo passo é combinar os **.gvcf** corrigindo o índice de *missing data* ou dados faltantes, pois às vezes *missing data* pode ser codificado no **.gvcf** como um homozigoto verdadeiro 0|0. Para a correção o programa avalia o genoma de referência novamente. Essa etapa é mais rápida. O ideal seria usar o **GenomicsDBImport** do GATK que faz esses dois passos aos mesmo tempo, mas ela só está disponível a partir da versão 4.1. A versão instalada no *cluster* do ITV é a versão 4.0.

```
gatk --java-options "-Xmx100g" CombineGVCFs -R REFERÊNCIA.fasta -V
individuo_1.gvcf -V individuo_2.gvcf -V individuo_3.gvcf -O
OUTPUT_COMBINADO.gvcf

gatk --java-options "-Xmx100g" GenotypeGVCFs -R REFERÊNCIA.fasta -V
INPUT_COMBINADO.gvcf -O OUTPUT.vcf
```

Após essa etapa o arquivo **.vcf** final está pronto. A etapa seguinte será a de filtragem de SNPs, utilizando outros programas como o BCFTools (DANECEK *et al.,* 2021) e o R (R CORE TEAM, 2022).

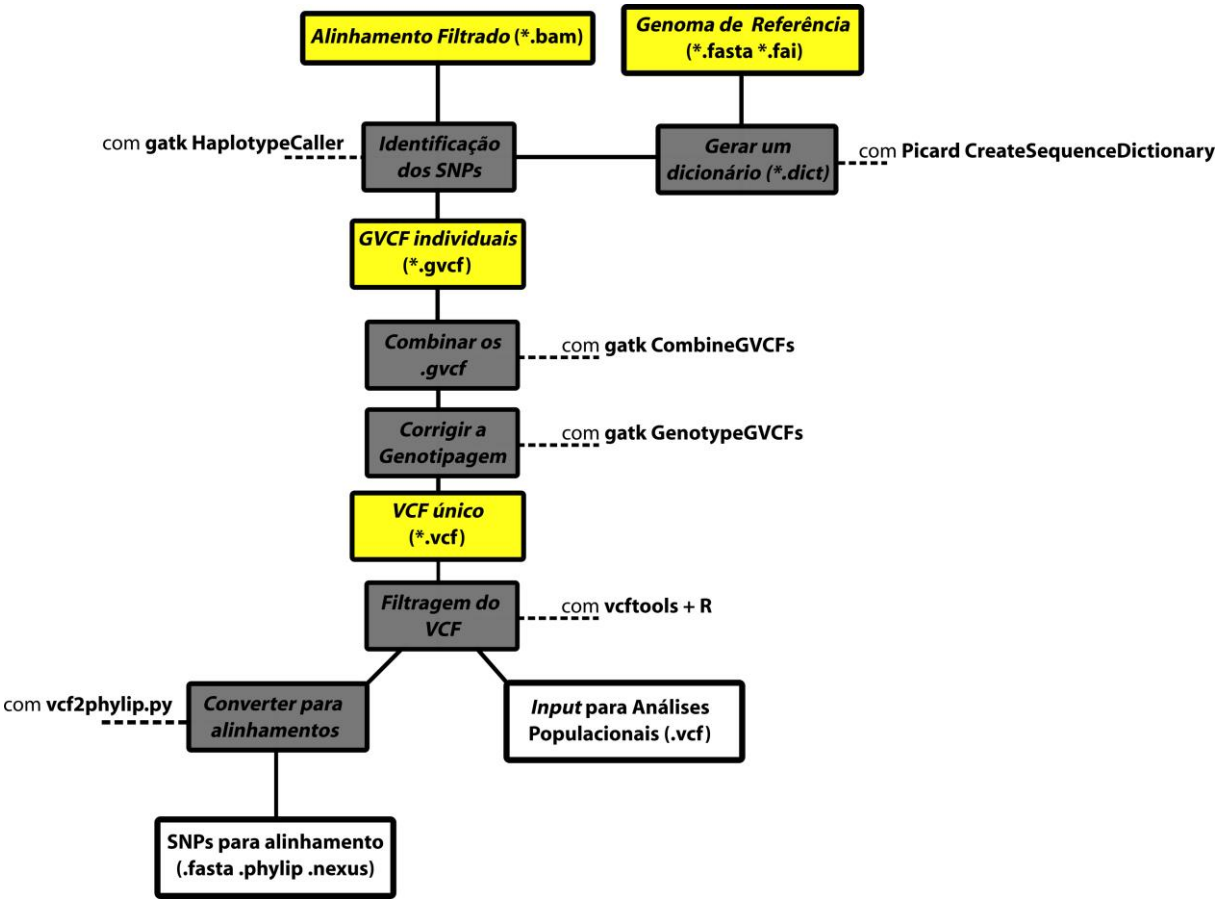

Figura 04: Resumo da etapa de identificação de SNPs com GATK.

4.2 POR QUE É NECESSÁRIO A FILTRAGEM DE SNPS?

O arquivo **.vcf** gerado pelo GATK contém todos os sítios variantes encontrados nas amostras, isto é, inclui além de SNPs os *indels*, sítios e SNPs multi-alélicos e variantes que podem ser erros de genotipagem. Para as análises populacionais, o interesse é usar apenas SNPs e se for um organismo diploide, apenas SNPs bialélicos, por isso é necessário uma filtragem dos sítios variantes.

Normalmente o arquivo **.vcf** é muito grande, assim, é recomendado fazer esta etapa de filtragem, ou pelo menos os passos iniciais dela dentro do *cluster.*

Essa etapa de filtragem seguirá as recomendações de Lou *et al.* (2021) e das recomendações da filtragem rígida do GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants). Os passos iniciais usando o programa BCFTools, chamada de **filtragem rígida,** podem ser efetuados para todos os conjuntos de dados, variando os parâmetros de filtragem se for necessário. Os demais passos da filtragem com GATK podem ser modificados, removidos ou alterados de acordo com o objetivo do estudo, são feitos em *script* R e essa etapa é chamada de **filtragem flexível**.

5.2 DICAS PARA ACELERAR A FILTRAGEM

Como foi dito anteriormente, todas as etapas foram automatizados em *scripts* **.sh** e **.R** e estão disponíveis na pasta de **SCRIPTS** ou no *cluster* na pasta da pipeline. Para aumentar a velocidade das filtragens pode-se converter e comprimir o arquivo **.vcf** para **.bcf.gz** e assim poupar espaço e memória durante as análises. A maioria das análises aceita o arquivo **.gz**, mas algumas análises não e por isso em alguns passos será necessário descomprimi-lo.

Para verificar os resultados da filtragem pode-se contabilizar os números de SNPs e de outros sítios variantes por meio de um comando que faz as estatísticas sumárias dos SNPs retidos. O *output* é um arquivo de texto com várias informações, sendo as mais importantes: *number of samples, number of records, number of no-ALTs, number of SNPs, number of MNPs, number of indels, number of others, number of multiallelic sites, number of multiallelic SNP sites.* A cada etapa da filtragem, pode-se analisar como as mudanças nos parâmetros de filtragens alteram o números e qualidade dos SNPs.

```
#Compactar e indexar o arquivo .vcf:
bgzip -c INPUT.vcf > OUTPUT.vcf.gz
tabix -p vcf OUTPUT.vcf.gz

#Converter o arquivo .vcf para .bcf:
bcftools view INPUT.vcf.gz -Ob -o OUTPUT.bcf.gz

#Estatísticas sumárias para o arquivo .vcf ou .bcf:
bcftools stats -s - INPUT.bcf.gz > stats.txt
```

5.3 FILTRAGEM RÍGIDA

São os filtros de qualidade do mapeamento e do sequenciamento e também alguns filtros específicos como o de frequência alélica, cobertura e dados faltantes por sítio. Os filtros de QD, MD, FS e SOR e POSRANK só funcionam com **.vcf** ou **.bcf** criados com GATK e não com outro *softwares* como o ANGSD.

```
bcftools filter -e "QD<2 || MQRankSum<-12.5 || FS>60 || SOR>3 ||
ReadPosRankSum<-8 || QUAL<20 || MQ<20 || MAF<0.05 || MEAN(FORMAT/DP)<0.8
|| MEAN(FORMAT/DP)>50 || F_MISSING>0.5" --SnpGap 10 INPUT.bcf.gz -Ob -o
PARCIAL.bcf.gz

bcftools view -c1 -v snps -m2 -M2 PARCIAL.bcf.gz -Ov -o PARCIAL.vcf

vcfutils.pl varFilter -1 0.000001 PARCIAL.vcf > FINAL.vcf

bcftools stats -s - FINAL.vcf > stats.hardfilters.txt
```

Abaixo há o detalhamento de cada parâmetro utilizado:

QUALIDADE POR COBERTURA (QD): Qualidade por cobertura (*Quality by Depth*; QD) é a confiança da identificação de uma variante em uma *base call* (ou seja, o campo QUAL do arquivo **.vcf**), dividida pela cobertura não-filtrada. Esta anotação destina-se a normalizar a qualidade da variante para evitar a inflação causada quando há cobertura muito alta (elemento repetitivos). Para fins de filtragem, é melhor usar QD do que apenas QUAL ou DP diretamente como é recomendado por Lou *et al.* (2021). O GATK recomenda a remoção de variantes com QD abaixo de 2.

QUALIDADE DO MAPEAMENTO (MQRankSum): O *Mapping Quality Rank Sum Test* (MQRankSum) compara as qualidades de mapeamento de leituras que suportam a referência e o alelo alternativo. Se todas as leituras mapeiam corretamente, o **MQRankSum** esperado é 0. Valores negativos indicam um MQ inferior das leituras que suportam o alelo alternativo, em comparação com as leituras

que suportam o alelo de referência. Os valores de **MQRankSum** geralmente variam de cerca de -10,5 a 6,5. O GATK recomenda a remoção de variantes extremas com **MQRankSum** inferior a -12,5.

VIÉS DE FITA DE LEITURA (FS): *Fisher Strand* (FS) é a probabilidade em escala de *Phred* de que haja polarização de fita naquela posição. Indica se o alelo alternativo foi visto com mais ou menos frequência na fita R1 ou R2 do que o alelo de referência. Um valor próximo de 0 indica pouca ou nenhuma polarização da fita. O GATK recomenda a remoção de variantes com FS maior que 60. Alguns autores indicam que esse é um valor conservador, que filtrará apenas casos extremos e que pode deixar muitos falsos positivos no conjunto de dados, recomendando um valor de 40.

VIÉS DE FITA DE LEITURA (SOR): *Strand Odds Ratio* (SOR) também é uma medida para o viés de fita de leitura (*strand bias*), mas leva em consideração as proporções de leituras que cobrem ambos os alelos e, portanto, é uma medida melhor para sítios que têm mais leituras em uma direção do que na outra. Os valores de SOR variam de 0 a mais de 9, um valor próximo a 0 indica pouca ou nenhuma polarização da fita. O GATK recomenda a remoção de variantes com SOR maior que 3.

SOR e FS estão relacionados, mas não são iguais. Todos eles medem o viés de fita de leitura (um tipo de viés de sequenciamento no qual uma fita de DNA é favorecida em relação à outra, R1 *vrs* R2), o que pode resultar em avaliação incorreta da quantidade de evidência observada para um alelo *versus* o outro. SB é o campo/anotação do **.vcf** que fornece as contagens brutas de leituras que suportam cada alelo na fita direta e reversa. FS é o resultado do uso dessas contagens em um teste *Fisher*. SOR é uma anotação relacionada que aplica um teste estatístico diferente (usando as contagens SB) que é melhor para dados de alta cobertura.

VIÉS POR POSIÇÃO DE LEITURA (ReadPosRankSum): O teste *Read Position Rank Sum* (ReadPosRankSum) mede a posição relativa da referência *versus* o alelo alternativo dentro das leituras. Por exemplo, os SNPs no final das leituras podem ser mais provavelmente um erro de sequenciamento do que um SNP verdadeiro. Um valor em torno de 0 significa que há pouca ou nenhuma diferença em onde os alelos são encontrados em relação às extremidades das leituras. O GATK recomenda a remoção de variantes com ReadPosRankSum inferior a -8,0.

QUALIDADE DO SEQUENCIAMENTO (QUAL): Qualidade de sequenciamento gerada pelo sequenciador em uma escala *Phred*. Lou *et al.* (2021)

recomenda remover todas as variantes com qualidade abaixo de 20. Alguns trabalhos usam um limiar mais restritivo, como 30 ou 33, com dados de alta cobertura. A versão atual do GTAK recomenda o uso do QD ao invés do QUAL para dados de alta cobertura.

QUALIDADE DO MAPEAMENTO (MQ): Lou *et al.* (2021) recomenda remover variantes com qualidade de mapeamento abaixo de 20 para dados de baixa cobertura (abaixo de 5× o genoma), enquanto o GATK recomenda remoção de MQ abaixo de 40 para dados de alta cobertura, mas recomenda o uso do QD ao invés do MQ para dados de alta cobertura.

FREQUÊNCIA ALÉLICA MÍNIMA (MAF): Etapa para remover SNPs com frequência baixa (alelos raros), eles podem ser erros de sequenciamento ou de genotipagem e não tem muito peso em análises populacionais globais. Um MAF de 0.05 indica que o alelo tem que aparecer em pelo menos 5% das amostras (indivíduos) para ser mantido.

COBERTURA MÍNIMA (DP): É necessário remover SNPs com poucas cópias, eles podem ser erros de sequenciamento ou de genotipagem. Lou *et al.* (2021) recomenda remover SNPs com cobertura média menor que 0,8× entre os indivíduos (após filtrar em qualidade de mapeamento). Pode-se testar e ver se são perdidos muitos SNPs com o filtro DP e testar um número diferente, menos restritivo. Há dois campos de DP no arquivo **.vcf**: **(*i*) INFO/DP =** cobertura combinada de todas as amostras, isto é, a soma de cópias de todas as amostra para aquele SNP. INFO/DP=154. **(*ii*) FORMAT/DP =** cobertura do SNP (número de cópias sequenciadas desse SNP), esse número de cópias tem que ser maior ou igual a 1. Para o padrão RAD-Seq um valor mínimo de 4 ou 5 cópias é recomendado (PARIS *et al.,* 2017). Para análises populacionais envolvendo lcWGS, com 10 amostras por população, é possível ter média de 0.125× (LOU *et al.,* 2021).

COBERTURA MÁXIMA (DP): Também é necessário remover SNPs com muitas cópias, eles podem representar áreas repetitivas do genoma ou erros de genotipagem ou alinhamento. Normalmente é usado a mediana do valor de cobertura de todos os SNPs + 2*desvio-padrão da cobertura. Para calcular a mediana e o desvio-padrão é mais fácil e rápido usando o VCFtools (DANECEK et al. 2021):

```
vcftools --gzvcf INPUT.vcf.gz --site-mean-depth
cut -f3 out.ldepth.mean | csvstat --median > MEDIAN.txt
cut -f3 out.ldepth.mean | csvstat --stdev > SD.txt
```

Infelizmente o BCFtools não permite chamar outras funções dentro da filtragem de SNPs, então é necessário olhar o valor da mediana e do desvio-padrão e alterar esses número no comando de filtragem.

VIÉS POR PROXIMIDADE DE INDEL (--SnpGap): SNPs e outras variantes próximas a *indels* podem ser artefatos e não variantes verdadeiras devido a erros no alinhamento e por isso devem ser removidos. O valor da distância mínima de SNP e um *indels* varia em diferentes artigos, sempre variando de 3 (CONG et al., 2022) a 10 pb de distância (LUDINGTON; SANDERS, 2021).

DADOS FALTANTES POR SNPS (F_MISSING): é importante remover SNPs com mais de 50% de dados faltantes (LOU *et al.,* 2021). A remoção de indivíduos com muito *missing data* será feita na próxima etapa com *scripts* em R. Também será possível na próxima etapa de filtragem remover indivíduos e SNPs com *missing data* maior que 30% (HUANG; KNOWLES, 2016).

SNPs MONOMÓRFICOS (-c1): Os SNPs retidos podem não ser informativos para análises populacionais, sendo considerados monomórficos. SNPs monomórficos são aqueles com o mesmo genótipo em todos os indivíduos e podem ser gerados após remoção de bases de baixa qualidade e/ou indivíduos inteiros que tinham previamente algum polimorfismo. Os SNPs monomórficos não atrapalham os cálculos de diversidade genética ou estruturação, pois não são informativos, mas acabam por exigir mais memória e espaço dos computadores e *clusters*. A remoção dos SNPs monomórficos é recomendada principalmente após a remoção de indivíduos por dados faltantes.

MANTER APENAS SNPs (-v): O arquivo **.vcf** possui vários tipos de variantes e para as análises de Genômica da Conservação é necessário analisar apenas os SNPs e remover *indels* e variantes não-SNPs.

SNPs BIALÉLICOS (-m -M): Os SNPs retidos até o momento podem ter mais de dois alelos se o organismo-alvo for poliploide ou, se forem diploides, eles podem apresentar mais dois alelos devido a erros de alinhamento. Por exemplo, quando três *contigs* alinham na mesma área do genoma, quando o esperado seria apenas dois *contigs*. Como os organismos-alvo são diploides é necessário manter apenas os SNPs

bialélicos. Caso sejam organismos poliploide pode-se ajustar a poliploidia esperada alterando os números de alelos mínimos e máximos em -m e -M.

VIÉS POR DETECÇÃO DE POLIMORFISMO (varFilter -1): A capacidade da variante ser verdadeira e não um viés do sequenciamento, baseado em um *p*-valor de detecção, calculado com outras métricas. É necessário manter bases com *p*-valor muito significativo, menores que 10e-6. O **vcfutils.pl** já vem instalado como BCFtools e ele só trabalha com arquivos **.vcf** descompactado.

5.4. FILTRAGEM FLEXÍVEL

Os seguintes filtros podem não ser recomendados de acordo com a pergunta do projeto. Por exemplo, se estudo for sobre adaptação ou estudos filogenéticos, as filtragens por *outlier* SNPs e por Equilíbrio de Hardy-Weinberg (HWE) podem não ser necessárias. Esses filtros estão implementados em um *script* R usando o arquivo **.vcf** e principalmente os pacotes “vcfR” (KNAUS; GRÜNWALD*,* 2017), “SNPfiltR” (DERAAD, 2022), “dartR” (GRUBER;GEORGES*,* 2019) e “pcadapt” (LUU; BAZIN; BLUM, 2017).

QUALIDADE DO GENÓTIPO (gq): Qualidade mínima do genótipo para que ele seja mantido no seu conjunto de dados. Por exemplo, 'gq = 30' removerá todos os genótipos com um índice de qualidade iguais ou menores que 29. Para a designação de um genótipo é necessário usar as chamadas de base (*base calls*) e pontuações de qualidade (*quality scores*) em processo dividido em duas etapas: chamada ou identificação de genótipo (*genotype calling*) e chamada ou identificação de SNP (*SNP calling* ou *variant calling*). A identificação de SNP visa determinar em quais posições existem polimorfismos ou em quais posições pelo menos uma das bases difere de uma sequência de referência. A chamada de genótipo é o processo de determinação do genótipo para cada indivíduo e normalmente é feito apenas para posições nas quais um SNP ou uma 'variante' foi identificada (NIELSEN *et al.*, 2011). Se os SNPs já foram filtrados por alguma métrica de qualidade anterior, esse passo é facultativo.

COBERTURA BRUTA (depth): seria o equivalente ao filtro **FORMAT/DP**, a cobertura do SNP (número de cópias sequenciadas desse SNP). A diferença é que neste caso, os SNPs com número de cópias inferiores ao limiar serão convertidos em *missing data* e não excluídos como no GATK.

EQUILÍBRIO ALÉLICO (Allele Balance): remove os genótipos heterozigotos que estão foram do equilíbrio alélico (AB). Em espécies diploides espera-se que os

genótipos heterozigotos neutros apresentem um AB próximo a ~0,5 pelo HWE. Pedersen *et al.* (2022) recomenda uma filtragem entre 0,2 e 0,8 e o padrão da função é 0,25 e 0,75. SNPs fora desses limites indicariam erros na chamada de genótipos e são excluídos.

DADOS FALTANTES (missing data): Um passo extra para remover indivíduos com muito *missing data* (*missing_by_sample*) e também SNPs (*missing_by_snp*). Para estudos populacionais, o ideal seria ter indivíduos e SNPs com menos de 30% de dados faltantes, mas não há um número mágico, mas acima de 40% pode ser problemático (CERCA *et al.*, 2021). Para estudos filogenéticos, *missing data* em torno de 50% é aceitável (STREICHER;SCHULTE;WEINS, 2016). Se poucos SNPs forem recuperados com esses filtros, pode-se fazer uma imputação dos dados faltantes a partir das frequências alélicas, mas esse tema não será abordado neste guia.

DESEQUILÍBRIO DE LIGAÇÃO (LD): SNPs muito próximos ou em um mesmo loci podem ter frequências similares por estarem segregando em conjunto por gerações. Por isso, é importante que os SNPs sejam independentes. Duas métricas podem ser calculadas para quantificar o LD, a estatística *D* e a correlação r^2, ambas usando as frequências alélicas do par de loci analisado. A maioria dos trabalhos usam o r^2 para as filtragens e a maioria dos trabalhos só eliminam SNPs e loci com r^2 muito alto e para análises de estruturação genética. SNPs em LD são importante para a cálculo de tamanhos populacionais baseados em LD, como o NeEstimator (DO *et al.,* 2014). Normalmente se cria dois bancos de dados: um para análises de estruturação e diversidade, onde não pode conter SNPs ligados, e outro banco de dados sem filtro ou com filtro mais relaxado para LD a fim de calcular do tamanho populacional por meio do LD.

Para dados de RAD ou GBS, o importante é que seja selecionado 1 SNP por loci/*contig* especialmente em organismos não-modelos com alinhamento *de novo,* pois não se sabe a posição do SNP dentro do genoma (O'LEARY *et al.*, 2018). No caso de lcWGS, com o mapeamento conhecido, há várias maneiras, mas normalmente escolhe-se 1 SNP a cada 10 mil pares de base (FRIEL *et al.,* 2021). No BCFTools pode-se calcular o valor de r^2 a cada janela de 10.000 pb e eliminar SNPs com valores de correlação maiores que um limiar, de por exemplo, 0.8, 0.6, 0.4 ou 0.2 dentro daquela janela. No PLINK (PURCELL *et al.,* 2007) pode-se escolher além do tamanho da janela, uma *sliding window* chamada *step* que é a mudança de leitura

dentro do tamanho do fragmento escolhido. No R, o pacote “SNPfiltR” (DERAAD, 2022) pode-se especificar apenas a distância mínima entre os SNPs, sem o cálculo do r^2. A melhor estratégia depende dos objetivos do estudo.

A forma de definir a janela para manter 1 SNP é arbitrária, não há uma explicação para uma janela de 10,000 pb ou 150 pb. Uma opção é a criação de um gráfico de decaimento de LD (*LD Decay*) que mostra qual distância mínima para a janela retornar apenas SNPs não-ligados. Usando o programa PopLDdecay (ZHANG *et al.*, 2019) é possível construir o gráfico (Figura 05) com os seguintes comandos:

```
#para criar os output usados no gráfico:
./PopLDdecay -i INPUT.vcf.gz -o OUTPUT.stat.gz -MaxDist 300 -s POP_A_SAMPLES.list

-MaxDist # distância máxima entre 2 SNPs em Kb para calcular o r².
-s POP_A_SAMPLES.list # usar apenas um grupo de amostras para o cálculo, uma coluna com o nome das amostras em arquivo de texto.

#para construir o gráfico
perl Plot_OnePop.pl -inFile OUTPUT.stat.gz -output FIG_LD_DECAY
```

Para filtrar SNPs a uma distância mínima no R, este guia usa a função “*distance_thin()*”. Para o BCFTools é necessário ter o *plugin* **+prune** para remover os SNPs por desequilíbrio de ligação. É possível no BCFTools remover os SNPs pela distância combinado com os valores de correlação entre as suas frequências alélicas (r^2). Ao limitar uma janela para o cálculo do r^2 ou para escolha de 1 SNP por janela, o programa removerá aquele SNP com menor frequência alélica quando for necessário eliminar um ou mais SNPs classificados como ligados. Se os dois possuem a mesma frequência, o primeiro SNP é mantido. Caso seja necessário usar exemplos da literatura para justificar tamanhos de janelas e valores de r^2 há alguns exemplos:

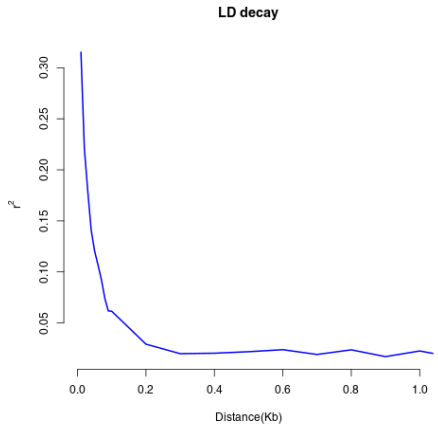

Figura 05: Gráfico de decaimento de LD a uma distância máxima de 1Kb em *Anodorhynchus hyacinthinus*. SNPs com distância de 100 pb já tem pouca correlação em suas frequências alélicas (r^2 ~ 0.05), podendo usar um limiar de 100 pb para as filtragens.

#Filtragem de 1 SNP em uma janela de 10.000 pb (FRIEL *et al.,* 2021):

```
bcftools +prune -w 10000bp -n 1 INPUT.vcf.gz -Ov -o OUTPUT.vcf
```

#Filtragem por r^2 > 0.8 em uma janela de 20.000 pb (CURRAN *et al.,* 2020).

```
bcftools +prune -l 0.8 -w 20000bp INPUT.vcf.gz -Ov -o OUTPUT.vcf
```

#Filtragem por r^2 > 0.8 em uma janela de 1.000 pb mantendo 1 SNP a cada 150 pb para certificar que foram selecionados SNPs ao longo de mais porções do genoma (GALLA *et al.,* 2018):

```
bcftools +prune -l 0.8 -w 1000bp INPUT.vcf.gz -Ov -o PARCIAL.vcf
bcftools +prune -w 150bp -n 1 PARCIAL.vcf -Ov -o OUTPUT.vcf
```

#Filtragem por r^2 > 0.85 em uma janela de 2.000 pb mantendo 1 SNP por janela (AGUILAR-ORDOÑEZ *et al.*, 2021):

```
bcftools +prune -l 0.85 -w 2000bp INPUT.vcf.gz -Ov -o PARCIAL.vcf
bcftools +prune -w 2000bp -n 1 PARCIAL -Ov -o OUTPUT.vcf
```

#Filtragem de 1 SNP em uma janela de 150 pb (tamanho do contig do RAD) (O'LEARY *et al.,* 2018):

```
bcftools +prune -w 150bp -n 1 INPUT.vcf.gz -Ov -o PARCIAL.vcf
```

#Filtragem por r^2 > 0.25 em uma janela de 800 pb, comprimento máximo das *reads* R1 + R2 mais 150 bp de cada lado (MANUZZI *et al.*, 2022). Seria 600 pb no caso de sequenciamento de 150 pb por *read*.

```
bcftools +prune -l 0.25 -w 800bp INPUT.vcf.gz -Ov -o OUTPUT.vcf
```

DESVIOS NO EQUILÍBRIO DE HARDY-WEINBERG (HWE): Essa etapa serve para remover SNPs que apresentam desvios no modelo de Equilíbrio de Hardy-Weinberg (HWE). O HWE descreve o estado de uma população ideal na ausência de forças evolutivas, onde as frequências alélicas são previsíveis, uma vez que permanecem constantes ao longo das gerações. A remoção de loci que não seguem o HWE é frequentemente usada para remover erros de genotipagem e loci que estão potencialmente sob seleção. A remoção de erros de genotipagem é, em geral, benéfica para análises, enquanto a remoção de loci sob seleção pode ser necessária para análises que assumem neutralidade dos loci. No entanto, muitos outros fatores podem causar desvios do HWE, como seleção purificadora, endogamia ou estruturação populacional e eliminar muitos loci desviantes pode acabar por homogeneizar demais suas amostras.

Pearman, Urban e Alexander (2022) concluíram que, apesar de ser uma abordagem de filtragem amplamente utilizada, a filtragem em que se remove qualquer SNP com desvio de HWE da população global é inadequada e leva a níveis reduzidos de estrutura populacional inferida – especialmente quando a estrutura populacional é alta. Uma opção é dividir as amostras por localidade ou populações (se elas já foram definidas previamente) e remover qualquer loci exibindo desvios de HWE em todas as populações ou na maioria delas. Outra opção é não fazer essa etapa de filtragem, caso não precise filtrar loci neutros, como em análises filogenéticas. Segundo os autores é ideal fazer análises exploratórias, aplicar os filtros de HWE por localidade e considerar o *trade-off* entre o número de loci perdidos pela aplicação desse filtro em relação às informações obtidas.

É comum que dentro de uma mesma população não seja eliminado poucos ou nenhum SNP. Os valores de corte para considerar um desvio do HWE são bem variados de acordo com os estudos, vai de 0.05 (LARSON *et al.,* 2014), até $1x10^{-4}$ (CARVALHO *et al.,* 2019), $1x10^{-6}$ (REED *et al.,* 2015) até $1x10^{-10}$ (ZHANG *et al.,* 2022). Neste guia será demonstrado como fazer isso no BCFTools e no R com o pacote “dartR”.

No BCFTools é necessário criar um arquivo **.txt** com duas colunas separadas por tabulação. Uma coluna contendo os nome das amostras e a outra a localidade ou população correspondente àquela amostra. Nomes não podem conter espaços. O BCFTools usa o teste de HWE de Wigginton *et al.* (2005), que é indicado para grandes banco de dados de SNPs, corrigindo os erros do tipo I, e com *p*-valor corrigido pela

frequência alélica mínima das amostras. Wigginton *et al.* (2005) calcularam *p*-valor (bilateral) como a probabilidade da amostra observada mais a soma de todas as probabilidades de casos mais extremos, um valor considerado conversador por Graffelman e Moreno (2013).

```
bcftools +fill-tags INPUT.vcf.gz -Oz -o OUTPUT.vcf.gz -- -S POP_INFO.txt
-t HWE
```

Para a análise em R, alguns parâmetros precisam ser escolhidos:

```
#Converter arquivo .vcf para to genlight
gl_vcf = vcfR2genlight(INPUT)
gl_vcf

#Adicionar informações sobre os grupos ou populações no genlight
pop(gl_vcf) = popmap$pop
popNames(gl_vcf)

# Filtrar com as recomendações de Pearman et al. (2022) no genlight
gl_vcf_fil = gl.filter.hwe(
  gl_vcf,
  subset = "each",
  n.pop.threshold = round(length(popNames(gl_vcf))/2,0), # para o SNP ser
mantindo ele não pode desviar do HWE na maioria das populações
  method_sig = "Exact", # (WIGGINTON et al., 2005)
  multi_comp = FALSE,
  multi_comp_method = "BH", # (BENJAMINI; HOCHBERG, 1995)
  alpha_val = 0.05,
  pvalue_type = "midp",# (GRAFFELMAN; MORENO, 2013)
  cc_val = 0.5,
  min_sample_size = 5, #população/grupo tem que ter 5 indivíduos
  verbose = NULL
)

#Manter apenas SNPs que restaram no objeto genlight filtrado:
filtered_snps = which(gl_vcf@loc.names %in% gl_vcf_fil@loc.names)
OUTPUT= INPUT[filtered_snps, ]
OUTPUT
```

REMOÇÃO DE LOCI ADAPTATIVOS (ADAPT): SNPs com sinais de seleção possuem frequências bem diferentes dos loci neutros o que pode enviesar os resultados. É comum removê-los para análises populacionais, mas não para as análises filogenéticas. Há vários programas que removem esses SNPs outliers, mas a maioria precisa da definição de populações *a priori*. O pacote “pcadapt” não precisa dessa definição de populações, ele seleciona por meio de uma PCA (LUU; BAZIN; BLUM, 2017). Os comandos em R para essa etapa estão disponíveis com os demais *scripts*.

5 IDENTIFICAÇÃO DOS SNPS COM ANGSD

A vantagem do ANGSD sobre o GATK é que ele é mais rápido, pois utiliza diretamente os arquivos **.bam** para a identificação dos SNPs ou para a estimativa das frequências alélicas para as análises, sem a necessidade de criação de arquivos individuais ou de identificar os genótipos em um arquivo **.vcf**. Além disso, ao mesmo tempo em que identifica os SNPs, o ANGSD pode filtrá-los por qualidade e outros parâmetros (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014). O fluxo de trabalho é dividido em duas etapas: (*i*) ANGSD gera a partir dos arquivos **.bam** os dados de entrada específicos para as análises populacionais ou salva os SNPs em formato **.bcf** ou **.beagle**; e a outra etapa (*ii*) um programa associado secundário usa esses arquivos como *inputs* para realizar as análises genômicas. Esses programas secundários podem ser um *script* em R ou Python, ou mesmo métodos computacionais intensivos (FUMAGALLI *et al.*, 2014). O foco deste guia é demonstrar a primeira etapa que corresponde à geração dos *inputs* para as análises genômicas.

O ANGSD assume que as amostras são diploides, não aceita *indels* para as análises e suporta vários modelos diferentes para calcular GL como o modelo SOAPsnp; modelo GATK; modelo SAMtools (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014). As taxas de erro de sequenciamento nesses modelos GL podem ser fixas e obtidas a partir da qualidade de sequenciamento e mapeamento ou podem ser estimadas a partir dos dados. No ANGSD, os SNPs são inferidos com base na estimativa de frequência alélica usando um teste de razão de verossimilhança (LTR; *Likelihood Ratio Test*) que pode rejeitar que a frequência alélica daquele sítio seja 0 (NIELSEN *et al.*, 2011).

No futuro, provavelmente todas as análises genômicas serão feitas usando GL diretamente dos arquivos **.bam** ou arquivos similares e não haverá uma etapa de identificação dos SNPs e seus genótipos. No entanto, atualmente muitas análises não foram generalizadas para se basear em GL, e elas ainda usam genótipos identificados e portanto é necessário incluir a identificação de SNPs quando se usa o ANGSD (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014).

Pode-se criar os *inputs* e **.vcf** no ANGSD a partir de uma lista de SNPs preexistente ou a partir dos arquivos **.bam.** No ANGSD alguns filtros importantes não estão disponíveis dentro do programa como os referentes ao desequilíbrio de ligação (LD). Por isso, faz sentido usar outros programas auxiliares, que lidam com o LD,

selecionar os SNPs e depois rodar novamente o ANGSD usando apenas a lista final com as posições dos SNPs de boa qualidade.

Os parâmetros do ANGSD pode ser divididos em três grupos: (*i*) parâmetros de entrada de dados, (*ii*) parâmetros de filtragem e (*iii*) parâmetros de saída dos dados. Para a lista completa de parâmetros, acesse http://www.popgen.dk/angsd/

Exemplos de parâmetros de entrada de dados:

`angsd` ou `./angsd` #execução do programa, dependendo de como ele foi instalado
`-P 10` #número de processadores ou *threads*
`-b` **`bam_pop1.filelist`** #arquivo com caminhos para os arquivos **.bam** dos indivíduos
`-ref` #caminho para referência em **.fasta**
`-anc` #caminho para genoma ancestral, na ausência pode-se repetir a referência em **.fasta**
`-r 11` # restringir a análise para o *scaffold* 11 ou cromossomo 11, código referente ao **Chrom**
`-GL 2` #especifica o modelo usado para o cálculo do GL. 1= SAMTools; 2=GATK
`-sites` #usa uma lista de sítios/SNPs *a priori* para as análises

Exemplos de parâmetros de filtragem:

`-remove_bads 1` #remove *reads* com baixa qualidade
`-uniqueOnly 1` #remove *reads* com múltiplos *hits* no mapeamento
`-only_proper_pairs 1` # remove *reads* que não estão em pares (R1 e R2)
`-baq 1` #remove falsos SNPs que estão próximos de INDELS
`-C 50` # reduz o efeito de *reads* com incompatibilidades excessivas
`-trim 0` #número de bases que serão removidas em cada ponta da *reads*
`-minMapQ 20` #remove bases com qualidade de mapeamento menor que 20
`-minQ 20` #remove bases com qualidade de sequenciamento menor que 20

`-doSNPstat 0` # 0=remove os SNPs;1=não remove os SNPs, retorna uma tabela com as métricas; esse comando permite as filtragem por SNPs
`-doHWE 1` # esse comando permite as filtragem por SNPs
`-hwe_pval 0.01` # remove sítios fora do HWE, limiar usado por Hager *et al.*, (2022)
`-sb_pval 0.01` #filtra por *strand bias;* limiar de Hager *et al.*, (2022)
`-edge_pval 0.05` # filtro por *edge bias*, limiar de Frei *et al.* (2022)
`-mapQ_pval 0.05` #filtro pelo *p*-valor do mapeamento; limiar de Frei *et al.* (2022)
`-qscore_pval 0.05` #filtro pelo *p*-valor do sequenciamento; limiar de Frei *et al.* (2022)

`-doMaf 1` #frequências alélicas fixadas para *Minor* e *Major,* permite filtrar por métricas baseadas em frequência alélica

`-doMajorMinor 1` #inferir frequências do *Major* e do *Minor* com GL

`-minMaf 0.05` #remover sítios que estão presentes em menos de 5% dos indivíduos

`-SNP_pval 1e-6` #limiar para a identificação de sítios polimórficos

`-skipTriallelic 1` #considerar SNPs bialélicos, lembrando que o ANGSD descarta os *indels*

`-doCounts 1` # permite filtrar por métricas de cobertura

`-setMinDepth 0.8` #se a cobertura for menor que 0.8 o sítio é removido

`-setMaxDepth 50` #se a cobertura for maior que 50 o sítio é removido

`-minInd 38` #*missing data* por sítio; 0= sem filtragem; número próximo de 30% seria o ideal

`-doGeno 2` #permite a filtragem por valor esperado de heterozigosidade, para explicação do número 2, veja parâmetros de output

`-hetbias_pval 0.05` #filtro para controlar o viés de heterozigotos; limiar por Balogh *et al.* (2020)

Exemplos de parâmetros de saída dos dados:

`-out test/Anodorhynchus_pop1` #caminho+nome dos resultados do ANGSD

`-doMajorMinor 1` #retorna as frequências alélicas; comando obrigatório para os filtros baseados em frequência alélicas

`-doMaf 1` #retorna as frequências alélicas baseadas em GL; comando obrigatório para os filtros baseados em frequência alélicas

`-doCounts 1` #retorna informação sobre cobertura; comando obrigatório para os filtros baseados em cobertura

`-doSNPstat 0` #retorna informação sobre os SNPs; comando obrigatório para os filtros baseados em qualidade dos SNPs

`-makeMatrix 1` #retorna uma matriz de distância entre os indivíduos; usada para PCA MDS

`-doIBS 2` #retorna a base consenso das reads para determinada posição; com essas bases são feitas a matriz para o PCA MDS

`-doCov 1` #retorna a matriz de covariância para o PCA

`-doGeno 2` #retorna arquivos de genótipo, faz o *Genotype Calling*

`-doBcf 1` #cria um **.bcf** com os SNPs filtrados

`-doPost 1` #estima o GL baseado em frequência alélica para fazer o *Genotype Calling*

`-doGlf 2` #cria arquivo **.beagle** com os dados de GL e PL para usar como *input* em outros programas

`-doSaf 1` #cria arquivos usados em análises demográficas e de FST; 1=estima GL com vários indivíduos

As análises com o ANGSD são feitas baseadas em frequências alélicas e por isso são realizadas em dois níveis: em (*i*) escala global e em (*ii*) escala populacional/local. Isso significa que a geração dos *outputs* deve ser feita considerando todos os indivíduos juntos e também cada população separadamente. Quando não há uma divisão *a priori* para as populações, pode-se fazer uma PCA e uma análise similar de *Admixture* com todos os indivíduos e a partir daí dividir as suas populações ou agrupamentos genéticos (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014; LOU *et al.*, 2021). Para qualquer um dos níveis (local ou global), são necessárias três etapas para gerar todos os *inputs* das análises genômicas: (*i*) a criação do arquivo **.geno** que será utilizado para o cálculo do desequilíbrio de ligação (LD); (*ii*) a remoção dos SNPs ligados e a criação de uma lista com as posições dos SNPs não-ligados ou *unlinked SNPs*; e (*iii*) a criação de todos os *inputs* necessários para as análises genômicas posteriores utilizando apenas a lista de SNPs não-ligados.

5.1 ETAPA 1: CRIAÇÃO DO ARQUIVO .GENO

O primeiro passo para rodar o ANGSD a partir dos **.bam** é criar uma lista com o caminho dos arquivos **.bam,** o **bam.filelist**, como no exemplo abaixo:

```
/bio_temp/share_bio/projects/Amazoomics/Anodorhynchus/mapped/RG_LGEMA12869.bam
/bio_temp/share_bio/projects/Amazoomics/Anodorhynchus/mapped/RG_LGEMA15979.bam
/bio_temp/share_bio/projects/Amazoomics/Anodorhynchus/mapped/RG_LGEMA18726.bam
```

É importante pensar bem na ordem das amostras, pois a ordem dos indivíduos nos *outputs* seguirá a ordem desse **.filelist** e o nome das amostras será o endereço completo informado nesse arquivo. Se a análise for local ou populacional, é necessário criar para cada população/localidade um **.filelist** diferente. Se a análise global for o objetivo, então deve-se listar todos os arquivos **.bam** em um mesmo **.filelist**.

Seguindo as recomendações de LOU *et al.* (2021) os filtros importantes para genomas de baixa cobertura seriam os seguintes:

```
#Para ativar o ANGSD no conda, dentro do cluster:
source activate /bio_temp/share_bio/softwares/miniconda3/envs/angsd
```

#Para rodar o ANGSD, argumentos em negritos são os que mais alteram o número de SNPs:

#Para análises globais:

```
angsd -P 10 -GL 2 -b /caminho/para/bam.filelist -ref /caminho/para/referência.fasta -anc /caminho/para/referência_ou_anc.fasta -remove_bads 1 -uniqueOnly 1 -only_proper_pairs 1 -baq 1 -C 50 -trim 0 -minMapQ 20 -minQ 20 -minMaf 0.05 -SNP_pval 1e-6 -skipTriallelic 1 -setMinDepth 0.8 -setMaxDepth 50 -minInd 2 -doSNPstat 1 -doHWE 1 -sb_pval 0.01 -edge_pval 0.05 -mapQ_pval 0.05 -qscore_pval 0.05 -hetbias_pval 0.05 -doMajorMinor 1 -doMaf 1 -doCounts 1 -doGeno 2 -out /caminho/para/resultados/STEP1/Nomedosoutputs
```

#Para análises por população e ou localidades é necessário um limiar para o p-valor do HWE:

```
angsd -P 10 -GL 2 -b /caminho/para/bam.filelist -ref /caminho/para/referência.fasta -anc /caminho/para/referência_ou_anc.fasta -remove_bads 1 -uniqueOnly 1 -only_proper_pairs 1 -baq 1 -C 50 -trim 0 -minMapQ 20 -minQ 20 -minMaf 0.05 -SNP_pval 1e-6 -skipTriallelic 1 -setMinDepth 0.8 -setMaxDepth 50 -minInd 2 -doSNPstat 1 -doHWE 1 -HWE_pval 0.01 -sb_pval 0.01 -edge_pval 0.05 -mapQ_pval 0.05 -qscore_pval 0.05 -hetbias_pval 0.05 -doMajorMinor 1 -doMaf 1 -doCounts 1 -doGeno 2 -out /caminho/para/resultados/STEP1/Nomedosoutputs
```

Trabalhos publicados usaram os seguintes comandos ou *scripts*:

#KORNELIUSSEN; ALBRECHTSEN e NIELSEN (2014):

```
./angsd -b bam.list -doMaf 1 -doMajorMinor 1 -snp_pval 0.01 -GL 1 -P 4 -baq 0 -ref hg19.fa -minQ 13 -minMapQ 10
```

THERKILDSEN E PALUMBI (2017):

```
-snp_pval 1e-6 -minQ 20 -setMinDepth 300 -setMaxDepth 3028
```
#(mean depth +2 standard deviations)

KONGSSTOVU *et al.* (2022):

```
-snp_pval 0.000001 -minMaf 0.05 -minInd 90
```

DE FERRAN *et al.* (2022):

```
-doFasta 2  -doCounts 1 -explode 1 -setMinDepth 10 -minMapQ= 20
```

Caso o número de SNPs retidos seja pequeno, pode-se modificar alguns parâmetros. Os parâmetros que mais alteram o número de SNPs são: **-minInd -hetbias_pval -minMaf -SNP_pval -setMinDepth** e **-hwe_pval** (lembrando que o último só deve ser usado para ANGSD em nível populacional).

5.2 ETAPA 2: FILTRAGEM DE SNPS NÃO-LIGADOS

Após a etapa 1, além de outros arquivos, será formado um arquivo **.geno.gz.** Esse arquivo deve ser criado com o comando **-doGeno 2** na Etapa 1, pois é o formato aceito nas próximas etapas. É necessário descompactar esse arquivo para conseguir as informações sobre as posições e número de SNPs e com esse arquivo calcular o r^2 para o desequilíbrio de ligação (LD), pode ser feito da seguinte forma:

```
#Descompactar arquivo .gz:
gzip -dk INPUT.geno.gz

#Criar o arquivo com as posições do SNPs
cut -f 1-2 INPUT.geno > POSITIONS_SNPS.txt

#Número de SNPs:
wc -l POSITIONS_SNPS.txt

#Número de indivíduos:
awk '{print NF}' INPUT.geno | sort -nu | tail -n 1
```

Com essas informações é possível calcular o LD usando os GL no programa ngsLD (FOX *et al.*, 2019) e remover os SNPs ligados. Para usar os seguintes comandos é preciso ter instalado alguns pacotes do R e módulos do Python, veja as instruções de instalação dos autores em https://github.com/fgvieira/ngsLD. Para a construção do gráfico de decaimento do LD é necessário informar os arquivos de métricas de r^2 como uma lista de endereços em um arquivo **.txt**. Todos os arquivos listados estarão no mesmo gráfico. O argumento **--fit_level** pode ser alterado para melhor o encaixe da curva de decaimento. Para remover os SNPs ligados, a etapa de *prune*, basta informar a distância máxima entre os SNPs em bp para assumir que eles estão conectados (**--max_dist**) e peso mínimo de uma aresta para assumir que os SNPs estão conectados (**--min_weight**). No caso, quanto maior o peso, mais SNPs são mantidos.

```
#Calcular o r² entre os SNPs:
./ngsLD --geno INPUT.geno.gz --n_ind 54 --n_sites 18062 --pos
POSITIONS_SNPS.txt --out METRICS_SNPS.txt

#Fazer o gráfico de decaimento do LD:
Rscript --vanilla --slave ./scripts/fit_LDdecay.R --ld_files
list_METRICS_SNPS.txt --fit_level 2 --out LDDecay.pdf > LDDecay_R2.txt

#Remover um dos SNPs ligados de acordo com a distância do gráfico de decaimento do LD:
python3 ./scripts/prune_ngsLD.py --input METRICS_SNPS.txt --max_dist 200
--min_weight 0.5 --out UNLINKED_SNPS.pos

#Número de SNPs restantes:
wc -l UNLINKED_SNPS.pos
```

Para rodar o ANGSD a partir de uma lista de SNPs que já existe, ou comparar os resultados com um **.vcf** já existente, é necessário fazer um arquivo de texto com as posições dos SNPs-alvo. Pode-se utilizar o R e o pacote *"vcfR"* com os seguintes comandos:

```
#A. Instalar ou carregar o pacote vcfR:
if(!require('vcfR')) install.packages("vcfR"); library('vcfR')

#B. Carregar o arquivo .vcf:
snpsR = read.vcfR("Anodorhynchus_Neutral_lcWGS.vcf", verbose = T)

#C. Extrair as informações sobre a posição e CHROM dos SNPs:
snps_pos = snpsR@fix[,1:2]

#D. Verificar o objeto:
head(snps_pos)
length(snps_pos[,1])

#E. Salvar o resultado em uma tabela sem nome de colunas ou linhas, separadas por tabulação:
write.table(as.data.frame(snps_pos),"UNLINKED_SNPS.pos", sep = "\t", quote
= F, col.names = F, row.names = F)

#F. O arquivo UNLINKED_SNPS.pos fica assim:
WOUG01000001.1 93434
WOUG01000002.1 29
```

5.3 ETAPA 3: CRIAR *INPUTS* COM SNPS NÃO-LIGADOS

Com esse arquivo contendo as posições dos SNPs filtrados e não-ligados é possível rodar o ANGSD só para a geração de *outputs*, sem a necessidade de parâmetros de filtragem, otimizando o tempo. Para isso é necessário antes de rodar o ANGSD, criar índices para os SNPs selecionados:

Se os coeficientes de endogamia estimados forem muito maiores que 0, pode-se corrigir seus genótipos e frequências alélicas pela endogamia com o ngsTools (FUMAGALLI *et al.*, 2014), mas neste caso é necessário indicar uma sequência ancestral com **-anc.**

```
#Para ativar o ANGSD no conda, dentro do cluster:
source activate /bio_temp/share_bio/softwares/miniconda3/envs/angsd

#Para indexar os SNPs selecionados:
angsd sites index /caminho/para/UNLINKED_SNPS.pos

#Código para análise:
angsd -P 50 -GL 2 -b /caminho/para/bam.filelist -ref
/caminho/para/referência.fasta -anc /caminho/para/referência_ou_anc.fasta
-sites /caminho/para/UNLINKED_SNPS.pos -doMajorMinor 1 -doMaf 1 -doCounts
1 -makeMatrix 1 -doIBS 1 -doCov 1 -doGeno 2 -doBcf 1 -doPost 1 -doGlf 23
-doSaf 1 -out /caminho/para/resultados/Nomedosoutputs
```

Serão gerados 12 arquivos de saída, dos mais variados formatos que serão utilizados por programas secundários para análises, como o PCangsd (MEISNER; ALBRECHTSEN, 2018; 2019) ou ngsTools (FUMAGALLI *et al.*, 2014).

6 CONSIDERAÇÕES FINAIS

Esse relatório abrange um guia para as três etapas principais das identificação de SNPs em trabalhos de Genômica da Conservação: (*i*) avaliação da qualidade do sequenciamento; (*ii*) mapeamento do genoma de baixa cobertura com um genoma de referência; e (*iii*) filtragem dos SNPs para as análises.

As interpretações dos resultados oriundos dos SNPs dependem da quantidade e qualidade dos dados gerados. Em um projeto como o Amazoomics que visa o sequenciamento e trabalhos de Genômica da Conservação de várias espécies amazônicas ameaçadas de extinção, é necessário uma padronização das ferramentas utilizadas e uma explicação clara das ferramentas utilizadas.

Todos os *scripts* e exemplos citados neste guia estão depositados no cluster do ITV, no endereço: *</bio_temp/share_bio/projects/Amazoomics/pipeline/>* com acesso permitido a qualquer usuário. Este guia será de fundamental importância para as análises futuras dos projetos e para os futuros alunos e bolsistas que desenvolverão trabalhos similares dentro e fora do ITV.

REFERÊNCIAS

AMITEYE, S. Basic concepts and methodologies of DNA marker systems in plant molecular breeding. **Heliyon**, v. 7, n. 10, p. e08093, 1 out. 2021.

ANDREWS, S. **FastQC: a quality control tool for high throughput sequence data**. . [S.l: s.n.]. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>. , 2010

BALOGH, A. et al. Population genomics in two cave-obligate invertebrates confirms extremely limited dispersal between caves. **Scientific Reports 2020 10:1**, v. 10, n. 1, p. 1–11, 16 out. 2020. Disponível em: <https://www.nature.com/articles/s41598-020-74508-9>. Acesso em: 17 nov. 2022.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 57, n. 1, p. 289–300, 1 jan. 1995. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.x>. Acesso em: 21 nov. 2022.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014. Disponível em: <https://academic.oup.com/bioinformatics/article/30/15/2114/2390096>. Acesso em: 17 ago. 2022.

BROAD INSTITUTE, T. **Picard toolkit**. . [S.l: s.n.]. , 2019

BUERKLE, C. A.; GOMPERT, Z. Population genomics based on low coverage sequencing: how low should we go? **Molecular Ecology**, v. 22, n. 11, p. 3028–3035, 1 jun. 2013. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.12105>. Acesso em: 17 nov. 2022.

CARVALHO, C. S. et al. Habitat loss does not always entail negative genetic consequences. **Frontiers in Genetics**, v. 10, p. 1101, 13 nov. 2019. Disponível em: <https://www.frontiersin.org/article/10.3389/fgene.2019.01101/full>. Acesso em: 7 jul. 2020.

CERCA, J. et al. Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. **Methods in Ecology and Evolution**, v. 12, n. 5, p. 805–817, 1 maio 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13562>. Acesso em: 21 nov. 2022.

CONG, P. K. et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. **Nature Communications 2022 13:1**, v. 13,

n. 1, p. 1–15, 26 maio 2022. Disponível em: <https://www.nature.com/articles/s41467-022-30526-x>. Acesso em: 21 nov. 2022.

DANECEK, P. et al. Twelve years of SAMtools and BCFtools. **GigaScience**, v. 10, n. 2, p. giab008, 29 jan. 2021. Disponível em: <https://academic.oup.com/gigascience/article/10/2/giab008/6137722>. Acesso em: 17 ago. 2022.

DE FERRAN, V. et al. Phylogenomics of the world's otters. **Current Biology**, v. 32, n. 16, p. 3650- 3658.e4, 22 ago. 2022.

DERAAD, D. A. snpfiltr: An R package for interactive and reproducible SNP filtering. **Molecular Ecology Resources**, v. 22, n. 6, p. 2443–2453, 1 ago. 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13618>. Acesso em: 21 nov. 2022.

DO, C. et al. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. **Molecular Ecology Resources**, v. 14, n. 1, p. 209–214, 1 jan. 2014.

EBBERT, M. T. W. et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. **BMC Bioinformatics**, v. 17, n. 7, p. 491–500, 25 jul. 2016. Disponível em: <https://link.springer.com/articles/10.1186/s12859-016-1097-3>. Acesso em: 17 nov. 2022.

FOX, E. A. et al. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. **Bioinformatics**, v. 35, n. 19, p. 3855–3856, 1 out. 2019. Disponível em: <https://academic.oup.com/bioinformatics/article/35/19/3855/5418793>. Acesso em: 18 nov. 2022.

FREI, D. et al. Genomic variation from an extinct species is retained in the extant radiation following speciation reversal. **Nature Ecology & Evolution 2022 6:4**, v. 6, n. 4, p. 461–468, 24 fev. 2022. Disponível em: <https://www.nature.com/articles/s41559-022-01665-7>. Acesso em: 17 nov. 2022.

FRIEL, J. et al. Comparative analysis of genotyping by sequencing and whole-genome sequencing methods in diversity studies of olea europaea l. **Plants**, v. 10, n. 11, p. 2514, 1 nov. 2021. Disponível em: <https://www.mdpi.com/2223-7747/10/11/2514/htm>. Acesso em: 21 nov. 2022.

FUENTES-PARDO, A. P.; RUZZANTE, D. E. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. **Molecular Ecology**, v. 26, n. 20, p. 5369–5406, 1 out. 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.14264>. Acesso em: 17 nov. 2022.

FUMAGALLI, M. et al. ngsTools: methods for population genetics analyses from next-generation sequencing data. **Bioinformatics**, v. 30, n. 10, p. 1486–1487, 15 maio 2014. Disponível em: <https://academic.oup.com/bioinformatics/article/30/10/1486/267009>. Acesso em: 17 nov. 2022.

GALLA, S. J. et al. Reference genomes from distantly related species can be used for discovery of Single Nucleotide Polymorphisms to inform conservation management. **Genes**, v. 10, n. 1, p. 9, 22 dez. 2018. Disponível em: <https://www.mdpi.com/2073-4425/10/1/9/htm>. Acesso em: 18 ago. 2022.

GRAFFELMAN, J.; MORENO, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. **Statistical Applications in Genetics and Molecular Biology**, v. 12, n. 4, p. 433–448, 1 ago. 2013. Disponível em: <https://www.degruyter.com/document/doi/10.1515/sagmb-2012-0039/html?lang=de>. Acesso em: 21 nov. 2022.

GRUBER, B.; GEORGES, A. **dartR: importing and analysing SNP and silicodart data generated by genome-wide restriction fragment analysis**. . [S.l: s.n.]. Disponível em: <https://cran.r-project.org/package=dartR>. , 2019

HAGER, E. R. et al. A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. **Science**, v. 377, n. 6604, p. 399–405, 22 jul. 2022. Disponível em: <https://www.science.org/doi/10.1126/science.abg0718>. Acesso em: 17 nov. 2022.

HAINS, T. et al. The complete genome sequences of 22 parrot species (Psittaciformes, Aves). **F1000Research 2020 9:1318**, v. 9, p. 1318, 12 nov. 2020. Disponível em: <https://f1000research.com/articles/9-1318>. Acesso em: 17 nov. 2022.

HELYAR, S. J. et al. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. **Molecular Ecology Resources**, v. 11, n. SUPPL. 1, p. 123–136, mar. 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1755-0998.2010.02943.x>. Acesso em: 17 nov. 2022.

HUANG, H.; KNOWLES, L. L. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. **Systematic Biology**, v. 65, n. 3, p. 357–365, 1 maio 2016. Disponível em: <https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syu046>. Acesso em: 7 fev. 2019.

KNAUS, B. J.; GRÜNWALD, N. J. vcfr: a package to manipulate and visualize variant call format data in R. **Molecular Ecology Resources**, v. 17, n. 1, p. 44–53, 1 jan. 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.12549>. Acesso em: 21 nov. 2022.

KONGSSTOVU, S. et al. Atlantic herring (Clupea harengus) population structure in the Northeast Atlantic Ocean. **Fisheries Research**, v. 249, p. 106231, 1 maio 2022.

KORNELIUSSEN, T. S.; ALBRECHTSEN, A.; NIELSEN, R. ANGSD: Analysis of Next Generation Sequencing Data. **BMC Bioinformatics**, v. 15, p. 356, 25 nov. 2014. Disponível em: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0356-4>. Acesso em: 17 nov. 2022.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 4 mar. 2012. Disponível em: <https://www.nature.com/articles/nmeth.1923>. Acesso em: 17 nov. 2022.

LARSON, W. A. et al. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). **Evolutionary Applications**, v. 7, n. 3, p. 355–369, 1 mar. 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/eva.12128>. Acesso em: 8 out. 2020.

LEACHÉ, A. D.; OAKS, J. R. The utility of Single Nucleotide Polymorphism (SNP) data in phylogenetics. **Annual Review of Ecology and Systematics**, v. 48, p. 69–84, 6 nov. 2017. Disponível em: <https://www.annualreviews.org/doi/abs/10.1146/annurev-ecolsys-110316-022645>. Acesso em: 17 nov. 2022.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 15 jul. 2009. Disponível em: <https://academic.oup.com/bioinformatics/article/25/14/1754/225615>. Acesso em: 17 ago. 2022.

LIU, J.; SHEN, Q.; BAO, H. Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. **PLOS ONE**, v. 17, n. 1, p. e0262574, 1 jan. 2022. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262574>. Acesso em: 17 nov. 2022.

LIU, X. et al. Variant callers for Next-Generation Sequencing data: a comparison study. **PLOS ONE**, v. 8, n. 9, p. e75619, 27 set. 2013. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075619>. Acesso em: 17 nov. 2022.

LO, E. et al. Selection and utility of Single Nucleotide Polymorphism markers to reveal fine-scale population structure in human Malaria parasite *Plasmodium falciparum*. **Frontiers in Ecology and Evolution**, v. 6, p. 145, 26 set. 2018.

LOU, R. N. et al. A beginner's guide to low-coverage whole genome sequencing for population genomics. **Molecular Ecology**, v. 30, n. 23, p. 5966–5993, 1 dez. 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.16077>. Acesso em: 18 ago. 2022.

LUDINGTON, A. J.; SANDERS, K. L. Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences. **Molecular Ecology**, v. 30, n. 2, p. 545–554, 1 jan. 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.15726>. Acesso em: 21 nov. 2022.

LUU, K.; BAZIN, E.; BLUM, M. G. B. pcadapt: an R package to perform genome scans for selection based on principal component analysis. 1 jan. 2017, [S.l.]: Blackwell Publishing Ltd, 1 jan. 2017. p. 67–77. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.12592>. Acesso em: 23 ago. 2020.

MARTIN, E. R. et al. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. **Bioinformatics**, v. 26, n. 22, p. 2803–2810, 15 nov. 2010. Disponível em: <https://academic.oup.com/bioinformatics/article/26/22/2803/227284>. Acesso em: 17 nov. 2022.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, v. 17, n. 1, p. 10–12, 2 maio 2011. Disponível em: <https://journal.embnet.org/index.php/embnetjournal/article/view/200/479>. Acesso em: 17 nov. 2022.

MCKENNA, A. 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, 1 set. 2010.

MEISNER, J.; ALBRECHTSEN, A. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. **Genetics**, v. 210, n. 2, p. 719–731, 1 out. 2018. Disponível em: <https://academic.oup.com/genetics/article/210/2/719/5931101>. Acesso em: 17 nov. 2022.

_____. Testing for Hardy–Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. **Molecular Ecology Resources**, v. 19, n. 5, p. 1144–1152, 1 set. 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13019>. Acesso em: 18 nov. 2022.

MORIN, P. A.; MARTIEN, K. K.; TAYLOR, B. L. Assessing statistical power of SNPs for population structure and conservation studies. **Molecular Ecology Resources**, v. 9, n. 1, p. 66–73, 1 jan. 2009. Disponível em: <http://doi.wiley.com/10.1111/j.1755-0998.2008.02392.x>. Acesso em: 18 maio 2019.

NIELSEN, R. et al. Genotype and SNP calling from next-generation sequencing data. **Nature Reviews Genetics**, v. 12, n. 6, p. 443–451, 18 maio 2011. Disponível em: <https://www.nature.com/articles/nrg2986>. Acesso em: 17 nov. 2022.

O'LEARY, S. J. et al. These aren't the loci you'e looking for: principles of effective SNP filtering for molecular ecologists. **Molecular Ecology**, v. 27, n. 16, p. 3193–3206, 1 ago. 2018. Disponível em: <http://doi.wiley.com/10.1111/mec.14792>. Acesso em: 9 jun. 2020.

PARIS, J. R.; STEVENS, J. R.; CATCHEN, J. M. Lost in parameter space: a road map for stacks. **Methods in Ecology and Evolution**, v. 8, n. 10, p. 1360–1373, 1 out. 2017. Disponível em: <http://doi.wiley.com/10.1111/2041-210X.12775>. Acesso em: 7 ago. 2018.

PEARMAN, W. S.; URBAN, L.; ALEXANDER, A. Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. **Molecular Ecology Resources**, p. 1–15, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13646>. Acesso em: 18 ago. 2022.

PEDERSEN, B. S. et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. **bioRxiv**, p. 2020.08.13.249532, 26 ago. 2020. Disponível em: <https://www.biorxiv.org/content/10.1101/2020.08.13.249532v3>. Acesso em: 21 nov. 2022.

POPLIN, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. **bioRxiv**, p. 201178, 24 jul. 2017. Disponível em: <https://www.biorxiv.org/content/10.1101/201178v2>. Acesso em: 8 jun. 2020.

PURCELL, S. 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, 1 set. 2007.

R CORE TEAM, T. **R: a language and environment for statistical computing**. . [S.l: s.n.]. Disponível em: <https://www.r-project.org/>. , 2022

REED, E. et al. A guide to genome-wide association analysis and post-analytic interrogation. **Statistics in Medicine**, v. 34, n. 28, p. 3769–3792, 10 dez. 2015. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.6605>. Acesso em: 21 nov. 2022.

SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics**, v. 27, n. 6, p. 863–864, 15 mar. 2011. Disponível em: <https://academic.oup.com/bioinformatics/article/27/6/863/236283>. Acesso em: 17 nov. 2022.

STRATTON, M. Genome resequencing and genetic variation. **Nature Biotechnology 2008 26:1**, v. 26, n. 1, p. 65–66, jan. 2008. Disponível em: <https://www.nature.com/articles/nbt0108-65>. Acesso em: 17 nov. 2022.

STREICHER, J. W.; SCHULTE, J. A.; WIENS, J. J. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. **Systematic Biology**, v. 65, n. 1, p. 128–145, 1 jan. 2016. Disponível em: <https://academic.oup.com/sysbio/article/65/1/128/2461451>. Acesso em: 21 nov. 2022.

THERKILDSEN, N. O.; PALUMBI, S. R. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. **Molecular Ecology Resources**, v. 17, n. 2, p. 194–208, 1 mar. 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.12593>. Acesso em: 18 nov. 2022.

VIGNAL, A. et al. A review on SNP and other types of molecular markers and their use in animal genetics. **Genet. Sel. Evol**, v. 34, p. 275–305, 2002.

WIGGINTON, J. E.; CUTLER, D. J.; ABECASIS, G. R. A note on exact tests of Hardy-Weinberg Equilibrium. **The American Journal of Human Genetics**, v. 76, n. 5, p. 887–893, 1 maio 2005.

ZHANG, C. et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. **Bioinformatics**, v. 35, n. 10, p. 1786–1788, 15 maio 2019. Disponível em: <https://academic.oup.com/bioinformatics/article/35/10/1786/5132693>. Acesso em: 18 ago. 2022.

ZHANG, Z. et al. The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in Duroc pigs. **Genomics**, v. 114, n. 1, p. 340–350, 1 jan. 2022.