Marcadores Moleculares na Era Genômica: Metodologias e Aplicações

ORGANIZADORAS

Andreia Carina Turchetto-Zolet

Caroline Turchetto

Camila Martini Zanella

Gisele Passaia

SBG Sociedade Brasileira de Genética

Marcadores Moleculares na Era genômica: Metodologias e Aplicações / Andreia Carina Turchetto-Zolet, Caroline Turchetto, Camila Martini Zanella e Gisele Passaia (organizadores). – Ribeirão Preto: Sociedade Brasileira de Genética, 2017.
181 p.

ISBN 978-85-89265-26-3

1. DNA. 2. Biologia molecular. 3. Genética. I. Turchetto-Zolet, Andreia Carina; Turchetto, Caroline; Zanella, Camila Martini; Passaia, Gisele, orgs.

Sociedade Brasileira de Genética

Rua Cap. Adelmio Norberto da Silva, 736
14025-670 - Ribeirão Preto - SP
16 3621-8540 | 16 3621-3552

Capítulo 8

Polimorfismo de Nucleotídeo único (SNP): metodologias de identificação, análise e aplicações

Dra. Andreia Carina Turchetto-Zolet, Dra. Caroline Turchetto, Dr. Frank Guzman, Dr. Gustavo Adolfo Silva-Arias, Dra. Fernanda Sperb-Ludwig, Msc. Nicole Moreira Veto

Considerações gerais

Polimorfismos de Nucleotídeo Único (SNPs - do inglês *Single Nucleotide Polymorphisms*) podem ser originados de mutações pontuais no DNA como as transições e transversões. As transições ocorrem entre trocas de bases purínicas (A/G) ou entre bases pirimidínicas (C/T); as transversões, onde há a troca entre bases purínicas por pirimidínicas (A/T, G/C, T/A e C/G). Alguns autores consideram *Indels* (adição de nucleotídeos extras ou a exclusão de um nucleotídeo) como SNPs, embora eles certamente ocorram por um mecanismo diferente (Kahl et al., 2005). Embora, em princípio, em cada posição da sequência de DNA seja possível ocorrer as quatro bases nucleotídicas, na prática os SNPs são geralmente considerados bialélicos. Uma das razões para isso é a baixa frequência de substituições de nucleotídeo único que originaram os SNPs, estimado estar entre 1×10^{-9} e 5×10^{-9} por nucleotídeo por geração nas posições neutras em mamíferos (Li et al., 1981; Martínez-Arias et al., 2001; Vignal et al., 2002). Dessa forma, a probabilidade de ocorrer duas mudanças idependentes da base nucleotidica em uma única posição é muito baixa. (Vignal et al., 2002). Por serem considerados bialélicos os SNPs são menos informativos por *locu*s examinado quando comparados a outros marcadores, como por exemplo, os microssatélites (SSRs – do inglês *Simple Sequence Repeats;* ver Capítulo 6). Entretanto, eles são abundantes e amplamente distribuídos nos genomas, podendo estar presentes em praticamente todos os *loci* gênicos, o que representa grande vantagem nas análises genéticas (Perkel, 2008).

Os SNPs são a classe mais abundante de variação genética encontrada em genomas eucarióticos, representando aproximadamento 90% do genoma humano (Brookes, 1999). Cerca de 15 milhões de SNPs já foram identificados em humanos pelo projeto 1000 genomas (Durbin et al., 2010; Mills et al., 2011). A densidade de SNPs pode variar substancialmente entre diferentes regiões de um genoma e entre diferentes espécies. A densidade de SNPs para humanos foi observada em 1.07 SNP / kb, enquanto para macaco (*Macaca mulatta*) a densidade de SNP foi calculada em 2.82 SNP / kb (Yuan et al., 2012). Em plantas a densidade de SNPs também é alta e pode variar entre espécies (Ching et al., 2002). Enquanto 0.64 SNP / kb foi encontrado em arroz (*Oryza sativa*) variedade Nipombare (Jeong et al., 2013) uma média de 6.1 SNP / kb foi observado em tomate (*Solanum lycopersicum*) (Kim et al., 2014).

Os SNPs são amplamente distribuídos no genoma e estão presentes em regiões codificadoras (éxons) e não codificadoras (íntrons e regiões intergênicas). Neste aspecto, para compreender o impacto da presença de um SNP nas regiões codificadoras é importante relembrar o conceito de mutações sinônimas e não sinônimas. As mutações sinônimas não alteram o aminoácido traduzido enquanto que as mutações não sinônimas

resultam na alteração da composição de aminoácidos, ausência ou modificações do produto proteico. Desta forma, os SNPs podem ter diferentes classificações, associadas: (1) a sua localização no genoma (éxons, íntrons ou espaçadores intergênicos) e; (2) ao impacto da sua presença dentro de regiões codificadoras ou reguladoras para o produto proteico e/ou o fenótipo (Kahl et al., 2005) (Figura 8.1).

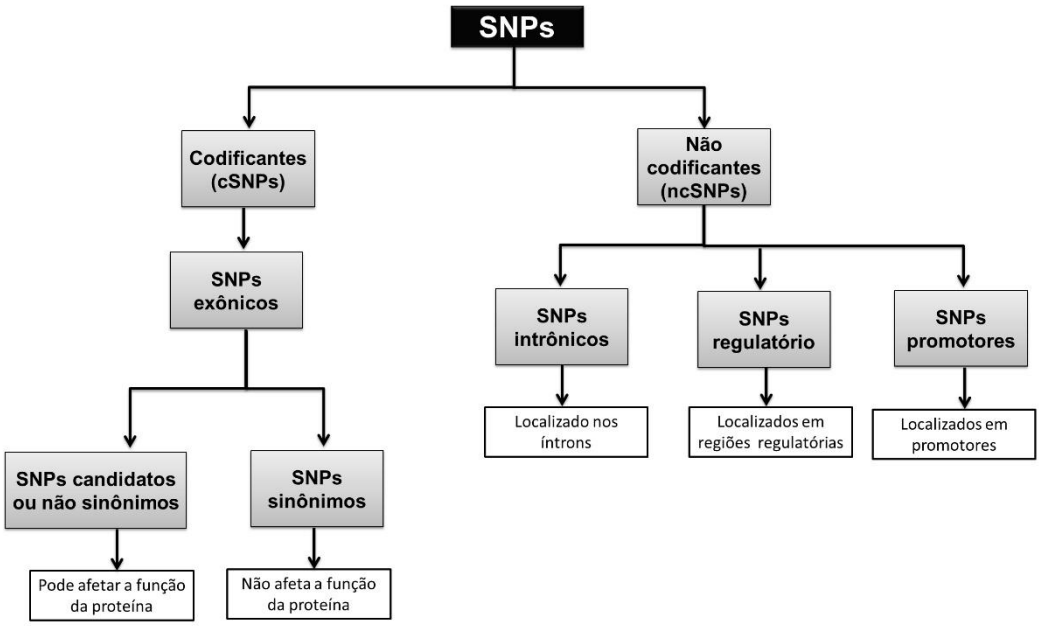

Figura 8.1 - Classificação dos SNPs quanto à localização no genoma e quanto ao impacto causado na proteína ou fenótipo.

A frequência de SNPs é geralmente maior em regiões não codificadoras do que em regiões codificadoras. Fatores tais como a taxa de mutação, recombinação genética e seleção natural podem influenciar a densidade de SNPs (Nachman, 2001; Barreiro et al., 2008). SNPs em regiões não codificadoras são chamados de SNPs não codificantes (ncSNPs), e os ncSNPs localizados dentro de íntrons são chamados de SNPs intrônicos. Já os SNPs encontrados em regiões codificadoras são chamados de SNPs codificadores (cSNPs), como por exemplo, em éxons (SNPs exônicos). Qualquer SNP em um éxon de um gene que pode ter impacto sobre a função da proteína codificada é chamado de SNP candidato, pelo fato de poder estar associado a alguma caraterística fenotípica. Outros ocorrem em regiões promotoras ou em regiões regulatórias do genoma e são chamados SNPs reguladores e SNPs promotores (pSNPs), respectivamente. Um SNP promotor pode influenciar drasticamente a atividade do gene dirigido por este promotor, por exemplo, um pSNP pode impedir a ligação de um fator de transcrição na sua sequência de reconhecimento, alterando a expressão do gene. Os ncSNPs são bastante utilizados para estudos de associação e mapeamento de desequilíbrio de ligação de todo o genoma, além de estudos evolutivos. O aparecimento de um SNP em um éxon pode ser totalmente neutro, ou seja, não altera a composição de aminoácidos do domínio ou da proteína codificada e, por isso não apresenta qualquer efeito sobre a sua função. Nestes casos, o SNP é chamado de SNP sinônimo (sSNP). Por outro lado, um SNP não

sinônimo (nsSNP) irá alterar o aminoácido codificado, podendo alterar a função da proteína correspondente. Apesar dos SNPs resultantes de mutações sinônimas, não modificarem a composição de aminoácidos, eles podem acometer o dobramento de proteínas, afetando o posicionamento de seus domínios de ligação, por exemplo (Figura 8.1) (Kahl et al., 2005).

Descobertos primeiramente no genoma humano, os SNPs provaram ser universais, sendo as formas mais abundantes de variação intraespecífica. Estudos têm mostrado que os SNPs podem ter efeitos biológicos importantes, tais como a associação com doenças complexas e reações e respostas à tratamentos em humanos. Os SNPs podem ser utilizados como marcadores moleculares em diversas áreas de estudo, tais como estudos evolutivos, filogenéticos, ecológicos, no melhoramento genético animal e vegetal, no mapeamento genético. Além de ser uma importante ferramenta para diferentes áreas que envolvem a análise genética do DNA humano, como, por exemplo, no diagnóstico e tratamento de doenças, em estudos antropológicos, e na identificação humana (análises forenses ou na determinação de paternidade).

Até pouco tempo atrás o uso dos marcadores SNPs era restrito a organismos modelo com genomas sequenciados, devido ao elevado custo de descoberta e genotipagem. Atualmente, com o avanço das ferramentas de bioinformática e o barateamento no sequenciamento, tem-se expandido o uso dos marcadores SNPs para espécies não modelo. Além disso, diversas metodologias para identificação e genotipagem destes marcadores já foram estabelecidas. Muitas dessas metodologias permitem a identificação e genotipagem de SNPs em uma única etapa, o que proporciona associar rapidez na obtenção dos dados e baixo custo. As tecnologias de Sequenciamento de Nova Geração (NGS- do inglês *Next Generation Sequencing*) (ver Capítulo 2 para detalhes) associadas às ferramentas computacionais existentes são altamente eficientes e robustas na descoberta de SNPs sem um genoma de referência.

Mais recentemente com o uso das plataformas de NGS, foram implementadas tecnologias que garantem a descoberta e genotipagem de variantes em um único passo, como, por exemplo, as técnicas que envolvem a redução genômica, RNA-seq e captura de sequências. Além disso, os dados de acesso à informações genômicas das espécies são um excelente recurso para o processo de procura e identificação de marcadores SNPs em genes candidatos ou espalhados pelo genoma (Nielsen et al., 2011; Kumar et al., 2012).

Neste capítulo mostraremos as principais metodologias de identificação, genotipagem e análise de SNPs, bem como as principais aplicações destes marcadores em diferentes áreas e organismos. Daremos maior enfoque às metodologias que utilizam tecnologias de NGS.

Metodologia de Identificação e Genotipagem

O estudo de marcadores SNPs basicamente envolve duas etapas principais: a identificação (descoberta) dos SNPs no genoma da espécie de interesse e a genotipagem destes marcadores na população desta espécie ou no indivíduo de interesse para posterior análise. A diferença entre essas duas etapas é que na descoberta dos SNPs pode ser utilizado um número pequeno (representativo) de indivíduos da espécie estudada enquanto na genotipagem é utilizado um número maior de indivíduos, os quais representem uma ou mais populações, dependendo do objetivo do estudo. Tanto a identificação quanto a genotipagem de marcadores SNPs pode ser realizada utilizando metodologias de pequena ou larga escala.

O procedimento de identificação dos SNPs pode ser realizado através de metodologias tais como o sequenciamento de produtos de PCR; a identificação eletrônica de SNPs (eSNP) utilizando como base, por exemplo, bibliotecas de EST (*expressed sequence tags*) ou bibliotecas genômicas (Picoult-Newberg et al., 1999; Panitz et al., 2007; van Oeveren e Janssen, 2009) disponíveis para a espécie em estudo. Nos últimos anos, o sequenciamento de alto rendimento também vem sendo utilizado para a identificação de SNPs em genomas inteiros ou trasnscritomas, por exemplo (Barbazuk et al., 2007; De Wit 2016, Boutet et al., 2016).

A genotipagem pode envolver diferentes categorias de métodos e técnicas, onde podemos destacar os métodos baseados em hibridização, como a hibridização alelo específica (Saiki et al., 1986; Howell et al., 1999; Prince et al., 2001), hibridização de sondas (eg. Sistema TaqMan por PCR em Tempo Real - McGuigan e, 2002) e hibridização em arranjos (SNP *array*) (Hehir-Kwa et al., 2007); métodos de ligação de oligonucleotideo baseado em PCR (Newton et al., 1991; Drenkard et al., 2000; Macdonald, 2007; Podder et al., 2008) e métodos baseados em NGS (van Orsouw et al., 2007; Baird et al., 2008; Torkamaneh et al., 2016).

Antes do advento das tecnologias de NGS, as etapas de identificação e genotipagem de SNPs eram sempre realizadas separadamente. Agora elas podem ser realizadas concomitantemente. Existem diversas abordagens que permitem realizar as duas etapas em um único passo e algumas delas serão abordadas no tópico seguinte.

Identificação e genotipagem usando tecnologias baseadas em NGS

As plataformas atuais de sequenciamento em larga escala (conforme Capítulo 2) permitem a descoberta de centenas ou milhares de SNPs que cobrem todo o genoma ou grande parte dele em um único experimento. Uma grande vantagem da utilização das plataformas de NGS é que possibilitou, através de diferentes abordagens, a descoberta e genotipagem de milhares de marcadores SNPs em um único passo, sendo possível a utilização em qualquer espécie de interesse, incluindo aquelas com pouca ou nenhuma informação genética previa disponível (Stapley et al., 2010). Esse aumento da eficiência e os benefícios de baixo custo foram realizados através da incorporação de uma estratégia de sequenciamento multiplex que usa um sistema de código de barras relativamente barato.

Dentre os métodos de genotipagem baseados em NGS muitos tem como componente principal o uso de enzimas de restrição específicas para reduzir a complexidade genômica do organismo de interesse; enquanto outros utilizam iscas de oligonucleotídeos de regiões conhecidas para ligar nas regiões de interesse (captura de sequências); ainda outros utilizam o sequenciamento de genes candidatos através do uso de oligonucleotídeos específicos para amplificar as regiões de interesse, bem como o sequenciamento completo de RNA de determinado tecido ou condição experimental (RNA-seq).

Os métodos de redução do genoma foram desenvolvidos como abordagens rápidas e robustas que combinam a descoberta de marcadores moleculares e a genotipagem dos mesmos simultaneamente. Além disso, estas técnicas permitem o sequenciamento simultâneo de vários indivíduos devido a combinação de adaptadores de código de barras de DNA em cada amostra. Isto permite posteiormente identificar e agrupar as sequências de acordo com o indivíduo. O método de redução genômica foi descrito pela primeira vez em humanos usando sequenciamento capilar para gerar um mapa de SNPs (Altshuler et al. 2000). Posteriormente, van Tassel et al., (2007)

adaptaram a técnica para o NGS: sequenciamento de bibliotecas de representação reduzida (RRLs, do inglês *reduced-representation libraries*). Outras técnicas que utilizam abordagem de redução genômica já foram descritos na literatura associadas ao NGS. Dentre esses métodos, podemos destacar o sequenciamento de fragmentos de DNA associados a sítios de restrição (*RAD-seq*, do inglês *restriction site-associated DNA sequencing*) (Miller et al., 2007; Baired et al., 2008) e a genotipagem por sequenciamento (GBS, do inglês *Genotyping by sequencing*) (Davey et al., 2011b), além de outras. Atualmente, estas técnicas estão sendo empregadas para uma gama de estudos genéticos e genômicos em diversas espécies, tais como em estudo de estrutura de populações em uma espécie arbustiva da Amazônia da família Violacea (Nazareno et al., 2017); descoberta de SNPs para construção de mapas genéticos em oliva (Ipek et al., 2016), em estudo de hibridação em espécies de peixe do gênero *Potamotrygon* do rio Paraná (Cruz et al., 2017) e para estudar a diversidade genética do mosquito *Anopheles moucheti*, vetor da malária (Fouet et al., 2017), além de diversos outros estudos que podem ser encontrados na literatura.

Embora cada método baseado em enzimas de restrição tenha suas particularidades no processo de preparo da biblioteca para o sequenciamento, eles compartilham um número de etapas comuns: Extração, quantificação e qualidade do DNA genômico; fragmentação do DNA genômico de todas as amostras com enzimas de restrição e ligação dos adaptadores; amplificação por PCR (RAD-seq e GBS) e seleção de tamanho dos fragmentos (RRL e RAD-seq); sequenciamento e análise das sequências com ou sem suporte de um genoma de referência, e identificação dos SNPs (Davey et al., 2011b; Poland and Rife, 2012).

Basicamente, os diferentes protocolos iniciam com a digestão do DNA com uma ou mais enzimas de restrição. Quando essas amostras são digeridas, diferentes tamanhos de fragmentos são gerados de acordo com a presença ou não do sítio de reconhecimento da(s) enzima(s) de restrição utilizada(s). A Figura 8.2, bem como os passos descritos a seguir mostram resumidamente as principais etapas das técnicas para construção de bibliotecas genômicas usadas nos métodos de RRL, GBS e RAD-seq. (1) **RRL** – todos os fragmentos de todas as amostras são agrupados num único pool, é realizada uma seleção de tamanho de fragmento (300-700pb) e em seguida a ligação do adaptador padrão de acordo com a plataforma de sequenciamento. Esta metodologia permite a detecção de polimorfismos dentro de uma população, mas não para cada indivíduo; (2) **RAD-seq** - Os fragmentos de cada amostra são ligados a adaptadores P1, posteriormente todos os fragmentos são agrupados, cortados aleatoriamente e selecionado por tamanho de fragmento (300-700pb), esta seleção pode ser realizada pelo corte diretamente do gel e purificação. Posteriormente são ligados adaptadores P2 com final divergente em todos os fragmentos com e sem adaptadores P1. Os fragmentos são amplificados por PCR com oligonucleotídeos específicos para P1 e P2, o que significa que apenas fragmentos com adaptadores P1 e P2 são amplificados, ou seja, os fragmentos que contém os sítios de restrição; (3) **GBS** – após a disgestão do DNA de cada amostra são ligados aos fragmentos de DNA adaptadores com código de barras e adaptadores comuns, produzindo fragmentos com três diferentes combinações de adaptadores: código de barras + comum, código de barras + código de barras e comum + comum. As amostras são agrupadas e amplificadas, e nesta etapa, apenas amostras curtas (<1 kb) são amplificadas com a combinação código de barras + adaptador comum e após são sequenciados (Davey et al. 2011).

Os SNPs encontrados nos fragmentos sequenciados podem ser utilizados como marcadores genéticos. Com a utilização do sequenciamento *paired-end*

(sequenciamento de ambas as extremidades do fragmento) na técnica de RAD-seq é possível montar para cada *locus* em um longo *contig* (conjunto de segmentos de DNA sobrepostos que juntos representam um consenso de uma região do DNA) com um comprimento médio de ~ 500 bases (Etter et al., 2011). Este *contig*, com cobertura suficiente, pode ser usado para identificar SNPs ao longo de todo o fragmento.

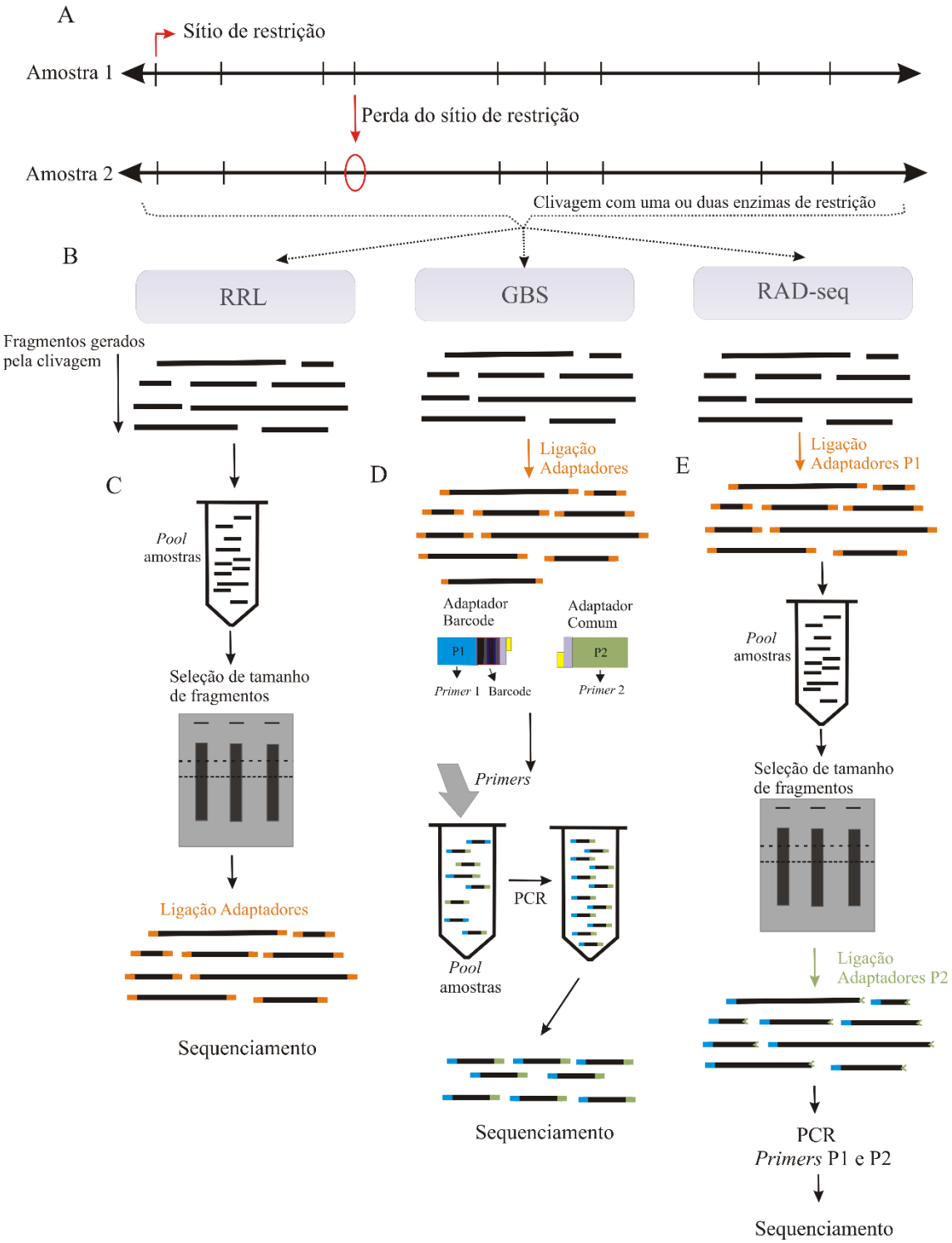

Figura 8.2 - (A, B) Visão geral do processo de clivagem por enzimas de restrição em uma região genômica de dois organismos; etapa comum aos três métodos (RRL, GBS e Rad-seq). A perda de sítios de restrição resulta na variação do número e tamanho dos fragmentos gerados com a clivagem. (C) Medotologia RRL: Após a clivagem os fragmentos das amostras são agrupados em um único pool

(mistura das amostras), em seguida é realizada a seleção do tamanho dos fragmentos, posterior ligação de adaptadores e sequenciamento. (D) Metodologia de GBS: após a clivagem do DNA as amostras individuais com uma enzima de restrição são ligados adaptadores com barcorde e comuns. Serão gerados fragmentos que terão a combinação de adaptadores, barcode + barcode, barcode + comum e comum + comum. Após a ligação dos adaptadores as amostras são agrupadas em um único pool e é realizado PCR com oligonucleotídeos P1 e P2, assim apenas os fragmentos com a combinação barcode + comum serão amplificados, selecionado assim fragmentos menores. Os fragmentos amplificados são sequenciados. (E) Metodologia original RAD-seq: após a clivagem são ligados adaptadores P1 aos fragmentos e as amostras são agrupadas em um único pool. É realizada a seleção do tamanho de fragmentos (geralmente entre 300-700 pb), em seguida são ligados adaptadores P2. O PCR é realizado com oligonucleotídeos específicos P1 e P2.

Os métodos que envolvem a captura de sequências são baseados no enriquecimento de um alvo (Mamanova et al., 2010) através da hibridização de "iscas" de DNA fita simples ou RNA (também chamadas sondas) a determinadas regiões do genoma, selecionando fisicamente estas regiões, e eliminando fragmentos de DNA indesejáveis, permitindo que os alvos sejam posteriormente sequenciados (Kandpal et al., 1994; Albert et al., 2007; Gnirke et al., 2009; Glenn e Faircloth, 2016). Por exemplo, regiões associadas com uma particular doença ou característica podem ser capturadas (Teer et al., 2010). A captura de sequências é uma tecnologia de DNA relativamente antiga, com início na década de 1990, quando muitos laboratórios estavam desenvolvendo métodos de identificação de regiões microssatélites de DNA (Tautz, 1989; Ellegren 2004) (ver Capítulo 6 para mais informações sobre microssatélites). Logo após a disponibilidade dos métodos de sequenciamento de nova geracão, a história de captura de DNA com sondas sintéticas foi recapitulada. Pesquisadores demonstraram que as sondas poderiam ser milhares de *oligonucleotídeos* sintetizados em microarranjos (Albert et al., 2007; Hodges et al., 2007; Porreca et al., 2007).

As abordagens de sequenciamento de genes candidatos também podem ser de particular interesse. Estas abordagens permitem que sejam analisados SNPs diretamente em gene com uma função conhecida relacionada a um processo particular, uma via metabólica, ou mesmo com um fenótipo, ou estarem sob seleção (Tabor et al., 2002; Cousin et al., 2003). As sequências dos genes candidatos podem ser obtidas através da metodologia de captura de sequências, em bancos de ESTs, ou também a partir do transcritoma ou genoma disponível para a espécie de interesse. Estas sequências são usadas para a projeção de oligonucleotídeos e posterior amplificação dos genes seguido do sequenciamento em plataforma de NGS (Hendre et al., 2012).

Embora muitas vezes usado para medir a expressão gênica, o sequenciamento de RNA (RNA-seq) em larga escala também tem sido bastante utilizado para a descoberta e genotipagem de marcadores SNPs. RNA-seq já foi usado para descobrir dezenas a centenas de milhares de SNPs em diferentes espécies modelos e não modelos. Isto pode ser feito a custos semelhantes aos métodos baseados em enzimas de restrição, sendo mais provável detectar SNPs relacionados com o fenótipo.

Com o avanço do NGS para produzir milhões de *reads* por corrida, a análise de dados para estas novas abordagens pode ser complexa nas metodologias baseadas em enzimas de restrição, multiplexação de amostras e comprimento de fragmento diferente. Por isso, fica evidente a necessidade do desenvolvimento de *pipelines* (fluxogramas de trabalho) avançados para filtrar, classificar e alinhar estas sequências. Citaremos como exemplo as etapas para analise de dados de GBS: um *pipeline* para GBS deve incluir etapas para limpar as *reads* de 'contaminação' de sequências de adaptadores e *barcodes*, filtrar *reads* de baixa qualidade, classificá-las por *pools* ou indivíduos com base no

código de barras da sequência, identificar lócus e alelos *de novo* ou alinhar as *reads* a um genoma de referência para descobrir polimorfismos e frequentemente determinar genótipos para cada indivíduo incluído no estudo. Geralmente, as pipelines para o tratamento de dados oriundos de uma abordagem por GBS são categorizadas em dois grupos: as baseadas em montagem *de novo* e as baseadas em um genoma de referência. Quando um genoma de referência está disponível, as *reads* do sequenciamento de redução genômica podem ser mapeados no genoma e os SNPs são identificados e genotipados. Alguns pipelines para GBS baseados em um genoma de referência estão disponíveis, tais como TASSEL-GBS (v1 e v2), Stacks, IGST, e Fast-GBS. Já na ausência de um genoma de referência, os pares de *reads* quase idênticas (presumidas para representar alelos alternativos de um *locus*) precisam ser identificados. Os pipelines mais usados nesse caso são UNEAK e Stacks (Davey et al., 2011; Torkamaneh et al., 2016).

Existem diversos outros programas para construir *pipelines* que podem ser usados para análise de dados provenientes de NGS e mineração de SNPs. Descrições de diversos desses programas podem ser encontrados em Altmann et al. (2012). Dentre eles destacamos o Pacote de programas SAMtools (Li et al., 2009), que é distribuído sob a licença MIT *open source*, livre para usos acadêmicos e comerciais e o GATK - *Genome Analysis Toolkit* (McKenna et al., 2010; DePristo et al., 2011). Um passo a passo do uso dos programas SAMtools e GATK para análise de dados provenientes de NGS está descrito a seguir.

No passo a passo abaixo está mostrado como realizar análise de dados provenientes de sequenciamento NGS, identificação e genotipagem de SNPs.

1. Instalação dos programas requeridos

Observação: Os seguintes programas foram instalados e testados em um sistema Ubuntu 16.04 (Xenial Xerus)

- **BWA**

Passo 1: Executar os seguintes comandos em um terminal:

```
$ sudo apt-get update
$ sudo apt-get install bwa
```

- **SAMtools**

Passo 1: Executar os seguintes comandos em um terminal:

```
$ sudo apt-get update
$ sudo apt-get install samtools
```

- **BCFtools**

Passo 1: Executar os seguintes comandos em um terminal:

```
$ sudo apt-get update
$ sudo apt-get install bcftools
```

- **Genome Analysis Tookit (GATK)**

Passo 1: Ir ao endereço https://software.broadinstitute.org/gatk/download/ e baixar a ultima versão do programa que está comprimido em um arquivo de extensão .bz2

Passo 2: Descomprimir o arquivo .bz2 usando o seguinte comando em um terminal:

```
$ tar xjf GenomeAnalysisTK-3.6-0.tar.bz2
```

O comando prévio vai gerar a pasta GenomeAnalysisTK-3.6-0 e vai conter o pré-compilado de Java executável GenomeAnalysisTK.jar.

Passo 3: Copiar e colar o executável GenomeAnalysisTK.jar na pasta onde se vão realizar as análises.

- **picard**

Passo 1: Ir ao endereço https://github.com/broadinstitute/picard/releases/ e baixar a ultima versão do programa que esta comprimido em um arquivo de extensão .zip

Passo 2: Descomprimir o arquivo .zip usando o seguinte comando em um terminal:

```
$ tar xjf picard-tools-2.4.1.zip
```

O comando prévio vai gerar a pasta picard-tools-2.5.0 e vai conter três pré-compilados de Java executáveis: picard.jar, picard-lib.jar e htsjdk-2.5.0-SNAPSHOT-all.jar.

Passo 3: Copiar e colar os três executáveis de extensão .jar na pasta onde serão realizadas as análises.

- **VCFtools**

Passo 1: Executar os seguintes comandos em um terminal:

```
$ sudo apt-get update
$ sudo apt-get install vcftools
```

- **vcflib**

Passo 1: Executar o seguinte comando em um terminal para baixar a fonte do programa:

```
$ git clone --recursive https://github.com/vcflib/vcflib.git
```

Passo 2: Entrar na pasta vcflib e executar o seguinte comando para compilar os executáveis do programa:

```
make
```

Passo 3: O comando prévio vai gerar a pasta de nome bin e vai conter diferentes compilados executáveis. Em seguida, executar os comandos abaixo para disponibilizar o executável vcffilter em todo o sistema operativo:

```
$ sudo cp vcffilter /usr/bin
$ cd /usr/bin
$ sudo chmod 775 vcffilter
```

- **Descrição dos dados que serão utilizados:** Nas próximas análises serão utilizados dados do sequenciamento do tipo *paired* e *single end* de 16 genes em dois indivíduos de *Arabidopis thaliana* obtidos com a tecnologia MiSeq da Illumina.

Sequência nucleotídica dos 16 genes em formato fasta: Athaliana_seqs.fasta

Bibliotecas *paired end* do indivíduo A1: A1_R1_paired.fastq
A1_R2_paired.fastq

Biblioteca *single end* do indivíduo A1: A1_single.fastq

Bibliotecas *paired end* do indivíduo B1: B1_R1_paired.fastq
B1_R2_paired.fastq

Biblioteca *single end* do indivíduo B1: B1_single.fastq

Observação: A etapa de identificar e excluir sequências dos adaptadores foi realizada previamente. Além disso, também foi realizado um *trimming* para excluir as bases da extremidade 3' de cada *read* com qualidade baixa. É muito importante fazer esses dois tipos de *trimming* antes de todo procedimento de identificação de SNPs.

2. Identificando SNPs com GATK

- **Alinhamento dos *reads* na referência**

Passo 1: Executar o seguinte comando em um terminal para criar o índex de bwa da referência:

```
$ bwa index Athaliana_seqs.fasta
```

Observação 1: O comando prévio vai gerar diferentes arquivos (índices) que permitirão um fácil aceso da referência pelo bwa.

Passo 2: Se proceder ao alinhamento dos *reads* da biblioteca *paired-end* para cada indivíduo:

No caso do indivíduo A1:

```
$ bwa aln -t 4 –f A1_R1_paired.sai Athaliana_seqs.fasta A1_R1_paired.fastq
$ bwa aln -t 4 –f A1_R2_paired.sai Athaliana_seqs.fasta A1_R2_paired.fastq
```

No caso do indivíduo B1:

```
$ bwa aln -t 4 –f B1_R1_paired.sai Athaliana_seqs.fasta B1_R1_paired.fastq
$ bwa aln -t 4 –f B1_R2_paired.sai Athaliana_seqs.fasta B1_R2_paired.fastq
```

Observação 1: No caso das bibliotecas *paired-end*, os *reads* das extremidades R1 e R2 são alinhados separadamente usando o módulo aln do bwa.

Observação 2: O parâmetro -t indica o número de processos do computador que serão usados para realizar o alinhamento e o parâmetro -f vai especificar o nome do arquivo de alinhamento de extensão .sai.

Passo 3: O seguinte comando do programa bwa permitirá agrupar cada alinhamento separado do R1 e R2 em um arquivo de alinhamento final em formato .sam:

No caso do indivíduo A1:

```
$ bwa sampe -r "@RG\tID:A1\tSM:A1" -f A1_paired.sam Athaliana_seqs.fasta A1_R1.sai A1_R2.sai A1_R1.fastq A1_R2.fastq
```

No caso do indivíduo B1:

```
$ bwa sampe -r "@RG\tID:B1\tSM:B1" -f B1_paired.sam Athaliana_seqs.fasta B1_R1.sai B1_R2.sai B1_R1.fastq B1_R2.fastq
```

Observação 1: O parâmetro -r permitirá adicionar o grupo de *reads* (RG) e o ID de cada amostra para cada um dos *reads*.

Observação 2: O parâmetro -f vai especificar o nome do arquivo de alinhamento de extensão .sam.

Observação 3: No arquivo de extensão .sam só estarão incluídos os *reads* R1 e R2 que ancoraram conjuntamente na referência

Passo 4: No caso das bibliotecas *single-end* os comandos para realizar o alinhamento e obtenção do arquivo .sam são os seguintes:

No caso do indivíduo A1:

```
$ bwa aln -t 4 –f A1_single.sai Athaliana_seqs.fasta A1_single.fastq
```

```
$ bwa samse -r "@RG\tID:A1\tSM:A1" -f A1_single.sam Athaliana_seqs.fasta A1_single.sai A1_single.fastq
```

No caso do indivíduo B1:

```
$ bwa aln -t 4 –f B1_single.sai Athaliana_seqs.fasta B1_single.fastq
```

```
$ bwa samse –r "@RG\tID:B1\tSM:B1" –f B1_single.sam Athaliana_seqs.fasta B1_single.sai B1_single.fastq
```

Passo 5: Para transformar os arquivos de extensão .sam em .bam utilizando o programa SAMtools executamos o seguinte comando no terminal:

No caso do indivíduo A1:

```
$ samtools view -bS –o A1_paired.bam A1_paired.sam
$ samtools view -bS –o A1_single.bam A1_single.sam
```

No caso do indivíduo B1:

```
$ samtools view -bS –o B1_paired.bam B1_paired.sam
$ samtools view -bS –o B1_single.bam B1_single.sam
```

Observação 1: O parâmetro –bS indica que o input é um arquivo de extensão .sam e o output será um arquivo de extensão .bam.

Passo 6: Agora se utilizará o modulo sort de SAMtools para classificar os *reads* em cada arquivo de extensão .bam de acordo com as coordenadas de ancoramento na referência:

No caso do indivíduo A1:

```
$ samtools sort A1_paired.bam A1_paired_sorted
$ samtools sort A1_single.bam A1_single_sorted
```

No caso do indivíduo B1:

```
$ samtools sort B1_paired.bam B1_paired_sorted
$ samtools sort B1_single.bam B1_single_sorted
```

Observação 1: Sort é um módulo de SAMtools e vai adicionar automaticamente a extensão .bam ao *output* pelo que não será necessário especificar essa extensão ao momento de executar o comando.

Passo 7: Finalmente se juntaram os dois arquivos classificados (*paired* e *single*) de extensão .bam em um só arquivo .bam:

No caso do indivíduo A1:

```
$ samtools merge A1_merged.bam A1_paired_sorted.bam A1_single_sorted.bam
```

No caso do indivíduo B1:

```
$ samtools merge B1_merged.bam B1_paired_sorted.bam B1_single_sorted.bam
```

- **Identificação de SNPs**

Passo 1: Os arquivos .bam gerados na etapa anterior contém os dados de informação de todos os *reads* da biblioteca que alinharam e não alinharam na referência. Neste último grupo de *reads*, existe um problema de compatibilidade entre eles e os diferentes módulos do programa picard, que é solucionado com o seguinte comando:

```
$ java -jar picard.jar CleanSam I=A1_merged.bam O=A1_merged_clean.bam
```

```
$ java -jar picard.jar CleanSam I=B1_merged.bam O=B1_merged_clean.bam
```

Observação 1: CleanSam é um módulo de picard, e os parâmetros I e O são os arquivos input e output em formato .bam, respectivamente.

Passo 2: Depois de solucionar o problema de compatibilidade se procederá em marcar e remover os *reads* duplicados que foram gerados a partir do mesmo fragmento utilizando o seguinte comando:

```
$ java -jar picard.jar MarkDuplicates VALIDATION_STRINGENCY=LENIENT
AS=true REMOVE_DUPLICATES=true I=A1_merged_clean.bam O=A1_markdup.bam
M=A1_markdup.metrics
```

```
$ java -jar picard.jar MarkDuplicates VALIDATION_STRINGENCY=LENIENT
AS=true REMOVE_DUPLICATES=true I=B1_merged_clean.bam O=B1_markdup.bam
M=B1_markdup.metrics
```

Observação 1: O arquivo gerado em formato .metrics vai conter as estatísticas do alinhamento do arquivo .bam.

Passo 2: Se procederá a possibilidade de adicionar informações sobre o nome das amostras e o tipo de sequenciamento ao arquivo .bam gerado previamente:

```
$ java -jar picard.jar AddOrReplaceReadGroups
VALIDATION_STRINGENCY=LENIENT I=A1_markdup.bam O=A1_rg.bam RGID=A1
RGLB=A1 RGPL=illumina RGPU=run RGSM=A1
```

```
$ java -jar picard.jar AddOrReplaceReadGroups
VALIDATION_STRINGENCY=LENIENT I=B1_markdup.bam O=B1_rg.bam RGID=B1
RGLB=B1 RGPL=illumina RGPU=run RGSM=B1
```

Passo 3: Para usar os arquivos .bam no programa GATK, se procederá como fazer um índex do mesmo com o SAMtools:

```
$ samtools index A1_rg.bam
```

```
$ samtools index B1_rg.bam
```

Passo 4: A primeira etapa no programa GATK é realizar o re-alinhamento local em torno dos *indels* para corrigir possíveis erros de alinhamento. Neste passo serão identificados sítios onde existe um *indel* verdadeiro:

```
$ java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -nt 4 -R
Athaliana_seqs.fasta –I A1_rg.bam --out A1.intervals
```

```
$ java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -nt 4 -R
Athaliana_seqs.fasta –I B1_rg.bam --out B1.intervals
```

Observação 1: O parâmetro -T é o tipo de análise do GATK, -nt é o número de processadores do computador, -R é a referência em formato .fasta, -I é o arquivo .bam criado previamente e --out é o output que vai conter os sítios re-alinhados.

Passo 5: Agora fazemos os realinhamento dos *reads* usando o arquivo *.intervals* gerado no passo 4:

```
$ java -jar GenomeAnalysisTK.jar -T IndelRealigner -R Athaliana_seqs.fasta –I A1_rg.bam –targetIntervals A1.intervals –o A1_realigned.bam
```

```
$ java -jar GenomeAnalysisTK.jar -T IndelRealigner -R Athaliana_seqs.fasta –I B1_rg.bam –targetIntervals B1.intervals –o B1_realigned.bam
```

Passo 6: Criar um index do arquivo .bam resultante com SAMtools:

```
$ samtools index A1_realigned.bam
```

```
$ samtools index B1_realigned.bam
```

Passo 7: Os arquivos .bam re-alinhados serão usados para identificar as variações existentes em cada amostra usando o HaplotypeCaller. Através desta análise, o GATK primeiro identifica regiões de interesse, determina haplótipos por re-montagem local das regiões, determina a probabilidade dos genótipos e designa genótipos para cada amostra.

```
$ java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R Athaliana_seqs.fasta -stand_emit_conf 10 -stand_call_conf 30 -ERC GVCF –I A1_realigned.bam –o A1.gvcf
```

```
$ java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R Athaliana_seqs.fasta -stand_emit_conf 10 -stand_call_conf 30 -ERC GVCF –I B1_realigned.bam –o B1.gvcf
```

Observação 1: O parâmetro -stand_emit_conf é o limite de confiança mínimo (em escala Phred) no qual o GATK reporta sítios que parecem ser possivelmente variáveis, -stand_call_conf é o limite de confiança mínimo (em escala Phred) no qual o GATK identifica sítios variáveis e -ERC permite especificar o tipo de formato do output.

Passo 8: Os arquivos .gvcf de cada amostra no passo anterior serão concatenados em um único arquivo ".vcf" que tem agregado as probabilidades dos genótipos de todas as amostras:

```
$ java -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -R Athaliana_seqs.fasta --stand_emit_conf 10 -stand_call_conf 30 --variant A1.gvcf --variant B1.gvcf -o Arabidopsis.vcf
```

Passo 9: Para visualizar o arquivo .vcf criado executamos o seguinte comando no terminal:

```
$ less Athaliana.vcf
```

Observação 1: O vcf criado reporta os SNPs e INDELs.

Passo 10: Com o seguinte comando vamos selecionar apenas SNPs e subsequentemente criar um novo arquivo .vcf:

```
$ java -jar GenomeAnalysisTK.jar -T SelectVariants -R Athaliana_seqs.fasta --variant Athaliana.vcf -selectType SNP -o Athaliana.snps.vcf
```

Passo 11: Ao novo arquivo .vcf criado aplicamos diferentes parâmetros para filtrar os SNPs e manter aqueles com maior confiabilidade de serem verdadeiros:

```
$ java -jar GenomeAnalysisTK.jar -T VariantFiltration -R Athaliana_seqs.fasta --variant Athaliana.snps.vcf --filterName "snpsfilter" --filterExpression "QD<2.0||MQ<40.0||FS>60.0||HaplotypeScore>13.0||MQRankSum<12.5||ReadPosRankSum<-8.0" --out Athaliana.snps.tagged.vcf
```

Observação 1: Este comando ira marcar com snpsfilter aqueles SNPs que não cumpriram com os requerimentos de filtragem e PASS aqueles que cumpriram. A descrição detalhada dos parâmetros utilizados e recomendados na filtragem pode ser consultada em https://software.broadinstitute.org/gatk/documentation/tooldocs/org_broadinstitute_gatk_tools_walkers_filters_VariantFiltration.php

Passo 12: Finalmente selecionamos os SNPs que passaram nos critérios de filtragem e criamos um novo arquivo .vcf:

```
$ java -jar GenomeAnalysisTK.jar -T SelectVariants -R Athaliana_seqs.fasta --variant Euniflora.snps.tagged.vcf -select 'vc.isNotFiltered()' -o Euniflora.snps.filtered.vcf
```

3. Identificando SNPs com SAMtools

- **Alinhamento dos *reads* na referência**
  Esta etapa da análise é a mesma que foi mostrada no caso do GATK. No caso de não ter realizado a identificação de SNPs com GATK é necessário repetir os passos 1 ao 7.

- **Identificação de SNPs**

Passo 1: Os arquivos .bam gerados na etapa anterior serão utilizados para identificar os SNPs e INDELs presentes nas amostras usando o seguinte comando:

```
$ samtools mpileup -D –u -f Athaliana_seqs.fasta A1_sorted.bam B2_sorted.bam | bcftools view -vcg - > Athaliana_candidates.vcf
```

Observação 1: Na primeira parte do comando, o mpileup do SAMtools calcula as probabilidades dos genótipos nas amostras, e na segunda parte o *output* dessa análise é processado pelo bcftools para identificar os SNPs e INDELs, baseado nas probabilidades identificadas inicialmente. O parâmetro -D indica ao programa para manter a cobertura em cada amostra no output, -u indica a geração de um arquivo .bcf não comprimido, -f é para indicar o arquivo .fasta da referência. O parâmetro -vcg indica ao bcftools identificar potenciais sítios variáveis, identificar SNPs e INDELs e identificar os genótipos de cada amostra, respectivamente.

Passo 2: Para visualizar o arquivo .vcf criado executamos o seguinte comando no terminal:

```
$ less Athaliana_candidates.vcf
```

Observação 1: O vcf criado reporta os SNPs e INDELs.

Passo 3: Com o seguinte comando do programa VCFtools vamos selecionar somente SNPs e criar um novo arquivo .vcf:

```
$ vcftools --vcf SNP_candidates_all.vcf --remove-indels --out Athaliana_candidates_snp.vcf --recode
```

Passo 4: No arquivo .vcf de SNPs obtido, filtraram-se os genótipos com uma cobertura menor de 20 *reads* utilizando o programa vcflib:

```
$ vcffilter -t --keep-info -g "DP > 20" Athaliana_candidates_snp.vcf > Athaliana_candidates_snp_dp20.vcf
```

Observação 1: O parâmetro -g especifica a característica do genótipo por filtrar, neste caso, DP representa a cobertura total dos *reads* nessa posição.

Passo 4: No novo arquivo de SNPs obtido .vcf, agora se filtraram os genótipos com uma qualidade de genótipo menor de 99:

```
$ vcffilter -t --keep-info -g "GQ > 98" Athaliana_candidates_snp_dp20.vcf > Athaliana_candidates_snp_dp20_gq99.vcf
```

Observação 1: O GQ representa a qualidade do genótipo.

Passo 5: A frequência alélica de cada SNP identificado pode ser calculada a partir do arquivo .vcf final com o seguinte comando:

```
$ vcftools --vcf Athaliana_candidates_snp_dp20_gq99.vcf --freq --out Athaliana_candidates_snp.freq
```

Passo 6: O número de missing data por *locus* pode ser calculada a partir do arquivo .vcf final com o seguinte comando:

```
$ vcftools --vcf Athaliana_candidates_snp_dp20_gq99.vcf –missing-site
```

Passo 7: O número de missing data por indivíduo pode ser calculada a partir do arquivo .vcf final com o seguinte comando:

```
$ vcftools --vcf Athaliana_candidates_snp_dp20_gq99.vcf –missing-indv
```

Métodos utilizados para análise de matrizes de SNPs

Nesta seção serão descritos alguns exemplos de métodos utilizados para a análise de matrizes de SNPs. Os exemplos apresentados são de análises dentro de um contexto evolutivo.

Estrutura populacional e fluxo gênico

A pergunta inicial numa análise de qualquer conjunto de dados genéticos de dados populacionais é estabelecer se há evidência de estrutura populacional. Os indivíduos amostrados pertencem a uma população geneticamente homogênea ou a uma

população que contém subgrupos com alguma descontinuidade genética? Podemos encontrar evidências de subestrutura nos dados e quantificá-la?

As estimativas de estrutura populacional são usadas principalmente para entender aspectos históricos e demográficos na evoluão das espécies, mas também é muito importante fazer uma boa caracterização da estrutura populacional para evitar falsas inferências em estudos de associação em escala genômica (GWAS- do inglês *Genome Wide Association*), na identificação de associações de SNPs a doenças em populações mixigenadas (*admixture mapping*), ou para detectar regiões do genoma sob processos de seleção recente.

A estrutura genética é avaliada com métodos de agrupamento ou atribuição com base em um conjunto de dados de genótipos *multilocos* individuais. Em geral, existem dois tipos de abordagens para inferir a estrutura genética de um conjunto de dados: 1) Análises exploratórias e, 2) Análises de agrupamento baseadas em modelos genéticos. A principal característica das análises exploratórias é que estas sintetizam os conjuntos de dados dentro de um número de variáveis reduzidas, e a partir destas novas 'variáveis sintéticas' ou 'componentes' é possível inferir estrutura populacional sem assumir nenhum processo evolutivo envolvido na geração dos dados. Em contraste, os métodos de agrupamento baseados em modelos desenvolvidos pela genética de populações explicam a distribuição das frequências alélicas em populações estruturadas. Estes métodos inferem grupos genéticos com base nos dados individuais, para depois atribuir um grupo a cada indivíduo ou calcular um coeficiente de ancestralidade que pode ser interpretado como as respectivas contribuições das populações ancestrais (ou grupos genéticos) para cada amostra particular. Maiores detalhes podem ser encontrados em François e Waits (2016).

Um dos métodos de análise exploratória mais usado para inferir estrutura genética com dados de SNPs é a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) (Jolliffe, 1986). A PCA apresenta algumas vantagens: 1) o tempo de análise é extremamente rápido, o qual se torna muito atrativo com grandes conjuntos de dados, enquanto que métodos baseados em modelos genéticos podem ser intratáveis; e 2) o método de PCA não tenta classificar todos os indivíduos em populações discretas, em vez disso o PCA fornece as coordenadas de cada indivíduo ao longo de eixos de variação que podem estar representando padrões de subdivisão discreta, mas também padrões graduais de diferenciação.

Segue abaixo uma descrição dos passos consduzidos durante uma análise de PCA utilizando uma matriz de dados de SNPs no pacote adegenet (Jombart and Ahmed, 2011) de R (R Development Core Team 2016).

- **Análise de componentes principais para matrizes de SNP com o pacote adegenet de R**

A análise será feita a partir de uma matriz de SNPs em formato vcf obtida com um dos procedimentos de filtragem, alinhamento e deteção de SNPs descritos na seção anterior.

Passo 1: Leitura da matriz de SNPs e criação do objeto tipo *genlight* na área de trabalho do R usando o pacote vcfR (Knaus and Grünwald 2016).

```
> library(vcfR)
```

```
> matrix_VCF <- read.vcfR("SNP_matrix.vcf")

> SNP_data <- vcfR2genlight(matrix_VCF)
```
Passo 2: Implementar a análise de componentes principais

```
> library(adegenet)

> pca1 <- glPca(SNP_data, parallel=TRUE, n.cores=NULL)

# Quando o argumento nf (número de fatores retidos) não é especificado, a função exibe o barplot de autovalores da PCA e pede ao usuário determinar o número de componentes principais a ser retidos

> barplot(pca1$eig, main="eigenvalues", col=heat.colors(length(pca1$eig)))

# Exibir o barplot de autovalores da PCA

> varPC1 <- round(pca1$eig[1]/sum(pca1$eig)*100, digits = 2)

> varPC2 <- round(pca1$eig[2]/sum(pca1$eig)*100, digits = 2)

# Exibir a porcentagem de variação retida no primeiro e segundo componente principal.
```

Passo 3:Fazer o gráfico dos resultados (3 opções diferentes)

```
> scatter(pca1, posi="topright")

> colorplot(pca1$scores,pca1$scores, transp=F, cex=2.5,
xlab=paste(paste("PC1", varPC1, sep = " - "),"%",sep = ""),
ylab=paste(paste("PC2", varPC2, sep = " - "),"%",sep = ""))

> plot(pca1$scores[,1], pca1$scores[,2],
col=c(rep("blue",7), rep("orange",7)),cex=2)

> text(pca1$scores[,1], pca1$scores[,2] + 0.7,
labels=rownames(pca1$scores), cex= 0.7)
```

Dado que a PCA é uma aproximação focada na síntese ou descrição da diversidade global da amostra, as inferências de padrões de agrupamento acabam se baseando em avaliações visuais subjetivas dos gráficos resultantes (scatterplots). Como alternativa, a análise discriminante de componentes principais (DAPC, do inglês *Discriminant Analysis of Principal Components*) (Jombart and Devillard 2010) tem sido implementada em dados genéticos com o objetivo de encontrar variáveis sintéticas (ou

funções discriminantes) que maximizam o componente de variação entre grupos, enquanto minimizam a variação dentro de cada grupo.

No DAPC o número de grupos tem que ser definido *a priori*. Considerando que na maior parte das análises o número de agrupamentos genéticos é desconhecido, e com frequência é umas das perguntas básicas de pesquisa, tem se desemvolvido um processo de otimização do 'melhor' número de agrupamentos genéticos (*k*) com base no algoritmo de agrupamento *k-means* que maximiza a variação entre grupos. Para identificar o ótimo valor de *k* o algoritmo é implementado sequencialmente incrementando os valores de *k*, ao final todas as alternativas de agrupamento são comparadas usando o Critério de Informação Bayesiana (BIC, do inglês *Bayesian Information Criterion*). O valor ótimo de *k* (melhor suportado pelos dados) é aquele que apresenta o menor valor de BIC, mas a realidade biológica é bem mais complexa para poder definir um 'melhor' *k*. Numa perspectiva diferente, e provavelmente mais realista, cabe melhor identificar um número de grupos úteis para descrever um conjunto de dados. Este valor pode ser identificado numa curva de valores BIC em função de *k*, no ponto onde a variação do BIC começa a ser muito baixa com o aumento do *k* (saturação da curva) estaria indicando que o ganho de poder explicativo com o incremento de *k* é muito baixo como para ser considerado como fonte importante na explicação da variação nos dados observados.

Segue abaixo a descrição dos passos para análise de DAPC utilizando matriz de dados de SNPs.

- **Passo a passo de uma análise discriminante de componentes principais para matrizes de SNP com o pacote adegenet de R**

A partir do mesmo objeto genlight de R obtido na PCA:

Passo 1: Identificar o 'melhor' *k*

```
> library(adegenet)

> grp <- find.clusters(SNP_data, n.pca=NULL, n.clust=NULL, glPca=pca1)
```

A execução do algoritmo é realizada utilizando os dados transformados usando PCA para reduzir o número de variáveis e acelerar a execução do algoritmo

A função exibe um gráfico de variância acumulada explicada pelos autovalores do PCA, e pede ao usuário determinar o número de componentes principais a ser retido.

Além do tempo computacional, não há razão para manter um pequeno número de componentes. Assim, neste passo podem ser retidos todos os componentes principais, e, portanto manter toda a variação dos dados originais.

Em seguida, a função exibe um gráfico de valores BIC para valores crescentes de *k*, e pede ao usuário determinar o número de grupos a ser analizados.

Passo 2: Executar o DAPC

```
> dapc1 <- dapc(SNP_data, pop=grp$grp, glPca=pca1)
```

O usuário tem que determinar o número de componentes principais e de funções discriminates para implementar a análise. O resultado do DAPC é especialmente sensível ao número de componentes principais usados, então análises preliminares são necessárias para estabelecer um número adequado de componentes principais para obter um resultado confiável.

```
> grp <- find.clusters(SNP_data, n.pca=6, n.clust=3, glPca=pca1)

> dapc1 <- dapc(SNP_data, pop= SNP_data@pop,
n.pca=6, n.da=2, glPca=pca1)
```

Passo 3: Plotar os resultados em gráficos

```
> scatter(dapc1)

> col <- colorRampPalette(c("green", "blue", "red"))( 4 )

> s.class(dapc1$ind.coord, matrix_VCF@pop, xax=1, yax=2,
sub="DAPC scatter plot axis 1x2", col=col, axesell=FALSE,
cstar=0, cpoint=1, grid=FALSE, cellipse=1, clabel = 0.8)

> compoplot(dapc1, posi="bottomright", leg=TRUE,
ncol=1, col=c("deepskyblue","darkorchid1","firebrick2"),
cleg = 0.01, space=0, cex.lab=1, cex.names=.1)
```

Existem outros métodos de análises exploratórias para dados genéticos que incluem de maneira explicita a informação geográfica para avaliar a influência do espaço na estrutura genética. Um exemplo é a análise espacial de Componentes Principais (sPCA, do inglês *spatial Princial Components Analysis*) (Jombart et al., 2008). De forma semelhante à PCA a sPCA cria variáveis que sintetizam a variância nos dados, mas também a autocorrelação espacial medida pela estatística do I de Moran (Moran, 1950).

O método de estruturação populacional baseado em modelos mais popular está implementado no programa STRUCTURE (Pritchard et al., 2000) que usa um algoritmo Bayesiano para identificar grupos de indivíduos em equilíbrio de *Hardy-Weinberg* e equilíbrio de ligação. Neste programa a estrutura genética pode ser avaliada sob um modelo de não-mistura no qual se assume que a amostra consiste em um número *k* de grupos genéticos divergentes. Os indivíduos são probabilisticamente atribuídos a um dos grupos e as probabilidades resultantes são chamadas coeficientes de atribuição. O programa também permite usar um modelo de mistura que assume que os dados genéticos foram originados da mistura de um número *k* de populações ancestrais que podem ou não ser observadas no estudo. Neste modelo para cada indivíduo é calculado

um coeficiente de ancestralidade que corresponde às proporções do genoma de cada indivíduo que provém de cada grupo (ou populacão ancestral) inferido na amostra. A análise de agrupamento ou atribuição implementado no STRUCTURE precisa que o número de grupos (*k*) seja definido *a priori*, no entanto sem ou com identificação da origem dos indivíduos em cada população ou espécie. O método mais usado para definir o 'melhor' número de *k* é o conhecido como o Δ*k* de Evanno que pode ser implementado nos servidores online Structure Harvester (Earl e vonHoldt, 2012) ou PopHelper (Francis, 2016). No entanto é importante considerar que o próprio desempenho do STRUCTURE, assim como o procedimento para identificar o 'melhor' *k* pode ser sensível aos tamanhos de amostragem desiguais entre populações e ao padrão real de estrutura da diversidade genética das populações avaliadas. Por isto tem se recomendado estabelecer o número de subpopulações ('melhor' *k*) com estatísticas menos sensíveis ao tamanho amostral ou padrões complexos de subestrutura (ver alternativas propostas por Puechmaille (2016)), avaliar a estrutura com vários métodos diferentes, avaliar se os resultados são robustos replicando as análises, subamostrando indivíduos do conjunto completo de dados para obter uma amostragem mais uniforme e comparar os resultados da estrutura populacional dessa subamostragens (mais uniformes) com o resultado obtido com o conjunto de dados completo (menos uniforme).

Várias abordagens vem sendo desenvolvidas recentemente para aperfeiçoar o desempenho das análises de agrupamento baseadas em modelos quando são implementadas com dados genômicos (por exemplo, milhares de SNPs). As aplicações fastSTRUCTURE (Raj et al., 2014) e ADMIXTURE (Alexander et al., 2009) usam o mesmo modelo estatístico do STRUCTURE, mas realizam os cálculos com maior rapidez usando algoritmos mais eficientes, criados especificamente para este fim. Também tem se desenvolvido aplicacões específicas para aperfeiçoar as análises de STRUCTURE paralelizando as replicas independentes em computadores com múltiplos núcleos, diminuindo assim o tempo total de corrida. Um exemplo é o pacote do R parallelStructure (Besnier and Glover, 2013).

Já estabelecido um padrão de estrutura genética num conjunto de dados genéticos, um segundo passo é obter um entendimento do nível de fluxo gênico que pode estar acontecendo entre populações ou nos agrupamentos identificados. Apesar das análises de estrutura genética permitirem fazer inferências do fluxo gênico através da comparação dos coeficientes de ancestralidade de cada indivíduo com informação independente como, distribuição geográfica ou características morfológicas, existem várias aproximacões que permitem quantificar especificamente o fluxo gênico. O modo mais comum de quantificar indiretamente o fluxo gênico entre populações é pelo meio de estatísticas de diferenciação genética. A métrica de diferenciação populacional mais antiga e utilizada é o F_{ST} de Wright, que é uma das três estatísticas *F* usadas para descrever a partição da variabilidade genética entre a população total ($_T$) ou a amostragem completa, a subpopulação ($_S$), e os indivíduos dentro de cada subpopulação ($_I$) (Wright, 1943). Então, o F_{ST} é uma medida de diferenciação genética em nível de subpopulação em relação à população total que varia de 0 (panmixia) até 1 (isolamento genético completo) e mede a divergência na frequência alélica entre subpopulações. O F_{ST} tem sido considerado como inversamente proporcional à medida do número de migrantes por geração ($4N_em$) entre populações, mas por causa das premissas irrealistas do modelo continente-ilha (em que está baseado as estatísticas *F* de Wright) a quantificação direta de *m* derivada do F_{ST} deve ser avaliada com cautela (Whitlock and McCauley, 1999).

Como tentativas para reduzir possíveis vieses nas estimativas de diferenciação genética relacionadas com as premissas do modelo continente-ilha foram desenvolvidas varias estatísticas derivadas do F_{ST} como o R_{ST} de Slatkin que assume um modelo mutacional passo a passo que é considerado mais adequado para marcadores microsatélites (Slatkin, 1995), o G_{ST} de Nei considerado adequado para medidas obtidas com *loci* de múltiplos alelos (Nei, 1973), o *Theta* (θ) de Weir e Cockerham derivado da análise de variância molecular, e as medidas G'_{ST}, G''_{ST} e *D* de Jost propostas como alternativas para conjuntos de dados com alta heterozigosidade, poucas populações, e provavelmente fora do equilíbrio de Hardy-Weinberg (Hedrick 2005; Jost, 2008). Adicionalmente, outras medidas de diferenciação genética baseadas na heterozigosidade ou composição alélica entre populações tem sido desenvolvidas como o D_C de Cavalli-Sforza (Cavalli-Sforza and Edwards, 1967), a distância genética (*D*) de Nei (Hattemer 1982), a proporção de alelos compartilhados (D_{PS}) (Bowcock et al., 1994) e a distância genética condicional (cGD) calculada a partir de redes de populações que estima a diferenciação entre pares de populacões considerando simultaneamente a covariância genética de todas as populações (Dyer et al., 2010). Uma revisão aprofundada destas medidas pode se encontrar em (Whitlock and McCauley, 1999; Meirmans and Hedrick, 2011).

Existem vários programas que são comumente usados para estimar diferentes estatísticas associadas ao fluxo de genes entre populações como ARLEQUIN (Excoffier and Lischer, 2010), SPAGeDi (Hardy and Vekemans, 2002), FSTAT (Goudet, 2013), GENEPOP (Raymond and Rousset, 1995). Entretanto, recentemente foram desenvolvidos vários pacotes de R que permitem um processamento mais eficiente dos dados genéticos especialmente quando se esta trabalhando com dados genômicos, entre estes incluem: adegenet, poppr (Kamvar et al., 2015), gstudio (Dyer, 2009), StAMPP (Pembleton et al., 2013), pegas (Paradis, 2010).

Abaixo segue alguns exemplos de análises:

Passo 1: obter um objeto genind do pacote Adegenet a partir de uma matriz de entrada de dados em formato de STRUCTURE

```
library(adegenet)

indvs = 113 # número de indivíduos na matriz

snps = 10547 # número de loci na matriz

SNP_matrix <-  read.structure("matrix.str", n.ind=indvs, n.loc=snps,

onerowperind=FALSE, col.lab=1, col.pop=2, NA.char="-9")
```

Passo 2: definir os nomes das populações no objeto de R

```
pop(SNP_matrix) <- c(rep("POP01",12),rep("POP02",12),rep("POP03",10),

rep("POP04",12), rep("POP05",11), rep("POP06",12), rep("POP07",10),

rep("POP08",8), rep("POP09",9), rep("POP10",7), rep("POP11",10))
```

Passo 3: converter o objeto genind em objeto genepop

```
SNP_matrix_pop <- genind2genpop(SNP_matrix)
```

Passo 4: obtenção de matrizes de diferenciação genética entre populacões

```
nei_dist <- dist.genpop(SNP_matrix_pop, method=1)
eucl_dist <- dist.genpop(SNP_matrix_pop, method=2)
fst_dist <- genet.dist(SNP_matrix_pop, method = "WC84")
```

Passo 5: obtenção de matrizes de distância ao nível de indivíduos

```
library(poppr)
library(ape)
library(pegas)

### número de diferenças de alelos entre indivíduos
allelic_diff_dist <- diss.dist(SNP_matrix, percent = FALSE, mat = FALSE)

### distância euclidiana
eucl_dist_indv <- dist(SNP_matrix, method = "euclidean",
diag = FALSE, upper = FALSE, p = 2)

### diferenças de loci entre indivíduos
matrix_loci <- genind2loci(SNP_matrix)
loc_dist <- dist.gene(matrix_loci, method="pairwise",
pairwise.deletion = FALSE, variance = FALSE)
```

No passo a passo seguinte será apresentado como calcular distâncias genéticas entre indivíduos e populacões com o pacote StAMPP:

Passo 1: leitura da matriz de dados em formato vcf

```
library(vcfR)
matrix_VCF <- read.vcfR("SNP_matrix.vcf")
matrix_VCF <- vcfR2genlight(matrix_VCF)
pop(matrix_VCF) <- pop(SNP_matrix) # mesmas populações do exemplo anterior
library(StAMPP)
matrix_stampp <- stamppConvert(matrix_VCF, type = "genlight")
fst_dist2 <- stamppFst(matrix_stampp, nboots = 10, percent = 95, nclusters = 4)
genomic_dist <- stamppGmatrix(matrix_stampp)
nei_dist2 <- stamppNeisD(matrix_stampp, pop = TRUE) # distância entre populações
nei_dist_indv <- stamppNeisD(matrix_stampp, pop = F) # distância entre indivíduos
```

Cálculo de distâncias genéticas condicionais entre populações com o pacote gstudio

Passo 1: salvar uma tabela simples de indivíduos versus *loci* a partir do objeto genind criado com o pacote adegenet

```
write.table(SNP_matrix, "SNP_matrix .txt", sep="\t")
```

Passo 2: instalar o pacote gstudio

```
require(devtools); install_github("gstudio", "dyerlab", ref = "develop")
```

Passo 3: carregar a matriz de dados no pacote gstudio

```
library(gstudio)
SNP_data <- read_population("SNP_matrix .txt", type = "snp",
sep="\t", header = T, locus.columns = c(2:9119))
```

Passo 4: definir os nomes das populações no novo objeto de R e calcular o diagrama de populações (population graph)

```
pop <- c(rep("POP01",12),rep("POP02",12),rep("POP03",10),
rep("POP04",12), rep("POP05",11), rep("POP06",12), rep("POP07",10),
rep("POP08",8), rep("POP09",9), rep("POP10",7), rep("POP11",10))
require(popgraph)

SNP_data_mv <- to_mv(SNP_data)
graph <- popgraph(x = SNP_data_mv, groups = pops)
V(graph)$name <- c(1:12)
```

Passo 5: Plotar o gráfico do diagrama das populações

```
plot(graph)
plot(graph, edge.color="black", vertex.label.color="darkred",
vertex.color="#cccccc", vertex.label.dist=1.5)
layout <- layout.fruchterman.reingold(graph)
plot(graph, layout=layout, edge.color="black", vertex.label.color="darkred",
vertex.color = c("red", rep("yellow", 2), "red", rep("green", 2),
rep("orange", 2), rep("red", 3)), vertex.label.dist=2)
```

Passo 6: Obter a matriz de distâncias genéticas condicionais entre populações

```
cGD <- to_matrix(graph, mode="shortest path")
as.dist(cGD)
```

O uso de matrizes de distância genética tem sido bastante útil na avaliação da influência de características topográficas, climáticas ou ecológicas, assim como mudanças temporais destas variáveis, nos processos de diferenciação populacional. Geralmente nessas abordagens são calculadas medidas de distância geográfica, dissimilaridade ambiental ou ecológica, caminhos de menor custo ou de resistência da paisagem à migração entre populações ou indivíduos, e se fazem diferentes tipos de

tratamentos estatísticos como correlações ou ajustes de modelos lineares para estabelecer que variáveis expliquem melhor os padrões de diferenciação genética e assim procurar suporte ou propor hipóteses de diferenciação ecológica, adaptação local nas populações de estudo (McRae, 2006; Wang and Bradburd, 2014).

Embora as estatísticas de diferenciação genética provem uma ideia do fluxo gênico entre populações, não são suficientes para quantificar objetivamente este parâmetro. O principal motivo é que populações ou linhagens separadas sempre têm algum nível de polimorfismo ancestral compartilhado, questão que dificulta distinguir entre divergência recente sem (ou pouco) fluxo gênico e divergência mais antiga com fluxo gênico recente. É por isto que duas populações podem chegar a ter uma determinada medida de diferenciação genética (por exemplo, F_{ST} = 0.17) por via de processos evolutivos diferentes, principalmente relacionados com o tamanho efetivo populacional, a taxa de migração e o tempo de divergência (Hey, 2006; Leaché et al., 2014). É por isto que uma estimativa confiável do fluxo gênico é fundamental para entender a importância deste parâmetro em diferentes processos evolutivos como a divergência de linhagens, adaptação e especiação.

Existem várias alternativas para estimar o fluxo gênico entre populações, mas é importante considerar que o fluxo gênico é um parâmetro que pode mudar ao longo da história das populações ou linhagens envolvidas. Entre os programas especializados para quantificar o fluxo gênico entre populações está o BAYESASS que aplica um algoritmo baseado em estatística Bayesiana para estimar taxas de imigração recente (nas últimas gerações) entre populações e distribuições de probabilidade posterior de ancestralidade migrante para cada indivíduo (Wilson and Rannala, 2003). O programa BIMr (*Bayesian Inference of imMigration rates*; (Faubet and Gaggiotti, 2008) é também um método Bayesiano que faz inferências de proporções recentes de genes migrantes entre populações e identifica fatores ambientais que possam estar potencialmente relacionados com a dinâmica de fluxo gênico observada.

Para a obtenção de estimativas de fluxo gênico numa escala temporal mais profunda, as abordagens baseadas em coalescência tem mostrado muita utilidade já que permitem fazer análises mais integrais levando em conta os processos estocásticos envolvidos na transmissão de genes ao longo das gerações, assim como outros processos evolutivos envolvidos na história das populações. Entre estas, estão as abordagens conhecidas como métodos *full-likelihood*, implementadas nos programas LAMARC (Kuhner, 2006), IMa (Hey and Nielsen, 2007) ou MIGRATE-N (Beerli and Felsenstein 1999). Uma desvantagem destas abordagens é a alta demanda computacional. Novas versões dos programas com suporte para utilizar múltiplos processadores em paralelo tem ajudado a superar o problema de analizar conjuntos de dados com alta quantidade de *loci*, como no caso de matrizes de SNPs.

Como alternativa, também tem sido desenvolvidas várias abordagens para analisar conjuntos de dados genômicos. Por exemplo, a análise baseada em estatística sumária conhecida como o test de ABBA/BABA que tem o potencial de discriminar entre padrões de compartilhamento de alelos derivados relacionados com processos de fluxo gênico pós-divergência vs. processos de sorteio incompleto de linhagens, o qual auxilia à estimativa do tempo e magnitude de um processo de fluxo gênico entre populações (Durand et al., 2011). Uma alternativa foi implementada no programa GphoCS (*Generalized Pylgenetic Coalescent Sampler*), que usa um algoritmo Bayesiano para inferir tamanho de população ancestral, os tempos de divergência populacional e taxas de migração em conjunto com a genealogia de conjuntos de sequências de multiplos *loci* separados ao longo do genoma (Gronau et al., 2011).

Também foram desenvolvidos métodos que exploram o espectro de frequência dos alelos (AFS; do inglês *Allele Frequency Spectrum*) que é a distribuição das frequências alélicas de um conjunto de *loci* determinado (frequentemente SNPs) numa população ou amostra. A utilidade do AFS para inferir parâmetros populacionais se baseia na ideia de que diferentes processos demográficos influenciam as distribuições de frequência alélica, por exemplo, um padrão de abundância de SNPs compartilhados pode ser relacionado com processos de fluxo genético, ou um processo de redução do tamanho das populações conduz à diminuição de SNPs de baixa frequência. Estes métodos usam uma extenção da teoria de verosimilhança conhecida como *composite-likelihood* que permite a aplicação do método de verosimilhança na análise de dados de enormes dimensões. Os métodos baseados no AFS dependem do cálculo da probabilidade de um AFS observado dado um vetor complexo de parâmetros que descrevem a história das populações em estudo. Estes métodos têm possibilitado a comparação de cenários demográficos e obter estimativas precisas de parâmetros genéticos populacionais, mesmo para modelos complexos de história populacional, usando conjuntos de dados compostos de milhares de *loci* de múltiplos indivíduos (Kern and Hey, 2016). Entre as abordagens para implementar estes métodos está o implementado no programa ∂a∂i (*Diffusion Approximation for Demographic Inference* (Gutenkunst et al., 2009) que usa uma aproximação de difusão para estimar o AFS de uma população e os implementados nos programas FASTSIMCOAL2 e o pacote de R Jaatha (Jsfs Associated Approximation of THe Ancestry) que usam simulações de coalescência para estimar o AFS esperado de uma população (Naduvilezhath et al., 2011; Excoffier et al., 2013). Jaatha considera modelos de mutação de sítios finitos, o que é necessário para evitar vieses na estimativa da taxa de mutação, tempos de divergência e taxas de migração.

Outra opção usada para a estimativa de parâmetros demográficos é a análise conhecida como Computação Bayesiana Aproximada (ABC, do inglês *Approximate Bayesian Computation*). Esta análise proporciona uma aproximação da distribuição posterior das probabilidades de um modelo demográfico estabelecido e os respectivos valores dos parâmetros populacionais. A análise é implementada através da simulação de diferentes modelos populacionais pre-estabelecidos cujos parâmetros são amostrados de distribuições *a priori* especificadas, seguido do cálculo de estatísticas sumárias informativas para os parâmetros avaliados. A obtenção da aproximação da probabilidade posterior é feita pela comparação das estatísticas sumárias das simulações com as obtidas com os dados observados. Esta análise pode ser implementada no programa DYABC que numa interface gráfica implementa todos os passos do ABC (construção de modelos, simulação, cálculo de estatísticas sumárias simuladas e observadas, rejeição de modelos, estimativa de parâmetros e avaliações do suporte do modelo e dos parâmetros), mas com a desvantagem de ser muito restritivo nos modelos que podem ser avaliados e também ser um programa fechado onde não pode se fazer um acompanhamento cuidadoso da análise (Cornuet et al., 2014). Alternativas que permitem maior flexibilidade e acompanhamento precisam do uso de diferentes programas em cada passo do modelo. Os programas mais comumente usados são FASTSIMCOAL2 (Excoffier et al., 2013) e ms (Hudson, 2002) para a simulação de modelos, "Arlsumstat" ou "sample_stats" para o cálculo de estatísticas sumárias e "msreject" ou "ABCestimator" para estimar a probabilidade posterior dos modelos. Existem também algumas aplicações que ajudam na implementação de cada passo como, por exemplo, ABCtoolbox (Wegmann et al., 2010), msABC (Pavlidis et al., 2010) e o pacote de R abc (Csilléry et al., 2012). Diversas revisões baseadas em dados simulados e empíricos mostraram que o desempenho das inferências aumenta

substancialmente com o aumento da quantidade e comprimento de *loci* sequenciados, enquanto que não se reporta benefício pela amostragem de grande número de indivíduos (Robinson et al., 2014).

Análises de Seleção

Com a crescente disponibilidade de sequenciamento de alto desempenho tornou-se possível o uso de uma alta densidade de marcadores genéticos para caracterizar a diversidade genética de indivíduos e populações, bem como identificar regiões do genoma que podem estar sob a influência de seleção natural. Um dos caminhos para identificar processos de seleção é por meio da abordagem de genômica populacional, no qual se baseia na estimativa da diferenciação genética entre as populações utilizando milhares de marcadores SNPs identificados ao longo do genoma para estabelecer um modelo nulo de diferenciação neutra e a partir deste identificar *loci outliers* que se presume estarem sob seleção ou ligados a regiões genômicas adaptativas. Este princípio pode ser implementado pelo algoritmo FDIST no software ARLEQUIN que identifica *outliers* que exibem fortes diferenças de uma distribuição nula da estatística F_{ST}. Aqueles *loci* com valores de diferenciação mais altos do que o esperado a partir da distribuição nula são presumidos estarem sob seleção diversificadora ou seleção local e aqueles valores de diferenciação menores do que o esperado são inferidos como sob seleção estabilizadora ou purificadora (Beaumont e Nichols, 1996).

O maior desafio nas abordagens de genômica populacional está no estabelecimento acurado da distribuição nula da diferenciação genômica neutra. Tem sido mostrado que processos demográficos como o *allele surfing* ou reduções do tamanho populacional (gargalo de garrafa) podem deixar padrões *outlier* semelhantes aqueles deixados por seleção. Além disso, padrões de estruturação espacial complexa podem aumentar a variação dos parâmetros genéticos no genoma acrecentando às altas taxas de surgimento de falsos positivos nos testes de *loci outlier*. Novos métodos Bayesianos, baseados em modelos populacionais, avaliam a probabilidade da hipótese nula (neutralidade) e alternativa (não neutral) dado um conjunto de dados, e estão implementados em programas como BayesFST (Beaumont and Balding, 2004) e BayeScan (Guillot, 2011), os quais foram desenvolvidos para corrigir possíveis vieses relacionados com estrutura populacional na amostra.

Mais recentemente, vem sendo desenvolvidas novas abordagens que consideram explicitamente padrões de covariância na frequência alélica entre as populações gerados pela história demográfica e efeitos espaciais, como a implementada no programa BayEnv (Günther and Coop, 2013). Neste programa, em um segundo passo da análise, é possível também avaliar a correlação entre as frequências alélicas em cada *locus* (ou apenas em *loci* de interesse) e variáveis ambientais. Outra abordagem que tem mostrado um bom desempenho considerando o poder de detecção e a taxa de erro em diferentes cenários de estrutura populacional e padrão de seleção é o Modelo misto de fatores latentes (LFMMs; *Latent factor mixed models*) (Frichot et al., 2013). Esta é uma abordagem muito geral e flexível que também apresenta a possibilidade de detectar relações entre frequências alélicas e variáveis ambientais levando em consideração a estrutura da população. O modelo pode ser visto como uma análise aproximada de componentes principais combinada com uma regressão, por isto tem a vantagem de ser computacionalmente mais rápida que as análises Bayesianas.

Uma revisão detalhada sobre métodos para identificar *loci* sob seleção pode ser encontrada em Pardo-Diaz et al. (2015).

Aplicações

Os marcadores SNPs representam a terceira geração de marcadores moleculares e são empregados com sucesso em diversos estudos. Podemos destacar a aplicação dos marcadores SNPs para investigar a variação genética e estrutura populacional em espécies nativas e cultivadas; para reconstrução filogenética e aplicação em taxonomia; para análise de seleção natural e evolução adaptativa; para a construção de mapas genético, em estudos de associação entre genótipo e fenótipo e para análises genéticas do DNA humano, incluindo associação com doenças. A Tabela 8.1 destaca alguns exemplos do uso e aplicação de marcadores SNPs envolvendo as tecnologias de NGS. Além disso, também apresentamos a descrição de alguns exemplos de estudos com dados de SNPs em diferentes áreas de conhecimento.

Tabela 8.1 – Exemplos de estudos com marcadores SNPs.

Espécie	Aplicação	Metodologia de identificação e/ou genotipagem	Plataforma de sequenciamento ou Genotipagem	Pipeline ou programa para filtrar SNPs	Nº de SNPs identificados	Referência
Athene noctua	Filogeografia e estrutura populacional	Identificação e genotipagem por GBS	Illumina HiSeq2000	Pipeline personalizado (ver descrição detalhada no artigo)	22,185	Pellegrino et all., 2016
Wyeomyia smithii	Filogeografia	Identificação e genotipagem por RAD-seq	Illumina GAIIX	Pipeline propria (ver descrição detalhada no artigo)	3,741	Emerson et al., 2010
Teleogramma	Filogeografia e taxonomia	Identificação e genotipagem por ddRAD	Illumina HiSeq 2500	Stacks 1.35	37,826	Alter et al., 2016
Brucella suis	Filogenia	Identificação e genotipagem por comparação de genomas obtidos de bancos de dados	na	kSNP	16,756	Sankarasubramanian et al., 2016
Olea europaea L	Filogenia	Identificação por sequenciamento de DNA genômico e. genotipagem em Fluidigm Dynamic Arrays	Illumina HiSeq 2000 (identificação), genotyping EP1 System (genotipagem)	na	145,974 identificados e 192 genotipados	Biton et al., 2015
Zea mays L.	Diversidade genética e estrutura populacional	MaizeSNP50 BeadChip da Illumina GenomeStudio	-	Illumina GenomeStudio	56,11	Zhang et al., 2016
Capra aegagrus hircus	Diversidade genética e estrutura populacional	Illumina goat SNP50 Bead chip)		Illumina GenomeStudio	15,105	Visser et al., 2016
Sarcophilus harrisii	Diversidade genética	Identificação por sequenciamento de genoma inteiro genotipagem por sequenciamento de genes candidatos	Illumina HiSeq 2000 e Miseq	SAMTOOLS e GATK	267	Wright et al., 2015
Phaseolus vulgaris	Estrutura populacional	Identificação e genotipagem por RAD-seq	Illumina HiSeq	Pipeline usada por Grattapaglia et al. (2011)	384	Valdisser et al., 2016
Helianthus annuus L.	Construção de mapa de ligação	Identificação e genotipagem por GBS	Illumina Genome Analyzer II,	TASSEL	46,278	Celik et al., 2016
Gasterosteus aculeatus	Mapeamento genético	Identificação e genotipagem por RAD-seq	Illumina Genome Analyzer sequencer	Scripts Perl	~13,000	Baird et al., 2008
Triticum aestivum L. e *Hordeum vulgare*	Mapeamento genético	Identificação e genotipagem por GBS	Illumina GAII e Illumina HiSeq2000	TASSEL	~20,000 (*T. aestivum*) e ~34,000 (*H. vulgare*)	Poland et al., 2012a

Capsicum spp.	Diversidade genética e mapeamento	Identificação por sequenciamento de genoma inteiro e genotipagem por hibridização em arranjos	Illumina Genome Analyzer II system e genotyping platforms in BGI-Shenzhen e,	GenomeStudio Genotyping sofware (v2011	~15,000	Cheng et al., 2016
Oryza sativa	Diversidade genética e associação genótipo/fenótipo	Identificação e genotipagem por RAD-seq	Ion Torrent PGM and Illumina Hiseq2500	TASSEL	22,682	Tang et al., 2016
Triticum aestivum L.	Melhoramento genético - seleção genômica	Identificação e genotipagem por GBS	Illumina HiSeq 2000	Population-based SNP calling. Uso do teste exato de Fisher para filtrar SNPs	41,371	Poland et al., 2012b
Solanum habrochaites (selvagem) e *S. lycopersicum (*cultivada*)*	Associação genótipo/fenótipo	Identificação eletrônica por comparação de transcritoma e de EST	Dados baixados de bancos de dados	SAMTOOLS; AutoSNP v 2	8,978	Bhardwaj et al., 2016
Humano	Estudo de câncer de pulmão	RNA-seq	Illumina GAII	GATK	85,028	Sathya et al. 2015
Lepus europaeus, Pallas 1778	Adaptação e especiação	RNA-seq	Illumina HiSeq 2000	GATK	66185	Amoutzias et al., 2016

Exemplos

Análises populacionais e filogeografia (diversidade genética e estrutura populacional)

O recente avanço das tecnologias de sequenciamento tem permitido a caracterização da diversidade e estrutura genética com milhares de SNPs em um amplo número de organismos, inclusive em espécies não modelo, as quais não dispõem de genomas de referência. Isto realça o potencial das tecnologias de sequenciamento genômico para abordar questões relacionadas com biologia evolutiva.

Os marcadores do tipo SNPs tem contribuído de forma significativa na obtenção de estimativas de variabilidade genética e estrutura populacional, medidas que são amplamente usadas em estudos evolutivos que buscam a reconstrução de processos históricos no nível intra- e inter-específico e têm sido aplicados em organismos silvestres tanto como em espécies domesticadas ou sob algum manejo antrópico.

A maior vantagem da utilização de grande quantidade de *loci* em estudos de biologia evolutiva está no aumento da precisão das estimativas de estrutura e divergência genética. Quanto maior a quantidade de *loci* avaliados maior a representatividade do genoma e assim uma variância reduzida entre *locus*. Além disso, o potencial de diferenciar efeitos *locus* específico (por exemplo seleção, acasalamento preferencial e recombinação) de efeitos genômicos (por exemplo, deriva, fluxo gênico, mudanças demográficas), o que gerou uma abordagem conhecida como genômica populacional (Black et al., 2001, Luikart et al., 2003). De forma semelhante, nas análises de atribuição (por exemplo, estimativas da ancestralidade individual a partir de conjuntos de dados de genótipos multilocus (Pritchard et al., 2000; Alexander et al., 2009) como descrito anteriormente, o poder de identificar padrões de estrutura genética, miscigenação e indivíduos migrantes aumenta conforme aumenta o número de *loci*.

Adicionalmente, a maior acessibilidade de genotipagem baseadas em marcadores SNPs tem permitido várias vantagens metodológicas. Entre estas, a possibilidade de obtenção de estimativas confiáveis de diversidade e estrutura genética com amostras populacionais pequenas, utilizando um grande número de *loci* (> 2 indivíduos (Willing et al., 2012); menor perda de informação genética a partir do uso de amostras de DNA degradado, já que o tamanho requerido do fragmento para obter um SNP é menor (~100 – 200 pb) em comparação a outros tipos de marcadores, tais como microssatélites (~100 – 400 pb) ou de sequências de genes, íntrons ou espaçadores intergênicos (~500 – 1500 pb). Isto permite a maior utilidade de amostras de DNA obtidas por meio de técnicas não invasivas como fezes e pelos de animais, assim como de espécimes de coleções de museus e herbários (Taberlet et al., 1999; Bi et al., 2013).

As técnicas de sequenciamento de alto desempenho recentemente se estabeleceram como ferramentas muito importantes para estudos populacionais e filogenéticos, uma vez que a análise da maior quantidade de *loci* possível se tornou um requerimento inevitável a partir dos fundamentos estabelecidos na filogeografia estatística (Knowles e Maddison, 2002) e no paradigma de árvore de espécies. Estes fundamentos proveram argumentos teóricos suficientes que apoiam a importância da incorporação de múltiplos *loci* para o estabelecimento de estimativas precisas de processos históricos de espécies e populações que considerem a estocasticidade nos processos de coalescência gene específicos (por exemplo, padrões aleatórios de herança genética).

Recentemente, diversos estudos com marcadores SNPs foram publicados. Estes estudos abordam diversas perguntas populacionais cujos dados têm sido obtidos com as metodologias descritas anteriormente. Por exemplo, usando como modelo biológico duas espécies de *Pleurodema* (rãs de quatro olhos) distribuídas na Caatinga e marcadores SNPs obtidos através do sequenciamento de regiões associadas a sítios de reconhecimento de enzimas de restrição (ddRADseq) mostraram as vantagens de implementar inferência filogeográfica baseada em modelos na qual se calcula o *composite-likelihood* dos dados observados (AFS calculado a partir da matriz de SNPs obtida) em relação a diferentes modelos demográficos propostos. A escolha do modelo melhor suportado pelos dados observados foi feita através da classificação deles usando o critério de informação de *Akaike* (Thomé e Carstens, 2016). Neste trabalho é mostrada a importância de primeiro obter um modelo demográfico apropriado para o grupo de estudo e os dados obtidos para assim conseguir uma determinação objetiva sobre quais parâmetros estimar.

Mapeamento genético

O mapeamento genético da variação genômica natural ou induzida é uma poderosa abordagem para entender a função dos genes em uma variedade de processos biológicos. A alta densidade dos marcadores SNPs nos genomas os torna ideais para estudar a herança de regiões genômicas. Mostrando pela primeira vez o uso de NGS com a técnica de RAD (RAD-seq), Baird et al. (2008) identificaram mais de 13.000 SNPs e foram capazes de mapear três características em dois organismos modelos. Neste estudo os autores reavaliaram alguns dos QTLs de um estudo anterior onde usaram a técnica original de RAD (Miller et al., 2007) e demonstraram a eficiência da técnica de RAD-seq, mapenado mais QTLs. Também demonstraram que diferentes densidades de marcador podem ser atingidas pela escolha da enzima de restrição. Além disso, desenvolveram um sistema de código de barras para multiplexação de amostras e revalidaram QTLs para perda de blindagem de placas laterais em *Gasterosteus aculeatus*, identificando pontos de interrupção recombinantes em indivíduos F2. A codificação de códigos de barras também facilitou o mapeamento de uma segunda característica, uma redução da estrutura pélvica, pela reclassificação *in silico* de indivíduos.

Em um estudo com girassol (*Helianthus annuus* L.), Celik e colaboradores (2016) identificaram SNPs nessa espécie usando a abordagem de genotipagem por sequenciamento (GBS) em uma população de mapeamento F2 intraspecífico. Um total de 46.278 SNPs foi identificado no genoma do girassol os quais estavam distribuídos em 17 grupos de ligação (LG1-LG17). Após as filtragens 9.535 SNPs foram mantidos e testados quanto ao polimorfismo nos parentais da população F2, sendo identificados 7.646 SNPs polimórficos. Muitos SNPs foram eliminados devido a grande quantidade de dados faltantes e no final um mapa genético de ligação foi construído baseado em 817 SNP distribuídos em 17 grupos de ligação. Os autores salientam que tanto os SNPs identificados quanto o mapa de ligação construído podem constituir ferramentas valiosas de genética molecular para a reprodução de girassol.

Estudos de associação genótipo/fenótipo

Os recentes avanços tecnológicos na descoberta e genotipagem de marcadores SNPs têm permitido uma maior precisão em estudos de associação genótipo e fenótipo. Tais estudos estão ganhando um grande destaque em diversas áreas. Nesse sentido, os

estudos de GWAS visam associar, direta ou indiretamente, marcadores moleculares do tipo SNP, a um determinado fenótipo, podendo se referir a uma ou mais características do indivíduo ou, até mesmo, uma doença (Reverter e Fortes 2013). Um estudo de GWAS em tomates silvestres (*Solanum habrochaites*) e cultivados (*Solanum lycopersicum*), Bhardwaj e colaboradores (2016) exploraram a consequência de substituições tanto em sequências nucleotidicas quanto no nível da estrutura proteica. Um total de 8.978 SNPs com taxa Ts / Tv (Transição / Transversão) de 1,75 foram identificados a partir de dados de ESTs (*Expressed Sequence Tag*) e de NGS de ambas as espécies disponíveis em bancos de dados públicos. Destes, 1.838 SNPs são não-sinônimo e distribuídos em 988 genes codificadores de proteínas. Entre estes, 23 genes contendo 96 SNPs estavam envolvidos em traços tais como, amadurecimento de frutos, resposta a frio, desenvolvimento de tricomas e textura de frutas. Além disso, haviam 28 SNPs deletérios distribuídos em 27 genes e alguns destes genes estavam envolvidos na interação planta patógenos e em rotas hormonais de plantas.

Análise genética do DNA humano

A análise genética de doenças sejam elas monogênicas ou multifatoriais, raras ou comuns, trazem a compreensão dos mecanismos envolvidos na expressão gênica e a correlação dos genótipos e fenótipos observados em pacientes, assim como a variação global observada entre diferentes populações decorrentes de pequenas variações no DNA.

A frequência observada de variação em uma única base no DNA genômico a partir de dois cromossomos é de 1/1000 pb. A taxa de diferença nucleotídica entre dois cromossomos escolhidos aleatoriamente é um índice denominado de diversidade de nucleotídeos. Isso significa que existe uma probabilidade média de 0,1% de qualquer base ser heterozigótica em um indivíduo, sendo que em éxons essa diversidade é cerca de quatro vezes mais baixa (Jobling et al., 2013). Entretanto algumas regiões do genoma apresentam variações amplas nestas taxas, como por exemplo, as regiões que envolvem os genes HLA, que apresentam variações de 5 a 10% em suas sequências (Brookes et al., 1999).

Os SNPs podem ser bi, tri ou tetra alélicos, no entanto, variações além da bialélica são extremamente raras em humanos. A frequência dos diferentes tipos de SNPs em humanos não é igualitária, onde aproximadamente 2/3 das alterações são C⇔T (G⇔A). Isto se deve às reações de deaminação de 5’ metilcitosina, que ocorrem em maior frequência, especialmente em ilhas CpG. O 1/3 restante é distribuído entre os outros três tipos de alterações. Ensaios de detecção de SNPs por pareamento de bases, como por exemplo, na utilização de sondas, devem ter uma estringência bastante alta, visto que o *mismatch* mais estável que ocorre neste tipo de ensaio é G:T, e coincidentemente afeta a distinção da variação mais abundante encontrada em humanos, C⇔T (G⇔A) (Brookes et al., 1999).

Vale lembrar que pseudogenes podem ser um desafio às análises convencionais de SNPs, uma vez que eles podem acarretar em um maior número de alterações em porções extremamente similares às regiões codificadoras de genes ativos (Robicheau et al., 2017).

A identificação de SNPs e o estabelecimento de sua possível correlação com fenótipos, como na atenuação ou exacerbação de condições clínicas é um grande desafio. Além de estarem diretamente relacionados com a evolução do genoma humano, os SNPs possuem uma íntima relação com a saúde humana, influenciando direta ou indiretamente nas doenças ou na forma como medicamentos são metabolizados.

Mutações e SNPs podem estar em opostos em um espectro fenotípico. SNPs sozinhos não são capazes de causar doença, mas junto com fatores ambientais, podem influenciar de maneira muito importante no fenótipo de doenças. Caracterizar a influência destes fatores sobre o fenótipo de um paciente, no estabelecimento de uma correlação genótipo-fenótipo é uma iniciativa de grande importância, pois se trata de ferramentas úteis no diagnóstico molecular. Seu uso como marcador molecular na investigação médica é de extrema relevância, pois as variantes podem estar associadas com o risco de ocorrência de determinadas doenças, apresentando-se em maior frequência entre pacientes quando comparados com indivíduos sem a doença.

A busca por SNPs no estabelecimento de bons marcadores moleculares é contínua na pesquisa em seres humanos. Estudos de associação buscando variantes em desequilíbrio de ligação com alelos patogênicos colaboram no estabelecimento de protocolos de diagnóstico de doenças. Vale ressaltar que populações menores, mais antigas e estáveis tendem a possuir relações de desiquilíbrio de ligação mais intrínsecas do que populações mais modernas e expandidas recentemente, havendo mais chance da identificação de marcadores moleculares em populações mais antigas, mas quando a formação da nova população se tratar de um efeito fundador, o mesmo pode não ocorrer.

Os estudos de associação consistem em determinar a frequência de um determinado genótipo entre pacientes e compará-los a controles saudáveis, podendo haver estratificação da população analisada, estabelecendo se há ou não risco relativo à determinados haplótipos. Um exemplo bem caracterizado é o do alelo ε4 do gene da Apolipoproteína E (ApoE). A ApoE é uma proteína plasmática envolvida no transporte de colesterol e outras moléculas hidrofóbicas. O gene está localizado no cromossomo 19 e apresenta três alelos: ε2, ε3 e ε4. O alelo ε4 está associado geneticamente com a doença de Alzheimer esporádica ou de início tardio, com risco aumentado em até 10 vezes para ocorrência da doença entre homozigotos para este alelo (Liu et al., 2013).

Outro exemplo de SNP que acarreta em alteração de fenótipo é a alteração C>T na posição -13910 (rs4988235), localizada a montante do gene da lactase-florizina hidrolase (LCT). Ela é a principal responsável pela persistência da atividade da enzima LCT, que permite que indivíduos adultos tolerem a ingesta de leite e seus derivados. O diagnóstico molecular da hipolactasia persistente em adultos, ou intolerância à lactose pode ser realizado pela amplificação da porção que inclui a posição -13910 seguida de clivagem com a enzima BsmFI (Figura 8.3) (Bulhões et al., 2007). Indivíduos homozigotos para o alelo C não toleram a ingesta de leite e seus derivados, apresentando sintomas como desconforto abdominal e diarréia, enquanto que homozigotos para o alelo T apresentam a enzima lactase persistente e os digerem de forma adequada. Heterozigotos C/T apresentam discordância de sintomas, com uma digestabilidade intermediária. Também existe a possibilidade de realizar o diagnóstico com o uso de ensaios de genotipagem por PCR em tempo real através de sondas marcadas fluorescentemente. A frequência global deste polimorfismo é de 84% C e 16% T, entretanto ao analisar populações específicas, diferenças marcantes são observadas, como por exemplo, 97% C e 3% T entre africanos, 49% C e 51% T entre europeus e 100% C entre indivíduos do leste asiáticos (1000 Genomes Project Consortium).

Os SNPs também são importantes ferramentas para a determinação de metabolizadores em farmacogenética. Hoje em dia sabemos que diferenças genéticas individuais encontradas nas enzimas metabolizadoras, transportadores e receptores em humanos podem influenciar a forma como determinados medicamentos são metabolizados pelo organismo, e consequentemente a forma como influenciam na

resposta ao tratamento e toxicidade. Atualmente é possível, através do uso da informação genética, adequar o tratamento aplicado ao paciente. São amplamente conhecidas as diferentes isoformas do gene do citocromo P450, que acarretam na classificação de metabolizadores lentos, intermediários e rápidos, de acordo com a atividade da enzima. Um dos exemplos bem conhecidos é o do gene *CYP2C9*, que apresenta mais de 50 polimorfismos (SNPs) que podem influenciar na resposta a medicamentos, incluindo a varfarina, importante anticoagulante associado à complicações hemorrágicas em determinados genótipos. Outro exemplo é o *CYP2D6*, que apresenta importante influência na metabolização de antidepressivos tricíclicos. A adequação da dose em decorrência do perfil genético do paciente pode variar de 28% a 180% da dose recomendada entre metabolizadores lentos e ultra-rápidos quando comparados a metabolizadores normais (Lynch et al., 2007; Zanger et al., 2013).

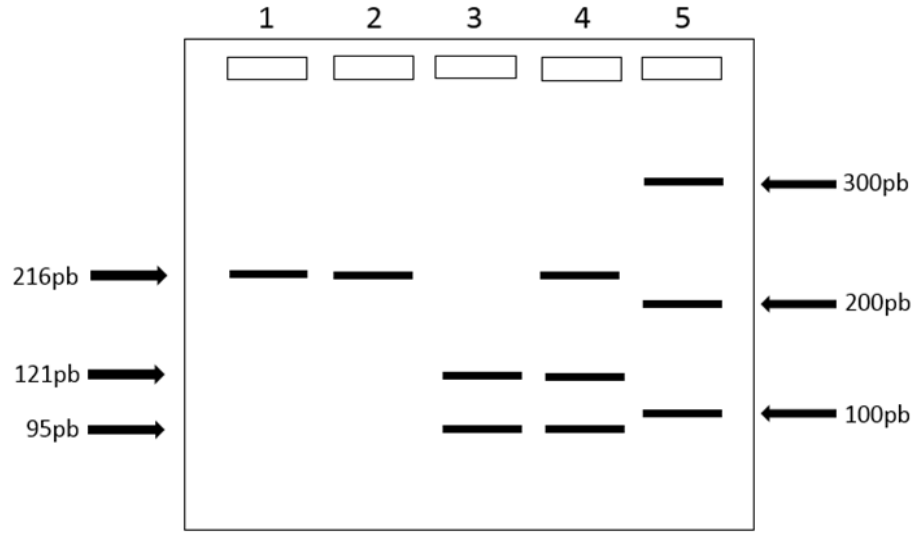

Figura 8.3 - Diagnóstico molecular de hipolactasia: eletroforese em gel de agarose 2,5% do produto de PCR da região à montante do gene LCT e a respectiva clivagem da posição -13910 com a enzima BsmFI. Produto não clivado (1), indivíduo homozigoto para o alelo C (2), indivíduo homozigoto para o alelo T (3) indivíduo heterozigoto C/T (4), marcador de peso molecular (5).

Os SNPs em humanos também são uma importante fonte de informação sobre a história evolutiva de populações. Eles são atualmente a ferramenta mais eficaz em estabelecer inter-relações entre populações, comportamentos migratórios e sua evolução. Conhecer a própria origem e história fascina a todos. A antropologia molecular é capaz de contar, através da construção de haplótipos a história de diferentes povos. A influência da seleção natural, de deriva genética, do fluxo gênico e de mutações permitiram que o genoma humano evoluísse de forma a gerar as particularidades físicas, culturais e comportamentais de diferentes populações. Além da caracterização genética, estudos antropológicos, arqueológicos e linguísticos se beneficiam da análise de variantes no genoma humano.

Vatsiou e colaboradores (2016) realizaram análises populacionais pelos métodos de XPCLR e iHS, utilizando como estratégia a pesquisa de genes envolvidos no metabolismo e sistema imune que possam ter sofrido pressão de seleção diferenciada nos ambientes ancestrais e atuais. Eles foram capazes de identificar 23 genes candidatos ligados ao metabolismo, 13 dos quais candidatos para seleção positiva.

Em outro trabalho bastante interessante, Sathya e colaboradores (2015) compararam indivíduos saudáveis que nunca fumaram, saudáveis fumantes, fumantes com câncer de pulmão e não fumantes com câncer de pulmão, na busca por SNPs através do uso de RNA-Seq, com a intenção de identificar possíveis marcadores associados ao câncer de pulmão.

A análise forense em amostras humanas tradicionalmente faz uso de regiões repetitivas do DNA, como VNTRs (*Variable Number Tandem Repeats)* e STRs (*Short Tandem Repeats)* (mais informações no capítulo 6), entretanto, com o avanço das tecnologias de sequenciamento e o maior conhecimento a cerca do genoma humano, o uso de SNPs com a finalidade de identificar individualmente seres humanos tem se tornado cada vez mais comum. Frequentemente amostras biológicas de cenas de crime ou de amostras fósseis estão misturadas, em pouca quantidade, ou com a conservação comprometida, e nesse sentido a possibilidade de trabalhar com amplicons de menor tamanho e com menor taxa de mutação é bastante interessante. A identificação do cromossomo Y, de diferentes haplótipos e de variantes do DNA mitocondrial através de SNPs são uma importante ferramenta em análises forenses, permitindo inclusive a identificação da origem geográfica dos indivíduos (Sobrino et al., 2013).

Para equivalerem às análises de STRs usuais, onde se analisam em média 10 *loci* no genoma, cerca de 60 diferentes SNPs precisam ser analisados para discriminar indivíduos. Nesse sentido, o uso de STRs ainda é mais vantajoso, levando-se em conta a ampla experiência acumulada ao longo dos anos neste tipo de análise e a existência de métodos muito bem estabelecidos para sua análise. Grandes estudos de associação de haplótipos, partindo de análises como o projeto "HapMap" (www.hapmap.org) permitirão que SNPs específicos, os mais informativos ao longo do genoma, sejam utilizados para esse tipo de identificação, assim como para outros diversos tipos de associações genótipo-fenótipo.

Importantes bancos de dados de livre acesso podem ser utilizados na pesquisa de SNPs. Entre os mais importantes estão o dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) e o HGBASE - Genic Human Bi-Allelic Sequences (Http://hgbase.interactiva.de/). Através do dbSNP é possível visualizar SNPs de diferentes espécies, assim como a sequência nas quais os mesmos estão inseridos. O HGBASE por sua vez descreve a localização e o componente gênico da alteração, assim como detalhes sobre ensaios e correlações fenotípicas.

Referências

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ et al. (2007) Direct selection of human genomic *loci* by microarray hybridization. *Nat Methods* 4:903–905. doi: 10.1038/nmeth1111

Alexander DH, Novembre J and Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664. doi: 10.1101/gr.094052.109

Alison G, Nazareno AG, Dick CW, Lohmann LG (2017) Wide but not impermeable: Testing the riverine barrier hypothesis for an Amazonian plant species. *Mol Ecol* 26: 3636 – 3648.

Alter ES, Munshi-South J, Stiassny MLJ (2017) Genome-wide SNP data reveal cryptic phylogeographic structure and microallopatric divergence in a rapids-adapted clade of cichlids from the Congo River. *Mol Ecol* 26: 1401–1419.

Altmann A, Weber P, Bader D, Preuß M, Binder EB and Müller-Myhsok B (2012) A beginners guide to SNP calling from high-Throughput DNA-sequencing data. *Hum Genet* 131:1541–1554. doi: 10.1007/s00439-012-1213-z

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L and Lander ES (2000) An SNP

map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513–6. doi: 10.1038/35035083

Amoutzias GD, Giannoulis T, Moutou KA, Psarra AMG, Stamatis C, Tsipourlianos A, Mamuris Z (2016) SNP Identification through Transcriptome Analysis of the European Brown Hare (*Lepus europaeus*): Cellular Energetics and Mother's Curse. *Plos One* DOI:10.1371/journal.pone.0159939

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA and Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*. doi: 10.1371/journal.pone.0003376

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant J* 51, 910–918

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nat Genetics*. 40 (3): 340–345.

Beaumont MA and Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 13:969–980. doi: 10.1111/j.1365-294X.2004.02125.x

Beaumont MA and Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proc R Soc B Biol Sci* 263:1619–1626. doi: 10.1098/rspb.1996.0237

Beerli P and Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152:763–773. doi: 10.1073/pnas.081068098

Besnier F and Glover KA (2013) ParallelStructure: A R Package to Distribute Parallel Runs of the Population Genetics Program STRUCTURE on Multi-Core Computers. *PLoS One*. doi: 10.1371/journal.pone.0070651

Bhardwaj A, Dhar YV, Asif MH, Bag SK (2016) In Silico identification of SNP diversity in cultivated and wild tomato species: insight from molecular simulations. *Scientific Reports* 6:38715

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R and Moritz C (2013) Unlocking the vault: Next-generation museum population genomics. *Mol Ecol* 22:6018–6032. doi: 10.1111/mec.12516

Biton I, Doron-Faigenboim A, Jamwal M, Mani Y, Eshed R, Rosen A, Sherman A, Ophir R, Lavee S, Avidan B, Ben-Ari G (2015) Development of a large set of SNP markers for assessing phylogenetic relationships between the olive cultivars composing the Israeli olive germplasm collection. *Mol Breeding* 35:107

Black WC, Baer CF, Antolin MF and DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annu Rev Entomol* 46:441–469. doi: 10.1146/annurev.ento.46.1.441

Boutet G, Alves Carvalho S, Falque M, Peterlongo P, Lhuillier E, Bouchez O, Lavaud C, Pilet-Nayel ML, Rivière N, Baranger A (2016) SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. *BMC Genomics* 17:121. doi: 10.1186/s12864-016-2447-2.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455–457. doi: 10.1038/368455a0

Brookes AJ (1999) The essence of SNPs. *Gene* 234.2: 177-186.

Bulhões AC, Goldani HA, Oliveira FS, Matte US, Mazzuca RB, Silveira TR. (2007). Correlation between lactose absorption and the C/T-13910 and G/A-22018 mutations of the lactase-phlorizin hydrolase (LCT) gene in adult-type hypolactasia. *Braz J Med Biol Res* 40: 1441-1446 (2007).

Cavalli-Sforza LL and Edwards a WF (1967) Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet* 19:233–257. doi: 10.1073/pnas.85.16.6002

Celik I. Bodur S, Frary A, Doganlar S (2016) Genome-wide SNP discovery and genetic linkage map construction in sunflower (Helianthus annuus L.) using a genotyping by sequencing (GBS) approach. *Mol Breeding* 36:133

Cheng J, Qin C, Tang X, Zhou H, Hu Y, Zhao Z, Cui J, Li B, Wu Z, Yu J, Hu K (2016) Development of a SNP array and its application to genetic mapping and diversity assessment in pepper (Capsicum

spp.). *Scientific Reports* 6:33293

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, and Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet* 3, 19.

Cornuet JM, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin JM and Estoup A (2014) DIYABC v2.0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* 30:1187–1189. doi: 10.1093/bioinformatics/btt763

Cousina E, Geninb E, Macea S, Ricarda S, Chansaca C, del Zompoc M, Deleuze JF (2003) Association Studies in Candidate Genes: Strategies to Select SNPs to Be Tested. Hum Hered 2003;56:151–159

Cruz VP, Vera M, Pardo BG, Taggart J, Martinez P, Oliveira C, Foresti F (2017) Identification and validation of single nucleotide polymorphisms as tools to detect hybridization and population structure in freshwater stingrays. *Mol Ecol Res* 17: 550–556.

Csilléry K, François O and Blum MGB (2012) Abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol Evol* 3:475–479. doi: 10.1111/j.2041-210X.2011.00179.x

Davey JW, Hohenlohe P a, Etter PD, Boone JQ, Catchen JM and Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12:499–510. doi: 10.1038/nrg3012

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–8. doi: 10.1038/ng.806

De Wit P (2016) SNP Discovery Using Next Generation Transcriptomic Sequencing. Methods Mol Biol 1452:81-95. doi: 10.1007/978-1-4939-3774-5_5

Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. *Plant Physiol*, 124, 1483–92.

Durand EY, Patterson N, Reich D and Slatkin M (2011) Testing for ancient admixture between closely related populations. *Mol Biol Evol* 28:2239–2252. doi: 10.1093/molbev/msr048

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

Dyer RJ (2009) GeneticStudio: A suite of programs for spatial analysis of genetic-marker data. *Mol Ecol Resour* 9:110–113. doi: 10.1111/j.1755-0998.2008.02384.x

Dyer RJ, Nason JD and Garrick RC (2010) Landscape modelling of gene flow: Improved power using conditional genetic distance derived from the topology of population networks. *Mol Ecol* 19:3746–3759. doi: 10.1111/j.1365-294X.2010.04748.x

Earl DA and vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 4:359–361. doi: 10.1007/s12686-011-9548-7

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev* 5:435–445. doi: 10.1038/nrg1348

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM (2010) Resolving postglacial phylogeography using high-throughput sequencing. *PNAS* 37:16196–16200

Etter PD, Preston JL, Bassham S, Cresko WA and Johnson EA (2011) Local de novo assembly of rad paired-end contigs using short sequencing reads. *PLoS One*. doi: 10.1371/journal.pone.0018561

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC and Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet*. doi: 10.1371/journal.pgen.1003905

Excoffier L and Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10:564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Faubet P and Gaggiotti OE (2008) A new Bayesian method to identify the environmental factors that

influence recent migration. *Genetics* 178:1491–1504. doi: 10.1534/genetics.107.082560

Fouet C, Kamdem C, Gamez S, White BJ (2017) Extensive genetic diversity among populations of the malaria mosquito Anopheles moucheti revealed by population genomics. Infection, Genetics and Evolution 48: 27–33.

Francis RM (2016) POPHELPER: An R package and web app to analyse and visualise population structure. *Mol Ecol Resour* n/a-n/a. doi: 10.1111/1755-0998.12509

François O, Waits LP (2016) Clustering and Assignment Methods in Landscape Genetics. In: Landscape Genetics: Concepts, Methods, Applications (eds Balkenhol N, Cushman SA, Storfer, AT, Waits, LP), John Wiley and Sons, Ltd, Chichester, UK. doi: 10.1002/9781118525258.ch07, 2016.

Frichot E, Schoville SD, Bouchard G and François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. Mol *Biol Evol* 30:1687–1699. doi: 10.1093/molbev/mst063

Glenn TC and Faircloth BC (2016) Capturing Darwin's dream. *Mol Ecol Resour* 16:1051–1058. doi: 10.1111/1755-0998.12574

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27:182–9. doi: 10.1038/nbt.1523

Goudet J (2013) FSTAT: a computer program to calculate F-Statistics. *J Hered* 104:586–590. doi: 10.1093/jhered/est020

Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ Jr (2011) High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biol* 11:65

Gronau I, Hubisz MJ, Gulko B, Danko CG and Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* 43:1031–1034. doi: 10.1038/ng.937

Guillot G (2011) On the Informativeness of Dominant and Co-Dominant Genetic Markers for Bayesian Supervised Clustering. *Open Stat Probab J* 3:7–12. doi: 10.2174/1876527001103010007

Günther T and Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics* 195:205–220. doi: 10.1534/genetics.113.152462

Gutenkunst RN, Hernandez RD, Williamson SH and Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. doi: 10.1371/journal.pgen.1000695

Hardy OJ and Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2:618–620. doi: 10.1046/j.1471-8278

Hattemer HH (1982) Genetic distance between populations. TAG *Theor Appl Genet* 62:219–223. doi: 10.1007/BF00276242

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* 59:1633–1638. doi: Doi 10.1554/05-076.1

Hehir-Kwa J, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, Veltman JA (2007) Genome-wide copy number profiling on high-density BAC, SNP and oligonucleotide microarrays: a platform comparison based on statistical power analysis. *DNA Res*, 14, 1–11.

Hendre PS, Kamalakannan R, Varghese M (2012) High-throughput and parallel SNP discovery in selected candidate genes in Eucalyptus camaldulensis using Illumina NGS platform. *Plant Biotechnology Journal* 10:646–656

Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. *Curr Opin Genet Dev* 16:592–596. doi: 10.1016/j.gde.2006.10.005

Hey J and Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A* 104:2785–2790. doi: 10.1073/pnas.0611164104

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ,

Hannon GJ et al. (2007) Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39:1522–7. doi: 10.1038/ng.2007.42

Howell WM, Jobs M, Gyllensten U, Brookes AJ (1999). Dynamic allelespecific hybridization. *Nat Biotechnol*, 17, 87–8

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338. doi: 10.1093/bioinformatics/18.2.337

Ipek A, Yılmaz K, Sıkıcı P, Tangu NA, Oz AT, Bayraktar M, Ipek M, Gülen H (2016) SNP Discovery by GBS in Olive and the Construction of a High-Density Genetic Linkage Map. *Biochem Genet* DOI 10.1007/s10528-016-9721-5

Jeong IS, Yoon UH, Lee GS, Ji HS, Lee HJ, Han CD, Hahn JH, An G, and Kim TH (2013) SNP-based analysis of genetic diversity in anther derived rice by whole genome sequencing. *Rice* 6, 6.

Jobling M, Hurles M, Tyler-Smith C (2013) Human evolutionary genetics: origins, peoples & disease. Garland Science.

Jolliffe, IT (1986) Principal Component Analysis. Springer Verlag, New York.

Jombart T, Ahmed I (2011) adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070–3071. doi: 10.1093/bioinformatics/btr521

Jombart T, Devillard S (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. doi: doi:10.1186/1471-2156-11-94

Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* (Edinb) 101:92–103. doi: 10.1038/hdy.2008.34

Jost L (2008) GST and its relatives do not measure differentiation. *Mol Ecol* 17:4015–4026. doi: 10.1111/j.1365-294X.2008.03887.x

Kahl G, Mast A, Tooke N, Shen R, Boom D. (2005). Single nucleotide Polymorphisms: Detection Techniques and Their Potential for Genotyping and Genome Mapping. In: The Handbook of Plant Genome Mapping: Genetic and Physical Mapping.

Kamvar ZN, Brooks JC and Grünwald NJ (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front Genet*. doi: 10.3389/fgene.2015.00208

Kandpal RP, Kandpal G and Weissman SM (1994) Construction of libraries enriched for sequence repeats and jumping clones, and hybridization selection for region-specific markers. *Proc Natl Acad Sci* U S A 91:88–92. doi: 10.1073/pnas.91.1.88

Kern A and Hey J (2016) Exact calculation of the joint allele frequency spectrum for generalized isolation with migration models. bioRxiv 65003. doi: 10.1101/065003

Kim JE, Oh SK, Lee JH, Lee BM, and Jo SH (2014) Genome-wide SNP calling using next generation sequencing data in tomato. *Mol Cells* 37, 36-42.

Knaus BJ and Grünwald NJ (2016) vcfr: A package to manipulate and visualize variant call format data in r. *Mol Ecol Resour*. doi: 10.1111/1755-0998.12549

Knowles LL and Maddison WP (2002) Statistical phylogeography. *Mol Ecol* 11:2623–2635. doi: 10.1046/j.1365-294X.2002.01637.x

Kuhner MK (2006) LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22:768–770. doi: 10.1093/bioinformatics/btk051

Kumar S, Banks TW and Cloutier S (2012) SNP discovery through next-generation sequencing and its applications. *Int J Plant Genomics*. doi: 10.1155/2012/831460

Leaché AD, Harris RB, Rannala B and Yang Z (2014) The influence of gene flow on species tree estimation: A simulation study. *Syst Biol* 63:17–30. doi: 10.1093/sysbio/syt049

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Li W-H, Gojobori T and Nei M (1981) Pseudogenes as a paradigm of neutral evolution. *Nature* 292:237–239. doi: 10.1038/292237a0

Liu, Chia-Chan, et al. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology* 9.2: 106-118 (2013).

Luikart G, England PR, Tallmon D, Jordan S and Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* 4:981–994. doi: 10.1038/nrg1226

Lynch et al. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am Fam Physician* 76: 391-6 (2007).

Macdonald SJ (2007) Genotyping by Oligonucleotide Ligation Assay (OLA).

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J and Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7:111–8. doi: 10.1038/nmeth.1419

Martínez-Arias R, Calafell F, Mateu E, Comas D, Andrés A and Bertranpetit J (2001) Sequence variability of a human pseudogene. *Genome Res* 11:1071–1085. doi: 10.1101/gr.GR-1677RR

McGuigan FEA, Ralston SH (2002) Single nucleotide polymorphism detection: allelic discrimination using Taqman. *Psychiat Genet*, 12, 133–6

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. doi: 10.1101/gr.107524.110

McRae BH (2006) Isolation by resistance. *Evolution* (N Y) 60:1551–1561. doi: 10.1111/j.0014-3820.2006.tb00500.x

Meirmans PG and Hedrick PW (2011) Assessing population structure: FST and related measures. *Mol Ecol Resour* 11:5–18. doi: 10.1111/j.1755-0998.2010.02927.x

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248 .

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* 37(1/2):17–23

Naduvilezhath L, Rose LE and Metzler D (2011) Jaatha: A fast composite-likelihood approach to estimate demographic parameters. *Mol Ecol* 20:2709–2723. doi: 10.1111/j.1365-294X.2011.05131.x

Nachman MW (2001) Single-nucleotide polymorphisms and recombination rate in humans". Trends in Genetics. 17 (9): 481–485.

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc *Nat Acad Sci* 70:3321–3323. doi: 10.1073/pnas.70.12.3321

Nielsen R, Paul JS, Albrechtsen A and Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 12:443–51. doi: 10.1038/nrg2986

Newton CR, Summers C, Heptinstall LE, Lynch JR, Finniear RS, Ogilvie D, Smith JC, Markham AF (1991) Genetic analysis in cystic fibrosis using the amplification refractory mutation system (ARMS): the J3.11 MspI polymorphism. *J Med Genet*, 28, 248–51

Panitz F, Stengaard H, Hornshøj H, Gorodkin J, Hedegaard J, Cirera S et al. (2007) SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation. *Bioinformatics* 23, 387–391

Pardo-Diaz, C., Salazar, C. and Jiggins, C. D. (2015), Towards the identification of the loci of adaptive evolution. *Methods Ecol Evol* 6: 445–464. doi:10.1111/2041-210X.12324

Paradis E (2010) Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* 26:419–420. doi: 10.1093/bioinformatics/btp696

Pavlidis P, Laurent S, Stephan W (2010) msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. Mol Ecol Res 10: 723–727. doi:10.1111/j.1755-0998.2010.02832.x

Pellegrino I, Boatti L, Cucco M, Mignone F, Kristensen TN, Mucci N, Randi E, Ruiz-Gonzalez A, Pertoldi C (2016) Development of SNP markers for population structure and phylogeography characterization in little owl (Athene noctua) using a genotyping- by-sequencing approach. *Conservation Genet Resour* 8:13–16

Pembleton LW, Cogan NOI and Forster JW (2013) StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol Ecol Resour* 13:946–952. doi: 10.1111/1755-0998.12129

Perkel J (2008) SNP genotyping: Six technologies that keyed a revolution. *Nat Methods* 5:575–575. doi: 10.1038/nmeth0608-575b

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA and Boyce-Jacino M (1999) Mining SNPs from EST databases. *Genome Res* 9:167–174. doi: 10.1101/gr.9.2.167

Podder M, Ruan J, Tripp BW, Chu ZE and Tebbutt SJ (2008) Robust SNP genotyping by multiplex PCR and arrayed primer extension. *BMC Med Genomics* 1:5. doi: 10.1186/1755-8794-1-5

Poland JA and Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Genome* J 5:92–102. doi: 10.3835/plantgenome2012.05.0005

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012a) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by Sequencing Approach. *Plos One* 2: e32253

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink J-L (2012b) Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant* Genome 5:103–113

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F et al. (2007) Multiplex amplification of large sets of human exons. *Nat Methods* 4:931–936. doi: 10.1038/nmeth1110

Prince JA, Feuk L, Howell WM, Jobs M, Emahazion T, Blennow K and Brookes AJ (2001) Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): Design criteria and assay validation. *Genome Res* 11:152–162. doi: 10.1101/gr.150201

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–59. doi: 10.1111/j.1471-8286.2007.01758.x

Puechmaille SJ (2016) The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Mol Ecol Resour* 16:608–627. doi: 10.1111/1755-0998.12512

R Development Core Team (2016) R: A Language and Environment for Statistical Computing. R Found Stat Comput Vienna Austria 0:{ISBN} 3-900051-07-0. doi: 10.1038/sj.hdy.6800737

Raj A, Stephens M and Pritchard JK (2014) FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* 197:573–589. doi: 10.1534/genetics.114.164350

Raymond M and Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* 86:248–249. doi: 10.1111/j.0021-8790.2004.00839.x

Reverter A and Fortes MRS (2013) Genome-Wide Association Studies and Genomic Prediction. Methods *Mol Biol.* doi: 10.1007/978-1-62703-447-0

Robicheau BM, Susko E, Harrigan AM, Snyder M (2017) Ribosomal RNA Genes Contribute to the Formation of Pseudogenes and Junk DNA in the Human Genome. *Genome biol evol* 9(2): 380-397.

Robinson JD, Bunnefeld L, Hearn J, Stone GN and Hickerson MJ (2014) ABC inference of multi-population divergence with admixture from unphased population genomic data. *Mol Ecol* 23:4458–4471. doi: 10.1111/mec.12881

Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA (1986). Analysis of enzymatically amplified beta globin and HLA-DQalpha DNA with allele-specific oligonucleotide probes. *Nature* 324, 163–6.

Sankarasubramanian J, Vishnu US, Gunasekaran P, Rajendhran J (2016) A genome-wide SNP-based phylogenetic analysis distinguishes different biovars of Brucella suis. Infection, *Genetics and*

Evolution 41:213–217

Sathya B, Dharshini AP, Kumar GR (2015) NGS meta data analysis for identification of SNP and INDEL patterns in human airway transcriptome: A preliminary indicator for lung cancer. Applied & Translational Genomics 4: 4–9

Sobrino B, Brión M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci Int* 154.2: 181-194.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457–462. doi: Article

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP and Slate J (2010) Adaptation genomics: The next generation. *Trends Ecol Evol* 25:705–712. doi: 10.1016/j.tree.2010.09.002

Taberlet P, Luikart G and Waits LP (1999) Noninvasive genetic sampling: Look before you leap. *Trends Ecol Evol* 14:323–327. doi: 10.1016/S0169-5347(99)01637-7

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 3:1-7

Tang W, Wu1 T, Ye J, Sun J, Jiang Y, Yu J, Tang J, Chen G, Wang C, Wan J (2016) SNP-based analysis of genetic diversity reveals important alleles associated with seed size in rice. *BMC Plant Biol* 16:93

Tautz D (1989) Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* 17:6463–6471. doi: 10.1093/nar/17.16.6463

Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Margulies EH, Green ED et al. (2010) Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 20:1420–1431. doi: 10.1101/gr.106716.110

The 1000 Genomes Project Consortium*. A global reference for human genetic variation, *Nature* 526, 68-74 (2015). doi:10.1038/nature15393.

Thomé MTC and Carstens BC (2016) Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proc Natl Acad Sci* 113:8010–8017. doi: 10.1073/pnas.1601064113

Torkamaneh D, Laroche J and Belzile F (2016) Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PLoS One*. doi: 10.1371/journal.pone.0161333

Valdisser PAMR, Pappas Jr. GJ, de Menezes IPP, Müller BSF, Pereira WG, Narciso MG, Brondani, Souza TLPO, Borba TCO, Vianello RP (2016) SNP discovery in common bean by restriction‐associated DNA (RAD) sequencing for genetic diversity and population structure analysis. *Mol Genet Genomics* 291:1277–1291

van Oeveren J and Janssen A (2009) Mining SNPs from DNA sequence data; computational approaches to SNP discovery and analysis. *Methods Mol Biol* 578:73–91. doi: 10.1007/978-1-60327-411-1_4

van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJT (2007) Complexity Reduction of Polymorphic Sequences (CRoPS): a novel approach for largescale polymorphism discovery in complex genomes. *PLoS One*, 2, e1172

van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* 5, 247–252.

Vatsiou AI, Bazin E, Gaggiotti OE (2016) Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol ecol* 25.1: 89-103

Vignal A, Milan D, SanCristobal M and Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genet Sel Evol* 275–305. doi: 10.1051/gse

Visser C, Lashmar SF, Marle-Köster EV, Poli MA, Allain D (2016) Genetic Diversity and Population Structure in South African, French and Argentinian Angora Goats from Genome-Wide SNP Data.

PLoS One 11(5): e0154353. doi:10.1371/journal.pone.0154353

Wang IJ and Bradburd GS (2014) Isolation by environment. *Mol Ecol* 23:5649–5662. doi: 10.1111/mec.12938

Wegmann D, Leuenberger C, Neuenschwander S and Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11:116. doi: 10.1186/1471-2105-11-116

Whitlock MC and McCauley DE (1999) Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). *Heredity* (Edinb) 82 (Pt 2):117–125. doi: 10.1038/sj.hdy.6884960

Willing EM, Dreyer C and van Oosterhout C (2012) Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. *PLoS One*. doi: 10.1371/journal.pone.0042649

Wilson GA and Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163:1177–1191. doi: Article

Wright S (1943) Isolation by Distance. *Genetics* 28:114–138. doi: Article

Wright B, Morris K, Grueber CE, Willet CE, Gooley R, Hoog CJ, O'Meally D, Hamede R, Jones M, Wade C, Belov K (2015) Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics* 16:791

Zanger UM, Schwab M (2013) Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics* 138.1:103-141

Zhang X, Zhang H, Li L, Lan H, Ren Z, Liu D, Wu L, Liu H, Jaqueth J, Li B, Pan B, Gao S (2016) Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. *BMC Genomics* 17:697

Yuan Q, Zhou Z, Lindell SG, Higley D, Ferguson B, Thompson RC, Lopez JF, Suomi SJ, Baghal B, Baker M, Mash DC, Barr CS, Goldman D (2012) The rhesus macaque is three times as diverse but more closely equivalent in damaging coding variation as compared to the human. *BMC Genetics* 13:52