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

ORGANIZADORAS

Andreia Carina Turchetto Zolet

Doutora em Genética e Biologia Molecular
Programa de Pós-Graduação em Genética e Biologia Molecular – PPGBM, Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul
aturchetto@gmail.com

Caroline Turchetto

Doutora em Genética e Biologia Molecular
Depto. de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul
carolineturchetto@gmail.com

Camila Martini Zanella

Doutora em Genética e Biologia Molecular
National Institute of Agricultural Botany, Cambridge, Reino Unido
milamzanella@gmail.com

Gisele Passaia

Doutora em Genética e Biologia Molecular
gisapassaia@gmail.com

REVISORES TÉCNICOS

Ana Lídia Variani Bonato
Empresa Brasileira de Pesquisa Agropecuária

Camila Campos Mantello
National Institute of Agricultural Botany, Reino Unido

Gabriel de Menezes Yazbeck
Universidade Federal de São João del-Rei

Fernanda Witt Cidade
Instituto Federal Catarinense

Graciela da Rocha Sobierajski
Centro APTA-Frutas/Instituto Agronômico de Campinas

Jucelene Fernandes Rodrigues
Universidade de São Paulo

Tiago Campos Pereira
Universidade de São Paulo

CAPA

A arte da capa desta obra foi desenhada a mão livre por Daniela Quinsani, mestre em Genética e Biologia Molecular (Programa de Pós-Graduação em Genética e Biologia Molecular – PPGBM/UFRGS). A arte representa a evolução das tecnologias na utilização dos marcadores moleculares, desde o uso de eletroforese em gel para a genotipagem de marcadores como o RFLP até as abordagens baseadas em sequenciamento em larga escala.

Lista de abreviaturas, siglas, acrônimos e símbolos

AFLP (*Amplified Fragment Length Polymorphism*): Polimorfismo de comprimento de fragmentos amplificados.

BLAST (*Basic Local Alignment Search Tool*): Ferramenta básica de pesquisa de alinhamento local

BSA (*bovine serum albumin*): Albumina de soro bovino

CAPS (*Cleaved Amplified Polymorphic Sequences*): Sequências Polimórficas Amplificadas Alivadas

dCAPS (*Derived Cleaved Amplified Polymorphic Sequences*): Sequências Polimórficas Amplificadas Clivadas Derivadas

CGH (*Comparative Genome Hibridization*): Hibridização Genômica Comparativa

CODIS (*Combined DNA Index System):* Sistema combinado de índices de DNA

DAPC (*Discriminant Analysis of Principal Components):* Análise discriminante de componentes principais

DArT (*Diversity arrays technology*)

ddNTP (*dideoxynucleotide*): Dideoxinucleotídeos

dNTP (*deoxynucleotide*): Deoxinucleotídeos

DNA (*desoxyribonucleic acid*): Ácido desoxiribonucleico

DMSO (*dimethyl sulfoxide*): Sulfóxido de dimetilo ou Dimetilsulfóxido

EST (*Exressed Sequence Tags*): Sequências Alvos Expressas

GBS (Genotyping by sequencing): Genotipagem por Sequenciamento

InDels (*insertion or deletion*): Inserções e/ou deleções de uma base ou um segmento de DNA

ISSR (*Inter-simple sequence repeats*): Entre sequências simples repetidas

IUPAC (*International Union of Pure and Applied Chemistry*): União Internacional de Química Pura e Aplicada

NGS (*Next Generation Sequencing*): Sequenciamento de nova geração

PCA (*Principal Component Analysis):* Análises de componentes principais

PCR (*Polymerase Chain Reaction*): Reação em cadeia da polimerase

PIC (*Polymorphism Information Content*): Conteúdo de informação polimórfica

P5CS *(Pyrroline-5-carboxylate synthase):* Pirrolina-5-carboxilato sintase

QTL (*Quantitative Trait Loci*): Lócus de Caracteres Quantitativos

RFLP (*Restriction Fragment Length Polymorphism*): Polimorfismo de comprimento de fragmento de restrição

RAPD (*Random Amplified Polymorphism DNA*): DNA polimórfico amplificado ao acaso

RAD-seq (*Restriction-site associated DNA sequencing*): Sequenciamento de fragmentos de DNA associados a sítios de restrição

RRLs (*Reduced-representation libraries*): Bibliotecas de representação reduzida

RNA (*Ribonucleic acid*): Ácido Ribonucleico

RNA-seq (*RNA sequencing*): Sequenciamento de RNA

RST (*Restriction Site Tagged Microarrays*): Microarranjo de sítio de restrição marcado

sDNA (*single-strand DNA*): DNA em fita simples

sPCA (*spatial Princial Components Analysis):* Análise espacial de componentes principais

SSR (*Simple Sequence Repeats*): Sequências simples repetidas

SSLP (*Simple Sequence Length Polymorphism*): Polimorfismos de comprimento de sequências simples

STRs (*Short Tandem Repeats*): Sequências curtas repetidas em tandem

SNPs (*Single Nucleotide Polymorphism*): Polimorfismo de nucleotídeo único

UPOV (União para Proteção das Obtenções Vegetais)

VNTRs (*Variable Number of Tandem Repeats*): Números variáveis de repetições em tandem

Kb - Kilo base

pb - pares de bases

ng - nanogramas

Símbolos

^{32}P - fósforo radioativo

λ - Lambda

ϕ - Phi

α -Alfa

μ - micro

Δ - Delta

SUMÁRIO

Prefácio

A detecção e a análise de polimorfismos genéticos são de interesse em diversas áreas, pois podem nos auxiliar a compreender a base molecular de vários aspectos biológicos. Marcadores moleculares são definidos como um segmento específico de DNA representativo das diferenças ao nível do genoma que permitem fazer inferências diretas sobre a diversidade genética e inter-relações entre os organismos ao nível do DNA. A publicação de Botstein et al. (1980) sobre a construção de mapas genéticos usando polimorfismo de comprimento de fragmentos de restrição (RFLP) foi a primeira técnica de marcador molecular relatada para a detecção de polimorfismos de DNA. Desde então, diversas técnicas foram desenvolvidas e têm agido como ferramentas versáteis em diversas áreas.

As técnicas para estudos com marcadores moleculares têm sido aprimoradas com os progressos fenomenais das metodologias que permitem a identificação de marcadores em escala genômica, tais como o surgimento de diferentes plataformas de sequenciamento de alto desempenho (hoje denominadas de plataformas de *Next Generation Sequencing* – NGS). Estas técnicas avançadas são uma fusão das características vantajosas de várias técnicas básicas, bem como a incorporação de modificações na metodologia para aumentar a sensibilidade e resolução na identificação e genotipagem. Tais tecnologias, aliadas às análises de bioinformática e estatística, tornaram possível identificar e genotipar diferentes indivíduos em um único passo para alguns tipos de marcadores moleculares.

O crescente avanço das técnicas de marcadores moleculares permitiu a geração de dados que estão disponíveis através de uma vasta literatura em forma de artigos científicos. Na tentativa de demonstrar uma evolução na utilização de marcadores moleculares paralelamente ao surgimento de diferentes metodologias de sequenciamento de alto desempenho, o presente livro traz um histórico sobre marcadores moleculares, inserindo as atualidades sobre as novas metodologias de identificação e genotipagem dos marcadores mais utilizados atualmente em diversas áreas. Esta obra reúne informações detalhadas das metodologias utilizadas para a identificação e genotipagem dos mesmos, métodos de análise e aplicações.

Esse livro oferece uma excelente introdução, visão geral e aplicação dos principais marcadores moleculares para estudantes de graduação, pós-graduação e pesquisadores que tenham interesse mais aprofundado sobre o assunto. As bases técnicas de uma série de diferentes marcadores moleculares são descritas em detalhe ao longo dos capítulos, o que é de extrema importância no momento da escolha do tipo de marcador e qual o mais apropriado para cada estudo em particular. Além disso, algumas metodologias de análises são descritas para cada marcador focando em diferentes aplicações. Nesta obra foi abordado um passo a passo de como proceder com dados de SNPs oriundos de metodologias que utilizaram plataformas de sequenciamento de alto desempenho.

Boa leitura!

As autoras

Capítulo 1

Marcadores genéticos baseados em DNA

Dra. Caroline Turchetto, Profª. Dra. Andreia Carina Turchetto-Zolet,
Dra. Gisele Passaia, Dra. Camila Martini Zanella

Um marcador genético é qualquer caráter visível ou um fenótipo que de alguma forma seja analisável, para o qual os alelos em *loci* individuais segregam de uma maneira mendeliana, tais como as características visíveis das ervilhas estudadas por Mendel. Os marcadores genéticos morfológicos foram os primeiros utilizados, e são ainda hoje, a base do melhoramento genético convencional, em que características desejáveis são selecionadas nos genitores para os cruzamentos. Os marcadores genéticos bioquímicos, tais como os terpenos e as isoenzimas, foram marcadores genéticos utilizados antes do surgimento dos marcadores moleculares. Esses marcadores bioquímicos foram aplicados a uma série de estudos. As primeiras moléculas a serem utilizadas como marcadores genéticos bioquímicos foram metabólitos secundários tais como antocianinas e compostos fenólicos, usados para distinguir entre diferentes variedades de plantas (Grover e Sharma, 2014). Marcadores enzimáticos (Alozimas) tiveram grande importância, apesar de apresentarem baixo grau de polimorfismo, tendo sido utilizados por um curto período de tempo, quando passaram a ser substituídos por marcadores de DNA capazes de detectar uma maior variabilidade entre indivíduos.

Polimorfismos de DNA surgem como resultado de uma variação (mutação) e são geralmente referidos pelo tipo de mutação que os criou. O desenvolvimento e uso de marcadores moleculares para a detecção e exploração desses polimorfismos do DNA é um dos avanços mais significativos no campo da genética molecular. Isto se deve ao fato de que a utilização de marcadores localizados no DNA fornece um número praticamente ilimitado de informação distribuídos aleatoriamente ao longo do genoma. Por isso, os marcadores moleculares são altamente valorizados em diversas áreas que envolvem genética, biologia molecular e biotecnologia, tais como genética de populações, filogeografia, filogenia molecular, mapeamento genético, diagnósticos de doenças genéticas, testes de paternidade e para quem investiga as relações entre genótipo e fenótipo.

Com o avanço das técnicas de biologia molecular, a manipulação do DNA em laboratório tornou-se uma técnica recorrente. No início dos anos 1980, o uso de marcadores moleculares passou a integrar rotineiramente a análise do DNA das mais diversas espécies. Desde então, eles vêm sendo aperfeiçoados e evoluindo juntamente com os avaços nas técnicas de sequenciamento em larga escala (Figura 1.1). A presença de vários tipos de marcadores moleculares e diferenças nos seus princípios, metodologias e aplicações requerem uma consideração cuidadosa na escolha de um ou mais desses métodos de acordo com a aplicação, bem como com os recursos (técnico, financeiro, equipamentos) disponíveis em cada centro de pesquisa.

Os marcadores de DNA são divididos em três categorias principais: os baseados em hibridização, os baseados em PCR (Reação em cadeia da Polimerase – *Polymerase Chain Reaction*) e por fim, marcadores baseados em sequenciamento. Os marcadores também podem ser classificados de acordo com o tipo de herança alélica em dominantes e codominantes. Os marcadores codominantes possibilitam diferenciar indivíduos homozigotos e heterozigotos, o que não é possível com marcadores dominantes, para os

quais apenas é possível identificar a presença ou ausência de um determinado alelo. Esta característica é bem importante dependendo do objetivo do estudo; por exemplo, não é possível realizar análise de paternidade com marcador dominante.

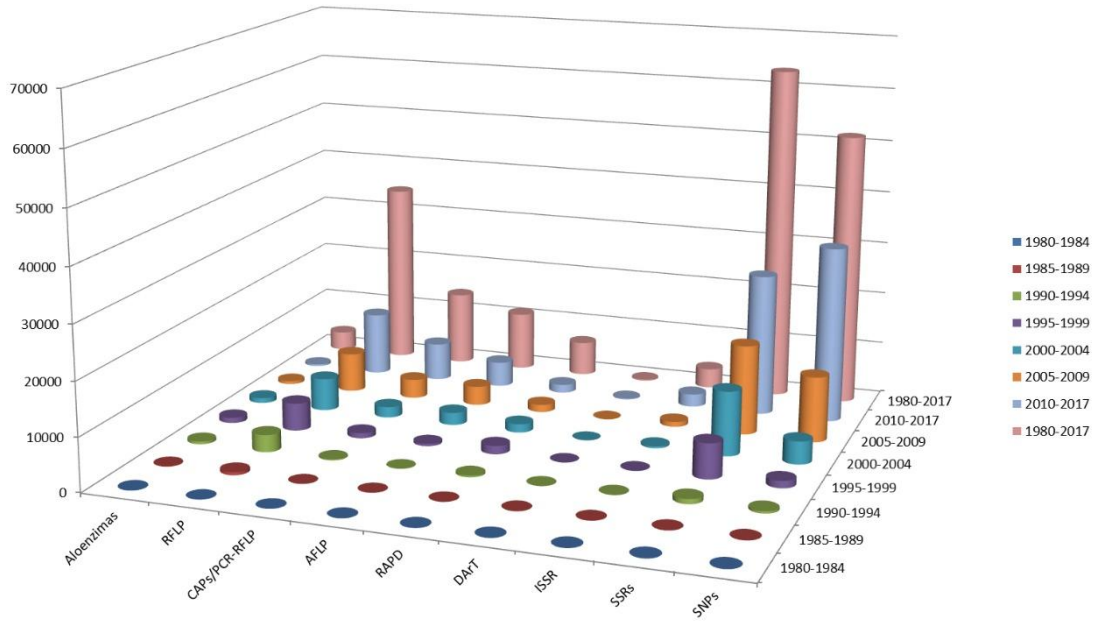

Figura 1.1 - Estimativa aproximada da popularidade dos marcadores moleculares nos últimos 37 anos. A construção deste gráfico foi baseada em dados obtidos através de buscas sistemáticas na base de dados do *Web of Science* usando como palavras chave o nome do marcador por extenso, sigla e o ano da publicação. Pesquisa realizada em Abril de 2017.

As técnicas de hibridização e PCR contribuíram para os avanços nos marcadores genéticos de DNA. O sistema de marcador genético baseado em hibridização é embasado na propriedade de pareamento de bases complementares. Este sistema permitiu o desenvolvimento de métodos que utilizam pequenos fragmentos de DNA como sondas para revelar polimorfismos apenas nas sequências homólogas à esta sonda. Um sistema de marcador genético derivado desta abordagem é a técnica de RFLP (*Restriction Fragment Length Polymorphism*) (Botstein et al., 1980) que foi o primeiro marcador genético baseado em DNA desenvolvido. Uma breve descrição do procedimento RFLP é mostrada no Capítulo 3. O polimorfismo deste marcador é baseado nos diferentes tamanhos de fragmentos gerados por enzimas de restrição. Este marcador foi utilizado para a construção do primeiro mapa molecular em humanos em 1980 (Botstein et al., 1980). Por ser codominante e identificar um *locus* específico, este tipo de marcador é informativo e capaz de discriminar genótipos individuais (Gebhardt et al., 1989). Este marcador foi e ainda hoje é utilizado em diferentes abordagens de estudo, tais como mapeamento genômico em plantas, marcação gênica, dinâmica populacional e relacionamento taxonômico (Figura 1.1), (Bonierbale et al., 1988; Yoshimura et al., 1992; Raybould et al., 1996; Jena e Kochert, 1991; Liu et al., 2016). Além disso, esta técnica foi também utilizada como base para a identificação e

isolamento de regiões repetitivas, tais como os microssatélites (por exemplo, Ali et al. 1986; ver Capítulo 6). Entretanto, esta técnica apresenta algumas limitações, como a necessidade de grandes quantidades de DNA (5-20 microgramas); o fato de espécies relacionadas e/ou indivíduos da mesma espécie poderem apresentar os mesmos alelos; o uso de radioatividade ou outras técnicas de coloração, além de tratar-se de uma abordagem intensiva em relação a trabalho e tempo, fatores estes que contribuíram para, em muitas aplicações, a escolha por outro tipo de marcador molecular. Entretanto, embora hoje em dia a utilização da técnica original do RFLP seja menor em comparação com outras classes de marcadores baseadas em PCR, ela foi a base de vários avanços subsequentes de marcadores relacionados (Chambers et al., 2014), como os CAPS (*Cleaved Amplified Polymorphic Sequences*), os quais também são chamados de marcadores PCR-RFLP, utilizados com frequência para uma detecção rápida de polimorfismos em sequências conhecidas (Figura 1.1), como detalhado no Capítulo 3.

Diferentes técnicas para a análise de marcadores de DNA foram desenvolvidas a partir do surgimento da PCR, permitindo a amplificação de uma grande quantidade de uma sequência específica de DNA sem necessidade de clonagem, começando com apenas algumas moléculas da sequência alvo. Uma vantagem dos métodos de marcadores baseados em PCR sobre os métodos de marcadores baseados em hibridização é que o último requer o isolamento de grandes quantidades de DNA. Entre os marcadores baseados em PCR, podemos citar, por exemplo: RAPD (*Random Amplified Polymorphism DNA*), ISSR (*Inter-simple sequence repeats*), SSR (*Simple Sequence Repeats*), AFLP (*Amplified Fragment Length Polymorphism*).

A técnica de RAPD foi descrita em 1990 independentemente por dois grupos de pesquisa (Welsh e McClelland, 1990; Williams et al., 1990). Marcadores RAPD foram os primeiros marcadores desenvolvidos baseados em PCR. Nesta técnica é utilizado um único *primer* de geralmente 10 bases de uma sequência arbitrária com 60% ou mais de conteúdo GC. A amplificação ocorre quando o sítio do *primer* (iniciador ou oligonucleotideo) está presente no DNA alvo em orientação oposta dentro de aproximadamente 2000 bases. Assim, a correspondência entre os fragmentos amplificados de diferentes espécies pelos mesmos iniciadores RAPD é natural (Rieseberg, 1996). Os polimorfismos RAPD resultam da variação da sequência nos locais de anelamento do *primer* e/ou variação de comprimento na sequência alvo situada entre os locais de ligação do *primer*. O RAPD é uma técnica não radioativa que requer uma pequena quantidade de DNA (15-25 ng), a maior vantagem desta técnica em relação a técnica de RFLP, e pode ser realizada em poucas horas. Outra vantagem é o grande número de marcadores obtidos distribuídos ao longo do genoma. Entretanto, o uso deste marcador é limitante por diversas características, como, por exemplo, o aparecimento de bandas não parentais na progênie restringindo o uso para mapeamento molecular (Riedy et al., 1992). Além disso, esse marcador se comporta como marcador dominante e apresenta baixa reprodutibilidade entre diferentes laboratórios e experimentos, principalmente associados ao uso de baixa temperatura de anelamento, que pode resultar em uma baixa especificidade ou até no não anelamento do *primer*. Pode ser influenciado, por exemplo, pela utilização de diferentes reagentes, incluindo *Taq* DNA polimerases, ou diferenças na qualidade do DNA (Schierwater e Ender, 1993; Skroch e Nienhuis, 1995).

Outra técnica descrita por Meyer et al. (1993), a qual essencialmente combina os benefícios do RAPD (grande número de marcadores obtidos distribuído sobre o genoma), aliados ao aumento na reprodutibilidade e especificidade, são os marcadores ISSR (*Inter-simple sequence repeats*). ISSR é uma técnica baseada em SSR em que a amplificação é realizada com um único *primer* consistindo de várias repetições (motivo

de SSR) e ancorado geralmente com 2 a 4 nucleotídeos arbitrários. A reprodutibilidade decorre do fato de serem utilizados *primers* mais longos para amplificação por PCR em comparação com RAPD, e a utilização de temperatura de anelamento mais altas na PCR. Como no caso dos RAPDs, praticamente nenhum conhecimento prévio de sequência alvo é necessária para os ISSRs, podendo assim ser aplicados com facilidade em espécies não-modelos. A ancoragem de sequências de iniciadores com sequências não repetidas garante que a amplificação seja iniciada na mesma posição de nucleotídeo em cada ciclo (Zietkiewicz et al., 1994). As amplificações são visualizadas em gel de agarose ou poliacrilamida. A principal vantagem deste método é o fato de que este tipo de marcador não requer etapas demoradas e caras, apesar do fato de os ISSRs apresentarem herança dominante. Este marcador ainda é utilizado atualmente em estudos principalmente de descrição de diversidade genética, como, por exemplo, genótipos de milho (Muhammad et al., 2017), genótipos de banana (Silva et al., 2017), *Campomanesia phaea*, espécie nativa da floresta Atlântica (Santos et al., 2016), entre outros (Figura 1.1).

Os marcadores AFLP (*Amplified Fragment Lenght Polymorphism* ou Polimorfismo de Comprimento de Fragmentos Amplificados) são resultado da combinação entre as técnicas utilizadas nos marcadores tipo RFLP e RAPD, onde enzimas de restrição e amplificação por PCR são as bases da técnica. AFLP baseia-se na amplificação seletiva por PCR de fragmentos de restrição gerados por enzimas de restrição específicas (geralmente uma de corte raro e uma de corte frequente) e ligados a adaptadores oligonucleotídicos (Vos et al., 1995; ver Capítulo 4). Assim como marcador RAPD, AFLP também estão distribuídas ao longo do genoma e, assim, são adequados para a construção de mapas de ligação genética (Becker et al., 1995); contudo apresentando uma maior reprodutibilidade que os RAPDs (Savelkoul et al., 1999), tornando-os mais confiáveis para estudos de genética de populações, por exemplo. Por outro lado, marcadores AFLPs mostram uma herança dominante, sendo genotipados pela presença ou ausência das bandas/alelos. Essa técnica é ainda relativamente cara, requer intenso trabalho e conhecimento técnico. AFLP tem sido uma técnica popular para estudos de genética de populações e diversidade, bem como em estudos ecológicos e evolutivos. Um considerável número de técnicas variantes do AFLP tem sido reportado na literatura, utilizando modificação no protocolo de clivagem por enzimas de restrição. Um exemplo é a técnica descrita por Suazo e Hall (1999), em que uma única enzima de corte raro é utilizada na mesma etapa de ligação de adaptadores. Além disso, a técnica de AFLP foi base para descrição de metodologias de isolamento de outros marcadores, como por exemplo, os microssatélites (ver Capítulo 6) e SNPs (*Single Nucleotide Polymorphism*) por meio do sequenciamento dos fragmentos em plataformas de sequenciamento de alto rendimento (técnica denominada de RAD-seq; ver Capítulo 4).

Marcadores do tipo microssatélites (SSR – *Simple Sequence Repeats*) foram desenvolvidos pela primeira vez para uso em mapeamento genético em humanos (Litt e Luty, 1989; Weber e May, 1989) e desde então vêm sendo amplamente utilizados em diversas áreas da ciência (Figura 1.1). Este tipo de marcador é baseado na amplificação por PCR de regiões específicas do genoma utilizando um par de *primers locus* específico. Os microssatélites são abundantes no genoma, fáceis de automatizar, codominantes, multialélicos, robustos e reprodutíveis. Os padrões de polimorfismo exibidos pelos SSR são maiores do que qualquer outro sistema de marcador contemporâneo. O método de bibliotecas enriquecidas para busca por SSR tem sido o mais popular, sendo que diferentes métodos de enriquecimento foram descritos. Outros procedimentos alternativos para identificação de SSR foram relatados baseado na

técnica de AFLP e RAPD (Zane et al., 2002), como mencionado acima. Inicialmente as sequências das bibliotecas eram determinadas pelo método de Sanger para posterior identificação de motivos de SSR nessas sequências e a possibilidade de projeção de *primers* nas regiões flanqueadoras, requerendo um considerável investimento para o desenvolvimento. Microssatélites têm sido um dos marcadores mais beneficiados com os avanços em técnicas de sequenciamento em larga escala. Com a disponibilidade de dados de sequências em domínio público, a identificação de motivos de SSR a partir de dados genômicos, de sequências de unigenes e de sequências alvos expressas (EST - *Expressed Sequence Tags*) utilizando ferramentas de bioinformática tornou-se uma abordagem extremamente atrativa e conhecida (Sharma et al., 2007). A popularidade dessa abordagem levou ao desenvolvimento de um número de *softwares* para identificação dos motivos de SSR, bem como predizer o potencial polimórfico dos microssatélites identificados. Para organismos não-modelos, foram também reportadas tentativas generalizadas de transferabilidade de marcadores de espécies taxonomicamente relacionadas. O advento da NGS (*Next Generation Sequencing* ou Sequenciamento de Nova Geração) simplificou muito o processo de isolamento de microssatélites, facilitando o desenvolvimento de alto rendimento de marcadores mesmo em espécies não-modelo em curto espaço de tempo e bom custo-benefício (Csenscics et al., 2010). Por exemplo, um estudo completo utilizando NGS pode revelar 100.000 *loci* de microssatélites em um genoma eucariótico (Abdelkrim et al., 2009) (ver Capítulo 6).

A identificação de uma sequência genômica específica analisada dentre vários indivíduos provê valiosa informação genética, pois fornece a completa informação da região investigada e a possibilidade de utilização de modelos evolutivos de mutação para as sequências analisadas. Apesar de não ser um marcador no sentido estrito, a análise da sequência de DNA também deve ser incluída nesta discussão devido a longa história de utilização em genética populacional e ainda tem sido a base para inferências filogeográficas e filogenéticas. Embora este método tenha sido inicialmente demorado e dispendioso, avanços recentes na tecnologia de sequenciamento permitem a análise de sequências de muitas regiões de DNA para muitos indivíduos (ver Capítulo 5).

Polimorfismo de nucleotídeo único (SNPs) são polimorfismos específicos a diferenças em uma única posição no genoma, um único nucleotídeo (substituição, deleção ou inserção). A maioria dos SNPs ocorre em regiões não codificadoras do genoma, um importante subconjunto corresponde a mutações em genes que estão associados a doenças ou outros fenótipos. A principal vantagem dos SNPs é o seu elevado potencial para uma análise automatizada de alto rendimento a custo moderado. O avanço nas plataformas de sequenciamento de alto rendimento tem contribuído para a descoberta de grande número de SNPs revolucionando projetos de avaliação da diversidade genética bem como estudos de associação genômica nos últimos anos (Figura 1.1). Outra possibilidade, muito importante para estudos com espécies não-modelo em que não se tem um genoma de referência disponível, é a identificação e genotipagem de SNP em um único passo, que pode ser combinado com a construção de uma biblioteca de representação genômica reduzida por meio da utilização de vários métodos disponíveis, como o sequenciamento em plataformas de alto rendimento (ver Capítulo 8).

Os marcadores tipo DArT (*Diversity arrays technology*) são baseados em hibridização para genotipar centenas de *loci* num único ensaio (Jaccoud et al., 2001) (Capítulo 7). Marcadores do tipo DArT geram impressões digitais de genoma marcando a presença versus ausência de fragmentos de DNA. Embora não muito popular comparado a outros marcadores (Figura 1.1), nos últimos anos, o DArT assumiu o

status de marcador altamente confiável, robusto e útil para a análise da diversidade genética, bem como o mapeamento genético usando estudos de ligação ou associação em uma variedade de culturas como o *Eucalyptus* (Sansaloni et al., 2010), trigo (Orabi et al., 2014), cevada (Lex et al., 2014), cenoura (Grzebellus et al., 2014), cana-de-açúcar (Aitken et al., 2014). A implementação das tecnologias de NGS ao método original de DArT, denominado DArTseq tem se demonstrado bastante eficiente e atualmente vem sendo utilizado em programas de melhoramento genético, como por exemplo, associação de características agronômicas em trigo (Mwadzingen et al., 2017).

Cada uma das tecnologias de marcadores moleculares tem suas vantagens e algumas limitações (Tabela 1.1). A escolha do marcador molecular depende tipicamente das questões biológicas que são abordadas, da quantidade de DNA disponível para o experimento, dos conhecimentos técnicos do investigador, das considerações monetárias e do equipamento disponível no laboratório. Além disso, uma série de outros fatores relacionados com o organismo alvo e a sua complexidade do genoma também desempenham um papel importante na seleção do marcador ou tecnologia a ser utilizada. Da mesma forma, a escolha do marcador deve ser adequada para atender os objetivos do estudo em questão.

Os marcadores moleculares proporcionaram ferramentas importantes para discriminar entre alelos. Em termos de genética clássica, as diferenças visíveis ou detectáveis no fenótipo são alelos. Em termos de genética molecular, diferentes sequências de DNA são denominadas alelos, que eventualmente também podem criar diferenças fenotípicas. Mesmo que os alelos detectados no nível do DNA não causem necessariamente variações fenotípicas, eles podem estar ligados a tais fenômenos.

O sequenciamento em plataformas de alto rendimento foi um marco importante e fundamental para a geração de dados moleculares de forma mais barata, rápida e em larga escala (ver Capítulo 2 para uma revisão). A partir de 2005 as tecnologias de sequenciamento em larga escala passaram a ser comercializadas e sua utilização para a identificação de marcadores moleculares em escala genômica passou a ser implementado rotineiramente no mundo todo. Essas tecnologias de sequenciamento de DNA permitem a triagem de *loci* múltiplos em muitos indivíduos simultaneamente. Estes métodos oferecem a vantagem de gerar informação explícita a partir de uma determinada região genômica levando frequentemente ao desenvolvimento simultâneo de milhares de marcadores adequados para a ampla gama de estudos.

Tabela 1.1. Comparações de diferentes marcadores moleculares em relação a algumas vantagens e limitações relacionadas a cada um deles.

Marcador	Vantagens	Limitações
RAPDs e derivados	• Produz um grande número de bandas que podem então ser usadas para caracterização individual; • Técnica simples e rápida.	• Baixa reprodutibilidade; • Principalmente dominante; • Difícil de analisar; • Comparações entre estudos é difícil.
CAPS	• Rápida identificação de genótipo fenótipo.	• Conhecimento prévio da sequência.
AFLPs	• Grande número de marcadores por reação; • Alta reprodutibilidade;	• Marcador do tipo Dominante (não identifica indivíduos

	• Rapidez; • Não há necessidade de conhecimento prévio do genoma.	homozigotos e/ou heterozigotos); • Homoplasia de tamanho das bandas geradas.
Microssatélites (SSR)	• Altamente informativo (multialélico); • Fácil de isolar; • Codominante; • Alta reprodutibilidade.	• Complexo comportamento mutacional; • Alelos nulos; • Comparações entre estudos requer preparação especial; • Espécie-especifico.
SNPs	• Baixa taxa de mutação; • Altamente abundante; • Novas abordagens analíticas estão sendo desenvolvidas atualmente; • Fácil comparação entre estudos.	• Substancial heterogeneidade da taxa entre sítios; • Baixo conteúdo informativo para um único SNP.
DArT	• Alta reprodutibilidade; • Alta acurácia; • Não há necessidade de conhecimento prévio do genoma; • Análises em paralelo; • Flexibilidade na aplicação.	• Essencialmente dominante.

Referências Bibliográficas

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2009) Fast, cost-effective development of species-specific microsatellite markers by genome sequencing. *Biotechniques* 46: 185–92.

Aitken KS, McNeil MD, Hermann S, et al. (2014). A comprehensive genetic map of sugarcane that provides enhanced map coverage and integrates high-throughput Diversity Array Technology (DArT) markers. *BMC Genomics* 15: 152.

Ali S, Muller CR, Epplen JT (1986) DNA fingerprinting by oligonucleotide probes specific for simple repeats. *Hum Genet* 74: 239–43.

Becker J, Heun M (1995) Mapping of digested and undigested random amplified microsatellite polymorphisms in barley. *Genome* 38: 991–8.

Bonierbale MW, Plaisted RL, Tanksley SD (1988) RFLP maps based on a common set of clones reveals models of chromosomal evolution in potato and tomato. *Genetics* 120: 1095–103.

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of genetic linkage map using restriction fragment length polymorphisms. *Am J Hum Genet* 32: 314–31.

Chambers GK, Curtis C, Millar CD, et al. (2014) DNA fingerprinting in zoology: past, present, future. *Investig Genet* 5: 3.

Csenscics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered dwarf bulrush (Typha minima) using next-generation sequencing technology. *J Heredity* 101: 789–93.

Gebhardt C, Ritter E, Debener T, et al. (1989) RFLP analysis and linkage mapping in Solanum tuberosum. *Theor Appl Genet* 78: 65–75.

Grzebellus D, Iorizzo M, Senalik D, et al. (2014). Diversity, genetic mapping, and signatures of domestication in the carrot (Daucus carota L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. *Mol Breed* 33: 625–37.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res* 29: e25.

Jena KK, Kochert G (1991) Restriction fragment length polymorphism analysis of CCDD genome species of the genus Oryza L. *Plant Mol Biol* 16: 837–9.

Lex J, Ahlemeyer J, Friedt W, Ordon F (2014). Genome-wide association studies of agronomic and quality traits in a set of German winter barley (Hordeum vulgare L.) cultivars using Diversity Arrays Technology. *J Appl Genet* 55: 295–305.

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 44: 397–401.

Liu Z, Furnier GR (1993) Comparison of allozyme, RFLP and RAPD markers for revealing genetic variation within and between trembling aspen and bigtooth aspen. *Theor Appl Genet* 87: 97–105.

Liu WL, Shih HC, Weng IS, et al. (2016) Characterization of Genomic Inheritance of Intergeneric Hybrids between Ascocenda and Phalaenopsis Cultivars by GISH, PCR-RFLP and RFLP. *PLoS One* 11(4): e0153512.

Meyer W, Mitchell TG, Freedman EZ, Vilgalys R (1993). Hybridization probes for conventional DNA fingerprinting used as single primers in the polymerase chain reaction to distinguish strains of *Cryptococcus neoformans*. *J Clin Microbiol* 31: 2274–80.

Mir RR, Hiremath PJ, Riera-Lizarazu O, Varshney RK (2013) *Evolving Molecular Marker Technologies in Plants: From RFLPs to GBS*. In Lübberstedt T, Varshney RK (Editors). Diagnostics in Plant Breeding. Springer: New York. Pp. 229-247.

Muhammad RW, Qayyum A, Ahmad MQ, et al. (2017) Characterization of maize genotypes for genetic diversity on the basis of inter simple sequences repeats. *Genetic and Molecular Resources* 16 (1.

Mwadzingeni L, Shimelis H, Rees DJG, Tsilo TJ (2017) Genome-wide association analysis of agronomic traits in wheat under drought-stressed and non-stressed conditions. *Plos One* 12 (2): e0171692.

Orabi J, Jahoor A, Backes G (2014) Changes in allelic frequency over time in European bread wheat (Triticum aestivum L.) varieties revealed using DArT and SSR markers. *Euphytica* 197: 447–62.

Raybould AF, Goudet J, Mago RJ, et al. (1996) Genetic structure of a linear population of Beta vulgaris ssp. maritime (sea beet) revealed by isozyme and RFLP analysis. *Heredity* 76: 111–17.

Riedy MF, Hamilton III WJ, Aquadro CF (1992) Excess of non parental bands in offspring from known primate pedigrees assayed using RAPD PCR. *Nucleic Acids Res* 20: 918.

Rieseberg LH (1996). Homology among RAPD fragments in interspecific comparisons. *Mol Ecol* 5: 99–105.

Sansaloni CP, Petroli CD, Carling J, et al. (2010). A high density diversity arrays technology (DArT) microarray for genome-wide genotyping in Eucalyptus. *Plant Methods* 6: 16.

Savelkoul PHM, Aarts HJM, de Haas J, et al. (1999) Amplified-fragment length polymorphism analysis: the state of an art. *J Clin Microbiol* 37: 3083–91.

Santos DN, Nunes CF, Setotaw TA, Pio R, Pasqual M, Cançado GM (2016) Molecular characterization and population structure study of cambuci: strategy for conservation and genetic improvement. *Genetic and Molecular Resources* 15(4).

Schierwater B, Ender A (1993) Different thermostable DNA polymerases may amplify different RAPD products. *Nucleic Acids Res* 19: 4647–8.

Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. *Trends Biotechnol* 25: 490–8.

Silva AV, Nascimento AL, Vitória MF, Rabbani AR, Soares AN, Lédo AS (2017) Diversity and genetic stability in banana genotypes in a breeding program using inter simple sequence repeats (ISSR) markers. *Genetic and Molecular Resources* (16 (1).

Skroch P, Neinhuis J (1995). Impact of scoring error and reproducibility of RAPD data on RAPD based estimates of genetic distance. *Theor Appl G*enet 91: 1086–91.

Suazo A, Hall HG (1999) Modification of the AFLP protocol applied to honey bee (Apis mellifera L.) DNA. *BioTechniques* 26: 704–9.

Vos P, Hogers R, Bleeker M, et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res, 23, 4407–14.

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* 44: 388—396.

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* 18 (24): 7213-7218.

Williams JGK, Kubelik AR, Livak KJ, et al. (1990) DNA polymorphisms *amplified* by arbitrary primers are useful as genetic markers. *Nucleic Acids Res* 18: 6531–5.

Yoshimura S, Yoshimura A, Saito A, et al. (1992) RFLP analysis of introgressed chromosomal segments in three near-isogenic lines of rice for bacterial blight resistance genes, Xa-1, Xa-3 and Xa-4. *Jpn J Genet* 67: 29–37.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Mol Ecol* 11: 1–16.

Zietkiewicz E, Rafalski A, Labuda D (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics* 20: 176–83.

Capítulo 2

Genômica e sequenciamento de nova geração

Dra. Ana Paula Christoff

Considerações gerais

A sequência nucleotídica que compõe o DNA genômico é um dos principais elos evolutivos entre os seres vivos que habitam o planeta Terra. O ácido desoxirribonucleico (DNA) possui uma estrutura complementar, em dupla-hélice, e é composto principalmente por nucleotídeos com quatro tipos de bases nitrogenadas, sendo duas purinas: adenina (A) e guanina (G), e duas pirimidinas: citosina (C) e timina (T) (Figura 2.1). A ordem em que estes diferentes nucleotídeos se combinam na molécula de DNA é o que determina a informação genética necessária para o funcionamento de um organismo. Assim como o DNA, o RNA (ácido ribonucleico) também possui a capacidade de armazenar e transmitir informações genéticas. A estrutura química destes ácidos nucleicos é muito importante para a estabilidade das moléculas, sendo que, ambas são formadas por nucleotídeos compostos por uma base nitrogenada, uma pentose e um ou mais grupamentos fosfato. O que diferencia o DNA do RNA é principalmente a estrutura de sua molécula de açúcar, que no DNA é uma desoxirribose e no RNA uma ribose. Além disso, uma das bases nitrogenadas pirimídicas que compõe o RNA difere do DNA, a uracila (U), ao invés da timina (T). Os nucleotídeos ligam-se entre si de forma covalente, onde o grupamento fosfato-5' de um nucleotídeo é ligado ao grupo hidroxila-3' do nucleotídeo seguinte, gerando uma ligação fosfodiéster no esqueleto das moléculas de DNA ou RNA. A interação entre duas cadeias complementares de ácidos nucleicos ocorre através de ligações de hidrogênio que se formam entre os grupos amino e carbonila das purinas e pirimidinas complementares. Enquanto o DNA é encontrado nas células em dupla fita, o RNA é encontrado em fita simples. Porém, uma fita de RNA geralmente apresenta a formação de estruturas secundárias devido à complementariedade de suas bases nitrogenadas, que acabam formando regiões de fita dupla.

A genética molecular ganhou força após a descrição da estrutura da molécula de DNA em 1953 por Watson e Crick (Box 2.1), que descreveram os padrões mais comuns de ligações de hidrogênio entre as bases nitrogenadas, onde A liga-se a T e G liga-se a C, formando interações específicas. A sequência em que estes nucleotídeos se arranjam nas moléculas de ácidos nucleicos é um dos principais alvos de estudos da genômica. Nas décadas subsequentes, entre 1970 e 1980, esta área teve avanços significativos com o surgimento das primeiras metodologias para amplificação e sequenciamento de DNA. Tais técnicas consistiam principalmente em métodos como o de degradação química (Maxam e Gilbert, 1977) ou o método Sanger com terminação de cadeia por dideoxinucleotídeos (Sanger et al., 1977). Desta forma, primariamente restrita a estudar pequenos fragmentos de DNA ou poucos genes, a genômica teve uma grande alteração de paradigmas com o aumento do conhecimento químico sobre os nucleotídeos e ácidos nucleicos, culminando no desenvolvimento de técnicas inovadoras para o sequenciamento de DNA em larga escala.

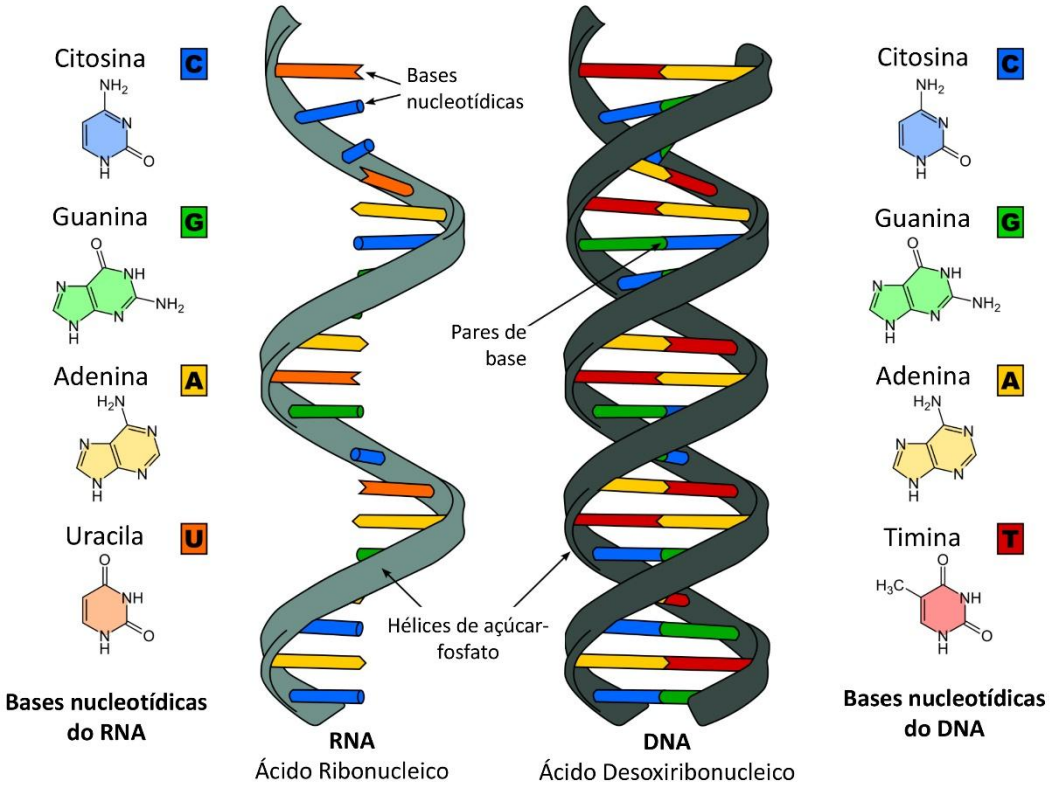

Figura 2.1 - Estrutura dos ácidos nucléicos, DNA e RNA, com suas bases nucleotídicas.

Entre as diversas técnicas de sequenciamento em larga escala que surgiram, destacam-se: o pirosequenciamento com detecção de pirofosfato, o sequenciamento por ligação da tecnologia SOLiD, a metodologia de semicondutores Ion e o sequenciamento por síntese na tecnologia Illumina. Ainda, novos métodos vêm sendo constantemente testados, aprimorados, ou ainda incorporados por outras tecnologias, tais como a tecnologia Helicos, o Pacific Biosciences e o Oxford Nanopore, que utilizam principalmente a premissa de sequenciamento de moléculas únicas em larga escala. Os eventos marcantes para o desenvolvimento dos métodos de sequenciamento de DNA, que culminaram nas principais técnicas conhecidas, podem ser cronologicamente organizados (Figura 2.2).

Box 2.1 - 1953

Ano marcante para a ciência, a genética e a biologia molecular. Artigos clássicos foram publicados descrevendo a estrutura em dupla hélice do ácido desoxirribonucleico (DNA) (Watson e Crick, 1953), com seu esqueleto externo de fosfato (Franklin e Gosling, 1953), com suas estruturas tridimensionais nas formas A e B e, ainda, demonstrando como os pareamentos de bases na dupla hélice permitiam a replicação do DNA (Watson e Crick, 1953). Também a detecção desta forma de DNA foi confirmada em sistemas biológicos (Wilkins, et al., 1953), sendo reconhecido como o material da hereditariedade (Avery et al.,1944).

Hoje em dia o chamado sequenciamento de nova geração (NGS – *Next Generation Sequencing*) permitiu aumentar consideravelmente a escala das análises genômicas, sequenciando e genotipando milhares de regiões e genomas de interesse em

um único passo, gerando um grande volume de dados biológicos (*high throughput sequencing* -HTS). Assim, os termos NGS, HTS ou sequenciamento massivo paralelizado, constituem um conceito genérico, englobando as diversas metodologias distintas para o sequenciamento de DNA que geram um grande volume de dados em comparação ao padrão da metodologia de Sanger, dispensando a necessidade de clonagem *in vivo* dos fragmentos de DNA. Mesmo em populações de organismos sem informações genéticas prévias, estes avanços tecnológicos em biologia molecular e sequenciamento em larga escala vêm permitindo inúmeras análises genômicas e transcritômicas, contribuindo para a compreensão e o estudo da variação biológica, através da descoberta de marcadores moleculares e genes importantes para rotas metabólicas e vias fisiológicas.

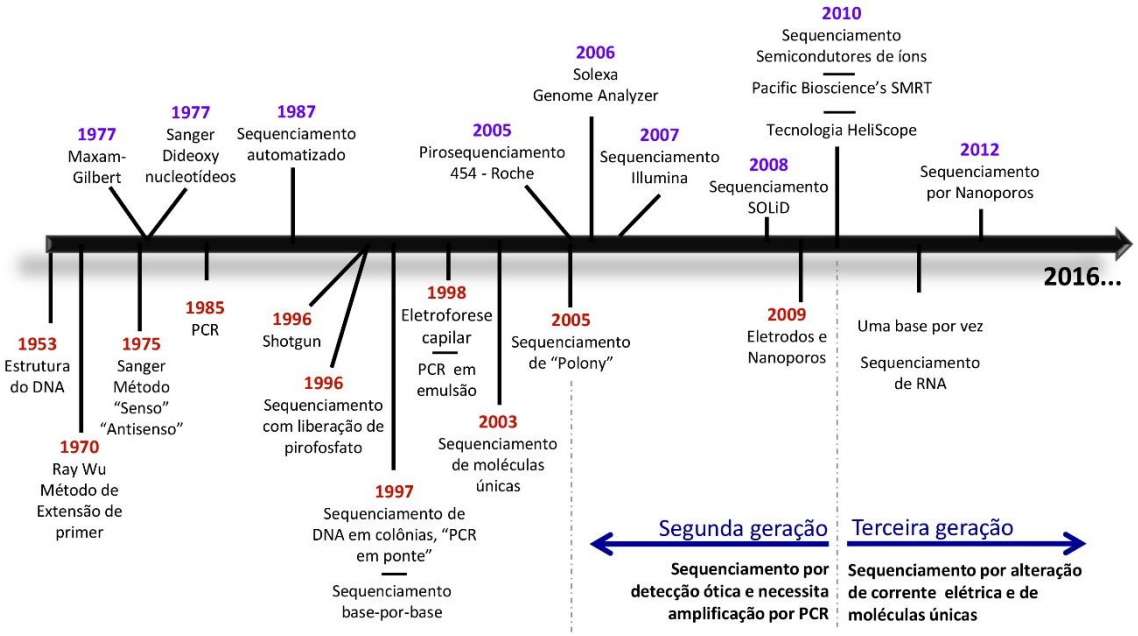

Figura 2.2 - Tecnologias de sequenciamento de DNA ao longo dos anos e as principais contribuições de metodologias moleculares para o desenvolvimento dos sequenciadores de DNA.

Historicamente, em 1975, Sanger introduziu o conceito de sequenciamento de DNA utilizando um método rápido, através da síntese com uma DNA polimerase e a utilização de pequenos oligonucleotídeos iniciadores da reação. Em 1977, foram publicados dois marcos históricos para o sequenciamento de DNA, o método de Sanger utilizando a terminação de cadeia com dideoxinucleotideos e o método de Maxam e Gilbert, pela degradação química de DNA. O método de Sanger por sua vez, tornou-se mais popular devido à sua alta eficiência a baixa radioatividade. Estes métodos culminaram no desenvolvimento dos primeiros sequenciadores de DNA automatizados, subsequentemente comercializados pela Applied Biosystems (ABI), pelo European Molecular Biology Laboratory (EMBL) e pela Pharmacia-Amersham, atual General Eletric (GE) healthcare. O primeiro sequenciador comercial foi introduzido em 1996, utilizando eletroforese em placas de gel pelo ABI Prism 310. Em 1998, a ABI novamente revelou o primeiro sequenciador de DNA utilizando um sistema automatizado de capilares, em substituição às placas de gel. Tais equipamentos foram utilizados no sequenciamento do primeiro genoma humano, em um esforço de 13 anos

do consórcio do projeto genoma humano, utilizando os refinamentos do método "dideoxi" estabelecido por Sanger em 1977.

No ano de 2005, foi introduzida no mercado a primeira plataforma de sequenciamento *high-throughput* NGS pela 454 Life Sciences, o GS 20. A grande novidade deste sequenciamento foi a ausência da necessidade de clonagem prévia dos fragmentos de DNA em vetores de multiplicação em bactérias ou leveduras. Com esta metodologia, todo o DNA presente em uma amostra era diretamente clivado por *shotgun*, e um PCR em emulsão de moléculas únicas era realizado para seu posterior pirosequenciamento. A tecnologia 454 foi adquirida pela companhia Roche, a qual aprimorou este método de sequenciamento por síntese, através da detecção de luz resultante da liberação de pirofosfato nas reações, o que ocorria quando um nucleotídeo era incorporado na molécula de DNA nascente.

Nos anos subsequentes, diversas tecnologias de NGS surgiram baseadas no princípio de detecção da fluorescência de terminadores reversíveis. O sequenciamento por síntese é realizado em paralelo para um grande número de amostras de DNA, as quais se encontram imobilizadas em uma superfície, minimizando os volumes de reação para um sistema miniaturizado. O princípio básico consiste na detecção de cada nucleotídeo incorporado na molécula crescente de DNA, este nucleotídeo marcado com fluorescência é detectado por uma câmera CCD. Subsequentemente, o fluoróforo ligado neste nucleotídeo terminador é removido e um novo ciclo se inicia para a incorporação de um novo nucleotídeo marcado e a determinação da próxima base na sequência de DNA. Em virtude deste processamento paralelo, uma molécula única pode ser "individualmente" sequenciada e ter sua proporção estimada no total de sequências obtidas. Assim, a grande combinação de informações quantitativas e qualitativas que são geradas possibilitou um enorme avanço nas análises genômicas que eram antes tecnicamente impossíveis, ou muito caras. Estas metodologias foram amplamente aplicadas nas plataformas de sequenciamento NGS de segunda geração (454 Roche, Illumina, SOLiD).

Apesar da amplificação de DNA ter revolucionado as metodologias de sequenciamento até então, em algumas circunstâncias ela pode introduzir erros de nucleotídeos, favorecer algumas sequências em detrimento de outras, ou ainda gerar um viés quantitativo alterando a relação de frequência e abundância de fragmentos de DNA existentes antes da amplificação e após a amplificação. Desta forma, o princípio básico da terceira geração de sequenciadores de alto rendimento, baseia-se na ausência da necessidade de amplificação prévia da amostra, podendo detectar a sequência de DNA diretamente de uma única molécula. Entre estas tecnologias destacam-se principalmente: a Heliscope, o SMRT da Pacific-Bioscience e o Oxford Nanopore, que utilizam desde polimerases modificadas imobilizadas em uma superfície até nanoporos que detectam variações de corrente elétrica, ou ainda, a liberação de íons resultantes do processo de polimerização da molécula nascente, como na tecnologia Ion. Contudo, a metodologia ion ainda utiliza bibliotecas enriquecidas, não sendo sequenciamento de moléculas únicas.

As metodologias de sequenciamento de DNA de nova geração, conhecidas como sequenciamento de alto desempenho, estão cada dia mais presentes em nosso cotidiano, revolucionando a pesquisa biológica em áreas como a genética, genômica, biotecnologia, medicina, etc. Podemos observar o incrível aumento na qualidade dos dados gerados, assim como no tamanho de sequências (*reads*), contando ainda com uma diminuição na quantidade inicial de amostra necessária, além de uma diminuição significativa no custo por base sequenciada. As mudanças que ocorrem neste campo são rápidas e inovadoras, resultando em técnicas de sequenciamento mais robustas e

acuradas, além de gerar uma grande quantidade de dados com uma velocidade incrível. Isto influencia diretamente a necessidade de desenvolvimento de novas metodologias para análises de dados, concentradas principalmente na área da bioinformática, que vem se tornando o grande desafio dos últimos anos. Assim, desde a descoberta da estrutura da molécula de DNA em 1953, o constante desenvolvimento das técnicas moleculares vem construindo marcos científicos para a evolução do sequenciamento de DNA, desde a metodologia de Sanger (1977) até os sequenciadores de segunda geração (2005), ou mais recentemente, e ainda em constante desenvolvimento, os sequenciadores de terceira geração (2010).

Os primeiros métodos de sequenciamento de ácidos nucleicos

Em 1965, uma das primeiras sequências de ácidos nucleicos que se teve conhecimento, revelou os 77 nucleotídeos do RNA transportador (tRNA) de alanina em leveduras (Holley et al., 1965). Nesta época ainda não existiam métodos bem desenvolvidos para o sequenciamento de ácidos nucleicos, o processo era realizado pela clivagem do RNA de interesse, com uma enzima do tipo RNAse, e os fragmentos obtidos eram submetidos a análises por cromatografia e eletroforese. Esta metodologia permitiu a identificação da composição da molécula do ácido nucléico, incluindo purinas, pirimidinas e também nucleotídeos não usuais, comumente encontrados em tRNAs. Entretanto, os pequenos fragmentos gerados pela clivagem não possuem uma ordem específica, nem uma sobreposição com tamanho adequado que permita a montagem de um fragmento maior, dificultando assim a caracterização da sequência nucleotídica final.

Alguns anos após, em 1970, Ray Wu desenvolveu um método que serviu como base para várias tecnologias que viriam no futuro, conhecido como "estratégia de extensão de primer" (Wu, 1970). Este método baseia-se na capacidade intrínseca de catálise da enzima DNA polimerase, utilizando uma fita molde de DNA e nucleotídeos específicos marcados radioativamente, resultando em uma sequência nucleotídica mais acurada, já com a ordem correta de organização dos nucleotídeos. Para realizar a determinação da sequência de deoxinucleotídeos em uma molécula de DNA, a região terminal desta molécula deve estar em fita simples para que haja o reconhecimento pela DNA polimerase (por exemplo, a enzima de *Escherichia coli*), que inicia o processo de síntese complementar adicionando nucleotídeos marcados na extremidade 3' da molécula, copiando a fita simples 5' complementar. Utilizando este procedimento, em 1970, foi possível realizar o sequenciamento de regiões fita-simples do bacteriófago lambda (ʎ) e também do bacteriófago DNA 186, gerando fragmentos sequenciados de aproximadamente 10 nucleotídeos. Para sequenciar outras regiões e fragmentos de interesse, podia-se também realizar a degradação limitada de uma das fitas através de enzimas com atividade de exonucleases, para formar extremidades de fita simples com até 20 nucleotídeos. Alguns avanços metodológicos foram realizados com essa técnica, uma vez que ainda era uma época difícil para se conseguir grandes quantidades de um DNA homogêneo. As moléculas de DNA eram muito grandes para o sequenciamento e não haviam muitas exonucleases específicas que possibilitassem a degradação de regiões do DNA para permitir o sequenciamento através da complementação da fita simples. Assim, em 1972, o método de Ray Wu foi aplicado para análises de sequências nucleotídicas de DNA de forma mais ampla, utilizando como base o DNA codificador de proteínas (Wu, 1972). Nesta abordagem, uma sequência proteica de aproximadamente quatro aminoácidos é selecionada e transformada em sequência nucleotídica através da melhor combinação possível com o código genético (ainda não

era possível a síntese de um oligonucleotídeo degenerado). A sequência deste oligonucleotídeo é sintetizada quimicamente para ser utilizada como um iniciador, que irá anelar na região específica do DNA molde e servirá para que uma DNA polimerase adicione os nucleotídeos complementares a partir da extremidade 3' do iniciador. Estes nucleotídeos a serem adicionados são marcados radioativamente para posterior detecção da sequência de nucleotídeos que foram incorporados pela DNA polimerase. Para a utilização deste método, entretanto, é necessário ter o conhecimento prévio da sequência completa ou parcial da proteína de interesse, porém, este era um método mais acurado para determinar a composição e a ordem dos nucleotídeos em uma sequência de DNA.

Outro método de crucial importância para o desenvolvimento das tecnologias de sequenciamento de DNA foi o método conhecido como "*plus and minus*" descrito em 1975 (Sanger e Coulson, 1975). Este método foi pioneiro na utilização de DNA em fita simples (ssDNA) juntamente com a adição combinada de nucleotídeos marcados e géis de poliacrilamida que separavam os produtos de reação. Para que se obtivesse uma sequência nucleotídica de aproximadamente 50 nucleotídeos (nt), o método de extensão de *primer* de Ray Wu era utilizado em duas reações de polimerização separadas. Na primeira reação, a utilização de apenas um *primer* resultava em uma síntese de DNA complementar lenta e dessincronizada, com diversos fragmentos de tamanhos diferentes e os nucleotídeos incorporados nesta etapa eram marcados com fósforo radioativo (^{32}P). O produto desta reação era então dividido em oito tubos para servirem como *primers* na segunda etapa. Nesta segunda reação, a mesma é terminada adicionando-se apenas um dos quatro nucleotídeos (nas reações "*plus"*) ou os três nulceotídeos restantes nas reações "*minus*". O resultado destas oito reações era submetido à eletroforese em gel de poliacrilamida de forma paralela, e as moléculas de diferentes tamanhos com as combinações de nucleotídeos terminadores conhecidas, eram comparadas de forma paralela, resultando na decodificação de uma sequência de DNA de aproximadamente 50 nucleotídeos em um único experimento.

Maxam e Gilbert: um método químico de sequenciamento de DNA

Este método de sequenciamento de DNA introduziu a ideia, praticamente ao mesmo tempo em que Sanger, de que sequências de DNA poderiam ser identificadas pela resolução de fragmentos marcados na última base nucleotídica em géis de poliacrilamida (Maxam e Gilbert, 1977). Um fragmento de DNA fita dupla (dsDNA), marcado com ^{32}P, pode ser clivado com reagentes químicos em bases nucleotídicas específicas. Estas reações específicas poderiam ser para ambas as purinas (A e G), ou somente para pirimidinas, também poderiam clivar ligações entre A e G, ou ainda clivar especificamente em citosinas (C). Quando estes fragmentos clivados em nucleotídeos específicos eram analisados, em paralelo, em gel de poliacrilamida, o padrão das bandas marcadas radioativamente e separadas por tamanho poderia ser analisado para a identificação da sequência de DNA. Este método permitia o sequenciamento de aproximadamente 100 bases a partir do nucleotídeo no qual a molécula foi marcada radioativamente.

Um período de novidades e o início dos grandes sequenciamentos de DNA com dideoxinucleotídeos

Em 1977, utilizando independentemente a abordagem de "extensão de *primer*" de Ray Wu, Frederick Sanger desenvolveu um novo método para a determinação de sequências nucleotídicas (Sanger et al., 1977). Esta técnica tornou-se amplamente

conhecida, e desde então passou a ser a metodologia mais utilizada para acessar as sequências de DNA por várias décadas. A grande novidade aplicada nesta técnica, conhecida como sequenciamento de "Sanger", foi a utilização de dideoxinucleotídeos (ddATP, ddGTP, ddCTP e ddTTP) como análogos aos deoxinucleotídeos padrão (dNTPs: dATP, dGTP, dCTP e dTTP). Os dideoxinucleotídeos atuam como inibidores da atividade da DNA polimerase, pois eles não possuem a extremidade 3'-OH (grupo hidroxila) necessária para a continuidade da extensão da cadeia nucleotídica. Assim, a reação da DNA polimerase termina exatamente na posição em que um destes análogos foi incorporado na molécula sendo sintetizada. Partindo-se deste princípio do método dideoxi de Sanger, o sequenciamento de DNA é realizado através da incubação de uma sequência de oligonucleotídeo iniciador (*primer)* com uma molécula de DNA molde e uma DNA polimerase, na presença de uma mistura de dNTPs e ddNTPS, onde os dideoxinucleotídeos eram marcados com ^{32}P. Inicialmente, o sequenciamento era manual, onde quatro reações de polimerização eram conduzidas separadamente, cada uma contendo um dos análogos de nucleotídeos terminadores de cadeia em concentrações específicas. Desta forma, utilizando cada dideoxinucleotideo em uma reação separada e correndo as amostras resultantes de forma paralela no gel, o padrão de bandas obtido pode ser analisado para revelar a identidade da sequência de DNA (Figura 2.3A).

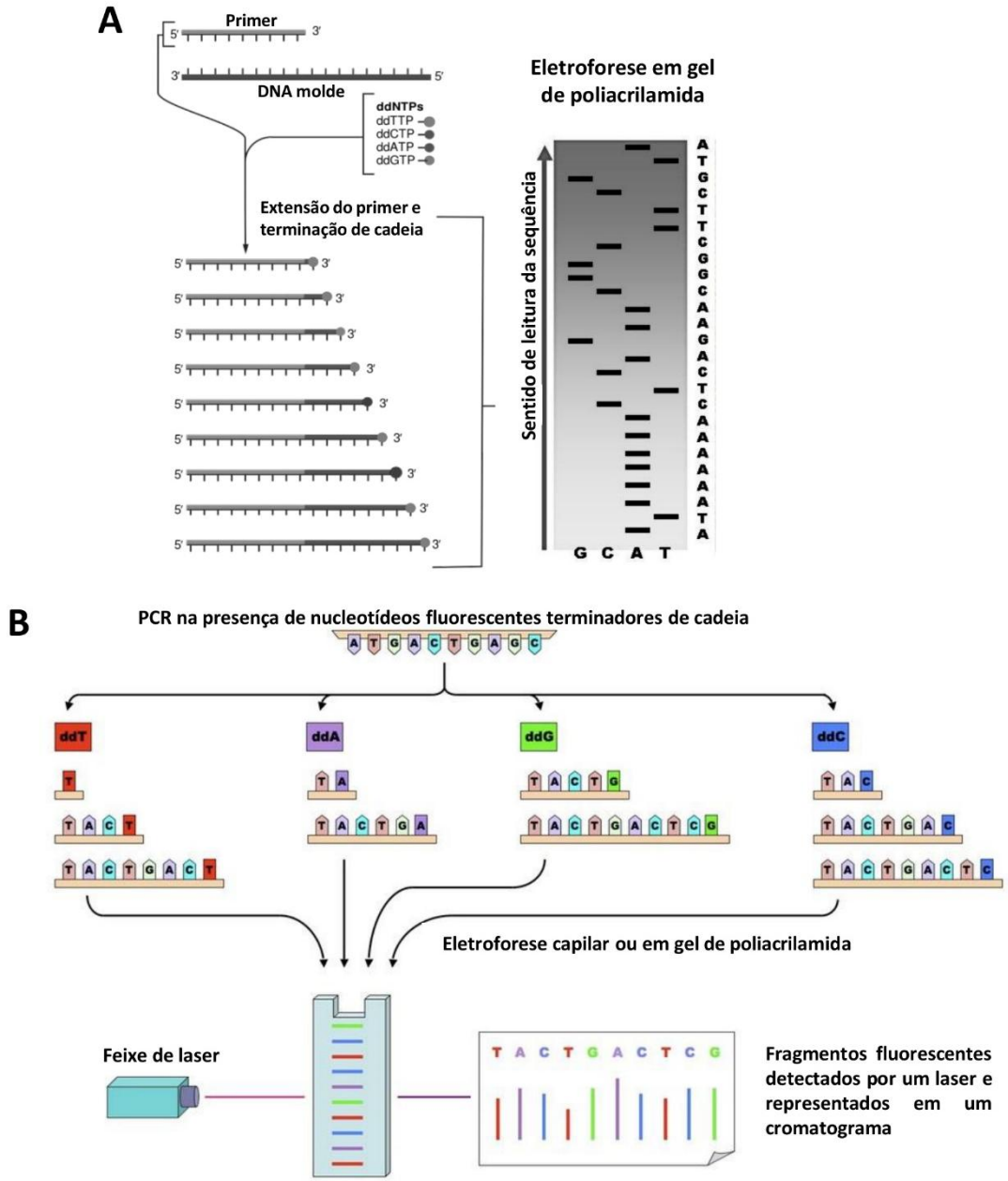

Figura 2.3 - Sequenciamento de DNA pelo método dideoxi. (A) O Método desenvolvido em 1977 por Sanger, onde após a síntese com os terminadores de cadeia as moléculas eram corridas e reveladas em um gel de poliacrilamida. (B) Automatização do método de Sanger, com a utilização de dideoxinucleotídeos marcados com fluorescência e detectados por um feixe de laser associado a um computador, que gerava a sequência nucleotídica final.

Mais recentemente, com o desenvolvimento das técnicas moleculares, a maior facilidade em sintetizar oligonucleotídes, produzir enzimas e reações *in vitro*, juntamente com o surgimento de técnicas como a reação de PCR (Box 2.2), favoreceu ainda mais a consolidação da tecnologia de sequenciamento de Sanger, que também passou por melhorias experimentais, principalmente deixando de utilizar nucleotídeos marcados com radioatividade (^{32}P). Logo em 1986, publicou-se uma metodologia de análise automatizada de DNA, baseando-se na detecção por fluorescência, ao invés de radioatividade (Smith et al., 1986) (Figura 2.3B).

Box 2.2 - 1985

Reação em Cadeia da Polimerase (PCR – *Polymerase Chain Reaction*)

Baseando-se na metodologia de extensão por *primers*, previamente descrita por Ray Wu (1970), Kary Mullis e colaboradores desenvolveram uma das técnicas mais amplamente utilizadas em biologia molecular, a reação de PCR. Através de uma amplificação enzimática por uma DNA polimerase, uma região de interesse em uma molécula de DNA pode ser amplificada de forma exponencial. Este aumento do número de moléculas-alvo de DNA é mediado pela utilização de um par de iniciadores (*primers*) específicos e complementares às fitas-molde de DNA. Desde então, a técnica de PCR tem sido aprimorada, com a utilização de uma ampla diversidade de polimerases, incluindo DNA polimerases termoestáveis, isotérmicas, de alta fidelidade, entre outras.

A aplicação da metodologia de PCR está presente no dia-a-dia da maioria dos laboratórios, sendo diretamente aplicada na análise e identificação de sequências e variações nucleotídicas, principalmente na clonagem e amplificação de sequências genômicas para o sequenciamento direto de DNA.

(Mullis e Faloona, 1987)

Diferentes fluoróforos são utilizados na reação com cada base específica (A,T,C,G), melhorando a capacidade de detecção dos nucleotídeos, reduzindo ruídos de fundo e sinais inespecíficos, além de não prejudicarem a reação de hibridização ou a mobilidade eletroforética. As reações separadas para cada dideoxinucleotídeo marcado eram então combinadas no processo de eletroforese em um único tubo contendo gel de poliacrilamida. As bandas de DNA com fluorescência eram então separadas por tamanho e detectadas perto do fundo do tubo, onde a informação da sequência poderia ser adquirida e armazenada por um computador associado, durante a eletroforese. A automatização desta metodologia de sequenciamento veio logo em seguida, uma vez que o método havia se tornado mais simples e uma grande quantidade de pesquisadores começavam a sequenciar mais moléculas de DNA. O primeiro sequenciador de DNA automatizado foi desenvolvido pela Applied Biosystems em 1987, entretanto, o auge da tecnologia de Sanger foi lançado em 1998, também pela Applied Biosystems. Esta companhia desenvolveu um novo sequenciador automatizado, utilizando um sistema de eletroforese em capilares, que continham uma matriz polimérica desnaturante para o sequenciamento de DNA. Diversos genes e vários genomas começaram a ser sequenciados em escalas maiores e em parcerias ao redor do mundo, incluindo o projeto genoma humano. Os sequenciadores automatizados dispensavam a utilização de géis, e integravam programas de computadores para a análise dos nucleotídeos sequenciados, fazendo com que a metodologia de Sanger fosse cada vez mais utilizada. Neste período, também houve a necessidade da criação dos primeiros bancos de dados, como o GenBank em 1982, para a centralização das informações genéticas de forma acessível. Programas de análises, buscas e comparações de sequências também começaram a ser desenvolvidos, como o BLAST (*Basic Local Alignment Search Tool*) e a implementação de formatos padrões de arquivos, como por exemplo, o formato FASTA.

Para o sequenciamento de pequenos fragmentos de DNA, ou alguns genes específicos, nesta época já era necessário menos esforço de laboratório, entretanto crescia o interesse em sequenciar fragmentos maiores, como os genomas completos. Para o sequenciamento destas regiões, era necessário o emprego de uma clonagem prévia da sequência de interesse em vetores como os YACs (cromossomos artificiais de levedura, que comportam insertos de 100-1000 kb), ou BACs (cromossomos artificias de bactérias, onde os fragmentos inseridos podem ter 150 – 350 kb). Também, poderia-

se utilizar plasmídeos (pequenas moléculas de DNA circular com capacidade de replicação independente do DNA genômico, usualmente encontrado em bactérias, que permitem insertos de até 15 kb). Os cosmídeos (plasmídeo híbrido com os sítios *cos* do fago lambda, que podem também ser encapsulados em capsídeos de fago, cujas sequências inseridas podem ter até 45 kb), ou os Fosmídeos (baseados no plasmídeo F de bactérias, também comportam insertos de até 40 kb), também poderiam ser utilizados. Quando a sequência inserida nestes vetores ainda era muito grande, ela passava por um processo adicional de clivagem com enzimas de restrição e era subclonada em vetores menores que poderiam então ser sequenciados, pois a capacidade de sequenciamento da metodologia de Sanger era de 500 a 750 pb por reação. Após o sequenciamento destes diversos subclones, a sequência completa poderia ser montada por sobreposição das sequências menores, e comparadas com um mapa físico de clivagem da sequência presente no vetor maior (BAC). Esta metodologia de clonagem foi amplamente utilizada no início do Projeto Genoma Humano, em 1990, e ficou conhecida como sequenciamento clone por clone (Lander et al., 2001).

A partir de 1996 a metodologia de *shotgun* (Box 2.3) foi introduzida como um método alternativo mais ágil e robusto para o sequenciamento e montagem de genomas, facilitando ainda mais os sequenciamentos baseados na metodologia de Sanger. Ainda, nesta época, viu-se que blocos gênicos poderiam ser confirmados através do sequenciamento de ESTs (*Expressed Sequence Tags*), obtidas através de bibliotecas de cDNA, resultantes do RNA total sendo expresso no organismo e tecido estudado (Adams *et al.*, 1991). Estas corridas por sequenciamentos genômicos resultaram em um efeito positivo para o desenvolvimento da metodologia de Sanger, além de baratear seu custo com o passar dos anos. Assim, durante as décadas subsequentes, o método de Sanger, em sua versão mais automatizada, tornou-se o padrão ouro para o sequenciamento de DNA.

Box 2.3 - 1996

Shotgun

Esta metodologia foi introduzida por Craig Venter e colaboradores, com participação da Celera genomics, sendo utilizada previamente no sequenciamento de genomas como o do bacteriófago lambda, do *Haemophilus influenzae*, do cromossomo 2 de *Arabidopsis thaliana* e também no genoma de *Drosophila melanogaster*. A maior visibilidade desta metodologia deu-se quando ela foi utilizada para acelerar o sequenciamento no projeto genoma humano. Esta metodologia permitia uma grande agilidade no preparo de amostras para o sequenciamento, uma vez que as sequências de DNA de interesse eram clivadas e inseridas aleatoriamente em vetores para a sua multiplicação (por exemplo, plasmídeos). Em seguida essas moléculas eram sequenciadas aleatoriamente, resultando em fragmentos de sequências (*reads*) que deveriam ser organizados, sobrepostos e montados através de programas de computador e análises de bioinformática. Os *reads* eram então montados em sequências maiores formando os *contigs*, ou mesmo, em blocos de sequências maiores e sobrepostos, formando os *scaffolds* de uma sequência genômica muito maior, por exemplo. (Venter et. al., 2001).

O sequenciamento de DNA e a detecção do pirofosfato

Novas ideias para o sequenciamento de DNA começaram a surgir, introduzindo metodologias que não exigissem um processo prévio de clonagem das sequências, ou eletroforese em gel de acrilamida, possibilitando o processamento paralelo de muitas amostras de uma forma automatizada. Uma destas novas metodologias demonstrou que era possível clivar por *shotgun* o DNA total de um organismo, e realizar o seu

sequenciamento direto e em tempo real. Nesta abordagem, durante o sequenciamento, a atividade da DNA polimerase pode ser monitorada durante a incorporação de cada nucleotídeo na molécula de DNA nascente, devido à liberação de uma molécula de pirofosfato inorgânico. Este pirofosfato liberado é detectado na forma de luz, por uma câmera, após o seu processamento por uma enzima ATP sulfurilase, e uma luciferase. Cada um dos quatro nucleotídeos (A, T, C e G) é adicionado de forma independente e sequencial na reação que contém uma molécula de DNA molde imobilizada com um pequeno fragmento complementar á sequência. Após a incorporação de cada nucleotídeo, um passo de lavagem é realizado, retirando os nucleotídeos e reagentes que não foram utilizados nesta etapa.

Assim, toda vez que um nucleotídeo é incorporado na molécula de DNA há a emissão de luz, que é registrada pelo sistema de um computador associado (Figura 2.4). A intensidade do sinal luminoso é proporcional ao número de bases que foram incorporadas naquela etapa do sequenciamento, podendo ser um único nucleotídeo (A), ou mais de um (AAAAA). Desta forma, ao final do processo, a sequência de DNA pode ser obtida pelos dados de luminosidade armazenados no programa de análises. Esta metodologia ficou conhecida como Pirosequenciamento (Ronaghi et al., 1996). Através dela, é possível aumentar em até 100x a escala de rendimento do sequenciamento comparando com a metodologia de Sanger e isto é obtido principalmente pela utilização do PCR em emulsão (Box 2.4), otimizado em um suporte sólido e com volumes em escalas de picolitros. Assim, o pirosequenciamento desenvolveu-se, reduziu tempo e custos, proporcionando um sequenciamento de DNA altamente escalável, com um maior volume de dados obtidos em um tempo relativamente menor ao do sequenciamento por capilaridade de Sanger (Margulies et al., 2005).

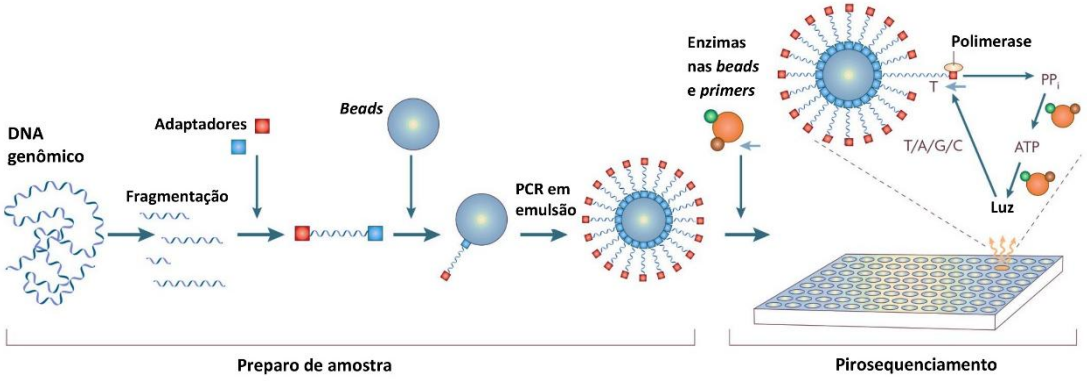

Figura 2.4 - Visão geral do método de pirosequenciamento. O DNA de interesse é fragmentado, ligado a adaptadores e *beads* magnéticas para amplificação em PCR de emulsão. Cada bead contendo cópias de uma molécula única de DNA é alocada em um pocinho da lâmina de sequenciamento, juntamente com os regentes necessários para a sua amplificação. Durante a síntese da fita de DNA complementar, os nucleotídeos incorporados podem ser sequencialmente identificados pela detecção de luz resultante da liberação do pirofosfato. Reproduzido com permissão de Macmillan Publishers Ltd, Nature Reviews Microbiology (Medini et al., 2008), copyright 2008.

Box 2.4 - 1998

PCR em emulsão

O princípio de que as células são eficientes por manterem juntos genes, RNAs e proteínas necessárias para certos processos relacionados, introduziu a ideia de que a reprodução deste mecanismo *in vitro* poderia representar um ganho experimental. De fato, a compartimentalização das reações em emulsões de gotículas de água em óleo, contendo uma molécula única de DNA molde juntamente com os reagentes e enzimas necessários para sua amplificação, foi um passo importante para a biologia molecular. Essa compartimentalização da reação de PCR reduz o número de artefatos quando existe uma grande complexidade de sequências, permitindo uma amplificação clonal mais eficiente das moléculas de DNA molde, sem os vieses de uma reação convencional. (Tawfik e Griffiths, 1998; Schütze et. al., 2011)

O pirosequenciamento, descrito detalhadamente em 2005, foi marcado pelo lançamento do sequenciador 454 Life Sciences, posteriormente adquirido pela Roche e comercializado como GS 20, capaz de sequenciar 20 milhões de pares de base em 4h. Em 2007, foi lançado o GS FLX, com capacidade de sequenciamento de 100 milhões de pares de base em 4h. Assim, unindo as metodologias de detecção de pirofosfato, com a reação de PCR em emulsão, o princípio básico da tecnologia 454 consistia na clivagem do DNA alvo de forma aleatória (*shotgun*), ligação de adaptadores e captura de moléculas individuais em microesferas magnéticas de 26µm com sequências complementares aos adaptadores. Uma grande quantidade destas microesferas é alocada nos 1.6 milhões de pocinhos em uma lâmina de fibra-ótica para que ocorra a amplificação das sequências em uma gotícula de emulsão. Neste arranjo, após a amplificação das moléculas de interesse, os quatro diferentes nucleotídeos são sequencialmente e independentemente adicionados. A quantidade de nucleotídeos que é incorporada em cada molécula de DNA complementar sendo sintetizada é registrada pela intensidade da luminescência na reação do pirofosfato, com uma câmera CCD acoplada ao sequenciador.

No início, esta metodologia produzia sequências (*reads)* com aproximadamente 80 – 120 nucleotídeos, através dos quais era possível realizar a montagem de genomas e outras sequências de interesse. Hoje em dia é possível que os fragmentos sequenciados atinjam até 1kb de tamanho, em um sequenciamento de 24 horas, e produzindo aproximadamente 1.000.000 de *reads*, com os novos modelos de sequenciadores e novas reações químicas do pirosequenciamento para a plataforma 454. Entretanto, um dos grandes vieses desta tecnologia é a formação de homopolímeros e a baixa acurácia de sequenciamento em regiões repetitivas, além do alto custo dos reagentes (Metzker, 2010). Entre as aplicações deste tipo de sequenciamento, pode-se destacar o sequenciamento de genomas, de *amplicons* e de transcritomas, gerando uma grande quantidade de dados e informações. A plataforma 454 veio como uma quebra de paradigmas e enfrentou bastante resistência entre os pesquisadores, já acostumados com a metodologia de Sanger. Em 2005, com esta metodologia, foi possível o sequenciamento do DNA de organismos já extintos partindo de quantidades ínfimas e já bastante degradadas de DNA, como por exemplo, o genoma do Mamute com 28.000 anos, ou ainda, os primeiros rascunhos do genoma dos Neandertais. O pirosequenciamento foi bastante utilizado em pesquisas com genômica microbiana, genética de plantas, análise de regulação gênica, transcritomas, pequenos RNAs (miRNA, siRNA), metagenômica e diversidade ambiental, além do contínuo sequenciamento do genoma humano e outros genomas completos.

Sequenciamento por ligação e a tecnologia SOLiD

Entre os sequenciadores de nova geração, a Applied Biosystems também desenvolveu um equipamento baseado no método altamente paralelizado de sequenciamento por hibridização e ligação, relacionado à amplificação por "*polonies*" (Box 2.5), o sequenciamento com a plataforma SOLiD (*Sequencing by Oligonucleotide Ligation and Detection*).

Neste processo, de forma geral (Figura 2.5), os fragmentos de interesse a serem sequenciados são amplificados na superfície de uma esfera magnética com 1 µm, em uma PCR de emulsão, para garantir a amplificação do sinal durante as reações de sequenciamento. As esferas com os *amplicons* são depositadas em uma lâmina (*flow cell*) e o sequenciamento por ligação se inicia com a hibridização de uma sequência de *primer* ao adaptador do fragmento, que é comum a todas as sequências amplificadas. Então, enzimas DNA ligase são fornecidas juntamente com combinações específicas de pequenas sondas (octâmeros de nucleotídeos) marcadas com fluorescência.

Box 2.5 - 1999

Polony (polymerase colony) amplification

Um método inovador, que permite clonar e amplificar sequências de DNA, a partir de uma molécula única, utilizando apenas uma reação de PCR com uma matriz de poliacrilamida imobilizada em uma lâmina de microscópio. Esta matriz retarda a difusão das moléculas de DNA amplificadas, que assim permanecem próximas de suas respectivas moléculas de DNA molde. No final da reação, estas colônias de amplificação (*polonies)* possuem em torno de 5 milhões de moléculas clones, oriundas de uma única molécula molde, e ficam dispostas de forma paralela na lâmina. Ainda, é possível fazer uma modificação na extremidade 5' dos *primers* utilizados e o DNA amplificado ficará covalentemente ligado à matriz de acrilamida, permitindo a utilização posterior desta lâmina para diversas aplicações, como por exemplo, o sequenciamento destas moléculas de DNA. Avanços mais recentes utilizam este método para a amplificação de moléculas únicas em PCRs de emulsão, ligados a *beads* magnéticas. A PCR em emulsão pode ser considerada uma amplificação em *polony*.

(Mitra e Church, 1999)

A bioquímica deste tipo de sequenciamento envolve a atividade de polimerases e ligases, onde o método de ligação degenerada reconhece um *primer* marcado por fluorescência que hibridizou na molécula de DNA molde, e faz a ligação desta molécula com o *primer* adjacente. Esta ligação emite um sinal de fluorescência que será detectado pelo programa do equipamento e determinará a sequência nucleotídica, uma vez emitido o sinal, este fragmento hibridizado é clivado e um novo ciclo se inicia. Em seguida, um passo de regeneração remove as bases da sonda ligada, incluindo o fluoróforo, e já se inicia outro ciclo de ligação. Estes ciclos de ligação são repetidos a partir de sequências de *primers* com tamanhos diferentes (n, n-1, n-2, n-3 e n-4), para que toda a sequência tenha uma boa cobertura.

Nesta metodologia, o DNA a ser sequenciado pode simplesmente ser clivado em fragmentos de até 100 pb aproximadamente e ligado aos adaptadores específicos para a amplificação em emulsão. Porém, outra metodologia que pode ser empregada é a estratégia de "*mate-paired*", ou sequenciamento de pares. Nesta segunda abordagem, os fragmentos podem ser clivados em tamanhos de até 10 kb sendo posteriormente purificados por tamanho em gel de agarose e ligados a adaptadores CAP que promovem a circularização destas moléculas. Esta região CAP possui um sítio específico de clivagem enzimática que gera, para ambos os lados da molécula circularizada, fragmentos de 25 pb aproximadamente em cada uma. Assim, apesar do tamanho do

fragmento ser pequeno, é possível sequenciar as extremidades de um fragmento maior, o que facilita a montagem das sequências de DNA após todo o processo.

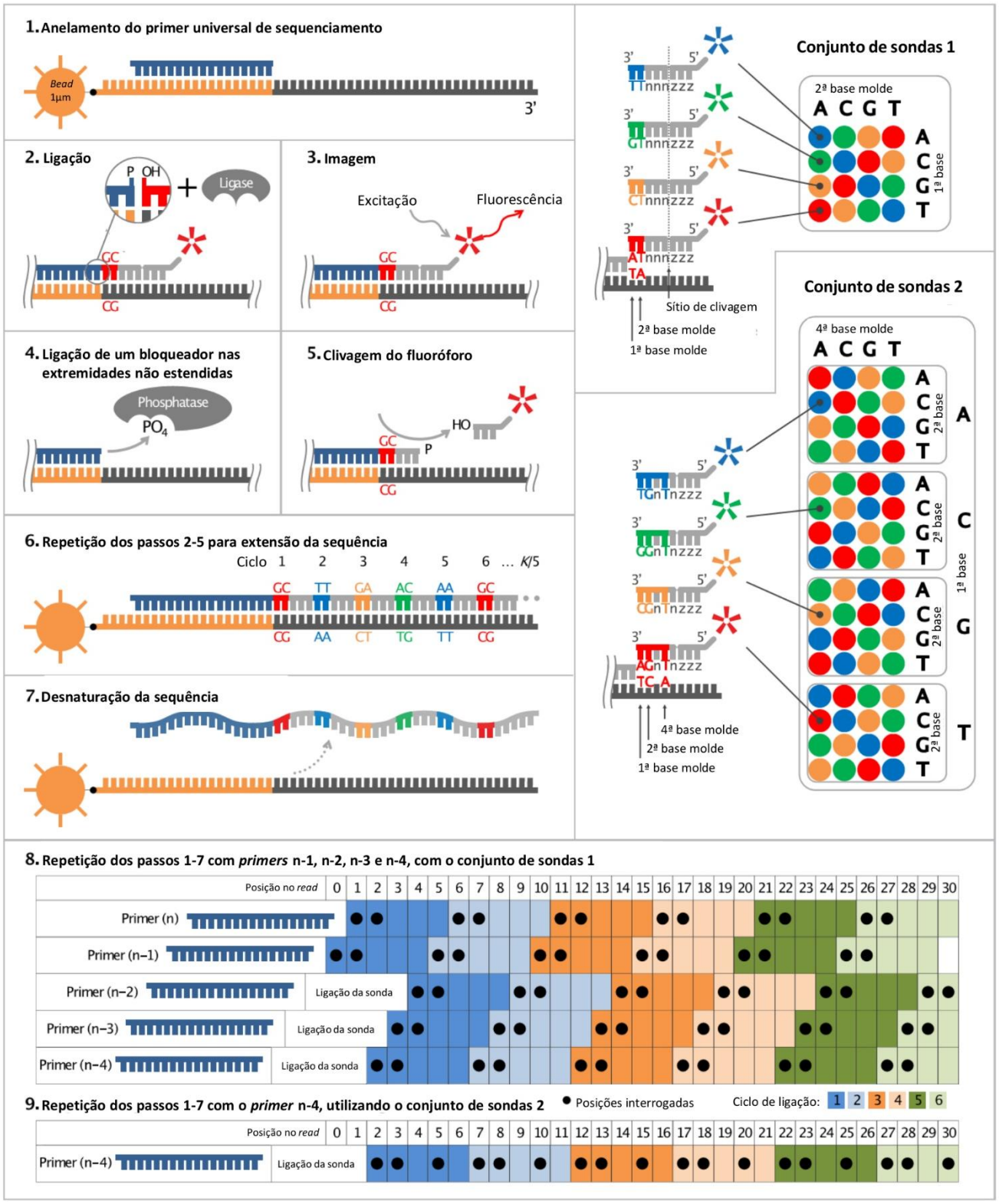

Figura 2.5 - Visão geral da metodologia de sequenciamento por ligação da plataforma SOLiD. Figura reproduzida de Applied Biosystems (2011), https://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf.

Mais recentemente, em 2010, o tamanho do fragmento chegava a cerca de 75 pb em um sequenciamento com ~30 Gb, com uma grande acurácia das bases nucleotídicas sequenciadas, devido ao sistema de sequenciamento de duas bases (Rusk, 2011; Liu et al., 2012). Esta é uma das grandes vantagens do sequenciamento SOLiD: a acurácia das bases sequenciadas, porém também requer um tempo de corrida muito longo.

O Sequenciamento SOLiD, já foi bastante utilizado para diversos projetos de re-sequenciamento de genomas, ou até mesmo em sequenciamentos inéditos (*de novo)*, que não possuem sequências de referência. Porém, com menor frequência neste, uma vez que os fragmentos pequenos dificultam a montagem precisa do genoma final, levando à necessidade de aumento da cobertura de sequenciamento (Box 2.6). Análises de transcritomas e sequências expressas também podem ser realizadas com esta metodologia, além de análises de padrão de metilação em genomas, de fragmentos oriundos de ChIP (*Chromatin immunoprecipitation* – imunoprecipitação de cromatina), descoberta de miRNAs, outros pequenos RNAs, etc. Este tipo de sequenciamento massivamente paralelo auxiliou na revolução das análises genéticas, reduzindo os custos e aumentando consideravelmente a quantidade de dados que podem ser gerados em um único experimento de sequenciamento (Mardis, 2008; Schuster, 2008; McKernan et al., 2009; Liu et al., 2012; Lee et al., 2015).

Box 2.6 - Cobertura

Cobertura média de sequenciamento

A cobertura de sequenciamento pode ser estimada de acordo o número de vezes em que cada nucleotídeo em particular é sequenciado, dado um certo número de *reads*, com determinado tamanho, e, assumindo que os *reads* estejam aleatóriamente distribuídos ao longo do genoma ou do fragmento de DNA de interesse. A cobertura média esperada para um sequenciamento pode ser calculada da seguinte forma:

Cobertura = Tamanho dos *reads* sequenciados x número de reads sequenciados / Tamanho do genoma haplóide.

Em geral a cobertura média varia para cada tipo de estudo, e deve ser observada em revisões bibliográficas, no planejamento dos experimentos, para obedecer aos critérios de análise estatística, representatividade e reprodutibilidade dos resultados. Por exemplo, análises de RNA-seq podem necessitar coberturas em torno de 100 x, ou mais, uma vez que os transcritos são expressos em diferentes níveis e a detecção dos menos abundantes requer um aumento de cobertura para alcançar o nível de sensibilidade esperado para o resultado. Análises de SNPs, ou mutações, em geral necessitam coberturas entre 10 - 30 x, dependendo do modelo de análise que será utilizado. O sequenciamento de genomas também pode utilizar coberturas variáveis, indo de 50 a 100 x para a identificação e caracterização do organismo, até coberturas bem mais profundas para obter o fechamento de um genoma completo, ou resolução de regiões mais difíceis, menos representadas na amostra. (Sims et al., 2014; Illumina Pub No. 770-2011-022 - 2014).

Amplificação em ponte e o sequenciamento Illumina

Durante o período de 1998-2006 é que uma das metodologias de sequenciamento de nova geração mais utilizadas atualmente começou a ser desenvolvida pela empresa Solexa, resultando em seu primeiro sequenciador o *Genome Analyzer* em 2006. Utilizando a técnica de "Amplificação em ponte" (Box 2.7), moléculas únicas de DNA eram amplificadas em aglomerados (*clusters*), que geravam um forte sinal ao incorporar cada base nucleotídica durante o processo de sequenciamento de DNA. Conhecido também como sequenciamento por síntese, esta tecnologia era capaz de sequenciar até 1

gigabase (Gb) de dados em uma única corrida. Logo, em 2007 a Solexa foi adquirida pela Illumina, juntamente com a sua metodologia de sequenciamento de ácidos nucléicos, que atualmente já possui uma capacidade muito maior, podendo gerar até 1 terabase (Tb) de dados em uma única corrida do equipamento.

Box 2.7 - 1998

Amplificação em ponte

Esta metodologia descreve como uma única molécula de ácido nucléico pode ser amplificada de forma clonal ao hibridizar-se em uma região complementar (*primer*) que está imobilizado em uma superfície. Esta molécula é estendida pela reação da polimerase em cadeia, formando uma região de dupla-fita, que é então desnaturada para anelar em outro *primer* complementar imobilizado, servindo de molde para a síntese de outra molécula (Figura 2.6). Durante a amplificação das moléculas complementares as mesmas mantêm-se formando uma "ponte" em forma de "U" entre as extremidades imobilizadas na superfície. Este processo gera grupos de moléculas idênticas, formando milhões de *clusters* de fragmentos clonais, oriundos da amplificação de uma molécula única de DNA. (Kawashima et al., 1998).

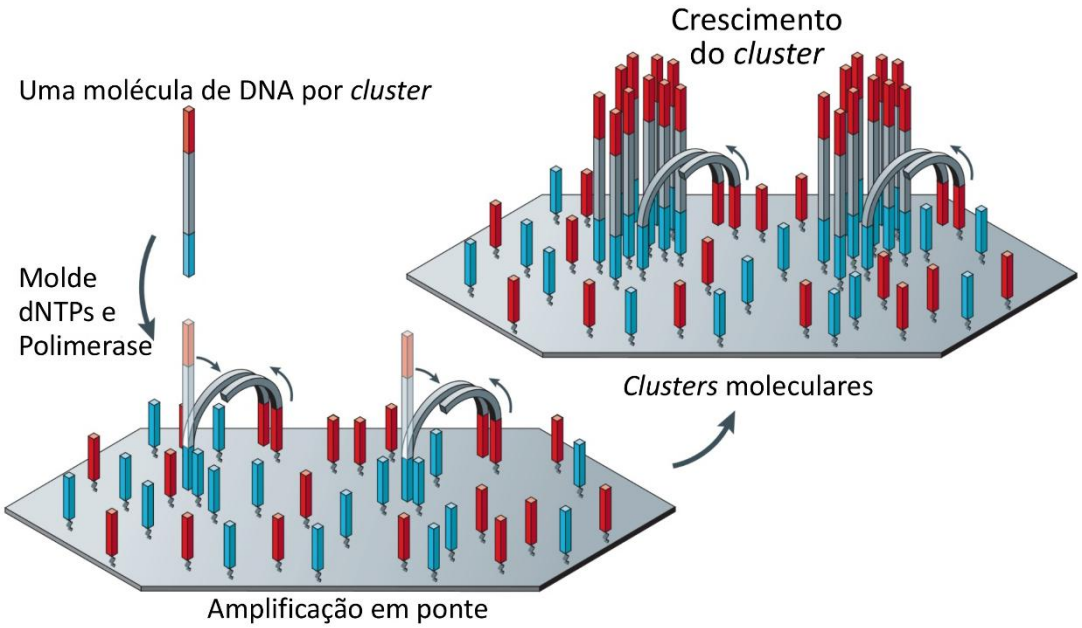

Figura 2.6 - Geração de grupos de sequências clonais (clusters), por amplificação em ponte. Reproduzido com permissão de Macmillan Publishers Ltd: Nature Reviews Genetics, Metzker ML 2010, *copyright* 2010.

Em geral, a metodologia Illumina (Figura 2.7) baseia-se no sequenciamento por síntese e consiste em fragmentar o DNA de interesse (ex: genoma total, metagenomas, *amplicons* de PCR, transcitomas, etc.), através de sonicação, nebulização, ou também por métodos químicos e enzimáticos. Estes fragmentos, ao final devem possuir um nucleotídeo com fosfato livre para a ligação dos adaptadores específicos "Illumina" em suas extremidades. Através da complementaridade destes adaptadores, os fragmentos a serem sequenciados hibridizam-se a pequenos *primers* imobilizados na célula de fluxo

(lâmina de vidro com canais de poliacrilamida) que compõe o *kit* de sequenciamento. Em seguida, inicia-se a formação dos agrupamentos (*cluster*s) clonais de sequências pela metodologia de amplificação em ponte, utilizando diversos *primers* adjacentes aos fragmentos imobilizados na lâmina de sequenciamento. Este processo resulta em um fragmento de DNA dupla-fita que é então desnaturado e pode ligar-se novamente a outros *primers* próximos, também imobilizados, para um novo ciclo de amplificação.

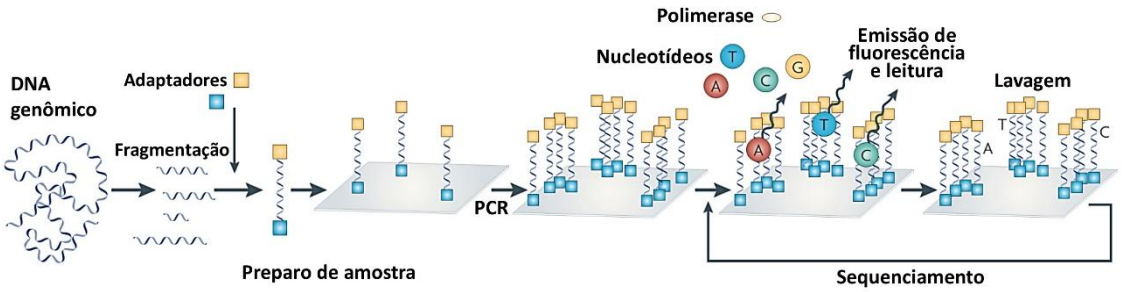

Figura 2.7 - Base metodológica do sequenciamento Illumina, utilizando amplificação em ponte e clusterização de moléculas únicas. Reproduzido com permissão de Macmillan Publishers Ltd, Nature Reviews Microbiology (Medini et al., 2008), *copyright* 2008.

Após a geração destes *clusters* densos de sequências idênticas, a sequência antisenso de cada molécula molde de DNA é clivada e o adaptador complementar é bloqueado, de forma que apenas as fitas senso permaneçam na lâmina de sequenciamento. Esta abordagem é conhecida como sequenciamento *single-end*, mas também existe a forma *paired-end*, onde após o sequenciamento das fitas senso, há uma regeneração das moléculas de DNA dupla-fita, através de novos ciclos de amplificação em ponte, e desta vez, as fitas senso é que são clivadas e bloqueadas, para que apenas as fitas antisenso sejam sequenciadas. O processo em si, ocorre pela metodologia de sequenciamento por síntese, onde o *primer* de sequenciamento hibridiza na extremidade livre das moléculas de DNA nos *clusters* da *flow cell*. Subsequentemente, uma DNA polimerase é utilizada para a incorporação complementar dos quatro nucleotídeos (A, T, G e C) marcados com fluorescência e reversivelmente bloqueados na extremidade 3'-OH. A cada ciclo, apenas um nucleotídeo é incorporado por vez em cada molécula, diminuindo grande parte dos homopolímeros que poderiam ser formados quando são sequenciadas regiões repetitivas. Assim, a cada vez que um nucleotídeo é incorporado, o sinal de fluorescência específico de cada *cluster* é detectado pelo equipamento e o bloqueador na extremidade 3' da molécula nascente é removido, para que o próximo nucleotídeo possa ser incorporado (Mardis, 2008; Medini et al., 2008; Shendure e Ji, 2008; Liu et al., 2012).

O sequenciamento Illumina é o mais amplamente utilizado no mundo atualmente, sendo que a cada ano a metodologia passa por otimizações, resultando em novos *kits* e equipamentos de sequenciamento com diferentes características, que se adaptam às necessidades experimentais. Atualmente, existem seis equipamentos diferentes sendo comercializados com esta metodologia. Três destes sequenciadores são conhecidos como "sequenciadores de bancada" por possuírem um tamanho menor e gerarem uma quantidade menor de dados, indo de 7.5Gb ou 15Gb para as versões do MiniSeq e MiSeq, e até 120Gb de dados para o NextSeq. Nos dois primeiros as corridas de sequenciamento podem levar de 4 a 55 horas, gerando até 25 milhões de *reads*,

enquanto que o NextSeq leva de 12-30h de sequenciamento, gerando até 400 milhões de *reads* por corrida. Estes sequenciadores de bancada são aplicáveis principalmente para o sequenciamento de *amplicons*, pequenos RNAs, painéis de genes-alvo específicos, pequenos genomas, exomas, transcritomas e re-sequenciamentos específicos. Ainda, os sequenciadores HiSeq e HiSeq X, possuem uma enorme capacidade de gerar dados (entre 1.500 e 1.800 Gb), que resultam em torno de 5 a 6 bilhões de *reads*, em corridas que podem levar de 1 a 6 dias. O mais recente lançamento, o NovaSeq pode gerar até 6 Gb com 20 bilhões de bases sequenciadas entre 19-40h. Estes sequenciadores maiores são mais utilizados para grandes genomas, exomas e transcritomas, também em escalas populacionais e comparativas. A possibilidade de multiplexagem de milhares de amostras em uma única corrida, juntamente com os menores custos e taxas de erro entre as tecnologias de sequenciamento disponíveis, fez com que o sequenciamento Illumina se tornasse o mais utilizado no mundo.

A plataforma Illumina possui uma ampla aplicação em pesquisas e desenvolvimento nas áreas de diagnóstico molecular, oncologia, genética microbiana, doenças complexas, genômica agrária, genômica forense, entre inúmeras outras. Dentre estas aplicações, o sequenciamento de DNA pode englobar o sequenciamento completo de um genoma, o sequenciamento direcionado de algumas partes do genoma (por exemplo, exoma), ChiP-Seq (imunoprecipitação de cromatina). Em sequenciamentos de RNA, onde as moléculas são previamente convertidas em cDNA, pode-se realizar o sequenciamento de RNA total, ou apenas de RNA mensageiro (mRNA), além de RNAs-alvo, pequenos RNAs ou genes específicos. Genotipagens em larga escala, além de análises de metilações também são possíveis utilizando o sequenciamento por síntese da Illumina (http://www.illumina.com). Abordagens como GBS (*genotyping-by-sequencing*) permitem a análise de organismos com genomas grandes e diversos, auxiliando, por exemplo, na reprodução assistida e melhoramento de plantas, possibilitando resultados mais rápidos e com custos reduzidos por amostra. Outras metodologias, como o ddRADseq (*double digestion Restriction Associated DNA sequencing*), que permite o sequenciamento com uma representação reduzida de um genoma grande e complexo, tornaram a genotipagem e comparação de organismos mais viáveis, escaláveis e baratas. A utilização das metodologias Illumina nestes contextos tiveram um impacto significativo, uma vez que, o custo do sequenciamento ficou bastante reduzido, mesmo em relação aos outros NGS já existentes.

Sequenciamento por semicondutores, a tecnologia Ion

Já no ano de 2010, o sequenciador Ion Torrent (PGM) foi lançado, baseando-se também no sequenciamento por síntese, mas com uma tecnologia de sequenciamento por semicondutores. Ao contrário das metodologias já existentes, este tipo de sequenciamento não utiliza nucleotídeos marcados com fluorescência, nem sistemas óticos de detecção, e é comumente referido como um pHmetro que sequencia DNA (Rusk, 2011). O princípio desta metodologia consiste em capturar a molécula de DNA a ser sequenciada, devidamente fragmentada, ligada aos adaptadores de sequenciamento e amplificada por PCR de emulsão. Estes fragmentos de DNA resultantes são distribuídos em um chip com um sistema micro fluídico, contendo micro poços, nos quais são liberados de forma individual os nucleotídeos a serem incorporados por uma enzima polimerase. Quando um nucleotídeo complementar é incorporado na fita-molde de DNA, um íon de hidrogênio (H^+) é liberado, alterando o pH na solução que é subsequentemente detectado por um sensor de íons. A cada ciclo em que o chip é inundado com algum dos quatro tipos de nucleotídeos (A, T, C ou G), a voltagem pode

ser detectada com uma determinada intensidade no caso da incorporação de um único nucleotídeo, ou com o dobro desta intensidade se, por exemplo, 2 nucleotídeos em sequência forem adicionados. Esta variação no pH é proporcional ao número de nucleotídeos incorporados, o que pode ser dificultado em alguns casos de sequências repetitivas, resultando em homopolímeros. Caso não hajam nucleotídeos incorporados, nenhum sinal de voltagem é detectado (Liu et al., 2012).

Uma calibração no início do sequenciamento é realizada utilizando os adaptadores específicos da tecnologia Ion, que estão ligados aos fragmentos de DNA a serem sequenciados, e possuem os quatro tipos de nucleotídeos para a validação da detecção do sinal de incorporação de cada uma das quatro bases de forma individual. Este tipo de sequenciamento não utiliza nucleotídeos modificados ou marcados, reduzindo as possíveis fontes de ruído na detecção da fluorescência dos nucleotídeos incorporados. Porém, esta metodologia também apresenta erros de sequenciamento relacionados à inserção ou deleção de nucleotídeos em regiões homopoliméricas, mais relacionados com o tamanho da região repetitiva em si, podendo se originar de uma saturação do detector de pH. A taxa de erro do equipamento é estimada em 01 base a cada 100 bases sequenciadas. De forma geral, esta tecnologia permite o sequenciamento em tempo real de moléculas, em uma alta velocidade e com um baixo custo, porém o tamanho dos fragmentos sequenciados e a quantidade de dados gerados ainda não são muito grandes, mas vêm sendo otimizados ao longo dos anos.

Atualmente, a plataforma Ion, já considerada como terceira geração de sequenciamento, pertencente à Thermo Fisher Scientific, disponibiliza quatro principais sequenciadores com diferentes capacidades e características. O primeiro dos sequenciadores a ser lançado, o Ion torrent (PGM) gera uma quantidade de até 2Gb de dados, podendo alcançar 5 milhões de *reads* com 200 nucleotídeos em algumas versões do chip de sequenciamento. O Ion Proton, lançado posteriormente possui uma escalabilidade de até 10Gb, podendo gerar 80 milhões de *reads* em sequências de até 200 nucleotídeos, com pelo menos 50 bases passando pelo filtro de qualidade. Ambos os sequenciadores têm um tempo estimado de corrida de até 2 horas dependendo do *kit* e chip utilizado. Mais recentemente, os sequenciadores Ion S5 e S5 XL foram lançados com corridas de sequenciamento que vão de 1 hora até 17,5 horas, gerando de 0,6 a 15 Gb de dados, dependendo do chip de sequenciamento utilizado. Os chips com capacidade de sequenciar até 15 Gb de dados (80 milhões de *reads*) geram fragmentos de 200 pb, enquanto que fragmentos maiores, de 400 pb, também podem ser obtidos, porém em versões de chips com capacidade de gerar até 8 Gb de dados (20 milhões de *reads*).

Os sequenciadores com a tecnologia Ion de semicondutores abrangem uma grande diversidade de aplicações, principalmente o sequenciamento de genomas e transcritomas bacterianos, regiões-alvo específicas em sequenciamentos de bibliotecas de *amplicons*, ou direcionadas para algumas regiões genômicas por captura e enriquecimento, como exomas. Também são bastante utilizados para detecção de variantes de DNA, como SNPs (*Single Nuleotide Polymorfism)*, porém, sequenciamentos completos de grandes genomas ainda são um desafio para esta tecnologia devido ao tamanho do fragmento gerado após a filtragem de qualidade da corrida. Porém, uma aplicação exemplar em que esta metodologia que é bastante utilizada, é na identificação de patógenos microbianos (Liu et al., 2012).

Sequenciamento de moléculas únicas e a tecnologia Helicos

A tecnologia Helicos tSMS, surgiu como uma promessa inovadora para o sequenciamento de DNA de terceira geração. Com a finalidade de sequenciar verdadeiramente moléculas únicas de DNA (SMS, *Single-molecule-sequencing*), possibilitando também o sequenciamento direto de moléculas de RNA. Esta metodologia de sequenciamento é bastante simples, constituindo-se de um preparo de biblioteca onde os ácidos nucléicos são fragmentados e ligados a caudas poli (A), que se hibridizam aos fragmentos poli (T) aderidos à célula de fluxo do sequenciamento (Figura 2.8). A extensão das moléculas de DNA ocorre com a adição sequencial de nucleotídeos com um único tipo de fluoróforo, que atua como terminador reversível marcado, permitindo que um único nucleotídeo possa ser incorporado e detectado pelo sequenciador (Heliscope) a cada ciclo do sequenciamento. Em seguida à incorporação de um nucleotídeo, ocorre a clivagem do fluoróforo e a reversão do nucleotídeo bloqueado o, que permite o início de um novo ciclo de adição de outro nucleotídeo fluorescente. Em 2011, esta tecnologia era capaz de produzir em torno de 28 Gb de dados em uma única corrida, que levava em torno de 8 dias e resultava em fragmentos de no máximo 55 bases (Pareek et al., 2011).

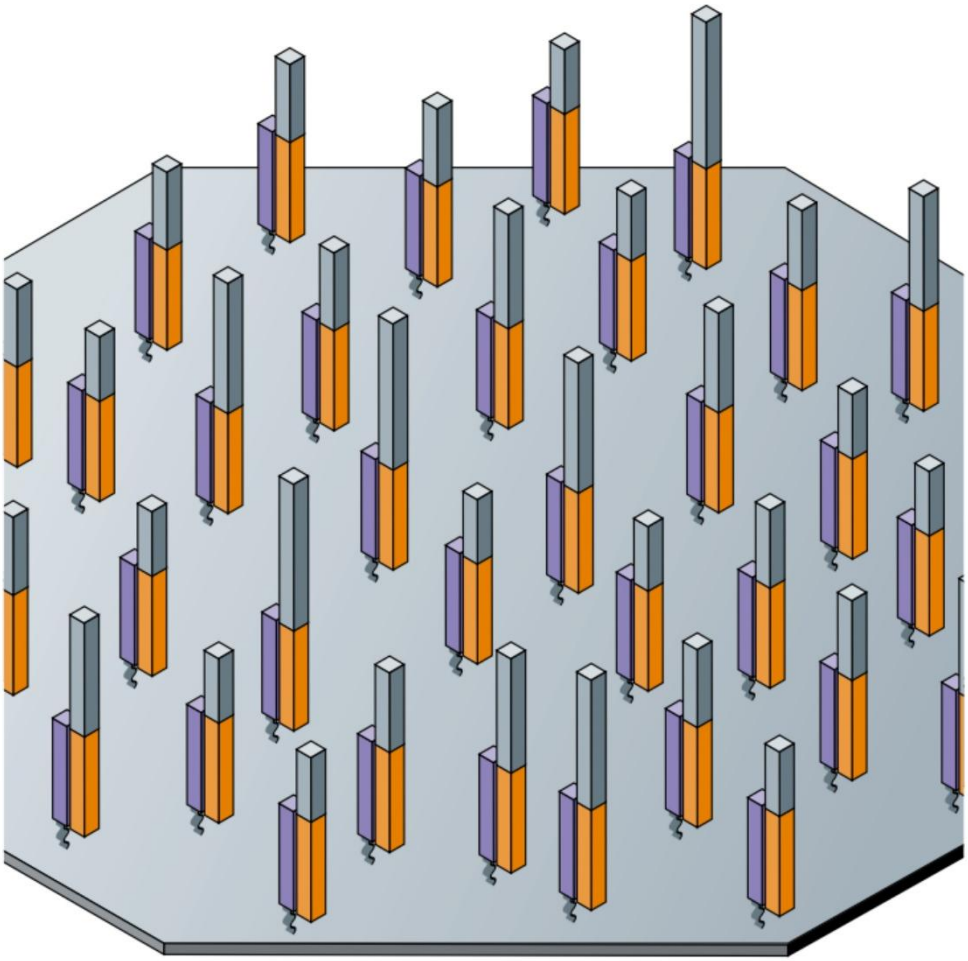

Figura 2.8 - Imobilização de moléculas pela cauda poli(A), para o sequenciamento. Reproduzido com permissão de Macmillan Publishers Ltd: Nature Reviews Genetics, Metzker ML 2010, *copyright* 2010.

Diferentemente de outras tecnologias, a Helicos foi inovadora na utilização de nucleotídeos fluorescentes como "terminadores virtuais". O próprio nucleotídeo em si era bloqueado de forma que apenas um fosse incorporado por ciclo de sequenciamento, porém este "bloqueio" era facilmente reversível para a extensão dos próximos ciclos. Esta abordagem reduziu a ocorrência de erros em regiões de homopolímeros durante o sequenciamento de ácidos nucléicos (Bowers et al., 2010). Inúmeras aplicações para este tipo de sequenciamento podem ser ressaltadas, entre elas: o sequenciamento de genomas, ou mesmo fragmentos menores como genes específicos, regiões-alvo específicas, *amplicons*, pequenos RNAs. Também era possível o sequenciamento direto de moléculas de RNA, quantificação de número de cópias, ChIP-seq, análise de regiões metiladas, entre outras inúmeras aplicações. Contudo, apesar desta tecnologia ser altamente promissora, os custos da mesma tornaram-se bastante elevados e a mesma foi descontinuada com o fechamento da empresa em 2012.

O sequenciamento direto de moléculas de RNA era uma das grandes apostas desta tecnologia, que havia começado a se desenvolver ainda em 2009 com o lançamento do Helicos®Genetic Analysis System. Todas as metodologias de análises de expressão gênica, pequenos RNAs, RNAs não codificadores, etc. requerem um passo de transcrição reversa, que converte as moléculas de RNA em cDNA, para que o mesmo possa ser sequenciado pelas metodologias conhecidas. Porém, sabe-se que esta conversão pode causar múltiplos vieses e artefatos na análise, interferindo com a caracterização e quantificação dos transcritos, principalmente transcritos com tamanho pequeno, degradados ou em baixas quantidades. Nesta nova abordagem, com a metodologia Helicos, foi proposto o sequenciamento direto das moléculas de RNA, sem necessidade de transcrição reversa, onde apenas a presença da cauda poli (A) dos transcritos era suficiente para sua hibridização na lâmina de sequenciamento, composta por oligos (dT). As moléculas de RNA, em quantidades mínimas (fentomoles) eram diretamente hibridizadas aos oligos (dT) na lâmina e então submetidas ao processo de sequenciamento por síntese (Ozsolak et al., 2009; Ozsolak e Milos, 2010; 2011).

Pacific Biosciences e o sequenciamento em tempo real de moléculas únicas

Em 2009, uma nova forma de sequenciamento de moléculas únicas de DNA em tempo real foi estabelecida pela Pacific Biosciences, culminando no lançamento de um sequenciador conhecido como SMRT (*single-molecule real time*). Esta metodologia baseia-se na atividade intrínseca das enzimas DNA polimerases, que catalisam a síntese de uma fita complementar de DNA com base na sequência nucleotídica de uma fita molde de DNA. Este processo ocorre através de uma única molécula de DNA polimerase fixada na parte inferior de um detector ZMW (*zero-mode waveguide detector*) com o tamanho de alguns nanômetros, feito de um filme de metal e depositado em um substrato de vidro. O tamanho deste orifício permite que a reação de polimerização seja feita em um volume de zeptolitros, assim, quando o nucleotídeo correto é detectado pelo sítio ativo da polimerase, ele é incorporado em um processo de milissegundos liberando um sinal de fluorescência que é detectado instantaneamente, mas que logo decai com a liberação dos fosfatos que contém o fluoróforo ligado (Figura 2.9) (Box 2.8). Para este processo, uma DNA polimerase específica (ϕ29) foi otimizada para aumentar sua afinidade por nucleotídeos fosfoligados, que possuem em torno de seis grupos fosfato, sendo que o último se encontra ligado à molécula do fluoróforo. Apesar de todos os nucleotídeos marcados serem liberados ao mesmo tempo na célula de sequenciamento, os mesmos ficam em constante processo de difusão pelos detectores ZMW, e apenas o nucleotídeo no sítio ativo da polimerase encontra-se próximo o

suficiente da região de detecção da fluorescência, de forma que a presença dos demais nucleotídeos marcados não interfere na detecção do sinal específico.

Box 2.8 - 2003

Sequenciamento de moléculas únicas de DNA

Através da utilização de microscopia de fluorescência é possível monitorar a atividade da enzima DNA polimerase para a obtenção das sequências. A incorporação de nucleotídeos marcados com fluorescência em fitas-molde individuais possuía a resolução de uma base única. A resolução necessária para este tipo de metodologia foi obtida através do controle do ruído de fluorescência no plano de fundo e outras impurezas. Para isto, uma combinação de microscopia de ondas infinintesimais e spFRET (*single-pair fluorescence energy transfer*) foi utilizada. Esta metodologia, porém, apresentava uma grande limitação no tamanho da sequência que pode ser definida, entre 5 a 15 pb. Ainda assim, é uma metodologia que permite grande paralelização de amostras, (Ex: 12 milhões de sequências em uma superfície de 25 mm, utilizando apenas alguns microlitros de reagentes. A importância inovadora desta metodologia introduziu a possibilidade de realizar a quantificação da expressão gênica de células únicas, por exemplo, inclusive sequenciando diretamente o mRNA de uma célula utilizando uma enzima transcriptase reversa. (Braslavsky et al., 2003).

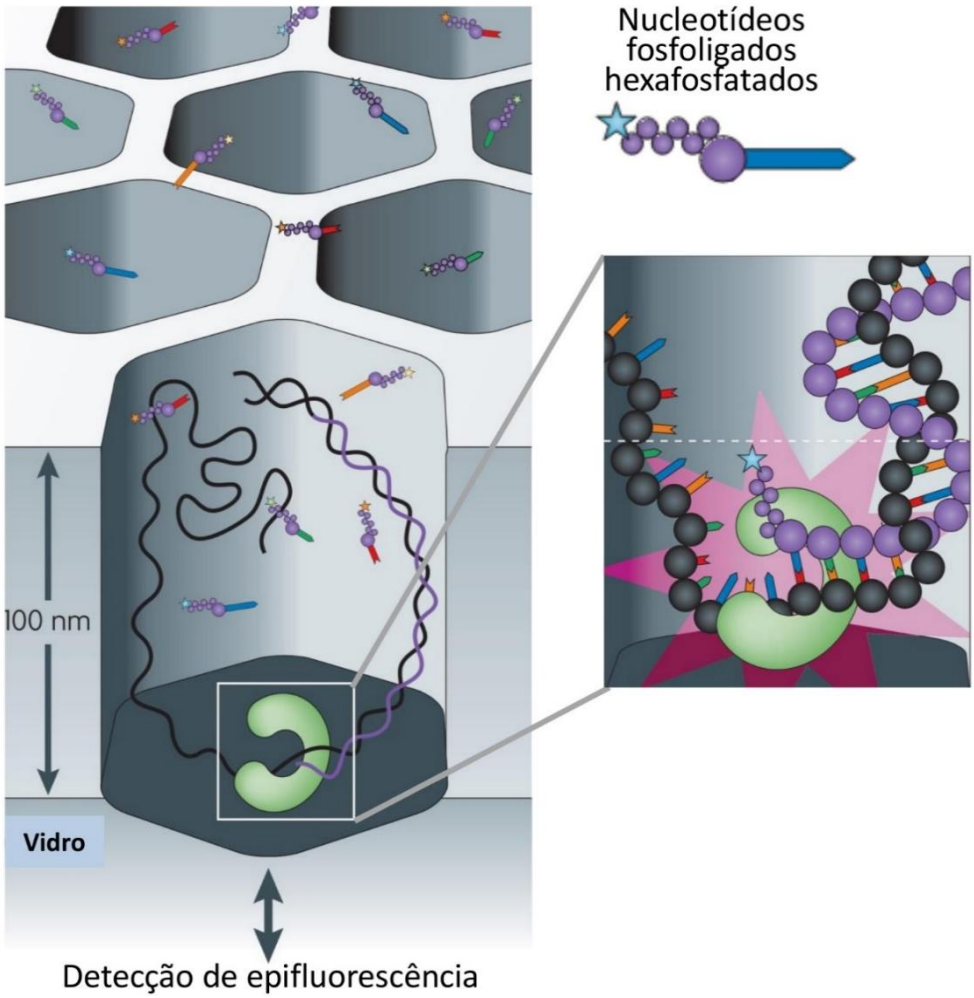

Figura 2.9 - Sequenciamento em tempo real pelo PacBio. A DNA polimerase imobilizada em uma lâmina de vidro incorpora os nucleotídeos hexafosfatados e a fluorescência desta reação de detectada em escalas muito pequenas. Reproduzido com permissão de Macmillan Publishers Ltd: Nature Reviews Genetics, Metzker ML 2010, *copyright* 2010.

A plataforma SMRT de sequenciamento requer quantidades mínimas de reagentes e preparo de amostra, além do rápido resultado por não possuir etapas de lavagem entre a incorporação de cada nucleotídeo marcado. Outra vantagem desta metodologia, que também a classifica como sequenciamento de terceira geração, é que não há necessidade de pré-amplificação das amostras para enriquecimento das bibliotecas de DNA. Uma vez que esta tecnologia se baseia na alta processividade da DNA polimerase, o SMRT tem os maiores tamanhos de *reads* entre as tecnologias de sequenciamento atuais, em média 1.000 pb, podendo chegar a um máximo de 10.000 pb. (Schadt et al., 2010). A taxa de erro das bases incorporadas neste sequenciamento é um pouco alta, porém, quando uma molécula é sequenciada com pelo menos 15 vezes de cobertura, a acurácia mediana das bases aumenta para 99.3% (Eid et al., 2009). Ainda assim, mesmo com esta taxa de erro elevada, o tamanho do fragmento sequenciando pelo Pacific Biosciences é uma boa metodologia para o sequenciamento de genomas, transcritomas e epigenomas completos ou, ainda, em complementação a outras tecnologias de sequenciamento para a reanálise e montagem de algumas regiões específicas (Metzker, 2010).

Uma das primeiras versões comerciais do sequenciador SMRT consistia em um arranjo de aproximadamente 75.000 ZMWs, cada um podendo conter uma DNA polimerase ligada a uma molécula de DNA molde a ser sequenciada. Isto resultava em uma potencial detecção de 75.000 moléculas em paralelo, porém devido ao processo de difusão aleatória pelo qual estas moléculas eram alocadas nos detectores ZMW, aproximadamente um terço dos detectores eram preenchidos com as moléculas em uma corrida de sequenciamento (Schadt et al., 2010). A previsão para esta tecnologia de sequenciamento, é que a mesma seria capaz de gerar 100 Gb de dados por hora, com fragmentos maiores que 1 kb (Pareek et al., 2011). Atualmente, o SMRT produz *reads* bastante longos entre 10 kb, podendo inclusive sequenciar alguns de 60 kb, gerando em torno de 500 Mb até 1 Gb por célula de corrida. Em uma corrida média de fragmentos de 20 kb, cada célula é capaz de gerar aproximadamente 55.000 *reads*. Estes longos *reads* são bastante importantes em abordagens como o sequenciamento *de novo* de genomas, catálogos de isoformas gênicas completas, alinhamento de sequências de forma não ambígua, resolução de regiões complexas com sequências repetitivas, entre outros (http://www.pacb.com).

Nanoporos e sua utilização no sequenciamento de DNA

Diversas metodologias para sequenciamento de DNA começaram a ser desenvolvidas em paralelo, baseadas em diferentes abordagens moleculares com utilização de nanoporos (Box 2.9), tanto sintéticos quanto biológicos (Deamer e Akeson, 2008; Branton et al., 2008), entretanto a metodologia mais promissora de todas foi desenvolvida pela *Oxford Nanopore Technologies* (Oxford, UK).

Em fevereiro de 2012, a primeira plataforma de sequenciamento com nanoporos foi anunciada pela *Oxford Nanopore*, introduzindo duas versões principais de sequenciadores: o GridION e o MinION, capazes de gerar grandes quantidades de dados, com um preparo de amostra simples resultando em longos *reads* a um baixo custo. O MinION, é um dispositivo portátil, pequeno, capaz de sequenciar 1 Gb de DNA, enquanto que o GridION, com um tamanho maior, é mais voltado para sequenciamentos maiores como grandes genes (Eisenstein, 2012). Entretanto, após o anúncio em 2012, a companhia levou mais dois anos para liberar a primeira versão beta do MinION, principalmente para testes da comunidade científica, com a finalidade de melhorar a metodologia de sequenciamento (Mikheyev e Tin, 2014). A média de

tamanho dos *reads* obtidos com os primeiros testes foi de 5.4 kb sendo alguns tão longos quanto 10kb. O MinION ainda não se encontra à venda, inicialmente ele faz parte de um projeto *early-access*, onde os pesquisadores inscrevem-se e podem pagar US$1.000, mais o envio para receber um MinION para teste em laboratório. Estes testes têm sido amplamente realizados e contribuem de forma significativa para o aprimoramento desta tecnologia de sequenciamento, corrigindo erros sistemáticos e melhorando a qualidade dos resultados obtidos (Check Hayden, 2014).

Box 2.9 - 1996 - 1997

Canais de membrana e nanoporos biológicos

Já na década de 90, tinha-se o conhecimento de que um campo elétrico era capaz de direcionar moléculas de DNA ou RNA por canais iônicos com 2.6 nm de diâmetro, em bicamadas lipídicas. Estas moléculas, estando em fita-simples, bloqueavam parcialmente o canal iônico, o que promovia uma alteração na corrente elétrica, podendo assim revelar o tamanho dos ácidos nucleicos que passavam pelo canal. Outras características também, como a composição de bases da sequência, poderiam vir a ser analisadas com o aprimoramento desta metodologia (Kasianowicz et al., 1996). Ao mesmo tempo, surgiu a necessidade de busca por sensores de canais iônicos que pudessem detectar uma variedade de moléculas, desde íons simples até compostos complexos, ou mesmo micro-organismos. Com este propósito, os poros de proteína demonstram uma grande aplicabilidade como componentes de biosensores de moléculas. Estes poros possuem diversas vantagens, uma vez que podem ser molecularmente modificados, possuem uma grande sensibilidade em escala de nanomolares, ligação rápida e seletiva a outras moléculas e uma estrutura reversível, entre outras inúmeras aplicações (Braha et al., 1997).

O princípio básico deste método de sequenciamento da Oxford Nanopore não requer nenhum tipo de marcação fluorescente, o que reduz os custos e aumenta sua velocidade. Em desenvolvimento por mais de 20 anos, esta metodologia, inicialmente baseava-se principalmente em um poro de α-hemolisina modificado, isolado de estafilococos, pelo qual uma molécula de DNA fita simples (ssDNA) passava sob uma diferença de potencial e a corrente iônica passando pelo poro era registrada. Cada base nucleotídica que passava pelo poro era registrada em sequência, de acordo com o decréscimo na amplitude da corrente iônica no poro (Kasianowicz et al., 1996; Deamer e Akeson, 2000; Bayley, 2006; Branton et al., 2008). Inicialmente, a velocidade com que as moléculas de DNA passavam pelo poro era um fator limitante para o sequenciamento, pois esta velocidade era muito alta (na ordem de microssegundos por base) para permitir que a alteração na corrente iônica fosse detectada pelo poro. Assim, em 2009, uma molécula adaptadora (ciclodextrina) foi covalentemente associada à região de barril da proteína, próximo ao centro do poro, de forma a otimizar a discriminação das bases nucleotídicas em taxas de aquisição de dados mais elevadas. Esta diferenciação de bases é mantida sob condições operantes, mesmo quando uma exonuclease próxima ao poro é utilizada. A exonuclease I, utilizada em associação ao complexo do nanoporo, permite uma redução da velocidade em que o DNA é translocado pelo poro, melhorando a qualidade da captura de sinal e o nanoporo por sua vez é capaz de detectar de forma contínua os nucleosídeos monofosfato (dNMPs), com auxílio da molécula de ciclodextrina que se liga a mononucleotídeos, fazendo com que os mesmos permaneçam no poro por aproximadamente 10 ms. Este nanoporo é capaz de discriminar os quatro dNMPs de forma acurada, com aproximadamente 99% de confiança, em condições ótimas, sendo capaz de reconhecer inclusive dCMPs metilados (Clarke et al., 2009; Schneider e Dekker, 2012).

A quantidade de dados brutos gerados pelo MinION é impressionante, incluindo o tamanho dos *reads*. Entretanto, a acurácia dos mesmos ainda é baixa, mas utilizando

programas de computador para a montagem de fragmentos longos, é possível extrair uma sequência consenso confiável quando uma cobertura de 16 x (*i.e.*, 16 vezes) é aplicada. A tecnologia atual, liberada para testes e experimentações na versão do MinION, ainda está em constante aprimoramento, para diminuir a taxa de erro, aumentar tamanho dos *reads* e melhorar a estabilidade dos nanoporos. O sequenciamento de moléculas únicas tem como vantagem um preparo de amostra mais simples, sem o custo dos fluoróforos e está entre as grandes promessas de tecnologias de sequenciamento. Outra grande vantagem desta metodologia de sequenciamento, implementada no MinION, é a possibilidade de levar esta técnica à campo, para estudos locais, epidemiológicos, de monitoramento ou de diversidade, uma vez que este sequenciador de bolso é de fácil conexão USB a qualquer computador, ou notebook, liberando os dados do sequenciamento em tempo real, conforme as moléculas de DNA vão sendo lidas pelos nanoporos (Hayden, 2015). Novas aplicações, além do desenvolvimento da metodologia já existente, estão constantemente sendo reveladas. Como ilustrado na própria página da *Oxford Nanopore*, não somente moléculas de DNA podem ser sequenciadas, como também os nanoporos podem ser adaptados para análise direta de RNAs, microRNAs, ou ainda, análise de proteínas.

Outras abordagens alternativas ao sequenciamento com nanoporos envolvem nanoporos em estados sólidos, com tamanhos de poros ajustáveis, sendo mais estáveis do que membranas biológicas, além de poderem ser reutilizados. Até o momento não foram demostradas análises de DNA com tais metodologias. Membranas com base em silicone são finas, correspondendo a aproximadamente 60 bases em uma única molécula de DNA, não sendo diretamente aplicáveis para o sequenciamento de DNA, mas que podem ser excelentes ferramentas em estudos biofísicos. Recentemente nanoporos de estado sólido feitos de grafeno, mostraram-se uma alternativa interessante, uma vez que a membrana de grafeno possui a espessura de um único átomo de carbono (Figura 2.10). Este material é um condutor elétrico, com possibilidade de utilização para detectar nucleotídeos únicos atravessando o poro, porém isto ainda é apenas uma ideia em um futuro talvez não tão distante (Schneider e Dekker, 2012). De fato, vários grupos de pesquisa já vêm trabalhando em diversas metodologias alternativas para sequenciamento de DNA utilizando nanodispositivos de grafeno, com algumas evidências experimentais que começaram a aparecer em 2016. Em geral, estas metodologias envolvem a passagem de DNA através de nanoporos de grafeno, pequenos orifícios (*nanogaps*), ou pequenas fitas *nanoribons*, ou ainda pela adsorção do DNA ao grafeno. Em conjunto, estas metodologias são bastante promissoras, constituindo propostas de sequenciamento de DNA bastante inovadoras, com grandes possibilidades de se tornarem reais (Heerema e Dekker, 2016).

Considerações finais

O NGS introduziu uma grande mudança de paradigmas científicos, refletindo diretamente em experimentos e análises, que passaram a ser feitos em larga escala, e com objetivos muito diversificados. Inúmeras amostras podem ser analisadas de forma paralela em uma única corrida de sequenciamento, com flexibilidade para a escolha do tamanho do fragmento gerado, ou a cobertura que cada amostra irá ter ao final do sequenciamento. A velocidade com que os dados, incontáveis sequências nucleotídicas, são gerados a cada dia, requer que também as análises destes dados sejam desenvolvidas por bioinformatas e computadores com grande capacidade de armazenamento. Estes avanços nas metodologias de sequenciamento levaram os cientistas à uma mudança de paradigma, saindo de um gargalo bioquímico de obtenção das sequências de DNA, para

um gargalo bioinformático. Este gargalo, muitas vezes fica de fora dos delineamentos experimentais e também da avaliação do custo da técnica, impactando significativamente a análise dos dados e o aproveitamento dos resultados.

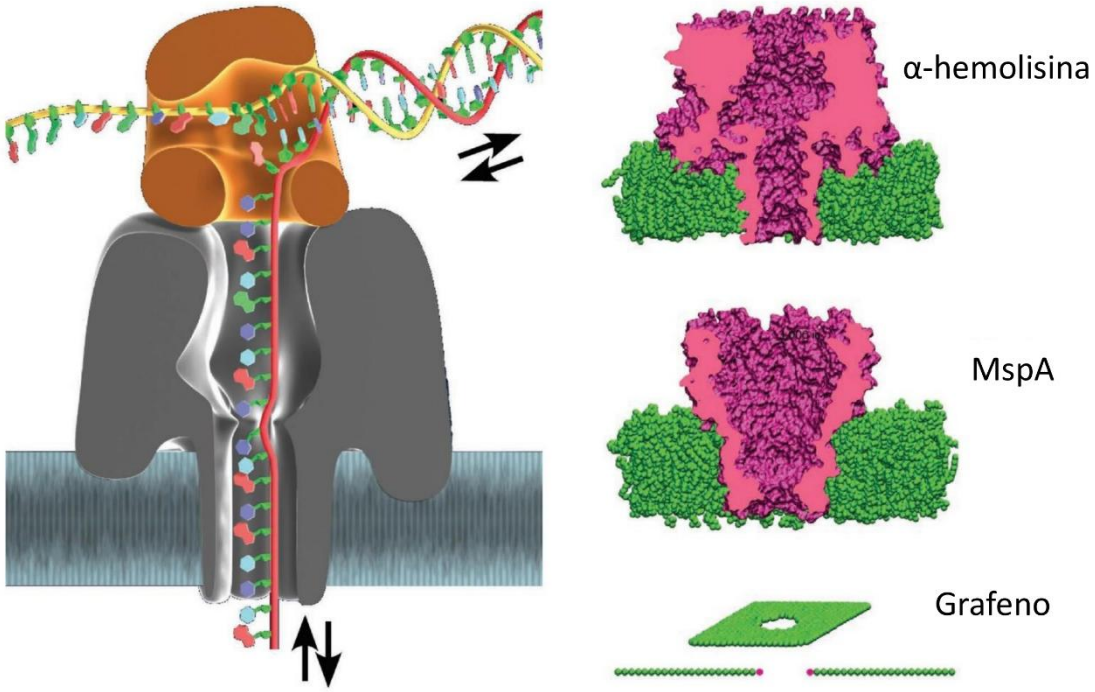

Figura 2.10 - Sequenciamento de DNA através de nanoporos proteicos (α-hemolisina e MspA), ou nanoporos de grafeno. A velocidade com que a molécula de DNA passa pelo poro biológico (cinza), é controlada por uma DNA polimerase (marrom) e a alteração de corrente elétrica é detectada pela camada onde o poro está embebido e resulta na definição da sequência nucleotídica que passa pelo poro. Alterações na corrente elétrica aplicada no sistema podem resultar em variações na velocidade em que o DNA passa pelo poro, e as mesmas podem ser controladas (indicadas pelas setas). Reproduzido com permissão de Macmillan Publishers Ltd: Nature Biotechnology, (Schneider e Dekker, 2012) *copyright* 2012.

As inúmeras aplicações do NGS em sequenciamentos de genomas *de novo*, resequenciamentos genômicos, metagenomas, RNA-seq, exomas, RNAs não codificantes, pequenos RNAs, *amplicons*, regiões marcadas, imunoprecipitadas, bibliotecas enriquecidas com fragmentos-alvo, entre diversas outras utilizações, requerem análises personalizadas dos dados. Cada plataforma de sequenciamento conhecida gera um tipo de dado com características próprias de cada metodologia. Essas diferenças vão desde o formato em que a sequência sai do equipamento, até o processo de pré e pós-filtragem das sequências avaliando a qualidade e confiabilidade das mesmas. Aglumas características das diferentes plataformas de sequenciamento estão sumarizadas na Tabela 2.1. Desta forma, não há um método único que possa ser seguido para analisar dados NGS, a quantidade de variáveis a serem consideradas em uma análise é enorme, sempre buscando a melhor reprodutibilidade e interpretação dos dados obtidos, de acordo com a finalidade do experimento que foi realizado.

De forma geral, o NGS, passando pela segunda, terceira e certamente entrando em uma quarta geração de metodologias de sequenciamento, proporcionou uma visão de um mundo ainda desconhecido pela nossa espécie. A possibilidade de desvendar nossa

biodiversidade e os mecanismos biológicos responsáveis por suas funções foi um avanço incrível proporcionado por este desenvolvimento científico-tecnológico. O que sabemos hoje certamente é a ponta de um iceberg do conhecimento, com muito mais sendo descoberto a cada dia pelos cientistas do mundo inteiro. Por isto, que a genética e a biologia molecular são áreas muito dinâmicas, com mudanças constantes de paradigmas e uma corrida tecnológica intensa, principalmente relacionada ao sequenciamento de ácidos nucléicos.

A aplicação das metodologias de sequenciamento de DNA em larga escala para análises de marcadores moleculares, levou a um aprimoramento destas técnicas resultando em estudos maiores, mais abrangentes, mais rápidos e gradativamente mais baratos. Novas tecnologias de sequenciamento têm sido aplicadas na identificação de microssatélites, polimorfismos (SNPs) ou genes em populações, na reconstrução filogenética e filogeográfica de espécies utilizando sequências de DNA marcadoras. Outras abordagens de genotipagem nas áreas biológicas, médicas, forenses, etc., empregando novos métodos de sequenciamento de DNA surgem a cada dia, buscando principalmente a utilização de marcadores moleculares de DNA na identificação dos organismos vivos e suas variações.

Tabela1. Caracteristícas de algumas metodologias NGS.

Plataforma	Tamanho máximo dos *reads*	Quantidade máxima de *reads*	Tempo de corrida	Desvantagens e erros conhecidos	Vantagens e aplicações
Roche 454 GS	700 pb	1 M	10 - 23 horas	Corridas muito caras; Erros com 1% indel; Formação de homopolímeros.	*Reads* mais longos e corridas rápidas. Aplica-se ao sequenciamento de genomas de bactérias e vírus, multiplex de produtos de PCR, validação e detecção de mutações pontuais.
ThermoFisher ABI SOLiD	75 pb	1.4 B	6 - 10 dias	Mais lento que outros métodos. Problemas em regiões palindrômicas. Erros ≤0.1% com viés AT.	Menor custo por base sequenciada, possibilidade de sequenciamento *paired-end*. Aplica-se à genomas complexos (ex: plantas, humanos), RNA-seq, multiplex de produtos de PCR, detecção de mutações somáticas.
Illumina MiniSeq MiSeq NextSeq HiSeq HiSeq X NovaSeq	2 x 150 pb 2 x 300 pb 2 x 150 pb 2 x 150 pb 2 x 150 pb 2 x 150 pb	25 M 25 M 400 M 5 B 6 B 20 B	4 - 24 horas 4 - 55 horas 12 - 30 horas 7 horas - 6 dias 3 dias 19 - 40 horas	Equipamentos mais caros. Erros ≤ 0.1% de substituição de bases.	Alto rendimento, possibilidade de sequenciamento *paired-end*. Aplica-se ao sequenciamento de genomas de bactérias, vírus, fungos, plantas, metazoários, com várias outras aplicações de abordagem genômica, metagenômica, RNA-seq, pequenos RNAs, multiplex de produtos de PCR, detecção de mutações somáticas, abordagens forenses e microbiológicas, testes pré-natais não invasivos.
ThermoFisher Ion PGM Ion Proton Ion S5 Ion S5 XL	400 pb 200 pb 600 pb 600 pb	5.5 M 80 M 80 M 80 M	3 - 23 horas 2 - 4 horas 2.5 - 4 horas 2.5 - 4 horas	Erros de formação de homopolímeros e em geral 1% de indel.	Equipamentos mais baratos. Aplica-se ao sequenciamento multiplex de produtos de PCR, microbiologia, doenças infecciosas, detecção de mutações somáticas e validação de mutações pontuais.

Pacific Biosciences	Até 40 kb	~0.35 M	1 - 6 horas	Rendimento moderado e o equipamento pode ser muito caro. Em média 13% de erros de sequenciamento.	Grande comprimento dos *reads*, com sequenciamento rápido. Utilizado para genomas complexos (como plantas e humanos), ou também genomas relacionados à microbiologia e doenças infecciosas. Detecção de fusão de transcritos e metilações.
Oxford Nanopore MinION PromethION	Até 200 kb	> 100.000	Até 48h horas	Menor rendimento, com taxa de erro em aproximadamente 12%. Custo elevado.	*Reads* longos, com sequenciadores portáteis, do tamanho da palma da mão. Aplica-se ao sequenciamento para vigilância de patógenos, detecção de mutações-alvo, metagenômica, genomas bacterianos e virais.

pb: pares de base; kb: kilobases M: milhões de *reads*; B: bilhões de *reads*; indel: inserção/deleção. Dados baseados em revisões de literatura (Quail et. al., 2012; Liu et. al., 2012, Goodwin et. al., 2016; Mardis, 2017) e nas informações disponibilizadas pelos fabricantes.

Referências Bibliográficas

Adams MD, Kelley JM, Gocayne JD, et al. (1991) Complementary sequencing: expressed sequence tags and human genome project. *Science* 252: 1651–1656.

Applied Biosystems (2011) SOLiD System accuracy with the Exact Call Chemistry module 1–8.

Bayley H (2006) Sequencing single molecules of DNA. *Current Opinion in Chemical Biology* 10: 628–637.

Bowers J, Mitchell J, Beer E, et al. (2010) Virtual Terminator nucleotides for next generation DNA sequencing. *Nature Methods* 6: 593–595.

Braha O, Walker B, Cheley S, Kasianowicz JJ, Song L, Gouaux JE, Bayley H (1997) Designed protein pores as components for biosensors. *Chemistry & biology* 4: 497–505.

Branton D, Deamer DW, Marziali A, et al. (2008) The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26: 1146–1153.

Braslavsky I, Braslavsky I, Hebert B, Hebert B, Kartalov E, Kartalov E, Quake SR, Quake SR. (2003) Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* 100: 3960–4.

Check Hayden E (2014) Data from pocket-sized genome sequencer unveiled. *Nature* 2–4.

Clarke J, Wu H, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology* 4: 265–270.

Deamer DW, Akeson M (2000) Nanopores and nucleic acids: Prospects for ultrarapid sequencing. *Trends in Biotechnology* 18: 147–151.

Eid J, Fehr A, Gray J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* (New York, N.Y.) 323: 133–138.

Eisenstein M (2012) Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology* 30: 295–296.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next generation sequencing technologies. *Nature Reviews Genetics* 17: 333-351.

Hayden EC (2015) Pint-sized DNA sequencer impresses first users Reality check for

fossil-fuel divestment. *Nature* 521: 15–16.
Heerema SJ, Dekker C (2016) Graphene nanodevices for DNA sequencing. *Nature Nanotechnology* 11: 127–136.
Holley RW, Everett GA, Madison JT, Zamir A (1965) Nucleotide sequences in the yeast alanine transfer ribonucleic acid. *The Journal of Biological Chemistry* 240: 2122–2129.
Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13770–13773.
Kawashima E, Farinelli L, Mayer P (1998) Method of Nucleic Acid Amplification. Patent US 7985565 B2.
Lander E, Linton, Lauren, Birren B, Nusbaum C (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
Lee JH, Daugharthy ER, Scheiman J, et al. (2015) Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols* 10: 442–458.
Liu L, Li Y, Li S, et al. (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* 2012.
Mardis ER (2008) Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics* 9: 387–402.
Mardis ER (2013) Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry* 6: 287–303.
Mardis ER (2017) DNA sequencing technologies: 2006-2016. *Nature Protocols* 12: 213-218.
Margulies M, Egholm M, Altman WE, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380.
Maxam a M, Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 74: 560–564.
McKernan KJ, Peckham HE, Costa GL, et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* 19: 1527–1541.
Medini D, Serruto D, Parkhill J, et al. (2008) Microbiology in the post-genomic era. *Nature Reviews Microbiology* 6: 419–430.
Metzker ML (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics* 11: 31–46.
Mikheyev AS, Tin MMY (2014) A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources* 14: 1097–1102.
Ozsolak F, Milos PM (2010) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12: 87–98.
Ozsolak F, Milos PM (2011) Transcriptome profiling using single-molecule direct RNA sequencing. (YM Kwon and SC Ricke, Eds.). *Methods in molecular biology* (Clifton, N.J.) 733: 51–61.
Ozsolak F, Platt AR, Jones DR, et al. (2009) Direct RNA sequencing. *Nature* 461: 814–818.
Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *Journal of Applied Genetics* 52: 413–435.
Quail MA, Smith M, Coupland P, et, al., (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* 242: 84–89.

Rusk N (2011) Torrents of sequence. *Nature Methods* 8: 44–44.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94: 441–448.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74: 5463–5467.

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. *Human Molecular Genetics* 19: 227–240.

Schneider GF, Dekker C (2012) DNA sequencing with nanopores. *Nature Biotechnology* 30: 326–328.

Schuster SC (2008). Next-generation sequencing transforms today's biology. *Nature methods* 5: 16–18.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature biotechnology* 26: 1135–1145.

Sims D, Sudbery I, Ilot NE, Heger A, Ponting CP (2014) Sequencing and depth coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15: 121-131.

Smith LM, Sanders JZ, Kaiser RJ, et al. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321: 674–679.

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. *Nature* 171: 737–738.

Wu R (1970) Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. *Journal of molecular biology* 51: 501–521.

Wu R (1972) Nucleotide Sequence Analysis of DNA. *Nature New Biology* 236: 198–200.

Capítulo 3

Marcador molecular CAPS - Sequências polimórficas amplificadas clivadas (*Cleaved Amplified Polymorphic Sequences*)

Dra. Ana Lúcia Anversa Segatto

Considerações Gerais

Origem do Marcador

Sequências Polimórficas Amplificadas Clivadas ou do *inglês Cleaved Amplified Polymorphic Sequences* (CAPS) são diferenças em comprimentos de fragmentos de restrição, causadas pela presença de mutações que modificam o sítio de reconhecimento de endonucleases de restrição, em sequências específicas, previamente amplificadas por PCR. Os marcadores CAPS também são chamados de marcadores PCR-RFLP, pois são derivados do RFLP- Polimorfismo de Comprimento dos Fragmentos de Restrição (do inglês *Restriction Fragment Length Polymorphism*). A técnica de RFLP consiste em digerir todo o DNA das amostras com uma ou mais enzimas de restrição. Os fragmentos obtidos são submetidos à eletroforese em gel de agarose ou poliacrilamida, desnaturados, transferidos para uma membrana e uma sonda de fita simples é hibridizada para detectar os polimorfismos (Figura 3.1), da mesma forma que na técnica de Southern Blot (Brown, 2001). As sondas podem ser detectadas por autorradiografia (Box 3.1), sendo previamente marcadas por elementos radioativos, porém, atualmente, esse tipo de sonda é pouco utilizado, sendo predominantes as sondas fluorescentes ou que participam de reações colorimétricas. Os RFLPs podem ser causados por mudanças de substituição de nucleotídeos no sítio de reconhecimento da enzima, rearranjo de DNA, inserção e/ou deleção. Por exemplo, se uma sonda hibridiza com um fragmento de dois mil pares de bases (2 Kb) em um indivíduo (A) e em outro indivíduo (B) a mesma sonda hibridiza com um fragmento de 2,3 Kb, provavelmente o indivíduo B não tem um sítio de restrição presente em A e o próximo sítio de restrição está a 0,3 Kb. Caso um indivíduo (C) apresente hibridização em fragmentos de 2 Kb e 2,3 Kb pode-se dizer que ele é heterozigoto (Figura 3.1 e Box 3.1). As vantagens da técnica são a codominância e a alta reprodutibilidade. Porém, esta técnica requer um maior trabalho laboratorial, grande quantidade de DNA com elevado grau de pureza. Além disso, necessita do conhecimento prévio da sequência para obter sondas específicas e tem um custo mais elevado quando comparada a outros marcadores, como marcadores de sequência e microssatélites.

Box 3.1

Sonda: fragmento de DNA de cadeia simples conjugado a um produto (radioisótopo, biotina, composto fluorescente) que permita sua visualização.

Marcadores dominantes: não têm a capacidade de diferenciar heterozigotos.

Marcadores codominantes: possuem a capacidade de identificar indivíduos heterozigotos.

Com o surgimento da técnica de PCR (Mullis e Faloona, 1987) foi possível desenvolver uma maneira mais rápida de detectar polimorfismos baseada nos princípios

da técnica de RFLP (Konieczny e Ausubel, 1993), os marcadores CAPS. Para a obtenção das CAPS, oligonucleotídeos iniciadores (*primers*) específicos são utilizados para amplificar uma sequência de interesse, em seguida o produto de PCR é digerido com uma ou mais enzimas de restrição e os fragmentos separados em gel de agarose ou poliacrilamida por eletroforese. No RFLP todo o DNA do indivíduo era clivado, agora são fragmentos amplificados por PCR. Os fragmentos são visualizados no gel utilizando-se os procedimentos padrões, como coloração por brometo de etídeo, GelRed™, ou nitrato de prata no caso de géis de poliacrilamida. As CAPS podem ser frutos dos mesmos processos mutacionais que causam os RFLPs, no entanto, a maneira de detecção é diferente. Por exemplo, se dois indivíduos tiverem a mesma sequência amplificada por PCR e clivada pela mesma enzima de restrição, um indivíduo (A) pode apresentar apenas um fragmento de 2 Kb no gel, provavelmente este não possui o sítio de restrição para a enzima utilizada, já um indivíduo (B) que apresenta dois fragmentos no gel, um de 0,5 Kb e um de 1,5 Kb, possui provavelmente um sítio de restrição. Se um indivíduo (C) apresentar os três tipos de fragmentos pode-se dizer que ele é heterozigoto. Apesar de ainda ser preciso ter conhecimento prévio das sequências, como no RFPL, nas CAPS são dispensados os procedimentos de confecção e detecção de sondas (Figura 3.2). Um dos primeiros trabalhos a utilizar esses marcadores, no qual eles foram chamados de CAPS pela primeira vez, foi publicado na revista *The Plant Journal* em 1993 por Konieczny e Ausubel, e descreve marcadores CAPS para mapeamento de genes em *Arabidopsis thaliana* L.

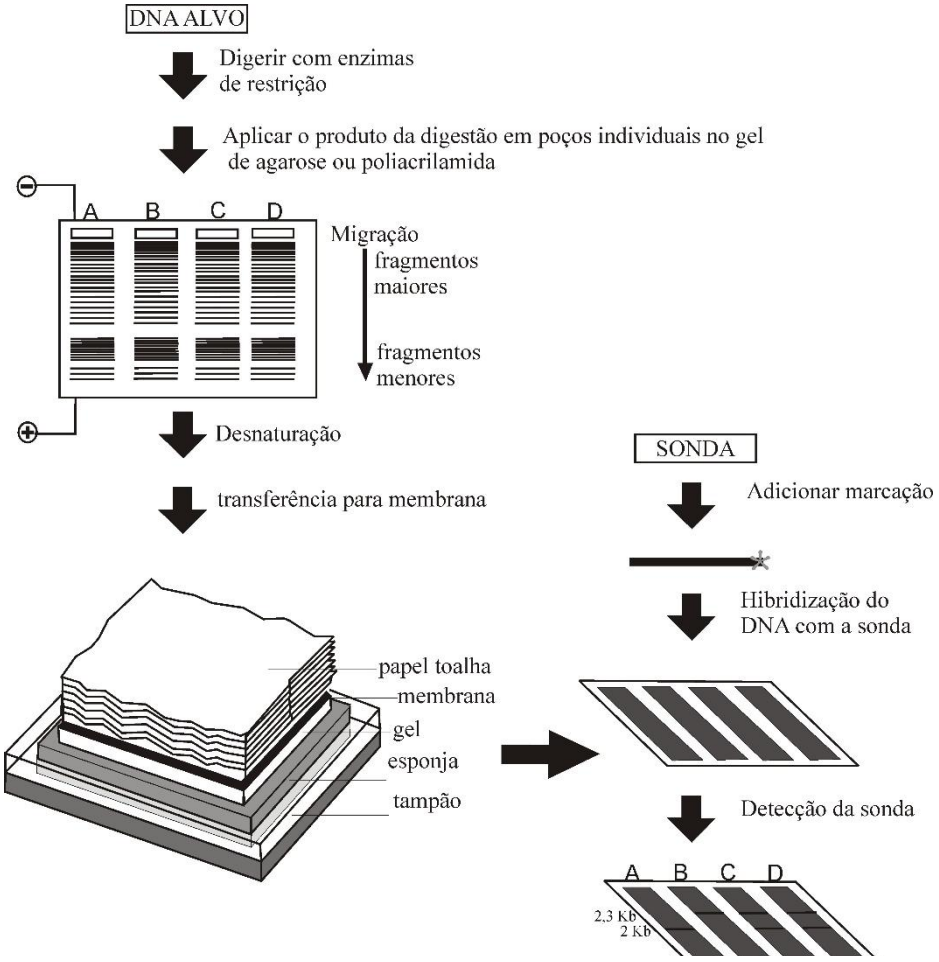

Figura 3.1 - Esquema representativo da técnica de RFLP (modificado de Brown, 2001). Os indivíduos A e B são homozigotos e os C e D são heterozigotos.

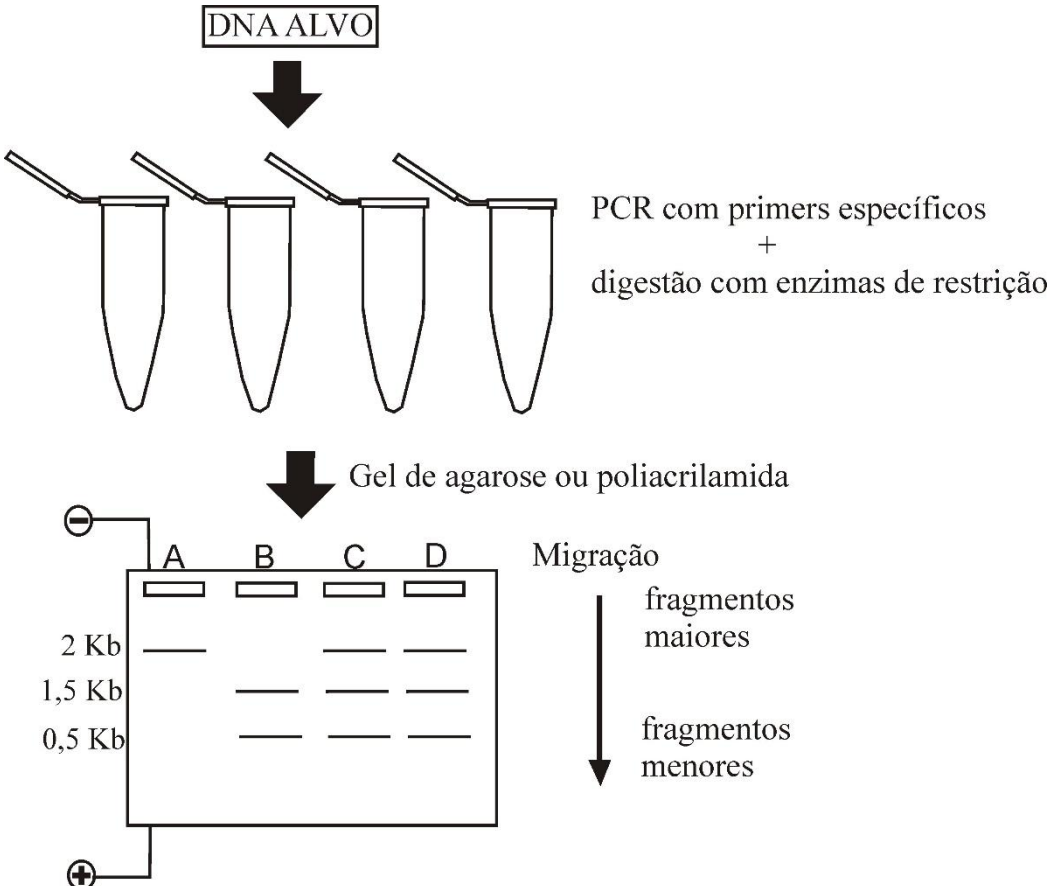

Figura 3.2 - Esquema representativo da técnica de CAPS. Os indivíduos C e D são heterozigotos.

Enzimas de restrição

A descoberta de enzimas de restrição revolucionou a biologia molecular. Existem quatro tipos de enzimas de restrição, classificadas de acordo com o número de subunidades, tipo de clivagem, grau de especificidade e requerimento de cofatores. As enzimas tipo II clivam a dupla hélice de DNA em sítios específicos e reconhecem sequências particulares de DNA, geralmente de 4, 5 ou 6 pares de bases de comprimento, chamados de sítio de reconhecimento (Pingoud e Jeltsch, 2001; Loenen et al., 2013). A clivagem ocorre em uma posição definida dentro dessa sequência. Algumas enzimas deixam extremidades cegas, outras deixam extremidades coesivas (Figura 3.3 e Box 3.2). A formação de extremidades coesivas é de grande importância para a ligação de sequências específicas de DNA por complementaridade. Os sítios de reconhecimento geralmente formam palíndromos (Box 3.2). As enzimas de restrição foram incialmente isoladas de bactérias, que utilizam essas enzimas como um meio de proteção para DNA invasor, atualmente muitas enzimas de restrição são sintetizadas em laboratório por técnica de engenharia genética. Algumas enzimas apresentam atividade fora de seu sítio de reconhecimento, conhecida como *star*

Box 3.2

Extremidades cegas: duplas fitas de ácidos nucleicos sem nucleotídeos não pareados nas extremidades.

Extremidades coesivas: duplas fitas de ácidos nucleicos com nucleotídeos não pareados na extremidade de uma das fitas.

Palíndromos: sequências que são iguais nas duas fitas quando são lidas na mesma direção. Por exemplo: 5' GAATTC 3'.

activity, por isso foram desenvolvidas enzimas de alta fidelidade em laboratório. Além disso, também foram desenvolvidas enzimas com novos sítios de clivagem e que clivam apenas uma das fitas de DNA. Uma enzima de restrição muito conhecida é a *EcoRI*, isolada da bactéria *Escherichia coli*, essa enzima reconhece o sítio 5'-GAATTC-3' e corta o DNA formando extremidades coesivas. (Figura 3.3).

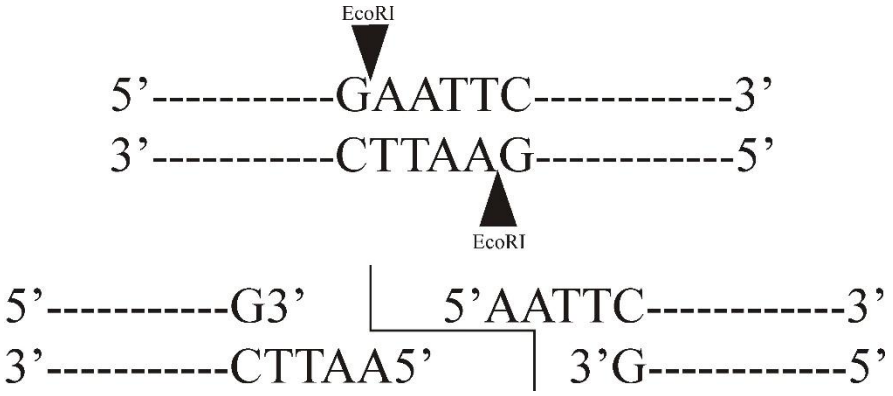

Figura 3.3 - Representação esquemática da clivagem de um fragmento dupla fita de DNA pela enzima EcoRI, o sítio de reconhecimento é indicado pelas letras representativas das bases e o triângulo preto representa o sítio de clivagem. A clivagem por EcoRI deixa extremidades coesivas.

Metodologia de isolamento/identificação e genotipagem

Na maioria dos casos marcadores do tipo CAPS são utilizados para uma detecção rápida de polimorfismos em sequências conhecidas, (por exemplo, um alelo que já sabemos causar um aumento desejado de tamanho de uma planta). A maneira pioneira de desenvolvimento de CAPS foi pela amplificação do fragmento nos indivíduos por PCR e tentativa de clivagem com diferentes enzimas de restrição, até se encontrar enzimas capazes de detectar polimorfismos (Konieczny e Ausubel, 1993). Atualmente, já se conhece o polimorfismo que se pretende identificar, por sequenciamento de mais de um indivíduo, ou por sequências já disponíveis em bancos de dados, como o GenBank (disponível em: http://www.ncbi.nlm.nih.gov/genbank/), e a identificação prévia da enzima adequada. O surgimento de novas tecnologias de sequenciamento, e sua popularização, principalmente depois de 2010, facilitou a geração de um grande número de sequências em um menor tempo e com menor custo. O procedimento mais utilizado para desenvolvimento de CAPS, atualmente, é o sequenciamento e identificação do polimorfismo de interesse, desenho dos *primers*, escolha das enzimas de restrição e genotipagem (Figura 3.4).

Existem programas de computador capazes de identificar marcadores CAPS a partir de sequências de interesse, por exemplo, o BlastDigester (Ilic et al., 2004), SGN CAPS designer e SNP2CAPS. Como arquivo de entrada é preciso fornecer sequências de interesse que apresentem polimorfismos. No programa BlastDigester o arquivo de entrada é o resultado do BLAST (Altschul et al., 1990) em formato de texto, e no SGN CAPS designer o arquivo de entrada é composto por sequências no formato FASTA ou um alinhamento no formato intervalado. O programa SNP2CAPS aceita formato de arquivos FASTA, sendo necessário fazer algumas modificações, e também vários tipos de alinhamentos múltiplos. Além disso, é necessário fornecer ao programa um banco de

dados de enzimas de restrição. Depois de identificados um ou mais polimorfismos de interesse, são desenhados *primers* que amplifiquem um fragmento, geralmente menor que 2 Kb, que contenha o polimorfismo. O polimorfismo em questão deve estar longe das extremidades 5' ou 3' dos fragmentos amplificados, isso porque a produção de fragmentos muito pequenos pelas enzimas de restrição é de difícil identificação. Além do tamanho dos fragmentos produzidos depois da restrição, outro fator importante é a quantidade de fragmentos produzidos, um padrão muito complexo é de difícil genotipagem e pode induzir erros.

Como resultado da genotipagem, é obtida uma matriz de alelos, representados por sinais, relativos à presença ou ausência do sítio de restrição (+/+; +/-; -/-), o tamanho dos alelos (1000/1000; 1000/500; 500/500) ou uma matriz representativa do padrão de bandas (1/1; 1/2; 2/2 ou a/a; a/b; b/b). Quando o fragmento amplificado possui mais de um sítio de restrição para uma determinada enzima podem ser encontrados mais de dois alelos por indivíduo. Isso também pode acontecer se mais de uma enzima de restrição for utilizada, pois cada enzima reconheceria um sítio de clivagem diferente. Também existem programas disponíveis que permitem a automatização da identificação de alelos por eletroforese, são programas de processamento de imagens, como, por exemplo, o Gelect (Intarapanich et al., 2015).

Sequências Polimórficas Amplificadas Clivadas Derivadas – dCAPS (*Derived Cleaved Amplified Polymorphic Sequences*) são uma variação da técnica de marcador CAPS. Nesse tipo de marcador mutações são introduzidas por meio de *primers* para criar o sítio de reconhecimento das enzimas. As dCAPS são utilizadas para polimorfismos que não podem ser detectados pelas enzimas de restrição disponíveis. Também são utilizadas quando uma mesma enzima identifica mais de um polimorfismo que estejam muito próximos. O programa dCAPS Finder realiza o desenho de *primers* para a inserção de sítios de reconhecimento de enzimas de restrição.

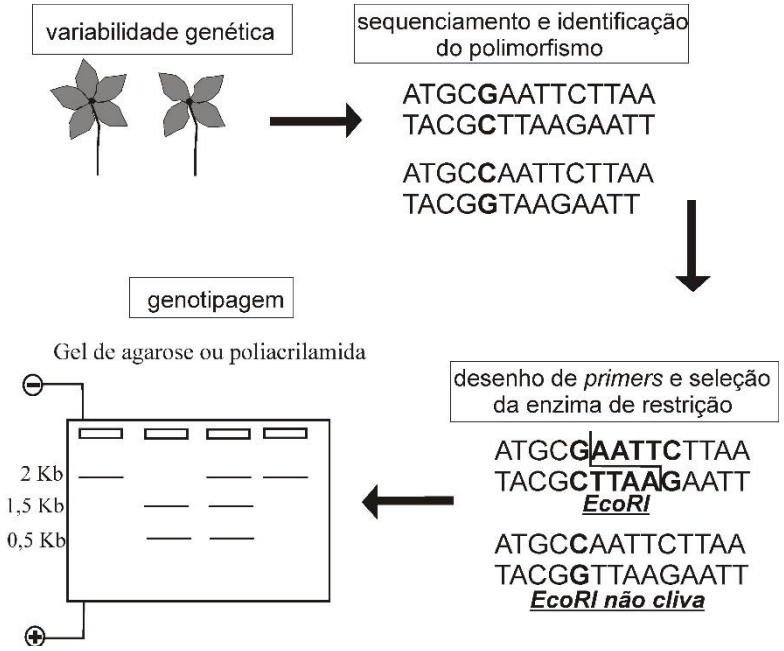

Figura 3.4 - Esquema representativo da identificação e genotipagem de marcadores CAPS.

Métodos utilizados para análise de matrizes de dados

CAPS são, na maioria das vezes, utilizados como ferramentas para mapeamento genético (Venail et al., 2010; Jaganathan et al., 2015) ou como marcadores em estudos populacionais (Segatto et al., 2014; Turchetto et al., 2014).

Quando utilizados para estudos populacionais, as primeiras análises da matriz de dados envolvem a caracterização dos marcadores, pode ser determinado o número de alelos por marcador, riqueza alélica, conteúdo informativo de polimorfismo (PIC- *Polymorphic Information Content*), heterozigosidade esperada e observada, entre outras estatísticas adequadas para este tipo de marcador. Não é possível determinar um modelo evolutivo para esses marcadores, assim as análises utilizadas devem se basear em frequências alélicas. Utilizando-se o cálculo de distância adequada, por exemplo, compartilhamento de alelos, é possível construir um agrupamento por UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*). Além disso, uma análise muito utilizada com indivíduos genotipados por marcadores CAPS é a análise de Componentes Principais (PCA - *Principal Component Analysis*), que permite a identificação de agrupamentos genéticos.

No mapeamento genético o desequilíbrio de ligação entre os *loci* segregantes é atribuído à ligação física entre eles. Existem programas de computador disponíveis que são capazes de analisar matrizes de fenótipos e marcadores moleculares permitindo o mapeamento genético e a identificação de *loci* de Características Quantitativas - QTLs (*Quantitative Trait Locus*).

Mapeamento genético e QTLs

Um mapa genético é construído utilizando-se marcadores relacionados a características fenotípicas. No processo de mapeamento são realizados cruzamentos entre indivíduos com fenótipos selecionados e conforme a frequência de recombinação entre os marcadores a distância entre eles vai sendo estimada, até que se consiga mapear todos os cromossomos de indivíduos eucariontes (Toledo et al., 2008). Em espécies modelo como, *Drosophila melanogaster* e *Arabidopsis thaliana* existem mapas genéticos bem refinados obtidos ao longo de muitos anos de estudo. Os mapas físicos, obtidos, por exemplo, de bandeamento cromossômico ou sequenciamento, são utilizados em combinação com os mapas genéticos propiciando um aumento na resolução dos mapas cromossômicos. Os principais programas utilizados para análise da ligação e ordenamento das marcas nos grupos de ligação são: Qgene, Mapmaker/EXP (Lander et al., 1987), Gmendel (Liu e Knapp, 1992), Map Manager QTX (Manly et al., 2001) e JoinMap (Van Ooijen e Voorrips, 2001).

QTLs são regiões do genoma associadas à fenótipos com distribuição contínua, tais como, altura e peso de plantas e de animais; produção de grãos; teor de óleo, entre outras (Miles e Wayne, 2008). Para o mapeamento de QTLs, são necessárias informações tais como: caráter quantitativo de interesse, dados de marcadores moleculares os quais são associados aos QTLs e uma metodologia adequada para associá-los (Toledo et al., 2008). Um QTL é chamado de *locus*, porém pode conter vários genes. Uma das grandes dificuldades do mapeamento de QTLs é determinar a posição dos *loci* identificados e a proporção do efeito de cada um (Broman e Sen, 2009). Além da posição e do número de *loci* é possível identificar o tipo de herança dos *loci*, classificando-os como aditivos, epistáticos, recessivos ou dominantes (Ferreira e Grattapaglia, 1998). Existem diversas metodologias propostas para a identificação de QTLs utilizando testes estatísticos clássicos, baseados em verossimilhança ou análise bayesiana (Toledo et al., 2008).

Aplicações

A primeira aplicação dos marcadores CAPS foi no mapeamento genético (Konieczny e Ausubel, 1993), porém com o passar do tempo muitas outras aplicações foram sendo introduzidas.

Os marcadores CAPS são principalmente utilizados para mapeamento genético e identificação de QTLs, em trabalhos com aspectos evolutivos, ou para identificação de *loci* relacionados a características de interesse econômico. Marcadores CAPS foram utilizados em combinação com microssatélites para detectar QTLs relacionadas ao comprimento do tubo da corola de flores de duas subespécies de *Petunia axillaris* (Lam.) Britton, Sterns & Poggenb com o objetivo de entender a evolução de características relacionadas a preferência de polinizadores (Venail et al., 2010). Outro estudo, visando identificar QTLs relacionadas com o peso do grão em arroz também utilizou marcadores CAPS em combinação com sequenciamento em larga escala para identificação dos polimorfismos de nucleotídeo único (SNPs - S*ingle Nucleotide Polymorphism*). Esses polimorfismos (SNPs) foram, posteriormente, identificados nos indivíduos por meio de CAPS (Xu et al., 2015). Um estudo recente, publicado na revista *Nature Genetics*, utilizou marcadores CAPS, desenvolvidos a partir de polimorfismos identificados por sequenciamento em larga escala (*RNA-seq*) para as espécies *Petunia axillaris* e *Petunia exserta* Stehmann. As CAPS foram utilizadas para genotipar populações naturais quanto a variabilidade de um *locus* determinante para a produção de flavonoides (Sheehan et al., 2016).

Um exemplo da utilização de marcadores dCAPS foi para permitir a identificação de indivíduos homozigotos e heterozigotos para uma substituição no gene que codifica a enzima *acetil-CoA carboxilase* em espécies de gramíneas (Kaundun e Windass, 2006). Essa substituição estava relacionada à resistência a herbicidas. A substituição não era reconhecida por nenhum tipo de enzima de restrição disponível e um primer foi utilizado para criar um sítio de restrição para a enzima *Nsi I* (Figura 3.5).

Outras utilizações frequentes de marcadores CAPS encontradas na literatura envolvem a identificação de híbridos (Mccleary et al., 2009; Segatto et al., 2014) e a determinação de sexo (De Kloet et al, 2011), bem como, a determinação de diversidade e estrutura genética em populações naturais (Turchetto et al., 2014). Em geral, atualmente, CAPS são utilizados em estudos de diversidade genética, principalmente em espécies que não possuem microssatélites descritos, para a obtenção de resultados rápidos, envolvendo poucos *loci*.

Considerações finais

Estamos vivendo na era da genômica e a facilidade de desenvolver marcadores moleculares para organismos não modelo vem crescendo. Muitas técnicas que diminuem a complexidade dos genomas, como Rad-seq, RRLs ou CRoPS são utilizadas permitindo a obtenção de um grande número de sequências para muitos indivíduos com menor custo e obtendo um grande número de polimorfismos (Davey et al., 2011). Depois de obtidas as sequências e identificados os polimorfismos, os CAPS ainda são uma alternativa utilizada para análises de mapeamento genético e diversidade genética, principalmente em plantas (Lui et al., 2015; Cheng et al., 2016), como uma ferramenta rápida e barata.

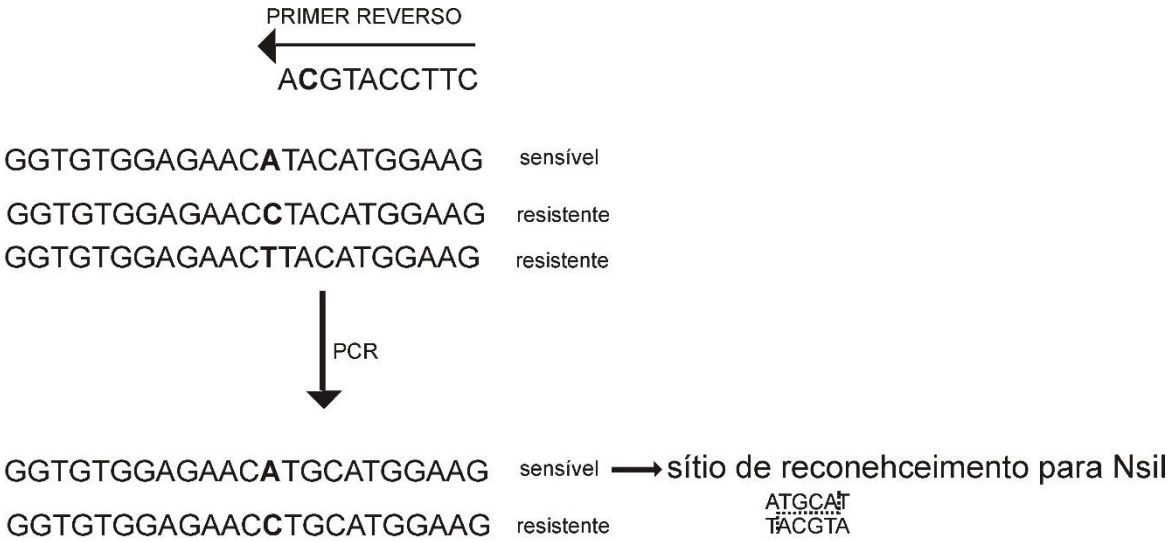

Figura 3.5 - Metodologia utilizada para o desenvolvimento dos marcadores dCAPS para a identificação de alelos relacionados a resistência a herbicidas em gramíneas.

Referências Bibliográficas

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.

Broman KW, Sen SAA (2009) Guide to QTL mapping with R/qtl. New York: Springer.

Brown TA (2001) Southern Blotting and Related DNA Detection Techniques. *Encyclopedia of Life Sceince.*

Cheng Y, Luan F, Wang X, et al. (2016) Construction of a genetic linkage map of watermelon (*Citrullus lanatus*) using CAPS and SSR markers and QTL analysis for fruit quality traits. *Scientia Horticulturae* 202: 25-31.

Davey JW, Hohenlohe PA, Etter PD, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews* 12: 499-510.

De Kloet SR (2011) Development of a CAPS (cleaved amplified polymorphic sequence) assay for sex identification of the emu (*Dromaius novaehollandiae*). *Molecular Ecology Notes* 1: 273-275.

Ferreira ME, Grattapaglia D (1998) Introdução ao uso de marcadores moleculares em análise genética. Brasília: EMBRAPA/CENARGEN.

Ilic K, Berleth T, Provart NJ (2004) BlastDigester - a web-based program for efficient CAPS marker design. *Trends in Genetics* 20: 280-283.

Intarapanich A, Kaewkamnerd S, Shaw PJ, Ukosaki K, Tragoonrung S, Tongsima S (2015) Automatic DNA Diagnosis for 1D Gel Electrophoresis Images using Bio-image Processing Technique. *BMC Genomics* 16: S15.

Jaganathan D, Thudi M, Kale S, et al. (2015) Genotyping-by-sequencing based intra-specific genetic map refines a ‘‘QTL-hotspot” region for drought tolerance in chickpea. *Molecular Genetics and Genomics* 290: 559-571.

Kaundun SS, Windass JD (2006) Derived cleaved amplified polymorphic sequence, a simple method to detect a key point mutation conferring acetyl CoA carboxylase inhibitor herbicide resistance in grass weeds. *Weed Research* 46:34–39.

Konieczny A, Ausubel FM (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant Journal* 4: 403-410.

Lander ES, Green P, Abrahmson J, et al. (1897) Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1: 174-181.

Liu BH, Knapp S (1992) GMendel, a program for Mendelian segregation and linkage analysis of individual or multiple progeny population using loglikelihooh ratios. *Journal of Heredity* 8: 407-418.

Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG, Murray NE (2013) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Research* 42: 3-19.

Manly KF, Cudmore JR, Meer JM (2001) Map Manager QTX, crossplatform software for genetic mapping. *Mammalian Genome* 12: 930-932.

Mccleary TS, Robichaud RL, Nuanes S, Anagnostakis SL, Schlarbaum SE, Severson, JR (2009) Four cleaved amplified polymorphic sequence (CAPS) markers for the detection of *the Juglans ailantifolia* chloroplast in putatively native *J. cinerea* populations. *Molecular Ecology Resources* 9: 525-527.

Miles C, Wayne M (2008) Quantitative trait locus (QTL) analysis. *Nature Education* 1: 208.

Mullis KB, Faloona F (1987) Specific synthesis of DNA in vitro via a polymerase chain reaction. *Methods in Enzymology* 155: 355-350.

Pingou A, Jeltsch A (2001) Structure and functions of type II restriction endonucleases. *Nucleic Acids Research* 29: 3705-3727.

Sheehan H, Moser M, Klahre U, et al. (2016) MYB-FL controls gain and loss of floral UV absorbance, a key trait affecting pollinator preference and reproductive isolation. *Nature Genetics* 48: 159-169.

Segatto ALA, Cazé ALR, Turchetto C, et al. (2014) Nuclear and plastid markers reveal the persistence of genetic identity: a new perspective on the evolutionary history of Petunia exserta. *Molecular Phylogenetics and Evolution* 70: 504-512.

Toledo ER, Leandro RA, Souza Junior CL, Souza AP (2008) Mapeamento de QTLs: uma abordagem bayesiana. *Revista Brasileira de Biometria* 26: 107-114.

Turchetto C, Fagundes NJR, Segatto ALA, et al. (2014) Diversification in the South American Pampas: the genetic and morphological variation of the widespread Petunia axillaris complex (Solanaceae). *Molecular Ecology* 23: 374–389.

Van Ooijen JW, Voorrips RE (2001) JoinMap version 3.0: software for the calculation of genetic linkage maps (software). Wageningen. *Plant Research International*.

Venail J, Dell'Olivo A, Kuhlemeier C (2010) Speciation genes in the genus Petunia. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 461-468.

Xu F, Sun X, Chen Y, Huang Y, Tong C, Bao J (2015) Rapid Identification of Major QTLs Associated with Rice Grain Weight and Their Utilization. *PlosOne* 27: 1-13.

Capítulo 4

Marcadores moleculares baseados em restrição: AFLP e suas variações

Dra. Márcia Goetze, Dra. Gisele Passaia, Dra. Fernanda Sperb-Ludwig

Considerações gerais

Neste capítulo os marcadores moleculares do tipo AFLP (*Amplified Fragment Lenght Polymorphism*) ou Polimorfismo de Comprimento de Fragmentos Amplificados serão descritos. Essa classe de marcadores é baseada na associação dos polimorfismos gerados por enzimas de restrição com posterior amplificação por PCR (*Polymerase Chain Reaction*). A técnica foi desenvolvida em Wageningen, Holanda e descrita em uma publicação científica pela primeira vez em 1995 por Vos e colaboradores (Vos et al., 1995). A técnica envolve basicamente quatro passos: 1) clivagem do DNA genômico, 2) ligação de adaptadores aos fragmentos gerados durante a clivagem do DNA, 3) amplificação seletiva de fragmentos via PCR, e 4) separação dos fragmentos selecionados em gel de alta resolução. Estima-se a geração de 40 a 200 fragmentos por reação de PCR (par de *primers* durante a amplificação seletiva), sendo que a detecção do polimorfismo se dá devido às mutações nos sítios reconhecidos pelas enzimas de restrição utilizadas (Grover e Sharma, 2016). Os fragmentos gerados são avaliados quanto à presença e ausência. Inicialmente, os fragmentos da amplificação seletiva eram analisados em gel de poliacrilamida corado com nitrato de prata, ou os fragmentos eram marcados radioativamente. Com o avanço e diminuição do custo dos métodos de sequenciamento, atualmente a análise dos produtos de amplificação é realizada via eletroforese capilar usando sequenciador, o que torna a análise mais rápida e confiável. Algumas características que tornam a técnica de AFLP uma ferramenta interessante para a análise de populações segregantes são: alta reprodutibilidade, rápida geração dos fragmentos e alta frequência de identificação de polimorfismos (Alexander et al., 2012). Os fragmentos estão distribuídos ao longo de todo o genoma, são de herança dominante (alelos dominantes são aqueles que se manifestam no fenótipo em dose simples, ou seja, mesmo que outro alelo seja diferente dele), não sendo necessário conhecimento prévio do genoma para a utilização da técnica. Entre os marcadores moleculares, AFLPs são considerados mais reprodutíveis do que RAPDs (*Random Amplified DNA Polymorphism*), portanto, mais confiáveis para estudos genéticos e de populações (Savelkoul et al., 1999). Por outro lado, a informação de polimorfismos gerada por AFLP é menor do que aquela obtida com marcadores do tipo microssatélite (Pejic et al., 1998).

A análise do DNA pela técnica de AFLP pode ser aplicada para qualquer organismo vivo, desde microrganismos, fungos, algas, animais e plantas. Nesse capítulo iremos abordar as aplicações da técnica na análise do DNA de plantas de importância agronômica, no estudo da genética de populações, com ênfase em populações de plantas nativas, visando a conservação de espécies. Além disso, casos clínicos aonde a técnica vem sendo aplicada com sucesso serão discutidos. O capítulo tratará das metodologias de isolamento dos marcadores, identificação dos fragmentos e genotipagem. Adicionalmente, métodos para análise das matrizes de dados gerados serão abordados. Por fim, apresentaremos uma série de exemplos em que esse marcador e suas derivações

(cDNA-AFLP, CRoPS – *Complexity Reduction of Polymorphic Sequences* e RAD-seq – *Restriction-site associated DNA sequencing*) foram usados com sucesso, respondendo a importantes perguntas científicas, tanto para a área de melhoramento genético e biotecnologia quanto para a conservação de espécies.

Metodologia de identificação e isolamento

A técnica de AFLP consiste em quatro principais etapas: 1) clivagem do DNA genômico, 2) ligação de adaptadores aos fragmentos gerados durante a clivagem do DNA, 3) amplificação seletiva de fragmentos via PCR, e 4) separação dos fragmentos selecionados em gel de alta resolução.

1) Clivagem do DNA genômico: essa etapa é realizada utilizando-se duas enzimas de restrição, uma de corte raro, que reconhece de 6 a 8 pares de base (pb) no genoma do organismo alvo, e outra de corte frequente, que reconhece 4 pb (Figura 4.1a). Três classes de fragmentos são gerados após a clivagem do DNA: (i) fragmentos cortados em ambas as extremidades pela enzima de corte raro; (ii) fragmentos cortados em ambas as extremidades pela enzima de corte frequente; (iii) fragmentos cortados por ambas as enzimas, de corte raro e frequente. A escolha das enzimas de restrição a serem utilizadas durante o desenvolvimento de algum estudo depende de algumas características do genoma do organismo em questão, como, por exemplo, a complexidade e composição de bases do genoma, ocorrência de sítios metilados, etc. O DNA de muitos eucariotos é rico em sequências AT, fazendo da enzima *Mse*I (que tem o sítio de reconhecimento TTAA) a preferida como endonuclease de corte frequente. Exemplos de enzimas de restrição de corte raro incluem a *Eco*RI, *Ase*I, *Hind*III, *Apa*I e *Pst*I (Vos et al., 1995). As enzimas *Eco*RI e *Mse*I são as mais utilizadas nos estudos de AFLP, com algumas exceções (Meudt e Clarke, 2007). Uma clivagem de DNA bem sucedida requer altas concentrações de DNA (~ 50 - 500 ng), mas, mais importante, um DNA de boa qualidade (não degradado e livre de contaminantes que possam inibir a clivagem do DNA, a ligação dos adaptadores e amplificação dos fragmentos).

2) Ligação dos adaptadores: adaptadores (oligonucleotídeos sintéticos) de fita dupla, entre 10 a 30 pb de comprimento, são ligados às extremidades correspondentes aos sítios de restrição das enzimas utilizadas durante a clivagem, através de uma DNA ligase. Essa etapa não restaura o sítio de reconhecimento das enzimas de restrição devido a uma alteração incorporada na sequência dos adaptadores (Figura 4.1b, representada por letras minúsculas). Essa mudança previne que ocorra uma nova clivagem após a ligação dos adaptadores, permitindo a realização das etapas 1 e 2 no mesmo tubo. A sequência dos adaptadores mais o sítio de reconhecimento da enzima de restrição do fragmento (representados em vermelho e verde, respectivamente para *Eco*RI, e em vermelho e laranja, respectivamente para *Mse*I na Figura 4.1b) servirão de molde para o anelamento dos *primers* na próxima etapa.

3) Amplificação seletiva dos fragmentos via PCR: o número de fragmentos gerados até essa etapa é muito alto, o que impossibilitaria obter-se uma resolução clara, mesmo em géis de alta resolução. Por isso, faz-se necessário uma etapa de seleção dos fragmentos a serem analisados. Dessa maneira, *primers* arbitrários (que não requerem nenhum conhecimento prévio do genoma) são utilizados na etapa 3. Eles apresentam em sua extremidade 5' a sequência complementar ao adaptador, na região central a sequência complementar ao sítio de restrição, e, na extremidade 3', de um a três (normalmente) nucleotídeos arbitrários (seletivos). A presença dos nucleotídeos seletivos na extremidade 3' dos *primers* exercerá uma pressão seletiva, pois o anelamento dos *primers* somente ocorrerá naqueles fragmentos onde a base arbitrária

(ou seletiva) for complementar a base presente na região interna do fragmento. A escolha de quantos nucleotídeos seletivos deve-se usar dependerá da complexidade do genoma do organismo alvo. O número de fragmentos gerados após a amplificação seletiva não deverá ser muito alto ao ponto de influenciar a interpretação e gerar perfis difíceis de serem genotipados, mas em um número suficiente para gerar informação. Quanto maior o número de bases arbitrárias utilizadas nos *primers*, menor a complexidade e número de fragmentos obtidos.

Para minimizar a geração de artefatos, e quando genomas complexos estão sendo analisados, duas amplificações seletivas podem ser incorporadas: uma pré-amplificação, onde *primers* com somente um nucleotídeo seletivo são utilizados (Figura 4.1c); e uma amplificação seletiva, com *primers* com duas ou três bases arbitrárias (Figura 4.1d). Na etapa de pré-amplificação somente os fragmentos que apresentarem complementariedade ao nucleotídeo seletivo do *primer* serão amplificados. Sendo assim, como a pressão de seleção ocorre em ambos os sítios de restrição, somente 1 em cada 16 fragmentos será amplificado (somente 1 em 4 x 4 bases). Os fragmentos selecionados nessa reação de PCR são diluídos e utilizados como molde na amplificação seletiva. Quando *primers* com 3 nucleotídeos seletivos são utilizados somente 1 em cada 4096 (1 em cada 4^3 x 4^3) fragmentos são amplificados. Dessa maneira, somente uma subpopulação dos fragmentos gerados é amplificada e genotipada. Altas temperaturas de anelamento são utilizadas na etapa 3, garantindo o emparelhamento perfeito das sequências (*primers* mais molde), diminuindo a geração de artefatos inespecíficos. Na amplificação seletiva, a extremidade 5' dos *primers* que se anelam ao adaptador do sítio da enzima de restrição de corte raro são normalmente marcados com radioisótopos, ou mais comumente nos dias atuais, fluoróforos (Figura 4.1d). Isso faz com que os fragmentos amplificados a partir desses *primers* sejam marcados com fluorescência, permitindo a visualização dos fragmentos na etapa 4. Dessa forma, somente as classes de fragmentos (i) e (iii) gerados durante a clivagem do DNA são visualizados (os fragmentos de corte raro-raro e corte raro-frequente).

Um grande número de fragmentos consegue ser gerado com a utilização de várias combinações de *primers* na etapa de amplificação seletiva, a partir da mesma clivagem e pré-amplificação. Recomenda-se a realização de um teste piloto com várias combinações de *primers* com representantes da amostra global, antes da genotipagem do número total de indivíduos, para a escolha da combinação de *primers* com o melhor resultado (melhor perfil). Perfis com alta qualidade apresentam bandas (no gel) ou picos (no sequenciador) bem separados, uma alta relação sinal ruído, poucas ou nenhuma presença de bandas ou picos fantasmas (*stutter*), fragmentos distribuídos ao longo da variação estabelecida para a genotipagem e polimorfismos claros (Meudt e Clarke, 2007).

4) Separação dos fragmentos em gel de alta resolução ou em eletroforese capilar: os fragmentos obtidos durante a amplificação seletiva podem ser submetidos a eletroforese em gel de poliacrilamida de alta resolução e os padrões obtidos são visualizados por autorradiografia (quando *primers* marcados com radioisótopos são utilizados). Em muitos casos também é possível fazer uso de géis corados com nitrato de prata para a visualização dos fragmentos (nesse método os *primers* utilizados durante a amplificação seletiva não precisam de nenhum tipo de marcação). Entretanto, nos dias atuais, cada vez mais se tem realizado a separação dos fragmentos em eletroforese capilar (sequenciador automático), utilizando-se *primers* marcados com fluoróforos na amplificação seletiva. Com esse método os *primers* podem ser marcados com até quatro fluoróforos distintos (nas plataformas da Applied Biosystems) e os produtos de cada um

desses *primers* (fragmentos amplificados) podem ser agrupados para a eletroforese capilar, em um sistema denominado *poolplexing*.

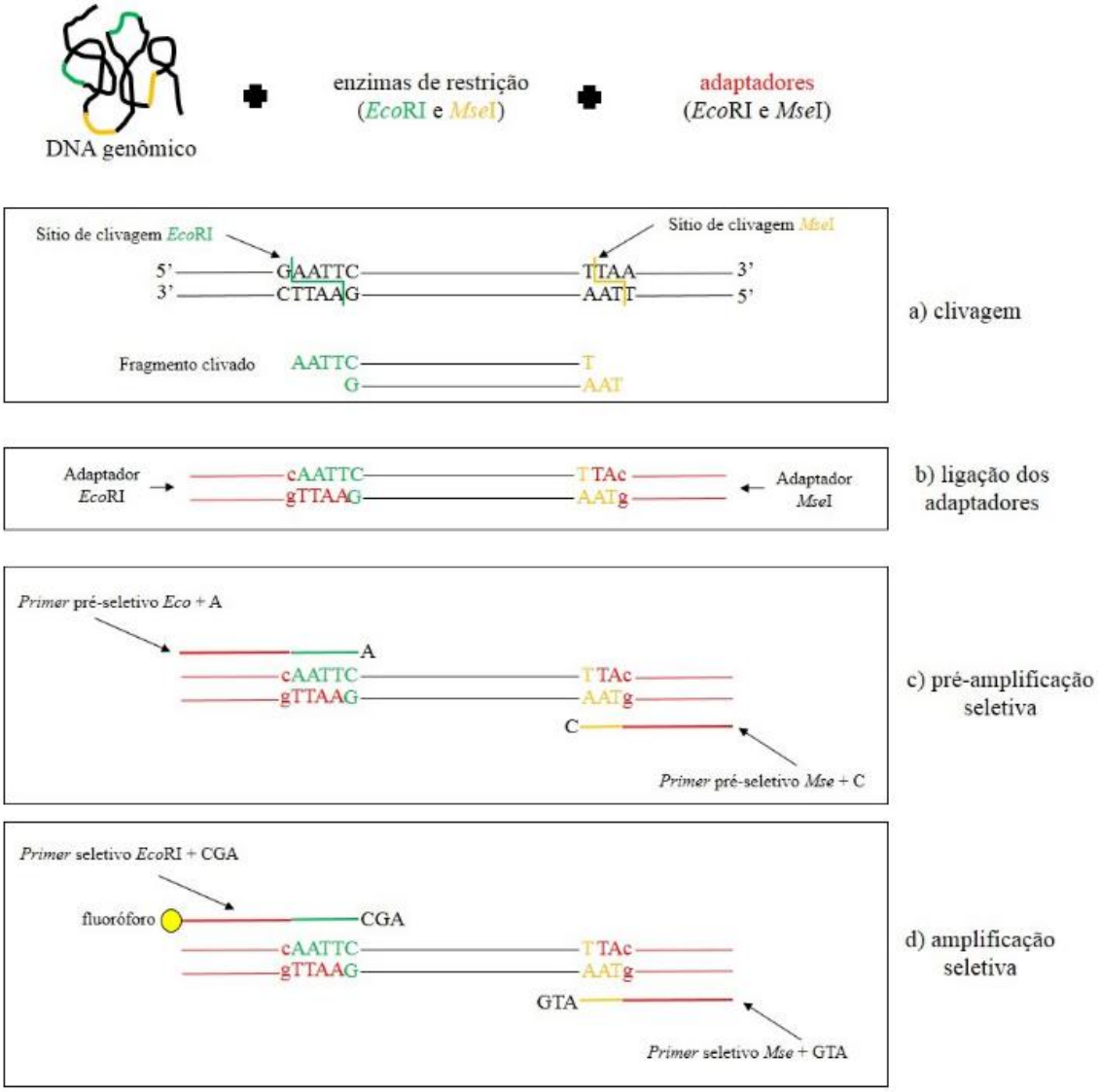

Figura 4.1 - Esquema mostrando três etapas da técnica de AFLP: digestão, ligação dos adaptadores e amplificação seletiva dos fragmentos. (a) O DNA genômico é clivado com enzimas de restrição, nesse caso EcoRI (sítios em verde) e MseI (laranja). (b) Adaptadores sintéticos (em vermelho) são ligados ao fragmento clivado. A sequência dos adaptadores é alterada (mostrado em letras minúsculas) de tal forma que o sítio de restrição não seja restaurado, permitindo que a clivagem e a ligação dos adaptadores seja realizada no mesmo tubo. (c) Pré-amplificação seletiva utilizando primers com um nucleotídeo arbitrário (seletivo) e (d) amplificação seletiva com primers com três nucleotídeos arbitrários.

Genotipagem

A genotipagem do perfil molecular obtido com o desenvolvimento da técnica de AFLP é uma das etapas mais importantes e muitas vezes a mais complexa de todo o processo. É nessa etapa que, a partir de um gel de poliacrilamida ou do eletroferograma do sequenciador (Figura 4.2), é extraída uma matriz binária de ausência (0) e presença

(1) de um determinado fragmento ou alelo (Figura 4.3). Uma das principais dificuldades nessa etapa é determinar quais são os fragmentos que realmente são homólogos quando se está analisando diversos indivíduos ou *taxa*. É necessário estabelecer se pequenas variações de tamanho (por exemplo, picos com tamanhos de 156,7 e 157,2 pb) correspondem ao mesmo fragmento (alelo) ou não. Uma vez que um padrão for estabelecido (por exemplo, considerar uma variação de 0,5 pb, como mencionado no exemplo, como o mesmo fragmento ou alelo), ele deve ser seguido até o final do projeto. Também é necessário estabelecer um ponto de corte mínimo (intensidade mínima do sinal do pico) a partir do qual um fragmento é considerado presente, e a variação do tamanho dos fragmentos que será analisada (por exemplo, somente serão considerados fragmentos de tamanho variando entre 80 e 400 pb). Atualmente existem muitos programas que fazem a genotipagem automática dos fragmentos, como GeneMapper®, GeneMarker®, Genographer, mas uma revisão manual normalmente se faz necessária.

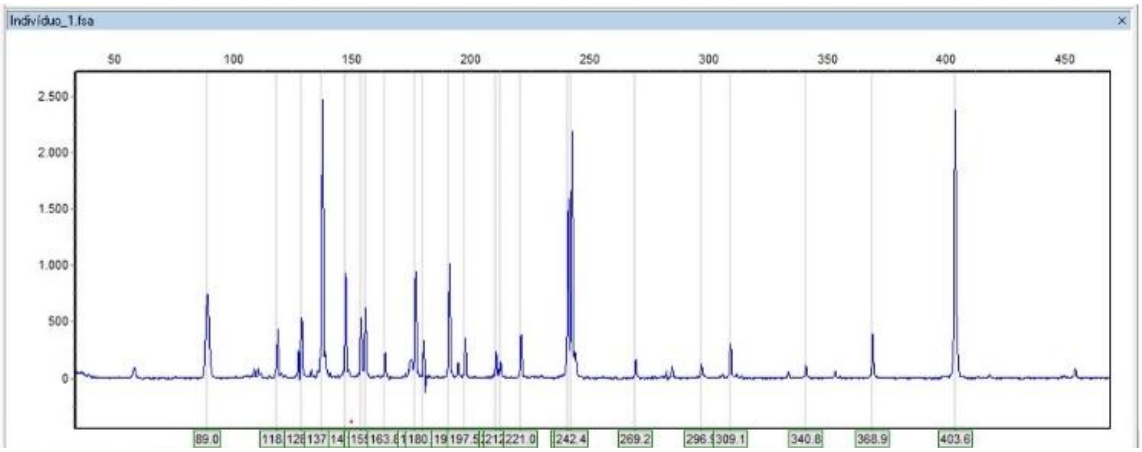

Figura 4.2 - Exemplo de eletroferograma obtido com o sequenciador automático para um loco de AFLP.

	92	101	107	109	111	113	116	117	120	122	127	129	132	136	137	138	147	149	150	153	155	157	159	161	162	163	165	173
Indivíduo_1	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
Indivíduo_2	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1
Indivíduo_3	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_4	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1
Indivíduo_5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_6	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1
Indivíduo_9	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_10	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0
Indivíduo_11	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1
Indivíduo_12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0
Indivíduo_13	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1
Indivíduo_14	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Indivíduo_15	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Indivíduo_16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Indivíduo_19	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_20	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Indivíduo_21	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_22	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
Indivíduo_23	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
Indivíduo_24	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Indivíduo_25	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1
Indivíduo_26	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
Indivíduo_27	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1

Figura 4.3 - Exemplo de matriz binária que deve ser montada a partir dos eletroferogramas para as análises dos dados de AFLP. Nas linhas estão representados os genótipos de 27 indivíduos e nas colunas os alelos.

Os polimorfismos observados entre as amostras analisadas (bandas ou picos presentes em algumas amostras, mas ausentes em outras) são originadas por mutações, as quais podem ser mutações pontuais que originam novos sítios de restrição ou causam a perda de um anteriormente existente. O polimorfismo também pode ser originado pela presença de eventos de deleção ou duplicação entre os sítios de restrição (causando a diminuição ou o aumento do fragmento amplificado, respectivamente), ou mutações (por exemplo SNPs – *Single Nucleotide Polymorphism*) no sítio(s) de anelamento(s) dos *primers* (Meudt e Clarke, 2007). Exemplos de alguns eventos que geram o polimorfismo observado com marcadores AFLP podem ser observados na Figura 4.4.

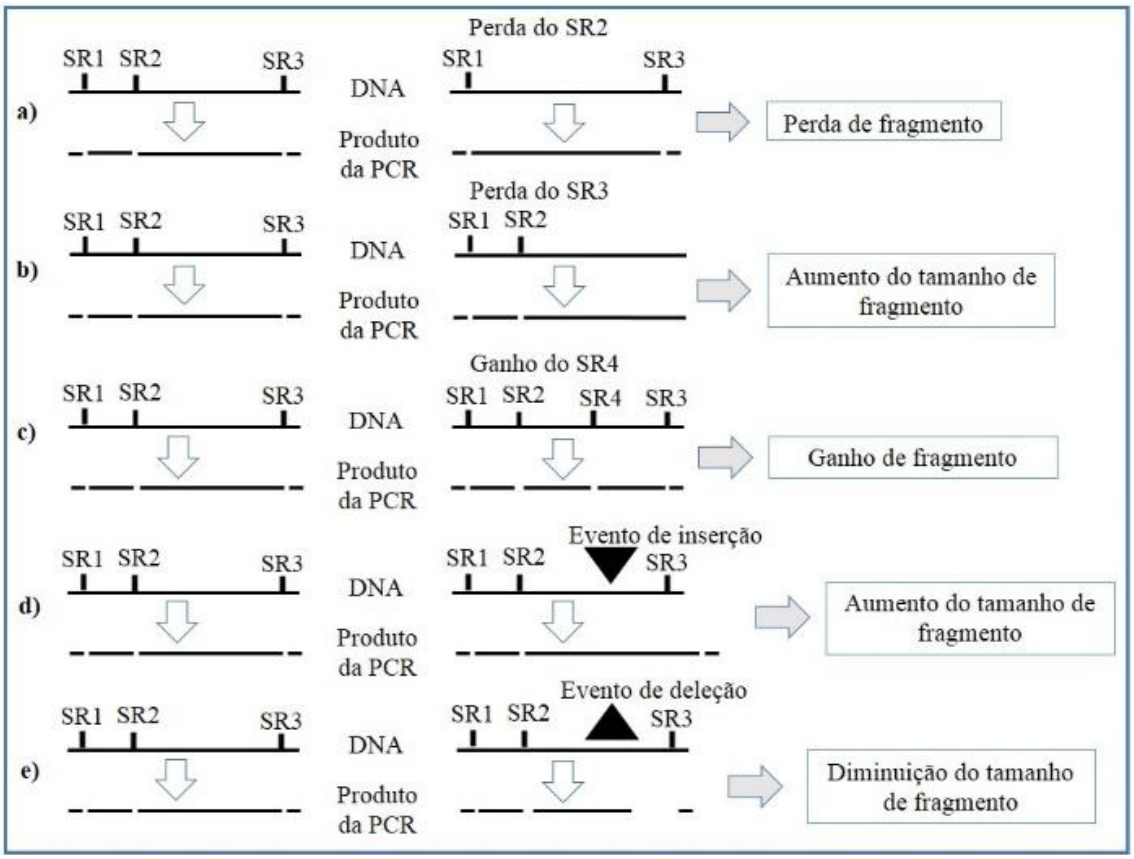

Figura 4.4 - Exemplos de mutações que podem ocorrer e ocasionar o polimorfismo detectado com os marcadores AFLP. (a) Perda do SR2, levando a perda do fragmento. (b) Perda do SR3, levando ao aumento do tamanho do fragmento. (c) Ganho do SR4, levando ao ganho de fragmento. (d) Evento de inserção, levando ao aumento do tamanho do fragmento. (e) Evento de duplicação, levando a diminuição do tamanho do fragmento. SR1 - sítio de restrição 1; SR2 - sítio de restrição 2; SR3 - sítio de restrição 3; SR4 - sítio de restrição 4.

Vantagens e Limitações da Técnica de AFLP

Com a utilização da técnica de AFLP é gerado um grande número de marcadores moleculares, de uma maneira rápida e com alta reprodutibilidade e robustez. Esse método combina o poder de amostragem da digestão enzimática do genoma, aliado a praticidade e velocidade de detecção dos polimorfismos via PCR. O alto número de fragmentos gerados com essa técnica é uma das principais vantagens dessa metodologia. Além disso, os fragmentos estão amplamente distribuídos ao longo do genoma nuclear,

apesar de às vezes concentrados em regiões centroméricas (Meudt e Clarke, 2007), sendo possível acessar polimorfismos ao longo de todo o genoma, aumentando a chance de detectar variações genéticas raras, por exemplo. Dessa forma, a rapidez, o baixo custo, a alta reprodutibilidade da técnica (pelo uso de *primers* mais longos e altas temperaturas de anelamento quando comparado com RAPD), e o fato de não ser necessário conhecimento prévio do genoma do organismo em estudo, podem ser apontadas como algumas das vantagens da utilização dessa técnica.

A principal limitação da técnica de AFLP é a dificuldade de distinção de indivíduos heterozigotos dos homozigotos, ou seja, o fato de serem marcadores dominantes (apenas um alelo é detectado, o fragmento amplificado). Isso acaba sendo problemático principalmente em estudos que precisam estimar a frequência dos alelos, como análises de heterozigosidade (Mueller e Wolfenbarger, 1999). Outra limitação relacionada ao uso de AFLPs é a homoplasia de tamanho, bandas de mesmo tamanho podem não ser homólogas e representar dois ou mais locus, o que pode influenciar estimativas de diversidade genética e reconstruções filogenéticas (Bensch e Åkesson, 2005). Para garantir a reprodutibilidade da técnica, o DNA utilizado deve ser de alta pureza, permitindo que o genoma seja completamente clivado pelas enzimas de restrição. A digestão parcial do DNA pode levar a ausência ou presença de alelos devido a outras razões que não o polimorfismo genético. Poucas bandas, ou fragmentos muito pequenos, podem indicar, por exemplo, que o DNA utilizado na clivagem estava degradado (Ferreira e Grattapaglia, 1995; Blears et al., 1998). Muitos autores ainda consideram a técnica de AFLP cara e laboriosa, além de necessitar de conhecimento técnico para a interpretação dos géis ou eletroferogramas (Grover e Sharma, 2016).

Métodos Utilizados para Análise de Matrizes de Dados e Aplicações

Análise da diversidade genética de populações ou espécies

Estimativas da diversidade genética de populações ou espécies podem contribuir para a elaboração de estratégias de conservação, através, por exemplo, da identificação de locais (populações) com maior diversidade genética, os quais devem ser priorizados quando da criação de unidades de conservação. Também é possível identificar e estimar o risco de extinção e o potencial evolutivo do organismo sob a análise. Parâmetros como heterozigosidade, Conteúdo Informativo de Polimorfismo (PIC, da sigla em inglês), porcentagem de locus polimórficos, índice de diversidade de Shannon, de Nei, etc., podem ser calculados a partir de matrizes de ausência e presença de marcadores AFLP. Programas normalmente utilizados para a obtenção desses parâmetros incluem o AFLP-SURV (Vekemans et al., 2002), NTSYSpc (Rohlf, 1997) e o POPGENE (Yeh e Boyle, 1997). Entretanto, muitos desses parâmetros levam em conta a estimativa das frequências alélicas, o que é dificultado no caso de marcadores dominantes como AFLP, pois a presença de uma banda (ou pico) pode indicar tanto um indivíduo heterozigoto quanto um homozigoto. Para resolver essa questão, duas principais metodologias são comumente utilizadas: i) uma abordagem Bayesiana onde a frequência do alelo ausente é estimada (Zhivotovsky, 1999) e, ii) a metodologia proposta por Lynch e Milligan (Lynch e Milligan, 1994), que assumindo que as populações ou espécies estejam em equilíbrio de *Hardy-Weinberg*, estima a frequência do alelo ausente através do cálculo da raiz quadrada da frequência de indivíduos sem a banda (0/0). Krauss (2000), avaliando essas metodologias concluiu que ambas levaram a resultados similares de heterozigosidade. Em um estudo realizado com gengibre Blanco e colaboradores (2016), utilizaram treze combinações de *primes* AFLP para a caracterização da

diversidade genética de 55 acessos brasileiros e seis originários da Colômbia. Os resultados mostraram que os acessos colombianos apresentaram maior diversidade genética. Entre as amostras brasileiras, as das regiões Sul e Sudeste apresentaram os maiores índices de diversidade. Os resultados obtidos poderão ser utilizados para o melhoramento genético do gengibre e para a manutenção do germoplasma da espécie.

Estrutura populacional

Estudos de estruturação populacional podem estimar o fluxo gênico e dispersão de uma espécie, além de fornecer dados sobre a diferenciação genética de populações ao longo da distribuição geográfica do organismo. Esses dados podem ser utilizados também em programas de conservação e podem elucidar eventos históricos ou geológicos responsáveis pelo padrão observado (estruturação). Matrizes de distância genética podem ser calculadas e utilizadas em testes de Mantel (que identificam se ocorre isolamento por distância). Valores de F_{ST} também podem ser obtidos a partir de dados de AFLP, utilizando, por exemplo, o programa Arlequin (Excoffier e Lischer, 2010). O programa Structure (Pritchard et al., 2000), que é muito utilizado para a identificação de estrutura genética entre populações, pode ser usado com dados de AFLP, produzindo resultados bem interessantes e informativos. AFLP é uma técnica muito comum em genética de populações (diversidade e estruturação) e diversos estudos com esse enfoque podem ser encontrados. Por exemplo, Rodrigues e colaboradores (2016), amostraram 11 populações de *Hypochaeris lutea*, uma erva perene que cresce em altitudes aproximadas de 1000 metros no Sul do Brasil, para elucidar a estruturação genética da espécie. Seis combinações de *primers* foram utilizadas, a partir dos quais 193 marcadores (fragmentos) foram obtidos. Alta diversidade genética dentro das populações e uma baixa estruturação genética foram encontrados para a espécie. Estudos em espécies animais com AFLP são menos frequentes. Entretanto, vários exemplos podem ser encontrados na literatura. Kneeland e colaboradores (2013) avaliaram a variabilidade genética da mosca *Stomoxys calcitrans*, o qual resultou em 387 bandas polimórficas a partir de quatro marcadores AFLP, que mostraram que a variação genética foi maior dentro das 12 populações avaliadas. Além disso, nenhuma diferenciação genética foi observada entre as localidades, evidenciando um alto fluxo gênico ao longo da distribuição geográfica da espécie. Marcadores AFLP também são utilizados para a caracterização do germoplasma de espécies cultivadas. Rapposelli e colaboradores (2015) utilizaram quatro combinações de *primer*s de AFLP, que geraram 165 bandas polimórficas, para caracterizar a diversidade genética e estruturação de *Salvia desoleana*. As folhas dessa espécie são utilizadas como fonte de óleos essencias na indústria farmacêutica e de cosméticos. Nesse estudo, tanto populações cultivadas como selvagens foram investigadas e os resultados mostraram que as populações cultivadas apresentam maior diversidade genética. Além disso, as análises do programa Structure indicaram a ocorrência de estruturação genética entre as populações cultivadas e selvagens. As informações geradas com o estudo poderão ser utilizadas na elaboração de estratégias de conservação para o germoplasma dessa espécie (especialmente dos acessos selvagens, que apresentaram uma menor diversidade), e também para o aumento da produtividade de óleo essencial para essa espécie.

Relacionamento filogenético interespecífico

Marcadores AFLP vêm sendo utilizados para acessar o relacionamento genético de organismos (Koopam, 2005), especialmente em espécies proximamente relacionadas, de divergência recente, onde outros tipos de marcadores (morfológicos ou sequências

nucleares e plastidiais, ver Capítulo 5) não revelaram polimorfismo suficiente ao nível taxonômico analizado (Després et al., 2003). Pelo fato dos marcadores AFLP estarem amplamente distribuídos no genoma, a chance de acessar variações genéticas raras entre os *taxa* em estudo é maior. Entretanto, para níveis taxonômicos mais altos (acima de gênero), inferências filogenéticas baseadas em AFLP se tornam problemáticas, devido a alta variabilidade apresentada por esse marcador, diminuindo a similaridade encontrada entre *taxa* distantes, e aumentando a chance de dois fragmentos de mesmo tamanho em espécies diferentes não serem homólogos (Mueller e Wolfenbarger, 1999). Testes para verificar se um determinado conjunto de dados tem realmente sinal filogenético (se não são muito divergentes) estão disponíveis para serem utilizados (Koopman e Gort, 2004). Cuidado deve ser tomado também na escolha das espécies ou *taxa* que serão utilizados como grupo externo nas inferências filogenéticas, uma vez que grupos muito distantes podem introduzir ruído na análise, baixando os valores de suporte estatístico da árvore, assim como mudando a topologia (Kirchberger et al., 2014). Para verificar a robustez dos resultados obtidos com marcadores AFLP em análises filogenéticas, muitos estudos os comparam àqueles que utilizaram outros tipos de marcadores (normalmente sequências ITS – *Internal Transcribed Spacer*). Além disso, muitos estudos combinam os dados gerados com AFLP à outras sequências de DNA, o que pode aumentar a robustez.

A matriz de presença e ausência gerada a partir da genotipagem dos marcadores AFLP pode ser usada diretamente nos métodos filogenéticos ou então utilizada para gerar matrizes de distância genética entre os indivíduos. Métodos baseados em distância genética, como *neighbour joining*, ou baseado em caracteres como parcimônia, máxima verossimilhança e análise Bayesiana podem ser utilizados nas inferências filogenéticas. Para as análises baseadas em distância, programas como o SplitsTree (Bryant e Moulton, 2004) geram uma matriz de distância (ou mais), e a partir dessa matriz o relacionamento dos indivíduos é inferido. O programa PAUP (Swofford, 2002) também é frequentemente utilizado. Análises bayesianas com AFLP podem ser realizadas em programas como MrBayes (Ronquist e Huelsenbeck, 2003), utilizando o modelo para sítios de restrição. Marcadores AFLP para inferências filogenéticas estão sendo utilizados com sucesso na família Bromeliaceae, por exemplo (Horres et al., 2007; Rex et al., 2007; Jabaily et al., 2013; Heller et al., 2015; Goetze et al., 2016; Pinangé et al., 2016; Cruz et al., 2017).

Identificação de híbridos

A partir dos dados gerados com a técnica de AFLP é possível fazer a identificação de indivíduos híbridos. A hibridação é considerada uma força evolutiva importante (Abbott et al., 2013). A hibridação pode gerar novidades fenotípicas e adaptativas, o que, consequentemente, pode resultar na formação de novas espécies. Por outro lado, a hibridação pode ser uma ameaça à biodiversidade, especialmente quando espécies raras estão envolvidas, levando à perda da identidade genética da espécie. Se os genótipos híbridos apresentarem *fitness* igual ou superior ao das espécies parentais, as zonas híbridas poderão expandir e levar ao deslocamento, ou até mesmo a extinção, de uma ou mais espécies parentais (Martin e Cruzan, 1999). Através do grande número de marcadores que são gerados com a técnica de AFLP é possível identificar alelos diagnósticos de cada um dos parentais no híbrido (Bensch e Åkesson, 2005). Metodologias normalmente utilizadas para a identificação de híbridos são a análise de componentes principais; a análise Bayesiana no programa Structure, que identifica indivíduos com um perfil molecular intermediário entre os parentais; e análises de agrupamento, como NeighborNet, geradas, por exemplo, no programa SplitsTree, que

baseadas em distâncias genéticas, identificam os principais grupos (parentais), ficando os híbridos em posição intermediária na rede. O programa NewHybrids (Anderson e Thompson, 2002) pode ainda ser utilizado para a identificação de diferentes classes de híbridos, como F1, F2, e retrocruzamentos com uma ou mais espécies parentais. No estudo desenvolvido por Certner et al. (2015), foram utilizadas três combinações de *primers* AFLP para acessar a frequência de híbridos interespecíficos entre *Knautia carinthiaca* e *K. arvensis*. Dos 128 indivíduos incluídos no estudo, 25 foram identificados como híbridos. O estudo conseguiu ainda concluir que a espécie mais rara, *K. carinthiaca*, não corre risco de extinção apesar da ocorrência de hibridação.

Melhoramento genético e agricultura

A seleção assistida de características de interesse econômico é uma das aplicações dos marcadores moleculares no melhoramento vegetal. A seleção assistida se baseia no mapeamento e associação de marcadores a genes que controlam uma determinada característica de interesse agrícola. Dessa forma, para que a seleção assistida por marcadores seja efetiva, o mapeamento de regiões do genoma associadas a estas características precisa ser saturado. As doenças fúngicas mais importantes para a cultura da videira no Brasil são o míldio (*Plasmopora Viticola*) e o oídio (*Uncinula necator*). O primeiro causa sérios prejuízos quando a precipitação pluviométrica é elevada. O segundo tem sua incidência em maiores níveis nas áreas tropicais durante o período seco. Pesquisadores da EMBRAPA Uva e Vinho (Dias de Oliveira et al., 2005) utilizaram a técnica de AFLP e analisaram nove cultivares de videira contrastantes para resistência ao míldio e ao oídio. A análise das bandas polimórficas obtidas pela técnica de AFLP foi determinada pelo escore de presença e/ou ausência de bandas. Os dados foram introduzidos em planilhas que são fonte de entrada para programas de análise como INDENTITY (Wagner e Sefc, 1999) e/ou NTSYSpc. O coeficiente de Jaccard foi utilizado para calcular a matriz de similaridade. Os autores identificaram 15 fragmentos polimórficos, os quais puderam associar com a resistência ao míldio e oídio. As cultivares em que esses fragmentos foram identificados foram então introduzidas no programa de melhoramento da videira para resistência fúngica.

Marcadores do tipo AFLP foram utilizados por Prinz e colaboradores (2001) para enriquecer a região cromossomal em torno do gene que confere resistência à ferrugem do trigo *Thinopyrum*-derived Lr19. Umas das bandas identificadas por AFLP foi convertida em marcador molecular que acompanha a região de interesse após diversos cruzamentos dentro do programa de melhoramento. Esse foi o primeiro caso de sucesso em que uma banda polimórfica identificada por AFLP foi transferida como marcador molecular na seleção assistida.

Aplicações na saúde humana e diagnóstico de doenças

A utilização de AFLP em amostras humanas não é tão frequente quanto seu uso em plantas e microrganismos, apesar de se mostrar como uma técnica bastante acessível e robusta. Atualmente sua utilização muitas vezes é combinada com outros métodos de biologia molecular, como o RT-qPCR para análise de expressão gênica (RT-AFLP), ou associada a diferentes métodos de detecção, como a eletroforese capilar (CE-AFLP), que torna o método ainda mais sensível.

A pesquisa de marcadores bialélicos em estudos de associação contribui de forma muito importante para a caracterização das bases genéticas de traços e doenças complexas. Neste sentido, o uso das técnicas de AFLP contribui de forma rápida e barata para identificação de diferentes variantes do genoma analisado.

O método já foi utilizado para aperfeiçoar a caracterização filogenética do cromossomo Y na determinação de linhagens de paternidade. Em importantes aplicações clínicas, a identificação de padrões de bandas geradas por diferenças encontradas em genes de grande impacto na saúde humana representaram por bastante tempo a melhor forma de estabelecer o diagnóstico de mutações patogênicas, mostrando-se como uma ferramenta extremamente útil na análise de mutações frequentes.

Uma das principais aplicações do uso de AFLP em saúde humana é na identificação de bactérias, fungos e parasitas patogênicos. Além de permitir a diferenciação de cepas e colônias, este método também permite a construção de mapas genéticos e caracterização de genes de resistência nestes organismos. As doenças infecciosas são responsáveis por mais de 17 milhões de mortes anualmente, e a correta identificação dos agentes infecciosos são determinantes para o estabelecimento de um tratamento eficaz. O advento de técnicas moleculares de identificação de microrganismos foi revolucionário para o diagnóstico e tratamento de doenças. Muitos agentes patogênicos foram primeiramente identificados através deste tipo de técnica, que permite caracterizar as relações filogenéticas, taxonômicas e epidemiológicas dos agentes patogênicos. Neste sentido, o AFLP é o método mais robusto para identificação de DNAs com sequência não determinada ou organismos não modelo.

O *fingerprinting* de DNA humano pode ser realizado através de técnicas de AFLP com muito êxito, permitindo a identificação de indivíduos (Vos et al., 1995), como por exemplo na análise de regiões como os locus 3' dos genes *APOB*, *PAH* e a região D1S80, utilizados com esta finalidade (Latorra et al., 1994). Já se demonstrou que a técnica de AFLP foi capaz de diferenciar tipos de tumores em situações onde aberrações genômicas não foram detectadas por métodos como CGH (*Comparative Genome Hibridization*) (Wong et al., 2004).

Variações da Técnica de AFLP

Entre as variações da técnica de AFLP publicadas, destacaremos o descobrimento de marcadores SNPs através de plataformas de Sequenciamento de Nova Geração (CRoPS e RAD-seq). O estudo de genes diferencialmente expressos utilizando a técnica de cDNA-AFLP é outra variação da técnica que foi muito utilizada antes de as plataformas de sequenciamento de larga escala tornarem-se economicamente acessíveis para os pesquisadores.

CRoPS (Complexity Reduction of Polymorphic Sequences)

Essa metodologia segue basicamente todas as etapas da técnica de AFLP, mas os fragmentos gerados durante a amplificação seletiva não são genotipados e sim sequenciados. Os fragmentos amplificados durante a PCR seletiva recebem uma *tag* de identificação (4 nucleotídeos de comprimento). Após essa identificação, os fragmentos das amostras são reunidos e sequenciados (Figura 4.5). As sequências obtidas são então agrupadas e alinhadas a algum genoma de referência, e posteriormente analisadas para pesquisa e identificação de SNPs (van Orsouw et al., 2007; Davey et al., 2010; Gompert et al., 2010). Essa metodologia já foi utilizada, por exemplo, em milho (van Orsouw et al., 2007), onde 1272 possíveis SNPs foram encontrados.

Na ausência de um genoma de referência, as sequências de leitura devem ser utilizadas para a montagem *de novo* do genoma. Sequências de leitura são aquelas obtidas no próprio sequenciamento do organismo em estudo, que são alinhadas umas as outras, à procura de regiões onde dois segmentos se sobrepõem. Essas sobreposições

podem ser incorporadas linearmente em um processo de montagem, que é contínuo. Quanto mais curtas as sequências, maior a quantidade de sobreposições necessárias para que essa tarefa possa ser executada. A cobertura genômica, isto é, o número de vezes que uma determinada região do genoma é coberta por segmentos de leitura, contribui para aumentar a acurácia de identificação da sequência de DNA na região considerada (Martins, 2013).

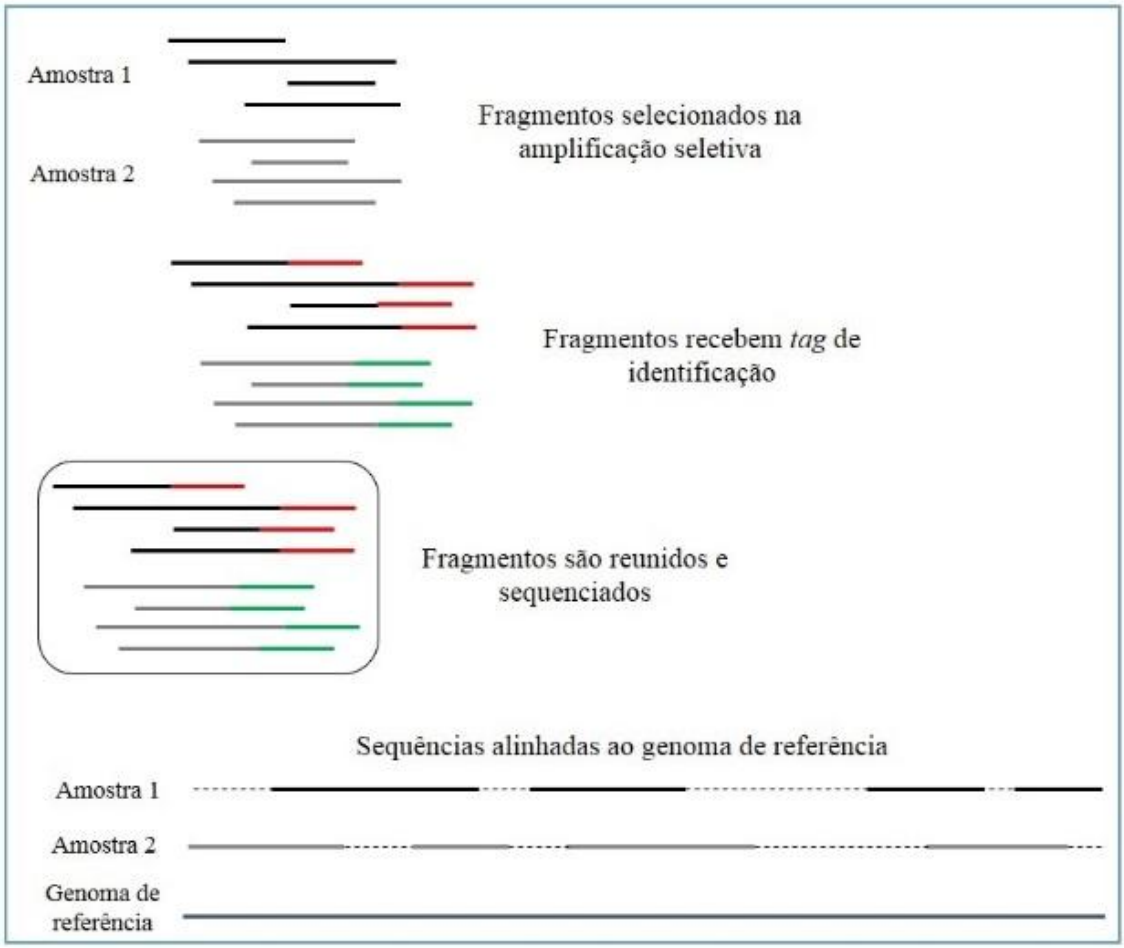

Figura 4.5 - Método de Complexity Reduction of Polymorphic Sequences (CRoPS): os fragmentos selecionados na etapa de amplificação seletiva da técnica tradicional de AFLP são marcados com uma tag de identificação e posteriormente reunidos e sequenciados. Após o sequenciamento, as tags são removidas e então as sequências são montadas de acordo com o genoma de referência.

RAD-seq (Restriction-site associated DNA sequencing)

Nessa metodologia o DNA genômico também é clivado com uma ou duas enzimas de restrição. Aos fragmentos gerados após a clivagem do DNA são ligados adaptadores P1 (em vermelho na Figura 4.6) que já contém uma *tag* de identificação (identificando cada indivíduo). Esses fragmentos com os adaptadores são reunidos e cortados aleatoriamente para a geração de fragmentos com um tamanho médio de algumas centenas de pares de base. Esses fragmentos cortados são então ligados a um segundo adaptador (P2, em azul na Figura 4.6) e usados como molde em uma reação de PCR com *primers* complementares aos adaptadores P1 e P2. Os adaptadores P2 têm uma estrutura em "Y" divergente que não irá permitir a ligação do *primer* P2, a menos

que o adaptador P2 tenha sido completado pela amplificação do adaptador P1. Isso garante que todos os fragmentos amplificados tenham o adaptador P1 com a *tag* de identificação, o sítio parcial de restrição, poucas centenas de bases flanqueando a sequência do sítio de restrição e o adaptador P2. Os fragmentos amplificados são então selecionados de acordo com o tamanho (em torno de 200 – 500 pb) e sequenciados. As sequências obtidas são agrupadas e alinhadas ao genoma de referência e posteriormente investigadas para a detecção de SNPs (Davey et al., 2010; Jones et al., 2013). Essa metodologia tem sido bastante utilizada em genética de populações, filogenia e filogeografia, entre outros, onde outros marcadores não conseguiram responder aos objetivos propostos (Davey e Blaxter, 2011; Jones et al., 2013).

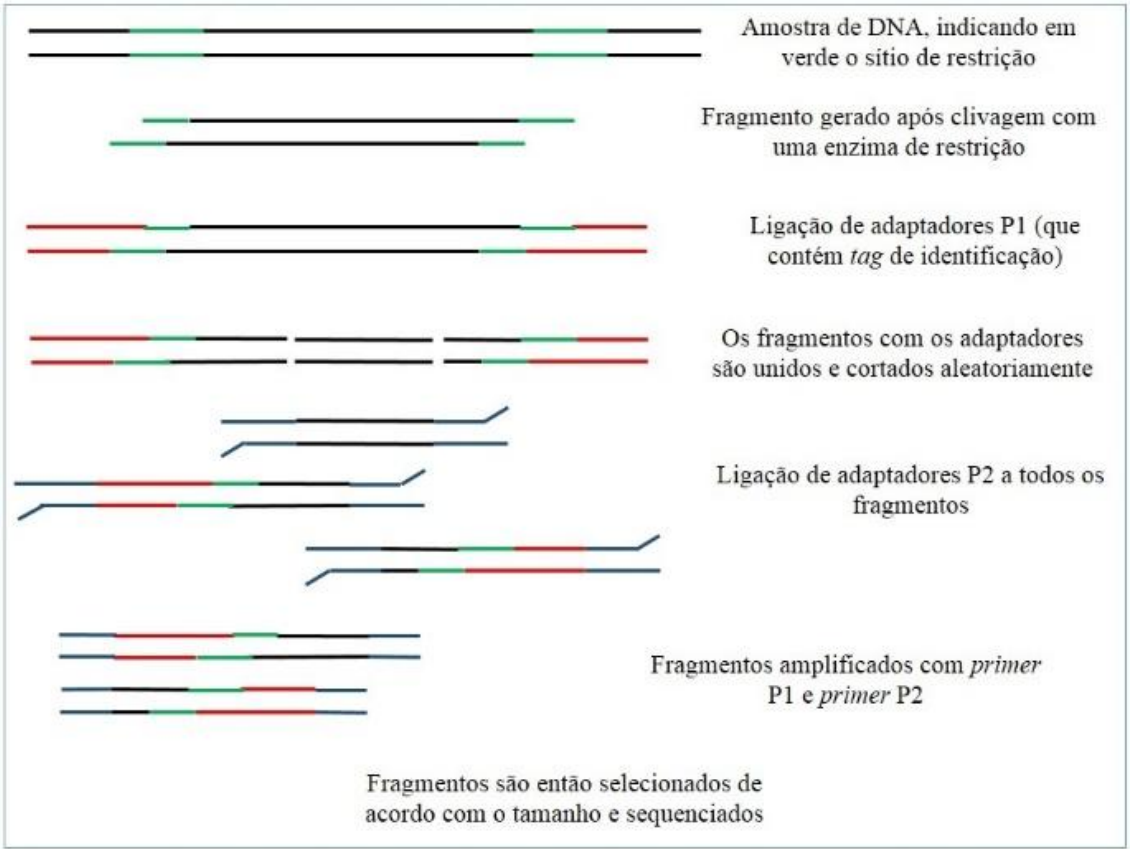

Figura 4.6 - Esquema mostrando as principais etapas da técnica de RAD-seq (*Restriction-site associated DNA sequencing*).

cDNA-AFLP

A metodologia da cDNA-AFLP permite a detecção de transcritos diferencialmente expressos. A técnica consiste das mesmas etapas do AFLP convencional, no entanto, ao invés do isolamento do DNA, faz-se o isolamento do RNA total e síntese de cDNA. Essa abordagem é utilizada com o objetivo de comparar o perfil transcricional de diferentes variedades, espécies ou mesmo tecidos do mesmo indivíduo. Atualmente o sequenciamento em larga escala de RNA (RNA-seq) substituiu em grande parte a técnica, por ser menos laborioso e com menos etapas técnicas.

Considerações Finais

Os marcadores do tipo AFLP podem ser considerados uma derivação dos marcadores baseados em restrição do DNA, mas com a adição de etapas de PCR seletiva que são capazes de detectar diferenças entre indivíduos de apenas um nucleotídeo. Assim, um número bastante significativo de bandas polimórficas, e acima de tudo reprodutível, é gerado. A técnica iniciou com a confecção de grandes géis de poliacrilamida, em que a preparação e visualização era trabalhosa, demandando tempo e ainda expondo o pesquisador à toxicidade da acrilamida ou de isótopos radioativos. Esses problemas foram superados quando a técnica foi adaptada para visualizar os resultados através de sequenciamento automático. Mas a evolução da técnica criada em 1995 não parou por aí. Como relatamos ao longo do capítulo, diferenças no transcritoma de diferentes indivíduos também podem ser detectadas através desta metodologia. No entanto, o aspecto mais importante é que o AFLP serviu de base para o desenvolvimento de outras técnicas baseadas em restrição como CroPs e RAD-seq implementados ao sequenciamento de alto desempenho. Estas técnicas empregam o sequenciamento convencional e/ou em larga escala como principal ferramenta para gerar uma quantidade imensa de dados. A técnica de AFLP não entrou em desuso, ainda é utilizada em vários laboratórios em que a quantidade de indivíduos analisados não é grande e a disponibilidade de recursos financeiros é limitada.

Referências Bibliográficas

Abbott R, Albach D, Ansell S, et al. (2013) Hybridizationa and speciation. *Journal of Evolutionary Biology* 26: 229-246.

Alexander LM, Kirigwi FM, Fritz AK, Fellers JP (2012) Mapping and quantitative trait loci analysis of drought tolerance in a spring wheat population using amplified fragment length polymorphism diversity array technology markers. *Crop Science* 52: 253-261.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160: 1217-1229.

Bensch S, Åkesson M (2005) Ten years of AFLP in ecology and evolution: Why so few animals? *Molecular Ecology* 14: 2899-2914.

Blanco EZ, Bajay MM, Siqueira MVBM, Zucchi MI, Pinheiro JB (2016) Genetic diversity and structure of Brazilian ginger germplasm (*Zingiber officinale*) revealed by AFLP markers. *Genetica* 144: 627-638.

Blears MJ, De Grandis AS, Lee H, Trevors JT (1998) Amplified fragmente length polymorphism (AFLP): a review of the procedure and its applications. *Journal of Industrial Microbiology & Biotechnology* 21: 99-114.

Bryant D, Moulton V (2004) NeighborNet: an agglomerative algorithm for the construction of planar phylogenetic networks. *Molecular Biology and Evolution* 21: 255-265.

Certner M, Kolár F, Schönswetter P, Frajman B (2015) Does hybridization with a widespread congener threaten the long-term persistence of the Eastern Alpine rare local endemic *Knautia carinthiaca*? *Ecology and Evolution* 5: 4263-4276.

Cruz GAS, Zizka G, Silvestro D, Leme EMC, Schulte K, Benko-Iseppon AM (2017) Molecular phylogeny, character evolution and historical biogeography of *Cryptanthus* Otto & A. Dietr. (Bromeliaceae). *Molecular Phylogenetics and*

Evolution 107: 152-165.
Davey JM, Blaxter ML (2011) RADSeq: next-generation population genetics. *Briefings in Functional Genomics* 5: 416-423.
Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2010) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499-510.
Després L, Gielly L, Redoutet B, Taberlet P (2003) Using AFLP to resolve phylogenetic relationships in a morphologically diversified plant species complex when nuclear and chloroplast sequences fail to reveal variability. *Molecular Phylogenetics and Evolution* 27: 185-196.
Dias de Oliveira PR, Scotton DC, Nishimura DS, Figueira A (2005) Genetic diversity and identification of AFLP markers associated with diseases resistance in grapevine. *Revista Brasileira de Fruticultura* 3: 454-457.
Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564-567.
Ferreira ME, Grattapaglia D (1995) Polimorfismo de Comprimento de Fragmentos Amplificados (AFLP). *In*: Ferreira ME, Grattapaglia D. Introdução ao Uso de Marcadores RAPD e RFLP em Análise Genética. EMBRAPA-CENARGEN, Brasília, pp 62-68.
Goetze M, Schulte K, Palma-Silva C, et al. (2016) Diversification of Bromelioideae (Bromeliaceae) in the Brazilian Atlantic rainforest: A case study in *Aechmea* subgenus *Ortgiesia*. *Molecular Phylogenetics and Evolution* 98: 436-357.
Gompert Z, Forister ML, Fordyce JA, Nice CC, Willianson RJ, Buerkle CA (2010). Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology* 19: 2455-2473.
Grover A, Sharma PC (2016) Development and use of molecular markers: past and presente. *Critical Reviews in Biotechnology* 36: 290-302.
Heller S, Leme EM, Schulte K, Benko-Issepon AM, Zizka G (2015) Elucidating phylogenetic relationships in the *Aechmea* alliance: AFLP analysis of *Portea* and the *Gravisia* complex (Bromeliaceae, Bromelioideae). *Systematic Botany* 40: 716-725.
Horres R, Schulte K, Weising K, Zizka G (2007) Systematics of Bromelioideae (Bromeliaceae) – evidence from molecular and anatomical studies. *Aliso* 23: 27-43.
Jabaily RS, Sytsma KJ (2013) Historical biogeography and life-history evolution of Andean *Puya* (Bromeliaceae). *Botanical Journal of the Linnean Society* 171: 201-224.
Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology* 22: 2986-3001.
Kirchberger PC, Sefc KM, Sturmbauer C, Koblmüller S (2014) Outgroup effects on root position and tree topology in the AFLP phylogeny of a rapidly radiating lineage of cichlid fish. *Molecular Phylogenetics and Evolution* 70: 57-62.
Kneeland KM, Skoda SR, Foster JE (2013) Amplified Fragment Length Polymorphism Used to Investigate Genetic Variability of the Stable Fly (Diptera: Muscidae) Across North America. *Journal of Medical Entomology* 5: 1025-1030.
Koopman WJM (2005) Phylogenetic signal in AFLP data sets. *Systematic Biology* 54: 197-217.

Koopman WJM, Gort G (2004) Significance tests and weighted values for AFLP similarities, based on *Arabidopsis in silico* AFLP fragment length distributions. *Genetics* 167: 1915-1928.

Krauss S (2000) Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. *Molecular Ecology* 9: 1241-1245.

Latorra D, Stern CM, Schanfield MS (1994) Characterization of human AFLP systems apolipoprotein B, phenylalanine hydroxylase, and D1S80. *Genome Research* 3: 351-358.

Lynch M, Milligan B (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology* 3: 91-99.

Martin LJ, Cruzan MB (1999) Patterns of hybridization in the *Piriqueta caroliniana* complex in Central Florida: evidence for an expanding hybrid zone. *Evolution* 53: 1037-1049.

Martins AM (2013) Sequenciamento de DNA, montagem *de novo* do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis*. Tese de Doutorado, Universidade de Brasília, Brasília, Brasil.

Meudt HM, Clarke AC (2007) Almost Forgotten or Latest Practice? AFLP applications, analyses and advances. *Trends in Plant Science* 12: 106-117.

Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. *Trends in Ecology & Evolotuin* 14: 389-394.

Pejic I, Ajmone-Marsan P, Morgante M, et al. (1998) Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. *Theoretical and Applied Genetics* 97: 1248-1255.

Pinangé DSB, Krapp F, Zizka G, et al (2016) Molecular phylogenetics, historical biogeography and character evolution in *Dyckia* (Bromeliaceae, Pitcairnioideae). *Botanical Journal of the Linnean Society* 10.1111/boj.12489.

Prinz R, Groenewald JZ, Marais GF, Snape JW, Koebner RMD (2001) AFLP and STS tagging of Lr19, a gene conferring resistance to leaf rust in wheat. *Theoretical and Applied Genetics* 103: 618-624.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multi-locus genotype data. *Genetics* 155: 945-959.

Rapposelli E, Melito S, Barmina GG, Foddai M, Azara E, Scarpa GM (2015) AFLP fingerprinting and essential oil profiling of cultivated and wild populations of Sardinian *Salvia desoleana*. *Genetic Resources and Crop Evolution* 62: 959-970.

Rex M, Patzolt K, Schulte K, et al. (2007) AFLP analysis of genetic relationships in the genus *Fosterella* L.B. Smith, Pitcairnioideae, Bromeliaceae. *Genome* 50: 90-105.

Rodrigues LA, Ruas EA, Ruas PM, et al. (2016) Population genetic structure of the South American species *Hypochaeris lutea* (Asteraceae). *Plant Species Biology* 31: 55-64.

Rohlf FJ (1997) NTSYSpc: numerical taxonomy and multivariate analyses system. Version 2.0. Exeter Publications, New York.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian inference of phylogeny. *Bioinformatics* 19: 1572-1574.

Savelkoul PHM, Aarts HJM, Haas J, et al. (1999) Amplified-Fragment Length Polymorphism Analysis: the State of an Art. *Journal of Clinical Microbioloy* 37: 3083-3091.

Swofford DL (2002) PAUP*: Phylogenetic Analysis using Parsimony (* and other methods) v. 4.0b10. Sinauer Assoc., Sunderland.

van Orsouw NJ, Hogers RCJ, Janssen A, et al. (2007) Complexity Reduction of

Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *PLoS One* 11: e1172.

Vekemans X, Beauwens T, Lemaire M, Roldán-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Molecular Ecology* 11: 139-151.

Vos P, Hogers R, Bleeker M, et al. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 21: 4407-4414.

Wagner HW, Sefc KM (1999) Identity 1.0. Centre for Applied Genetics - University of Agricultural Sciences, Viena.

Wong KY, Chuan YC, Aggarwal A, Tham L, Kong WM, Tan P (2004) Identifying patterns of DNA for tumor diagnosis using capillary electrophoresis-amplified fragment length polymorphism (CE-AFLP) screening. *Journal of Bioinformatics and Computational Biology* 2: 569-587.

Yeh FC, Boyle TJB (1997) Population genetic analysis of codominant and dominant markers and quantitative traits. *Belgian Journal of Botany* 129: 157.

Zhivotovsky L (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* 8: 907-913.

Capítulo 5

Marcadores moleculares baseados na análise de sequências: utilização em filogenia e filogeografia

Dra. Ana Lucia Anversa Segatto, Dra. Márcia Goetze, Dra. Caroline Turchetto

Considerações gerais

A obtenção de marcadores moleculares baseados em sequência envolve a determinação da sequência de nucleotídeos do fragmento de DNA que se pretende analisar e a detecção de diferenças nessas regiões (deleções, inserções, substituições), em diferentes organismos ou indivíduos da(s) espécie(s) em estudo. A determinação da sequência nucleotídica do fragmento é obtida, atualmente, através de sequenciadores automáticos (ver Capítulo 2 para uma revisão sobre técnicas de sequenciamento).

Marcadores de sequência são muito utilizados em estudos de Sistemática Filogenética; para a determinação da variabilidade genética intra e interpopulacional; e para compreender eventos demográficos passados que influenciou a atual distribuição geográfica da diversidade genética, uma abordagem filogeográfica. Os marcadores de sequências podem ser isolados a partir de genomas nucleares e citoplasmáticos (mitocôndria e cloroplasto), dependendo do objetivo do estudo; e de diferentes regiões, tanto DNA codificador como não codificador.

Os primeiros estudos filogenéticos utilizando marcadores de sequência de DNA foram publicados ainda na década de 1980 (Cann et al., 1987; Felsenstein, 1988; Cavalli-Sforza, 1998). Já estudos combinando genética de populações e filogeografia são mais recentes, com as primeiras publicações a partir de meados da década de 1990 (Avise, 1987; Beheregaray, 2008).

Para muitos marcadores de sequência é possível obter a taxa de mutação, o que permite fazer inferências sobre padrões históricos de diversificação de linhagens de um determinado organismo ou grupo de espécies, uma das principais vantagens e aplicações desse grupo de marcadores. Entre os aspectos limitantes podemos citar a necessidade da projeção de *primers* específicos para a espécie em estudo, principalmente quando se objetiva amplificar um gene nuclear de baixo número de cópias. Esse problema está sendo contornado com a utilização de novas metodologias de sequenciamento que permitem a geração de um grande número de sequências com menor custo.

Um cuidado indispensável quando se amplifica regiões gênicas nucleares é a necessidade da certificação de que estão sendo analisados genes ortólogos e não parálogos (veja Box 5.1 para maiores detalhes). Além disso, a obtenção de sequências que apresentam polimorfismos (variabilidade) ao nível taxonômico desejado para o estudo específico é outro aspecto importante. Por exemplo, para estudos filogeográficos, objetiva-se alcançar níveis de variabilidade entre indivíduos de uma mesma espécie, sendo necessário utilizar uma ou mais regiões do genoma com altas taxas de mutação; por outro lado, em estudos filogenéticos, objetiva-se encontrar variação que represente as relações evolutivas dentro de um grupo de espécies.

Box 5.1

Homologia: semelhança devido à herança de um caráter através de um ancestral comum.

Homoplasia: semelhança devido à convergência de um caráter sem ancestralidade comum.

Genes homólogos podem ser classificados como ortólogos ou parálogos.

Genes ortólogos: são genes que compartilham homologia como um resultado imediato de especiação. Por exemplo, um par de genes de cópia única entre duas espécies diferentes.

Genes parálogos: são genes que resultam da duplicação gênica dentro de um genoma, ocorrendo antes ou depois da especiação.

Evolução em concerto: processo de homogeneização de diferentes cópias de fragmentos (podem ser genes) do genoma causado por recombinação e conversão gênica.

Dentro do contexto evolutivo, a maioria dos estudos tem focado em analisar de forma simultânea o genoma nuclear e citoplasmático para a obtenção de informações complementares. Isto se deve ao fato do diferente modo de herança desses genomas. Por exemplo, na maioria das plantas, o genoma de cloroplasto é herdado maternalmente e apresenta taxas de mutação e recombinação menores que o genoma nuclear. Dessa forma, os padrões encontrados com o genoma de cloroplasto podem refletir processos históricos, enquanto os resultados obtidos com o genoma nuclear podem refletir padrões mais contemporâneos (Avise, 2009).

Em plantas, inicialmente, os estudos filogenéticos utilizaram sequências de genes, como por exemplo, *rbcL*, *ndhF*, *atpB*, *matK*, e regiões não codificadoras do genoma de cloroplasto (íntrons e espaçadores intergênicos), para responder questões em diferentes campos de pesquisa (ver Chase et al., 1993; Steele e Vilgalys, 1994; Clark et al., 1995; Hoot et al., 1995). As regiões não codificadoras foram utilizadas principalmente em estudos de Sistemática Filogenéticaa nível infragenérico, por apresentarem maior variabilidade quando comparadas às regiões gênicas.

As sequências plastidiais continuam tendo uma alta popularidade e utilidade devido a muitas características biológicas e técnicas que as tornam ideais para estudos ecológicos e evolutivos. Algumas dessas características incluem a ordem altamente conservada dos genes no cloroplasto, ausência ou baixa taxa de recombinação, baixos níveis de substituição de nucleotídeos, herança predominantemente uniparental (o que permite estudar o fluxo gênico via sementes, por exemplo), baixo tamanho efetivo populacional e tempos de coalescência curtos. Todas as características citadas tornam os marcadores de sequência de cloroplasto ideais para estudos de filogeografia em plantas. Além disso, essas características tornam o genoma do cloroplasto um alvo ideal para projeção de *primers* universais que amplificam locus homólogos em espécies filogeneticamente divergentes (Petit et al., 2005; Petit e Vendramin, 2007; Ness, 2016; Twyford e Ness, 2016). Os *primers* universais são ferramentas valiosas em estudos de inúmeras espécies, tanto em abordagens filogenéticas como filogeográficas (Tarbelet et al., 1991; Shaw et al., 2005, 2007; Jansen et al., 2007; Avise, 2009).

De maneira geral, o DNA total extraído é enriquecido de genoma de cloroplasto devido ao grande número de plastídios por célula (Bendich, 1987), o que torna mais

fácil o sequenciamento dessas regiões do que genes nucleares cópia única, principalmente a partir de amostras degradadas (Staats, 2013). Mais recentemente, com os avanços nas técnicas de sequenciamento (Capítulo 2), tem-se adotado a estratégia de sequenciamento de todo o genoma de cloroplasto, para obtenção de maior resolução, elucidando perguntas previamente não respondidas, bem como para identificação de locus mais variáveis para estudos filogenéticos e inferências em genética de populações (Hollingsworth et al., 2016).

O genoma mitocondrial de plantas é menos utilizado em estudos com abordagens evolutivas, principalmente por apresentar baixa variabilidade a nível de sequência de nucleotídeos. Entretanto, o genoma de mitocôndria apresenta variação considerável no tamanho e arranjo dos genes. Essas características permitiram que genes mitocondriais fossem utilizados em estudos para inferir eventos antigos, como a origem das angiospermas ou das plantas com sementes (Judd et al., 2009). A identificação de *primers* universais do genoma mitocondrial também segue, em normas gerais, o que foi descrito acima para marcadores de cloroplasto, sendo a projeção de *primers* realizada em regiões conservadas, os quais então têm maiores chances de amplificação em espécies distintas (Dumolin-Lapegue et al., 1997).

Em animais, os primeiros estudos filogenéticos e filogeográficos utilizaram genes mitocondriais. O genoma mitocondrial em animais é compactado, apresentando o conteúdo e ordem dos genes extremamente conservados entre espécies e quase toda a sequência apresenta função codificadora. A herança por via de regra é materna e a taxa de mutação é, em geral, mais alta que a de um gene cópia única do genoma nuclear (Solferini e Selivon, 2012). Os principais genes, ainda muito utilizados, são os que codificam as subunidades I e II da citocromo oxidase (COI, COII), o rRNA mitocondrial 12S, o citocromo b, além da região controladora mitocondrial (por exemplo, Randi et al., 2001; Chen et al., 2002; Zhang e Sota, 2007).

Diversos marcadores moleculares do genoma nuclear podem ser utilizados para estudos evolutivos. O espaçador interno transcrito (ITS, do inglês *Internal Transcribed Spacer*), localizado entre os genes ribossomais 18S e 26S é amplamente utilizado por apresentar várias cópias no genoma, sofrer evolução em concerto (Box 5.1) e por ter *primers* universais descritos (Álvarez e Wendel, 2003).

Entretanto, para muitos *taxa*, e em estudos a nível intraespecífico, a região ITS apresenta baixa variabilidade. Desse modo, regiões de cópia única no genoma nuclear começaram a ser investigadas para avaliação do potencial em estudos evolutivos. Muitos genes nucleares já foram utilizados em estudos filogenéticos e uma revisão sobre os marcadores comumente aplicados nessas abordagens, em plantas, foi publicada por Zimmer e Wen (2012).

Neste capítulo abordaremos a metodologia de isolamento e análise de marcadores de sequências com enfoque em estudos evolutivos, tais como filogenia de espécies e evolução de genes e famílias multigênicas, genética de populações, filogeografia e processos de diversificação e especiação com principal ênfase em plantas.

Metodologia de Isolamento

Diferentes níveis de variabilidade genética são observados entre os mais variados marcadores de sequência, os quais providenciam resolução em diferentes níveis taxonômicos. Tendo isso em mente, em estudos envolvendo marcadores de sequência a primeira etapa é a escolha criteriosa de qual locus utilizar. Para isso, algumas características gerais das sequências devem ser levadas em consideração para um determinado estudo obter um resultado informativo: (a) preferencialmente um

marcador que seja de cópia única, pois assim garante a amplificação de regiões ortólogas entre indivíduos (espécies) diferentes; (b) que seja fácil de alinhar, ou seja, que as sequências dos diferentes indivíduos apresentem similaridades suficientes para permitir a determinação das posições homólogas no alinhamento; (c) que tenha uma taxa de mutação adequada ao problema, ou seja, um nível de variação adequado ao nível taxonômico do estudo. Taxas de mutação variam entre as diferentes regiões do genoma. Diferentes níveis taxonômicos exigem a utilização de marcadores com taxas de mutação diferentes. Se uma determinada região tem taxas de mutação muito baixas ela pode ser utilizada para determinar as relações evolutivas de níveis taxonômicos mais elevados. Quando um marcador é muito variável e a intenção é estudar níveis taxonômicos mais elevados é provável que o alinhamento não esteja refletindo homologias verdadeiras, uma mesma base (por exemplo, guanina) pode estar em uma mesma posição por convergência e não por homologia. A melhor maneira de determinar qual o melhor marcador ou conjunto de marcadores de sequência a ser(em) utilizado(s) é analisando a variabilidade desses marcadores em organismos próximos. A ferramenta BLAST pode ser utilizada para encontrar sequências de determinada região nos bancos de dados, para a espécie ou para espécies próximas, e a partir disso verificar se o polimorfismo existente é ideal para o estudo. Caso essa informação não esteja disponível, experimentos pilotos devem ser realizados.

Uma maneira de isolamento de marcadores de sequência é baseada em genomas de referência, onde *primers* específicos são projetados em regiões conservadas para amplificação de determinada região de interesse. Entretanto, para espécies não-modelo, que não tem genoma sequenciado, são utilizados os chamados *primers* universais. Esses *primers* estão descritos na literatura e podem ser utilizados para amplificar regiões específicas de DNA. Eles são posicionados em regiões conservadas do genoma, potencializando a chance desses marcadores serem amplificados em um grande número de espécies como, por exemplo, alguns espaçadores intergênicos plastidiais descritos para plantas (*psbA-trnH*, *trnG-TrnS*, *rpL32-trnL*, *rpS16-trnK*); e os genes mitocondriais em animais (COI, D-loop, CYB). Quando se faz uso de *primers* universais, independente do genoma a partir do qual os *primers* foram isolados, normalmente é necessário a otimização do protocolo de amplificação da região sob estudo no grupo alvo. De maneira geral, além da qualidade e quantidade de DNA, outros parâmetros como a temperatura de anelamento dos *primers* e o número de ciclos durante a PCR são as principais condições que necessitam de ajustes. A adição, na reação de PCR, de reagentes que aumentam a especificidade de amplificação, como DMSO (*dimethyl sulfoxide*) e BSA (*bovine serum albumin*), também pode ser relevante, normalmente aumentando a qualidade do produto obtido (Box 5.2).

Box 5.2

Ação do DMSO e BSA na reação de PCR

DMSO: inibe a formação de estruturas secundárias na amostra de DNA ou nos *primers*. Esse reagente também pode melhorar o acesso da enzima *Taq* DNA polimerase às regiões ricas em GC.

BSA: pode melhorar a estabilidade da *Taq* DNA polimerase, reduzir a perda de reagentes devido a absorção pelas paredes dos tubos, e, ainda, superar alguns inibidores da PCR, como polissacarídeos e fenóis.

Outra maneira de obter sequências, sem a utilização de PCRs específicas para o gene de interesse é através do sequenciamento em larga escala, ou buscas em genomas e transcritomas disponíveis. O sequenciamento em larga escala tem sido muito utilizado para análise de todo o genoma do cloroplasto das plantas e da mitocôndria dos animais. Cada gene no genoma destas organelas pode ser identificado, e a conservação da sequência permite um fácil alinhamento e comparação entre espécies, além de ser

improvável o conflito entre genes parálogos e homólogos. O sequenciamento de genomas de cloroplasto ou mitocondriais permitem isolar um grande número de marcadores de sequência para as espécies de interesse (por exemplo, Bourguignon et al., 2015; Carbonell-Caballero et al., 2015; Leliaert et al., 2016).

A primeira sequência completa de genoma plastidial foi obtida para *Nicotiana tabacum* através de sequenciamento via Sanger, e sobreposição de clones de fragmentos gerados pela clivagem com enzimas de restrição (Shinozaki et al., 1986). Entretanto, nos dias atuais, o rápido desenvolvimento de técnicas de sequenciamento e ferramentas de bioinformática têm viabilizado a montagem do genoma plastidial completo para muitas espécies não-modelo (Nock et al., 2011), e a utilização desses dados em estudos evolutivos, sem a necessidade de clonagem.

Diferentes estratégias de preparação de bibliotecas, tecnologias de sequenciamento e abordagens de montagem já foram descritas. A preparação de bibliotecas consiste em duas principais abordagens: uma baseada no sequenciamento direto do DNA genômico (que inclui o plastidial) e outra baseada no enriquecimento para DNA plastidial, a qual pode ser realizada por diferentes métodos. Na primeira abordagem, é realizada a montagem do genoma plastidial a partir de uma biblioteca de NGS (*Next Generation Sequencing*) sem prévio enriquecimento ou isolamento de DNA plastidial (Nock et al., 2011). Neste caso, por exemplo, pode-se adotar a estratégia de sequenciamento do genoma com baixa cobertura (~ 0.1 – 10 x) o que poderá ser suficiente para a montagem do genoma plastidial completo (por exemplo, Coissac et al., 2016), visto que o DNA utilizado no experimento é uma mistura entre DNA nuclear e organelar.

Por outro lado, a metodologia baseada no enriquecimento foca no sequenciamento somente do DNA plastidial. Nessa abordagem, quatro principais métodos para enriquecimento do DNA plastidial têm sido descritos: (1) baseados no isolamento do plastídio a partir de folhas frescas; (2) baseados no princípio de que genomas de organelas de eucariotos têm muito menos metilação CpG do que genoma nuclear; (3) baseados na utilização de sondas curtas de oligonucleotídeos para isolar sequências complementares de DNA plastidial a partir de um DNA genômico extraído; 4) via PCR. O isolamento do plastídio a partir de folhas frescas poderá ser realizado via gradiente de sacarose utilizando métodos caseiros ou *kits* comerciais (por exemplo, Miflin e Beevers, 1974; e Sigma *Chloroplast Isolation Kit*), por precipitação com altas concentrações de sal, ou então proceder a extração de DNA genômico seguido do tratamento com DNaseI. Entretanto, os dois últimos métodos podem apresentar baixo rendimento e contaminação com DNA mitocondrial (Shi et al., 2012). Devido ao baixo rendimento no isolamento dos plastídios, uma etapa adicional de amplificação pode ser requerida antes do sequenciamento. Apesar da vantagem de conseguir uma completa montagem mesmo com um pequeno número de *reads* (Shi et al., 2012), a principal limitação deste método é a necessidade de grandes quantidades de tecido vegetal e a otimização do protocolo de isolamento do plastídio, específico para cada espécie, dificultando estudos comparativos em larga escala. Como o genoma das organelas apresenta muito menos metilação CpG do que o genoma nuclear, o DNA genômico pode ser particionado numa porção altamente metilada (CpG), e uma fração com baixa metilação, a qual é enriquecida com DNA plastidial (3,2 – 11,2 vezes; Yigit et al., 2014). Este segundo método de enriquecimento, entretanto, pode não ser viável para amostras de DNA degradado, como de herbário (Twyford, 2016). Um método promissor para o enriquecimento de bibliotecas de DNA plastidial é a utilização de sondas curtas de oligonucleotídeos para isolar sequências complementares de DNA plastidial a partir de um DNA genômico extraído. A biblioteca de sequências capturadas

é sequenciada por NGS. Este método é adequado para uma vasta gama de material vegetal, incluindo amostras de herbário (ver, por exemplo, Stull et al., 2013 e Comer et al., 2015 que usaram microarranjos para captura de sequências plastidiais). Outra forma de enriquecimento com sequência plastidial é via PCR. Este método consiste do uso de um conjunto de *primers* universais para amplificação de sequências curtas, seguido de sequenciamento Sanger, mais adequado para aplicações que requerem a sequência parcial do plastídio, como por exemplo, estudos de genética de populações (Whittall et al., 2010). Ou então a amplificação de sequências longas e sequenciamento em plataforma NGS, em que o tamanho grande dos *amplicons* permite a ancoragem de todos os *primers* em regiões com baixa variabilidade do genoma. Entretanto, a falha de um único PCR pode resultar em uma grande lacuna na montagem das sequências. Por exemplo, Dong et al. (2013), desenvolveram *primers* universais para amplificar todo o genoma de cloroplasto de angiospermas em 138 PCRs, com sequências de tamanhos entre 0,8 – 1,5 Kb (embora críticas a esses *primers* tenham sido feitas, ver Prince, 2015). Também já foram desenvolvidos *primers* para amplificação de sequências mais longas (4 – 23 Kb de comprimento; Yang et al., 2014; Uribe-Convers et al., 2014).

Uma vez escolhida uma abordagem de preparação da biblioteca, deve-se escolher uma estratégia de sequenciamento que corresponda com os objetivos. É recomendado usar longos *reads* e/ou dados *paired-end* (Straub et al., 2012). Leituras de sequências de plastídios podem ser montadas a partir de um genoma de referência ou montagem *de novo* (sem a utilização de um genoma de referência). Para uma revisão de estratégias de sequenciamento e montagem ver Twyford et al. (2016).

Os estudos iniciais para a determinação da sequência completa do genoma mitocondrial de animais também envolveram a clonagem das sequências mitocôndriais em plasmídeos e posterior sequenciamento dos fragmentos clonados via método de Sanger (Flook et al., 1995). Mais recentemente, a determinação da sequência completa do DNA mitocondrial tem sido feita utilizando-se a abordagem de amplificação de sequências curtas do genoma. A partir do sequenciamento dessas regiões são desenhados *primers* que permitem a amplificação do restante do genoma da mitocôndria (Yang et al., 2012; Wang et al., 2013; Zhou et al., 2017a, 2017b). Outra estratégia é utilizar tanto *primers* que amplificam regiões curtas quanto longas do genoma a partir de *primers* universais (Wang et al., 2015). Porém, essas abordagens estão sendo substituídas pelas novas tecnologias de sequenciamento (NGS), utilizando a estratégia baseada no sequenciamento direto de DNA genômico, mais práticas, por não envolver o desenho de *primers*, e cada vez mais viáveis economicamente (Anmarkrud e Lifjeld, 2017; Zhou et al., 2017a, 2017b).

Métodos utilizados para análise de matrizes de dados obtidos por PCR da região de interesse

Após a obtenção das sequências para as populações ou espécies em estudo, a partir de *primers* específicos, elas devem ser alinhadas, o que pode ser realizado, por exemplo, no MUSCLE (Edgar, 2004), implementado no programa MEGA (Kumar et al., 2016). O alinhamento das sequências então é inspecionado a procura de polimorfismos, os quais devem ser confirmados, se possível, através da análise dos electroferogramas de cada indivíduo, o que pode ser realizado no programa CHROMAS. Uma vez que os polimorfismos foram confirmados, as sequências alinhadas são então utilizadas para as diferentes análises, dependendo da abordagem do estudo.

Sequências de diferentes regiões do mesmo genoma normalmente são concatenadas e analisadas de maneira conjunta, utilizando diferentes modelos evolutivos para cada região, se necessário. Índices de diversidade genética, como diversidade nucleotídica, diversidade haplotípica, número de haplótipos, número de sítios polimórficos (transições, transversões e *indels*) são estimados utilizando-se, entre outros programas, o ARLEQUIN (Excoffier e Lischer, 2010) e o DNASP (Librado e Rozas, 2009). A identificação de haplótipos e a determinação de fase de sequências com sítios heterozigotos para marcadores nucleares, (veja Box 5.3 para detalhes), pode ser realizado no programa DnaSP. O relacionamento dos haplótipos identificados pode ser obtido no programa NETWORK. Análises demográficas, tais como detecção de eventos de gargalo de garrafa (*Bottleneck*) ou expansão populacional, podem ser obtidas através da utilização de programas como ARLEQUIN, DNASP, LAMARC (Kuhner, 2006) e BEAST (Drummond et al., 2012). Estimativas da estruturação genética das populações podem ser obtidas utilizando-se, por exemplo, o programa BAPS (Corander et al., 2008) e também através do cálculo de análogos da estatística *F*, como θ_{ST}, no programa ARLEQUIN. Árvores filogenéticas de espécies, haplótipos ou genes podem ser obtidas utilizando-se diferentes métodos como: análise por distância, como por exemplo, UPGMA e Neighbor-joinning, através do programa PHYLIP e MEGA; utilizando-se métodos de parcimônia no PAUP (Swofford, 2002); utilizando-se os métodosprobabilísticos de Máxima Verossimilhança no PHYML (Guindon et al., 2010) ou RAXML (Stamatakis, 2014) inferências Bayesianas no BEAST e MRBAYES (Ronquist e Huelsenbeck, 2003). Na Tabela 5.1 são listados os sítios na Internet onde encontrar os programas citados.

Tabela 5.1 - Programas utilizados na análise de marcadores de sequência.

Programa	Localização na Internet	Referência
Arlequin	http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html	Excoffier e Lischer, 2010
BAPS	http://www.helsinki.fi/bsg/software/BAPS/	Corander et al., 2008
BEAST	http://beast.bio.ed.ac.uk/	Drummond et al., 2012
CHROMAS	http://technelysium.com.au	-
DNASP	http://www.ub.edu/dnasp/	Librado e Rozas, 2009
JMODELTEST	http://www.molecularevolution.org/software/phylogenetics/jmodeltest	Darriba et al., 2012
LAMARC	http://evolution.genetics.washington.edu/lamarc/lamarc_download.html	Kuhner, 2006
MEGA	http://www.megasoftware.net/	Kumar et al., 2016
MRBAYES	http://mrbayes.sourceforge.net/	Ronquist e Huelsenbeck, 2003
MUSCLE	Implementado no MEGA	Edgar, 2004
NETWORK	http://www.fluxus-engineering.com/	-
PAUP	http://paup.sc.fsu.edu/downl.html	Swofford, 2002
PHASE	Implementado no DnaSP	Stephens et al., 2001; Stephens e Donnelly, 2003

PHYLIP	http://evolution.genetics.washington.edu/phylip.html	-
PHYML	http://www.atgc-montpellier.fr/phyml/	Guindon et al., 2010
RAXML	http://sco.h-its.org/exelixis/software.html	Stamatakis, 2014

Na análise das sequências obtidas com os marcadores nucleares é necessário ficar atento à ocorrência de sítios heterozigotos. Um sítio é classificado como heterozigoto quando mais de um pico é visualizado no mesmo local no electroferograma, sendo que o sinal mais fraco deve ser pelo menos 25% da intensidade do sinal mais forte. Além disso, picos duplos devem estar presentes no mesmo sítio em ambas as fitas (Fuertes Aguilar et al., 1999; Fuertes Aguilar e Nieto Feliner, 2003) (Figura 5.1). Uma vez identificados, os sítios heterozigotos podem ser codificados de acordo com o código de ambiguidade de nucleotídeos IUPAC (*International Union of Pure and Applied Chemistry*, disponível em https://iupac.org/). Essa é uma das maneiras de leitura dos dados nos programas específicos de análise (filogenia ou filogeografia, citados acima). Entretanto, também é possível estabelecer as fases dos haplótipos (Box 5.3), e considerar esses dados para as análises. Um programa muito utilizado para reconstruir os haplótipos a partir de matrizes de dados com sítios heterozigotos é o PHASE (Stephens et al., 2001; Stephens e Donnelly, 2003) que usa um método Bayesiano baseado em coalescência para inferir os haplótipos. O PHASE está implementado no programa DnaSP, que também apresenta outros métodos para a reconstrução da fase dos haplótipos. A clonagem dos fragmentos da PCR é uma estratégia para a identificação precisa dos heterezogitos, porém trabalhosa e de custo elevado.

Box 5.3

Determinação da fase dos haplótipos:

Para sequências nucleares, sítios heterozigotos podem ser encontrados. A determinação da fase dos haplótipos envolve a reconstrução, por métodos laboratoriais ou estatísticos, da combinação dos vários sítios heterozigotos presentes em um determinado alelo (sequência). Por exemplo, se um gene apresenta dois sítios heterozigotos, digamos A:G e C:T, teremos quatro possíveis haplótipos (alelos): AC, AT, GC ou GT. Ou seja, quatro combinações possíveis de terem sido herdadas dos parentais. Os métodos estatísticos calculam qual das fases (haplótipos) é a combinação mais provável.

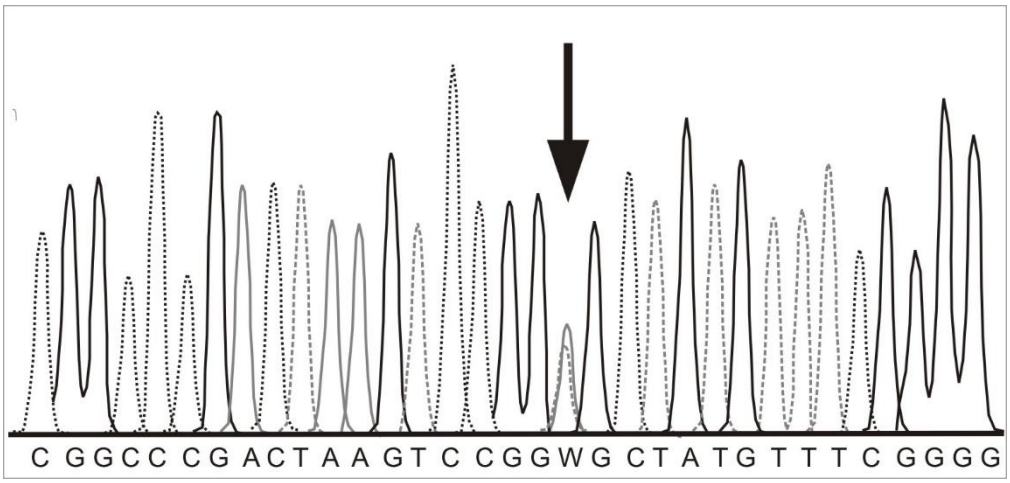

Figura 5.1 - Eletroferograma com um sítio heterozigoto indicado pela seta.

Como já mencionado, uma característica vantajosa dos marcadores de sequências é a possibilidade de se utilizar um modelo evolutivo, que permite a utilização de métodos probabilísticos para estimar as relações evolutivas entre espécies ou genes. Os modelos evolutivos levam em consideração, entre outras características, a possibilidade de mudanças múltiplas em um mesmo sítio, proporções diferentes de transições e transversões e proporções diferentes dos quatro nucleotídeos nas sequências (Salemi et al., 2009). O modelo evolutivo adequado para os marcadores e indivíduos analisados pode ser estimado, por exemplo, através do programa jModelTest (Darriba et al., 2012). A taxa de mutação de um gene pode ser calculada baseada em registros fósseis, eventos geológicos, a partir de taxas de espécies próximas com hábitos de vida similares, permitindo o cálculo de divergência entre as Unidades Taxonômicas Operacionais (OTUs) de uma filogenia ou entre haplótipos em abordagens filogeográficas (ver Lorenz-Lemke et al., 2010; Turchetto-Zolet et al., 2016).

Para uma revisão dos métodos de análise em uma abordagem filogeográfica com marcadores de sequência consultar o livro Guia Prático para estudos Filogeográficos (Turchetto-Zolet et al., 2013).

Exemplos de aplicações

Famílias gênicas estão presentes em grande número nos genomas dos organismos. Estudar a evolução dos genes durante a divergência das espécies revela particularidades das pressões seletivas que atuam em cada gene e do surgimento das rotas metabólicas durante a evolução. A partir de dados de sequência é possível estimar o tempo em que ocorreram os eventos de duplicação gênica, que geralmente são correlacionados ao surgimento de novidades evolutivas (Lynch e Force, 2000). Nesse tipo de estudo, onde normalmente são incluídos genes de muitas espécies filogeneticamente distantes, é normal a utilização de alinhamentos de aminoácidos, pois, devido à degeneração do código genético, são mais conservados que os nucleotídeos (por exemplo, Mei e Dvornyk, 2015; Segatto et al., 2016).

Um exemplo de estudo evolutivo utilizando sequências de genes foi descrito para o gene P5CS *(Pyrroline-5-carboxylate synthetase)* em 2009. Neste estudo os autores demonstraram que a maioria das espécies de plantas apresentam duas cópias deste gene, resultantes de eventos de duplicações evolutivas independentes, apresentando um padrão espacial e temporal de expressão (Turchetto-Zolet et al., 2009). Esta enzima regula a síntese do aminoácido Prolina, sendo demonstrada em estudos com trigo, cevada e soja, a expressão aumentada do gene codificador para esta enzima em condições de estresse hídrico e salinidade (Aprile et al., 2009; Deng et al., 2013; Khoshro et al., 2013).

As consequências evolutivas em genes que controlam, por exemplo, transições fenotípicas têm sido estudadas através das sequências desses genes com o método de filogenia molecular. Por exemplo, um estudo recente examinou como repetidas perdas de pigmentação floral (antocianinas), associada com a transição da cor da flor em espécies da família botânica Solanaceae, têm afetado a evolução molecular de três genes envolvidos na via de biossíntese de antocianina. Os autores demonstraram uma ampla conservação desses três genes através das linhagens com ou sem pigmentação floral (antocianina). Esses achados são consistentes com o consenso crescente de que as perdas de pigmentação floral são em grande parte alcançados por alterações na expressão do gene, em oposição a mutações estruturais (Ho e Smith, 2016).

Romero e colaboradores (2017) investigaram a evolução do gene que codifica a proteína pró-filagrina, com função importante na pele de mamíferos. O gene filagrina

apresenta regiões repetidas em sequência, e o número de cópias dessas repetições variam entre e dentro de espécies. Dois principais modelos explicam a evolução de genes com variação do número de repetições em tandem, o modelo de evolução em concerto e o nascimento e morte (*birth-and-death*). Para determinar qual desses modelos melhor explica a evolução desse gene, a sequência repetitiva de um dos éxons do gene foi determinada em vários primatas, entre eles, orangotangos, gorilas, chimpanzés, e comparada com a sequência encontrada em humanos. Os autores observaram alta diversidade nucleotídica entre regiões de repetição em tandem na filagrina, o que se encaixa no modelo de evolução de nascimento e morte. Análises filogenéticas indicaram que vários eventos de duplicação independentes, e eventos de perda das repetições, deram origem ao diferente número de repetições e aos tamanhos distintos dessas repetições encontradas nas diferentes espécies investigadas. Os autores concluíram que a variação no número de cópias das regiões repetitivas de filagrina é uma consequência de divergência espécie-específica e expansão.

Análises filogenéticas tem sido um elemento essencial para a biologia moderna. Marcadores de sequência têm auxiliado também em trabalhos de sistemática para classificação e caracterização da diversidade. O campo da sistemática, que foi tradicionalmente baseado em informações de morfologia, anatomia, comportamento, fisiologia e geografia, experimentou uma revolução com a introdução de sequenciamento Sanger. Esses dados permitiram que hipóteses anteriormente propostas apenas com base em dados morfológicos, por exemplo, fossem testados utilizando-se também sequências de DNA e/ou RNA, aumentando substancialmente a resolução e suporte filogenético. Os dados de sequências de DNA têm fornecido evidências para grandes rearranjos taxonômicos, assim, aumentando significativamente a nossa compreensão sobre a "Árvore da Vida". Mais atualmente, tem-se utilizado dados de todo genoma para a reconstrução de árvores filogenéticas, utilizando as tecnologias de NGS ao invés de método de sequenciamento Sanger, uma abordagem chamada de Filogenômica. Um exemplo em plantas foi publicado em 2014, onde os autores usaram genomas nucleares disponíveis para reconstruir uma árvore filogenética entre clados representativos de monocotiledôneas para a inferência de eventos de duplicação do genoma ancestral. O estudo identificou vários eventos de duplicação do genoma, incluindo um unindo as monocotiledôneas Comelídeas (palmeiras, gengibre, ervas), um clado de importância ecológica/econômica e alta diversidade (Jiao et al., 2014).

A filogenia das plantas verdes foi estimada recentemente por um estudo que obteve os dados a partir de genomas de cloroplastos disponíveis em bancos públicos. Nesse estudo foram utilizados 78 genes codificadores de proteínas do cloroplasto de 360 espécies de plantas. Os autores foram capazes de determinar relações filogenéticas até então não resolvidas, porém ressaltam os desafios de trabalhar com genomas inteiros. Particularmente, encontraram inconsistências entre análises utilizando aminoácidos e nucleotídeos e perceberam que variações na proporção de GC entre as linhagens e dentro dos genomas afetaram o posicionamento de vários *taxa* (Ruhfel et al., 2014).

O sequenciamento de todo cloroplasto tem sido utilizado para investigar o relacionamento filogenético de *Ficus*. O estudo utilizou a montagem *de novo* do genoma plastidial de 59 espécies do gênero e mais seis grupos externos obtendo um maior suporte para os relacionamentos (Bruun-Lund et al., 2017), visto que análises conduzidas com sequências nucleares de baixo número de cópias e regiões específicas de cloroplasto não foram capazes de resolver com confiança os relacionamentos do grupo. Entretanto, foram observados conflitos entre a nova topologia baseada em todo o genoma de cloroplasto e os estudos prévios baseados em sequência nuclear, tanto para

as espécies individuais como para as relações entre algumas seções em níveis mais profundos. Os autores sugerem que os conflitos podem ser causados pela falta de resolução nos dados nucleares ou podem indicar potencial discordância cito-nuclear (Cruaud et al., 2012; Bruun-Lund et al., 2017).

Na área de filogeografia e genética de populações esses marcadores têm sido utilizados principalmente para compreender os eventos demográficos populacionais passados, e a atual estruturação genética encontrada nas populações em relação à distribuição geográfica. Recentemente, as tecnologias de sequenciamento de nova geração têm permitido a identificação de milhares de marcadores, como SNPs (*single nucleotide polimorphism*), que são aplicados para responder perguntas sobre a história dos organismos, com a promessa de revolucionar este campo de estudo, formando o que chamamos de Genômica populacional. Entretanto, marcadores organelares herdados de forma uniparental (por exemplo, DNA mitocondrial para animais e DNA de cloroplasto para plantas), continuam a representar um importante componente de dados filogeográficos (Garrick et al., 2015). Isto se deve ao fato de haplótipos de sequências de DNA organelar serem informativos sobre eventos e processos históricos (Sunnucks, 2000; Brumfield et al., 2003). Esses dados têm sido utilizados também em conjunto com marcadores de sequência nuclear, ou outros marcadores nucleares, como SSRs ou SNPs. Por exemplo, Thomaz e colaboradores (2015) utilizaram marcador de sequência mitocondrial (COI e NADH) juntamente com modelagem climática para avaliar se os padrões de variação genética podem refletir o impacto histórico da mudança climática no nível do mar ou nas condições ambientais em populações de *Hollandichthys multifasciatus*, um peixe de água doce endêmico da Floresta Atlântica da costa Brasileira. Os autores encontraram que a estrutura genética da espécie estava associada com as paleodrenagens, sugerindo que as conexões passadas das populações devido às mudanças do nível do mar tiveram um papel significativo na diversificação da ictiofauna ao longo das drenagens costeiras brasileiras.

Outro estudo investigou a história de dispersão espacial do HIV-1C brasileiro, utilizando uma abordagem filogeográfica com as sequências dos genes *pol* e *env*. O subtipo C é responsável pela maior carga de infecção pelo HIV em todo o mundo. No Brasil o HIV-1C foi recentemente introduzido quando comparado com o HIV-1B, espalhando-se rapidamente pelo Sul, onde é a variante dominante. Neste estudo foi mostrado um papel central para a cidade de Porto Alegre, capital do Rio Grande do Sul, na epidemia brasileira de HIV-1C. Os autores hipotetizaram que a expansão para o norte poderia estar ligada a populações fonte com cargas mais elevadas e proporções maiores por HIV-1C (Gräf et al., 2015).

A utilização de dados de sequências em análises filogenéticas e filogeográficas vêm aumentando, sendo utilizados nas mais diversas áreas de conhecimento, desde as ciências básicas às aplicadas. Neste capítulo foram apresentados alguns exemplos, porém o leque de possibilidades de utilização e análise, para responder as mais diferentes questões, é enorme.

Referências Bibliográficas

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215:403–10.

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417-434.

Anmarkrud JA, Lifjeld JT (2017) Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Molecular Ecology Resources* 17: 334-341.
Aprile A, Mastrangelo AM, De Leonardis AM, et al. (2009) Transcriptional profiling in response to terminal drought stress reveals differential responses along the wheat genome. *BMC Genomics* 10: 279.
Avise JC, Arnold J, Ball RM, et al. (1978) Intraspecific phylogeography: the mitochondrial-DNA bridge between population-genetics and systematics. *Annual Review of Ecology and Systematics* 18: 489-522.
Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography* 36: 3-15.
Bendich AJ (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays* 6: 279–282.
Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology* 17: 3754-3774.
Bourguignon T, Lo N, Cameron SL, et al. (2015) The Evolutionary History of Termites as Inferred from 66 Mitochondrial Genomes. *Molecular Biology and Evolution* 32: 406-421.
Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution* 18: 249–256.
Bruun-Lund S, Clement WL, Kjellberg F, Rønsted N (2017) First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of Ficus L. (Moraceae). *Molecular Phylogenetics and Evolution* 109: 93-104.
Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325: 31-36.
Carbonell-Caballero J, Alonso R, Ibañez V, Terol J, Talon M, Dopazo J (2015) A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic Species within the Genus *Citrus*. *Molecular Biology and Evolution* 32: 2015-2035.
Cavalli-Sforza LL (1998) The DNA revolution in population genetics. *Trends in Genetics* 14: 60-65.
Chase MW, Soltis DE, Olmstead RG, et al. (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 3: 528-580.
Chen CA, Wallace CC, Wolstenholme J (2002) Analysis of the mitochondrial 12S rRNA gene supports a two-clade hypothesis of the evolutionary history of scleractinian corals. *Molecular Phylogenetics and Evolution* 23: 137-149.
Clark LG, Zhang W, Wendel JF (1995) A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Systematic Botany* 20: 436-460.
Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25: 1423–1428.
Comer JR, Zomlefer WB, Barrett CF, et al. (2015) Resolving relationships within the palm subfamily Arecoideae (Arecaceae) using plastid sequences derived from next-generation sequencing. *American Journal of Botany* 102: 888–899.
Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modeling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 9: 539.

Cruaud A, Rønsted N, Chantarasuwan B, et al. (2012) An extreme case of plant-insect codiversification: figs and fig-pollinating wasps. *Systematic Biology* 61: 1029–1047.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772.

Deng G, Liang J, Xu D, Long H, Pan ZH, Yu M (2013) The Relationship between Proline Content, the Expression Level of P5CS (Δ1-Pyrroline-5-Carboxylate Synthetase), and Drought Tolerance in Tibetan Hulless Barley (*Hordeum vulgare* var. *nudum*). *Russian Journal of Plant Physiology* 60: 693–700.

Dong W, Xu C, Cheng T, Lin K, Zhou S (2013) Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biology and Evolution* 5: 989–997.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969-1973.

Dumolin-Lapegue S, Pemonge M-H, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. *Molecular Ecology* 6: 393-397.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792-1797.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564-567.

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* 22: 521-565.

Flook PK, Rowell CHF, Gellissen G (1995) The Sequence, Organization, and Evolution of the *Locusta migratoria* Mitochondrial Genome. *Journal of Molecular Evolution* 41: 928-941.

Fuertes Aguilar J, Nieto Feliner G (2003) Additive polymorphisms and reticulation in an ITS phylogeny of thrifts (*Armeria*, Plumbaginaceae). *Molecular Phylogenetics and Evolution* 28: 430-447.

Fuertes Aguilar J, Rosselló JA, Nieto Feliner G (1999) Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artificial hybrids of *Armeria* (Plumbaginaceae). *Molecular Ecology* 8: 1341-1346.

Garrick RC, Bonatelli IA, Hyseni C, et al. (2015) The evolution of phylogeographic data sets. *Molecular Ecology* 24:1164–1171.

Gräf T, Vrancken B, Maletich Junqueira D, et al. (2015) Contribution of Epidemiological Predictors in Unraveling the Phylogeographic History of HIV-1 Subtype C in Brazil. *Journal of Virology* 89: 12341-12348.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59: 307-321.

Ho WW, Smith SD (2016) Molecular evolution of anthocyanin pigmentation genes following losses of flower color. *BMC Evolutionary Biology* 16: 98.

Hollingsworth P, De-Zhu L, Van der Bank M, Twyford A (2016) Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B* 371: 20150338.

Hoot SB, Culham A, Crane PR (1995) The utility of *atpB* gene sequences in resolving phylogenetic relationships: comparison with *rbcL* and 18S ribosomal DNA sequences in the Lardizabalaceae. *Annals of the Missouri Botanical Garden* 82: 194-208.

Jansen RK, Cai Z, Raubeson LA, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences* 104: 19369–19374.

Jiao Y, Li J, Tang H, Paterson AH (2014) Integrated Syntenic and Phylogenomic Analyses Reveal an Ancient Genome Duplication in Monocots. *The Plant Cell* 26: 2792–2802.

Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (2009) Sistemática Vegetal: um enfoque filogenético. 3ª. ed. 632 p.

Khoshro HH, Taleei A, Bihamta MR, Shahbazi M, Abbasi A (2013) Expression Analysis of the Genes Involved in Osmotic Adjustment in Bread Wheat (Triticum aestivum L.) Cultivars under Terminal Drought Stress Conditions. *Journal Crop Science Biotechnology* 16: 173-181.

Kuhner M (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22: 768-770.

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870-1874.

Leliaert F, Tronholm A, Lemieux C, et al. (2016) Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Scientific Reports* 6: 25-367.

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.

Lorenz-Lemke AP, Togni PD, Mäder G, et al. (2010) Diversification of plant species in a subtropical region of eastern South American highlands: a phylogeographic perspective on native *Petunia* (Solanaceae). *Molecular Ecology* 19: 5240-5251.

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459-473.

Mei Q, Dvornyk V (2015) Evolutionary History of the Photolyase/ Cryptochrome Superfamily in Eukaryotes. *PLoS ONE* 10: e0135940.

Miflin BJ, Beevers H (1974) Isolation of intact plastids from a range of plant tissues. *Plant Physiology* 53: 870–874.

Ness RW, Kraemer SA, Colegrave N, Keightley PD (2016) Direct estimate of the spontaneous mutation rate uncovers the effects of drift and recombination in the Chlamydomonas reinhardtii plastid genome. *Molecular Biology and Evolution* 33: 800–808.

Nock CJ, Waters DL, Edwards MA, et al. (2011) Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* 9: 328–333.

Petit RJ, Duminil J, Fineschi S, et al. (2005) INVITED REVIEW: comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* 14: 689–701.

Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction. In: Phylogeography of Southern European Refugia (eds Weiss S and Ferrand N), pp. 23–97. Springer, AA Dordrecht, The Netherlands.

Prince LM (2015) Plastid primers for angiosperm phylogenetics and phylogeography. *Applications in Plant Sciences* 3: apps.1400085.

Randi E, Lucchini V, Hennache A, Kimball RT, Braun EL, Ligon JD (2001) Evolution of the mitochondrial DNA control-region and cytochrome b genes, and the inference of phylogenetic relationships in the avian genus *Lophura* (Galliformes). *Molecular Phylogenetics and Evolution* 19: 187-201.

Romero V, Hosomichi K, Nakaoka H, Shibata H, Inoue I (2017) Structure and evolution of the filaggrin gene repeated region in primates. *BMC Evolutionary Biology* 17: 10.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian inference of phylogeny. *Bioinformatics* 19: 1572-1574.

Ruhfel BR,' Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG (2014) From algae to angiosperms – inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*. 14: 23.

Salemi M, Vandamme A, Lemey P (2009) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge, UK: Cambridge University Press.

Segatto ALA, Thompson CE, Freitas LB (2016) Molecular evolution analysis of *WUSCHEL-related homeobox* transcription factor family reveals functional divergence among clades in the homeobox region. *Development Genes and Evolution* 226: 259-268.

Shaw J, Lickey EB, Beck JT, et al. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142-166.

Shaw J, Lickey EB, Schilling EE, Small RL (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in Angiosperms: the tortoise and the hare III. *American Journal of Botany* 94: 275-288.

Shi C, Hu N, Huang H, et al. (2012) An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE* 7: e31468.

Shinozaki K, Ohme M, Tanaka M, et al. (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal* 9: 2043–2049.

Solferini VN, Selivon D (2012) Polimorfismos de isozimas. *In*: Biologia Molecular e Evolução. Matioli SR, Fernandes FM (eds.) Ribeirão Preto: Holos. pp 165-169.

Staats M, Erkens RHJ, van de Vossenberg B, et al. (2013) Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* 8: e69189.

Stamatakis A (2014) RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30: 1312-1313.

Steele KP, Vilgalys R (1994). Phylogenetic analyses of Polemoniaceae using nucleotide sequences of the plastid gene *matK*. *Systematic Botany* 19: 126-142.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics* 73: 1162-1169.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics* 68: 978-989.

Straub SC, Parks M, Weitemier K, et al. (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

Stull GW, Moore MJ, Mandala VS, et al. (2013) A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* 1:1200497.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution* 15: 199–203.

Swofford DL (2002) PAUP: Phylogenetic Analysis using Parsimony (and other methods) v. 4.0b10. Sinauer Assoc., Sunderland.

Tarbelet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 17: 1105-1109.

Thomaz AT, Malabarba LR, Bonatto SL, Knowles LL (2015) Testing the effect of palaeodrainages versus habitat stability on genetic divergence in riverine systems: study of a Neotropical fish of the Brazilian coastal Atlantic Forest. *Journal of Biogeography* 42: 2389–2401.

Turchetto-Zolet AC, Margis-Pinheiro M,·Margis R (2009) The evolution of pyrroline-5-carboxylate synthase in plants: a key enzyme in proline synthesis, Molecular Genetics Genomics 281: 87–97.

Turchetto-Zolet AC, Segatto ALS, Turchetto C, Palma-Silva C, Freitas LB (2013) Guia Prático para estudos Filogeográficos. Sociedade Brasileira de Genética, Brasil.

Turchetto-Zolet AC, Salgueiro F, Turchetto C, et al. (2016) Phylogeography and ecological niche modelling in *Eugenia uniflora* (Myrtaceae) suggest distinct vegetational responses to climate change between the southern and the northern Atlantic Forest. *Botanical Journal of the Linnean Society* 182: 670-688.

Twyford AD (2016) Will benchtop sequencers resolve the sequencing trade-off in plant genetics? *Frontiers in Plant Science* 7: 433.

Twyford AD, Ness RW (2016) Strategies for complete plastid genome sequencing. *Molecular Ecology*. doi: 10.1111/1755-0998.12626.

Uribe-Convers S, Duke JR, Moore MJ, Tank DC (2014) A long PCR-based approach for DNA enrichment prior to next-generation sequencing for systematic studies. *Applications in Plant Sciences* 2: 1300063.

Wang JP, Co TW, Xuan SB, Wang H, Zhang M, Ma EB (2013) The complete mitochondrial genome of *Sasakia funebris* (Leech) (Lepidoptera: Nymphalidae) and comparison with other Apaturinae insects. *Gene* 526: 336-343.

Wang Y-l, Chen Y-h, Xia C-c, et al. (2015) The complete mitochondrial genome of the Common Red Apollo, *Parnassius epaphus* (Lepidoptera: Papilionidae: Parnassiinae). *Journal of Asia-Pacific Entomology* 18: 239-248.

Whittall JB, Syring J, Parks M, et al. (2010) Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19: 100–114.

Yang MR, Zhou ZJ, Chang YL, Zhao LH (2012) The mitochondrial genome of the quiet-calling katydids, *Xizicus fascipes* (Orthoptera: Tettigoniidae: Meconematinae). *Journal of Genetics* 91: 141-153.

Yang JB, Li DZ, Li HT (2014) Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Molecular Ecology Resources* 14: 1024–1031.

Yigit E, Hernandez DI, Trujillo JT, Dimalanta E, Bailey CD (2014) Genome and metagenome sequencing: Using the human methyl-binding domain to partition genomic DNA derived from plant tissues. *Applications in Plant Sciences* 2:11.

Zhang AB, Sota T (2007) Nuclear gene sequences resolve species phylogeny and mitochondrial introgression in *Leptocarabus* beetles showing trans-species polymorphisms. *Molecular Phylogenetics and Evolution* 45: 534-546.

Zhou Z, Zhao L, Liu N, et al. (2017a) Towards a higher-level Ensifera phylogeny inferred from mitogenome sequences. *Molecular Phylogenetics and Evolution* 108: 22-33.

Zhou Z, Min Q, Cheng S, Xin T, Xia B (2017b) The complete mitochondrial genome of *Thitarodes sejilaensis* (Lepidoptera: Hepialidae), a host insect of *Ophiocordyceps sinensis* and its implication in taxonomic revision of Hepialus adopted in China. *Gene* 601: 44-55.

Zimmer EA, Wen J (2012) Using nuclear gene data for plant phylogenetics: Progress and prospects. *Molecular Phylogenetics and Evolution* 65: 774-785.

Capítulo 6

Microssatélites: Metodologias de identificação e análise

Dra. Camila Martini Zanella, Dra. Caroline Turchetto, Dra. Clarisse Palma-Silva,
Dra. Fernanda Sperb-Ludwig

Considerações gerais

Microssatélites, conhecidos também como *Short Tandem Repeats* (STRs), *Simple Sequence Repeats* (SSR) ou *Simple Sequence Length Polymorphism* (SSLP), são repetições em tandem de motivos de 1 a 6 nucleotídeos que podem ser classificados de acordo com seu tamanho e o tipo de unidade de repetição (Box 6.1) (Litt e Luty, 1989; Tautz, 1989; Edwards et al., 1991; Jacob et al., 1991). Essas regiões repetitivas são frequentes tanto em procariotos quanto em eucariotos e também são comuns nos genomas nucleares e organelares (Zane et al., 2002). Dada a sua natureza repetitiva, o tamanho do fragmento dos locos de SSR tende a aumentar ou diminuir devido ao escorregamento da DNA polimerase durante a replicação (*slippage*), bem como outros eventos mutagênicos tais como *crossing over* desigual e retrotransposição (Schlötterer e Tautz, 1992). Como consequência, esses locos têm altas taxas de mutação, variando de 1×10^{-7} a 1×10^{-3} mutações por loco por geração em eucariotos em geral (Buschiazzo e Gemmell, 2006); entretanto, uma heterogeneidade complexa de eventos mutacionais é observada frequentemente em níveis de alelos, lócus e taxon. O número de unidades de repetição pode ser variável entre genótipos, o que torna os SSRs marcadores altamente polimórficos e adequados para diversos tipos de análises genéticas. Normalmente, as sequências de DNA que flanqueiam os motivos de SSR são conservadas entre indivíduos de uma mesma espécie, permitindo a projeção de oligonucleotídeos iniciadores (*primers*) específicos para essas sequências adjacentes ao SSR. Assim, por meio da Reação em Cadeia da Polimerase (PCR) é possível amplificar um determinado loco e identificar os polimorfismos individuais relacionados ao número de repetições do SSR para aquele indivíduo (Figura 6.1). Portanto, cada loco de SSR tem grande conteúdo informativo, pois é multialélico e permite a diferenciação entre indivíduos homozigotos e heterozigotos, por isso é considerado um marcador codominante.

Box 6.1:

Modelo de Mutação *Stepwise*: Quando uma região contento um motivo de SSR sofre uma mutação, ela ganha ou perde uma unidade de repetição. Esse ganho ou perda da unidade de repetição apresentam a mesma probabilidade de ocorrência, sendo considerada uma taxa fixa de mutação. Isto implica que dois alelos diferindo por somente um motivo são mais relacionados (ex. compartilham um ancestral comum mais recente) que alelos diferindo por várias repetições. Este modelo de mutação é usualmente preferido quando se estima estrutura populacional e relações entre indivíduos, exceto na presença de homoplasia (ex. quando dois alelos são idênticos por estado, mas não por descendência).

Os SSRs apresentam segregação mendeliana e podem ser isolados tanto de regiões codificadoras como regiões não codificadoras, estando sujeito ou não a seleção natural. Tais características têm possibilitado a sua aplicação em uma ampla gama de estudos, sendo que nas últimas décadas vários trabalhos foram publicados com diferentes abordagens como, por exemplo, a distribuição genômica dos SSRs, a sua

dinâmica evolutiva, função biológica e aplicações, incluindo análise forense, epidemiologia molecular, parasitologia, genética de população e da conservação, mapeamento genético, delimitação de espécies, análise de parentesco, perfil molecular, padrões de hibridação, determinação do modo de reprodução, entre outros (Tautz e Schlötterer, 1994; Jarne e Lagoda, 1996; Schlötterer, 1998; Chambers e MacAvoy, 2000; Li et al., 2002; Dieringer e Schlötterer, 2003; Ellegren, 2004; Buschiazzo e Gemmell, 2006; Chistiakov et al., 2006; Oliveira et al., 2006; Selkoe e Toonen, 2006; Barbará et al., 2007; Subirana e Messeguer, 2008; Sun et al., 2009; Kalia et al., 2011; Wheeler et al., 2014; Merritt et al., 2015).

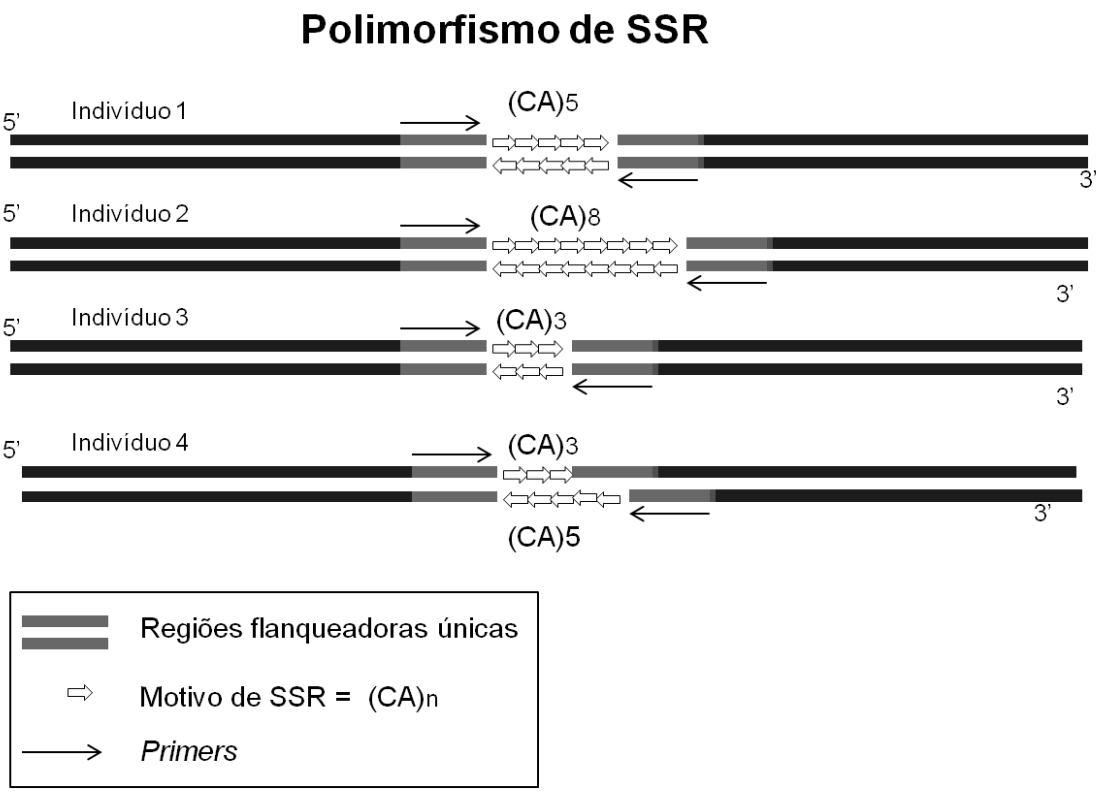

Figura 6.1 - Representação esquemática de uma região contendo um SSR e da variação de número de reptições entre indivíduos. A projeção de *primers* para amplificação por PCR em região flanqueadora única permite a identificação das diferenças nos números de repetição do SSR encontrado em cada indivíduo. Indivíduos 1 - 3 são homozigotos e indivíduo 4 é um heterozigoto.

Quadro 6.1 - Classificação dos microssatélites quanto ao tamanho e a unidade de repetição.

Tamanho da Repetição	Unidade de repetição
• Mononucleotídeo $(A)_n$ • Dinucleotídeos $(AC)_n$ • Trinucleotídeos $(ACG)_n$ • Tetranucleotídeos $(ACGT)_n$ • Pentanucleotídeos $(ACGTA)_n$ • Hexanucleotídeos $(ACGTAC)_n$	• Perfeita simples $(CA)_n$ • Imperfeita Simples $(AAC)_n$ TG $(AAC)_n$ • Interrompida $(CCA)_n$ TTGC $(CCA)_n$ • Perfeita composta $(CA)_n$ $(GAA)_n$ • Imperfeita composta $(CA)_n$ TG $(AGC)_n$

n = número de repetições

Apesar de todas as vantagens e aplicabilidades que os marcadores SSR apresentam, a grande limitação reside na necessidade do isolamento e desenvolvimento de *primers* específicos para cada espécie, ou ter disponível *primers* de espécies relacionadas para testar transferabilidade. As primeiras abordagens de isolamento e obtenção de marcadores SSRs polimórficos eram desenvolvidas através de técnicas laboriosas de construção de bibliotecas genômicas ou bibliotecas enriquecidas para obtenção dos marcadores para a espécie a ser estudadaEntretanto, nas últimas décadas têm ocorrido progressos significativos no desenvolvimento e genotipagem de marcadores SSR. Um deles se refere à emergência das tecnologias de sequenciamento de nova geração (*Next Generation Sequencing* – NGS), que têm possibilitado o desenvolvimento de centenas e milhares de marcadores SSRs, tornando possível a descoberta e uso de muitos locos polimórficos tanto para espécies modelo como para as não modelo. Além disso, a utilização do PCR multiplex ou simplesmente a genotipagem multiplex têm sido facilitadas por equipamentos de eletroforese capilar com base em tecnologia de DNA fluorescente induzida por laser, reduzindo significativamente os custos e tempo das análises genéticas (Chamberlain et al., 1988; Butler et al., 2001, 2004).

Metodologia de isolamento/identificação e genotipagem

Metodologias de Isolamento

Em uma época onde novas abordagens para análises genômicas surgem para estudar a variação genética, é importante lembrar que muitas questões biológicas podem ser eficientemente abordadas com um número limitado de marcadores altamente polimórficos, como os SSRs. Apesar dos diversos benefícios da utilização desses marcadores, uma das limitações da técnica é que os *primers* de microssatélites precisam ser isolados a partir da espécie a ser estudada ou, alternativamente, podem ser utilizados *primers* projetados para espécies ou gêneros evolutivamente próximos, caso estes estejam disponíveis na literatura. Contudo, Barbara et al. (2007) observou em sua pesquisa com espécies vegetais, fungos e animais, uma distribuição desigual do sucesso de transferência de marcadores SSRs polimórficos entre diferentes *taxa*, quanto maior a distância evolutiva dos *taxa* estudados menor o sucesso da amplificação e polimorfismo, além disso, o tamanho do genoma e sistema reprodutivo das plantas analisadas parecem também influenciar no sucesso da tranferibilidade desses marcadores. A transferência de marcadores pode facilitar a comparação entre espécies proximamente relacionadas, permitindo estudos de mecanismos envolvidos em divergência populacional, hibridação e especiação, bem como padrão de diversidade em uma comunidade.

Inicialmente, os locos de microssatélites eram identificados a partir da construção de uma biblioteca genômica para a espécie alvo (Zane et al., 2002), podendo esta biblioteca ser enriquecida com sequências contendo microssatélites (Figura 6.2), uma etapa crucial para espécies com grandes genomas. Para a construção de tais bibliotecas, o DNA genômico de alta qualidade precisa ser fragmentado usando enzimas de restrição, que clivam o DNA em regiões específicas, ou, menos comumente, por sonicação, um método físico de fragmentação do DNA via ultrassom (Karagyozov et al., 1993). A escolha das enzimas a serem usadas depende do comprimento médio desejado dos fragmentos de DNA, da repetição de microssatélites que podem ser encontrados e do tipo de extremidades (extremidade coesiva ou cega) dos fragmentos de restrição. O DNA clivado é então selecionado pelo tamanho dos fragmentos, para preferencialmente

obterem-se fragmentos pequenos de 300 a 700 pares de base (pb). Uma grande parte dos protocolos de enriquecimento envolve a ligação de sondas marcadas com biotina a esferas magnéticas recobertas por estreptavidina. A ligação estável formada por biotina e estreptavidina permite a utilização de sondas ligadas a esferas magnéticas para selecionar fragmentos de interesse através da utilização de um ímã. Dependendo do método de clivagem, os fragmentos de DNA podem ser ligados a um vetor de clonagem (plasmídeo), seja diretamente ou após a ligação de adaptadores específicos. Esta etapa é mais crítica, devido ao risco de obtenção de um número reduzido de recombinantes e a formação de concatâmeros entre fragmentos genômicos. A transformação de células bacterianas com produto de ligação geralmente produz milhares de clones recombinantes, que podem ser subsequentemente pesquisados quanto à presença de sequências de microssatélites. Após a identificação de clones contendo as repetições, estes são sequenciados e são projetados *primers* específicos para a amplificação de cada loco de SSR (Figura 6.2). As condições da PCR são otimizadas para permitir a amplificação dos locos em diferentes indivíduos de uma população (Zane et al., 2002).

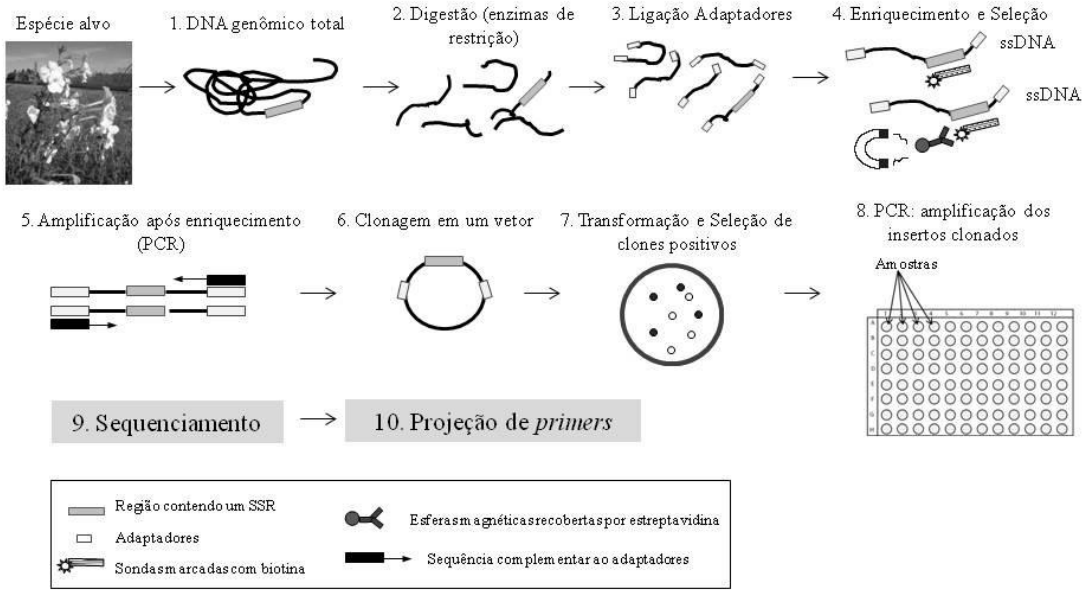

Figura 6.2 - Representação esquemática das principais etapas da construção de uma biblioteca genômica enriquecida de locos de microssatélites até a obtenção de primers para amplificação na espécie alvo. ssDNA – DNA de fita simples.

Várias estratégias alternativas foram desenvolvidas objetivando reduzir o tempo investido no isolamento de marcadores microssatélite e/ou aumentar a sua eficácia, como por exemplo, baseado em modificações da técnica de RAPD (*Random Amplified Polymorphism DNA*). Neste protocolo, vários *primers* RAPD são usados para obter fragmentos de DNA aleatoriamente amplificados do genoma da espécie alvo e as bandas do PCR são hibridizadas com sondas de motivos de SSR. Essa técnica era promissora devido ao fato de que os fragmentos de RAPD pareciam conter repetições de microssatélites mais frequentemente que clones genômicos aleatórios (Williams et al., 1990; Lunt et. al., 1999; Cifarelli et. al., 1995). Outra estratégia diferente foi proposta para a produção de bibliotecas altamente enriquecidas com repetições específicas de microssatélites usando uma reação chamada extensão de *primers*. Este método se baseia

na construção de uma biblioteca genômica primária com fragmentos de DNA de fita simples, que são usados na reação de extensão com oligonucleotídeos com repetições específicas, gerando um produto de dupla fita com a repetição desejada. Entretanto, este método envolve um elevado número de etapas (Ostrander et. al., 1992; Paetkau, 1999). Outro protocolo, particularmente muito simples, foi proposto baseado em hibridização seletiva. A primeira etapa deste protocolo é idêntica ao procedimento de isolamento tradicional, objetivando produzir pequenos fragmentos de DNA genômico que são então ligados a uma sequência conhecida, um vetor ou um adaptador. Posteriormente, o DNA é hibridizado com sondas contendo repetições de SSR. Após esta etapa de hibridização e várias lavagens para remover ligações não específicas, o DNA é eluído e amplificado por PCR e, finalmente, o DNA enriquecido é clonado. Dependendo da eficiência de todo o procedimento, os clones recombinantes podem ser diretamente sequenciados (Karagyozov et. al., 1993; Armour et. al., 1994; Hamilton et. al., 1999; Kijas et. al., 1994).

Nos últimos anos com a disponibilidade de bases de dados públicas, também está sendo possível identificar e projetar *primers* para locos de microssatélites a partir de sequências depositadas nessas bases de dados, por exemplo, GenBank. Bases de dados como de ESTs (*Expressed Sequences Tags*), por exemplo, têm sido alvo para a pesquisa e isolamento de locos SSRs. Esses microssatélites são geralmente chamados de EST-SSR ou microssatélites gênicos. Nessa abordagem, programas computacionais de identificação de motivos de SSR são utilizados para a pesquisa nos dados das sequências dos ESTs (Varshney et al., 2007). Vários programas computacionais têm sido desenvolvidos para reconhecer o padrão de SSR em arquivos de sequências, como por exemplo, MISA (MIcroSAtellite), SSRfinder, Sputnik, SSRIT (SSR Identification Tool), SSRSEARCH, TRF (Tandem Repeat Finder), etc. EST-SSR podem oferecem algumas vantagens sobre os SSRs genômicos neutros, dependendo do objetivo do estudo, porque eles podem também detectar variação em regiões expressas do genoma e, consequentemente, gerar marcadores diretamente associados a uma característica de interesse, além de poderem ser usados em um grande número de espécies relacionadas devido ao maior grau de transferabilidade (Gupta et al., 2003; Victoria et al., 2011).

Avanços nas tecnologias de sequenciamento de nova geração (NGS – *Next Generation Sequencing*; Capítulo 2) têm possibilitado o isolamento de milhares de locos SSRs, tornando acessível a identificação de muitos locos SSRs polimórficos, principalmente para espécies não modelo. Essas tecnologias não requerem obrigatoriamente a criação de bibliotecas enriquecidas para motivos de SSRs, sendo que o DNA ou RNA total podem ser sequenciados. A principal vantagem de isolar marcadores SSRs a partir de sequências do transcritoma é, assim como os EST-SSR, a possibilidade de encontrar associações com genes funcionais e, assim, com o fenótipo (Li et al., 2002; Kumar et al., 2015). Por outro lado, uma vez que a taxa de mutação do DNA em sequências codificadoras é menor que em outras regiões (por exemplo, íntrons, regiões intergências), espera-se que o número de SSR e o grau de polimorfismo também sejam menores comparados às regiões não codificadoras (Blanca et al., 2011).

Duas principais tecnologias de NGS têm sido amplamente utilizadas para o isolamento de marcadores microssatélites, uma delas é o pirosequenciamento na plataforma 454 e o sequenciamento por síntese na plataforma Illumina (ver Capítulo 2). A plataforma 454 pode ter algumas vantagens sobre a Illumina quando se analisa o genoma ou o transcritoma, principalmente pelo maior comprimento dos *reads* que são gerados, o que torna hábil projetar *primers* para amplificação de locos SSRs mesmo com baixa e média cobertura de sequenciamento (Zalapa et al., 2012).

Devido à quantidade de dados gerados a partir de NGS, os pesquisadores devem analisar cuidadosamente os dados de sequência para isolar locos SSRs de alta qualidade para futuros testes. Várias ferramentas de bioinformática têm sido desenvolvidas para caracterizar e detectar SSRs. Uma vez que foi encontrada variação significativa entre os algoritmos usados para a detecção de SSRs, é recomendado considerar o uso de mais de um programa para procurar motivos de SSRs em conjuntos de dados obtidos a partir de tecnologias de NGS (Merkel e Gemmell, 2008; Cavagnaro et al., 2010; Lim et al., 2013). Alguns programas úteis que podem ser usados de forma independente ou em conjunto durante a descoberta de SSRs incluem o já citado MISA, mreps (um programa capaz de também encontrar repetições imperfeitas), SSR Locator, WebSat, AS-SSR e CandiSSR (Kolpakov, et al., 2003; Thiel et al, 2003; Maia et al., 2008; Martins et al., 2009; Pickett et al., 2016; Xia et al., 2016).

Projeção de primers, otimização da reação de PCR e genotipagem

A projeção dos *primers* é uma etapa crucial para uma análise bem-sucedida de marcadores SSR. Para isto, é necessário o conhecimento prévio das sequências que flanqueiam os motivos de SSR, sendo possível projetar *primers* em regiões conservadas que amplifiquem um único loco em diferentes populações dentro de uma espécie, ou até mesmo que seja possível amplificar em espécies relacionadas, dependendo do objetivo do estudo.

A seleção dos locos SSRs e a projeção de *primers* devem ser realizadas com rigor, principalmente quando se deseja realizar PCR multiplex. Neste caso, deve-se dar preferência por selecionar pares de *primers* que amplifiquem fragmentos de tamanhos contrastantes (100pb, 200pb, 300pb), o que permitirá a genotipagem de vários marcadores com uma única fluorescência. Existem programas computacionais que simultaneamente identificam e projetam *primers* para multiplex, sendo que alguns até pesquisam combinações adequadas de pares de *primers* (Kaplinski et al., 2005; Rachlin et al., 2005; Kraemer et al., 2009; Shen et al., 2010). Para garantir o sucesso da co-amplificação, é fundamental eliminar os *primers* com interações potenciais de dímeros de *primers* (Vallone e Butler, 2004; van Asch et al., 2010). Um Blast local ou ferramentas específicas, tais como *Multiplex Manager* ou *NetPrimer* (Premier Biosoft International, EUA) podem ser utilizadas para esta finalidade (Holleley e Geerts, 2009). Pensando-se em PCR multiplex, os *primers* devem ter uma faixa similar de temperatura de anelamento (58–60 °C tem sido considerado ótimo). Também é importante evitar a presença de nanossatélites no amplicom. Isto tem sido levado em conta, por exemplo, no programa QDD designado para isolar locos SSRs a partir de bibliotecas com milhares de fragmentos de DNA (Butler, 2005; Hill et al., 2009; Meglécz et al., 2010).

Uma vez obtidos os SSRs candidatos, uma série de escolhas podem ser feitas para selecionar os melhores marcadores e economizar tempo em otimizações. Interessantemente, a grande quantidade de dados de sequências disponíveis obtidos a partir de NGS tem permitido esta pré-seleção de marcadores a serem utilizados. Três características principais devem ser observadas para a escolha do SSR: tipo, tamanho e número da unidade de repetição.

Deve ser dada preferência por motivos perfeitos, uma vez que se deve assegurar que os locos de microssatélites sigam, tanto quanto possível, o modelo de mutação *Stepwise* (Box 6.1) utilizado em métodos baseados em coalescência para inferir eventos demográficos (Estoup et al., 2001; Gusmão et al., 2006).

Repetições de mononucleotídeos de SSRs podem ser difíceis de genotipar com precisão, de modo que muitas vezes são eliminados já no início para evitar erros de genotipagem (Sun et al., 2006; Kim et al., 2008). Por outro lado, infelizmente, muitas

vezes as repetições de dinucleotídeos mostram uma ou mais bandas "*stutter*", ou seja, são visualizados vários produtos da PCR de um mesmo marcador, que são tipicamente mais curtos por uma ou algumas repetições, do que o produto de comprimento completo. Bandas "*stutter*" são atribuídas ao escorregamento da enzima DNA polimerase durante a amplificação tornando a designação do alelo difícil, especialmente para os heterozigotos com alelos adjacentes (Levinson e Gutman, 1987; Meldgaard e Morling, 1997; Chambers e MacAvoy, 2000;). Em contraste, repetições de tri-, tetra- ou pentanucleotídeos parecem ser significativamente menos propensos ao escorregamento. Por isso, por exemplo, esses motivos de SSR são muitas vezes preferidos para análises forenses e de parentesco. Já o número de repetições tem um efeito crítico sobre o comportamento da mutação, inclusive esta característica ajuda a definir que sequências realmente representam microssatélites. No geral, locos de SSR com mais repetições têm taxas de mutação mais elevadas, sendo mais polimórficos. Dessa forma, torna-se necessário selecionar locos com um número suficiente de repetições para assegurar que existam polimorfismos. Além disso, microssatélites com numerosas repetições são caracterizados por grande faixa alélica, de modo que um menor número de locos pode ser combinado num dado multiplex (Edwards et al., 1991; Hoffman e Amos 2005; Kelkar et al., 2010).

Os pares de *primers* de SSR podem ser testados e otimizados em PCR simples para assegurar a seleção de *primers* eficientes e que sejam informativos. Através dessa verificação inicial é possível identificar e descartar *primers* que apresentem picos duplos, excesso de bandas "*stutter*", alelos nulos, baixo polimorfismo, bem como outros artefatos. Se determinado loco é de extrema importância, é possível reprojetar o *primer* ou aperfeiçoar a reação de PCR. Para esses testes iniciais, é necessário utilizar um conjunto de amostras que seja representativo da diversidade genética de diferentes populações, por exemplo, tornando possível identificar os níveis de polimorfismo do marcador. A escolha deste conjunto de indivíduos representativos da diversidade para a pesquisa inicial pode diminuir o risco de se encontrar novos alelos que diferem amplamente no seu tamanho, podendo se sobrepor com alelos de outro marcador com a mesma fluorescência, caso seja utilizada a estratégia de genotipagem multiplex. Caso os testes iniciais sejam realizados em sequenciador automático, é necessário sintetizar o *primer* direto marcado com fluorescência, o que aumenta consideravelmente os custos destes testes. Uma estratégia de genotipagem que tem sido amplamente utilizada é adicionar um terceiro *primer* universal na reação de PCR denominado M13, o qual é marcado com fluorescência, além disso, é necessário sintetizar o *primer* direto com a mesma sequência universal M13 na extremidade 5', constituindo uma espécie de "cauda". Na PCR adicionam-se os três *primers*: direto com cauda M13 (*forward*), reverso (*reverse*) e *primer* M13 marcado com fluorescência, que pode ser 6-FAM, NED, VIC, PET ou HEX dependendo da fluorescência do padrão de peso molecular utilizado. A principal vantagem da utilização da estratégia do *primer* adicional M13, é que há a necessidade de comprar apenas três ou quatro *primers* marcados com a fluorescência, podendo utilizá-los com inúmeros locos de SSR, inclusive para diferentes espécies. Diversos estudos com diferentes organismos têm utilizado essa técnica para desenvolver todo o estudo e não apenas para os testes iniciais em função da redução dos custos. Neste caso as PCRs são realizadas separadamente e apenas a genotipagem é multiplex (Collins et al., 2000; Schuelke, 2000; Cryer et al., 2005).

Alguns autores têm defendido o uso de PCR multiplex, chamando de verdadeiro multiplex quando os marcadores são combinados numa única PCR. O objetivo desta técnica é combinar todos os marcadores a serem utilizados no estudo em um pequeno número de PCRs com cada loco com uma dada fluorescência. Esta técnica pode

diminuir tanto o trabalho de laboratório quanto os custos com reagentes e utilizar uma quantidade menor de DNA. Entretanto, potenciais problemas na PCR como falsos negativos, devido a uma falha de reação, ou falsos positivos devido à contaminação, marcam a PCR multiplex como uma técnica sensitiva. Por isso, para obter resultados reproduzíveis, é necessária uma padronização cuidadosa de todas as etapas. Por exemplo, a padronização da concentração do DNA é de extrema importância, sendo que pouco DNA pode resultar em má amplificação, bem como evasão de alelos, porém, muito DNA pode ser mais problemático, podendo levar a um sinal fluorescente fora da escala, bem como vários artefatos, desde o desbalanço entre locos, até bandas "*stutter*" de várias formas. Assim, rigoroso controle de qualidade deve ser usado para limitar os erros de genotipagem (revisado em Guichoux et al., 2011).

Em função da alta sensibilidade e necessidade de padronização do PCR multiplex, uma estratégia que tem sido muito utilizada é realizar as PCRs de cada loco separadamente, marcados com diferentes fluorescências, e realizar apenas uma genotipagem multiplex. Esses *primers* podem ser marcados diretamente com fluorescência ou utilizar a estratégia da cauda universal M13, como descrita previamente. Este tipo de abordagem tem grande aplicabilidade entre as espécies não modelo, principalmente quando se utiliza locos heterólogos, na qual não se podem ajustar as etapas para o desenvolvimento da PCR multiplex.

Depois de concluída a etapa da PCR, a genotipagem dos indivíduos pode ser realizada por eletroforese capilar em equipamento automatizado ou também pode ser realizada por géis de poliacrilamida corados com nitrato de prata. A genotipagem em gel de poliacrilamida tem uma precisão de identificação de diferença de alelos de até 2 pb. A genotipagem por eletroforese capilar em sequenciador automático é mais eficiente e precisa, sendo possível diferenciar fragmentos com apenas 1 pb de diferença, tornando mais fácil genotipar locos dinucleotídicos e com muitos alelos. O sequenciador automático também possibilita analisar placas com 96 ou 384 amostras e vários locos por leitura. A determinação do tamanho do alelo é realizada em comparação a um padrão de peso molecular corado com fluorescência (por exemplo, 350ROX, 400HD, 120LIZ, 500LIZ ou 600LIZ). O padrão de peso molecular é adicionado em cada uma das amostras na placa de corrida, e o programa de análise do sequenciador automático utiliza esse padrão para criar uma curva para cada capilar analisado, gerando uma análise individual e mais precisa para a determinação do tamanho do alelo.

Identificação dos alelos

Após a leitura dos locos SSRs multiplexados, os genótipos correspondentes de cada amostra devem ser lidos e identificados. Considerando a genotipagem realizada em sequenciador automático há duas etapas importantes: a leitura do verdadeiro tamanho (usando números decimais) e a conversão do alelo para um tamanho real de fragmento de DNA.

Para a primeira etapa é realizada a identificação do *amplicon* (picos/bandas) que correspondem ao alelo e o tamanho dos fragmentos correspondentes. Programas computacionais são fornecidos pelos desenvolvedores dos sistemas de eletroforese capilar e apresentam uma correção automática de problemas comuns de genotipagem (picos saturados, bandas "*stutter*", excessivo ruído de linha de base, etc), ou pode-se utilizar programas alternativos como Peak Scanner™ ou Genemarker, por exemplo. Entretanto, é recomendada a visualização de todos os alelos individualmente e a edição manual para obter um padrão de leitura dos alelos entre indivíduos e espécies. Uma vez que esta etapa pode ser trabalhosa e pode estar sujeita a erros, é importante selecionar

bem os marcadores que apresentam um bom padrão de genotipagem desde o início, como discutido anteriormente.

A etapa seguinte, a conversão do alelo, é bem crítica e pode ser a responsável pela maioria das discrepâncias encontradas entre laboratórios na anotação de alelos dinucleotídeos (Weeks et al., 2002), causadas por decisões arbitrárias na conversão do tamanho do alelo, principalmente quando se assume a edição automática fornecida pelos programas computacionais sem visualização individual dos alelos. Um procedimento que pode ser adotado é exportar os dados de tamanho de fragmento para uma planilha e usá-lo para compilar parcelas de frequência acumulativa de distribuição de tamanho. Assim, novas posições para o número inferido de repetições pode então ser construída em torno destas distribuições (veja Jayashree et al., 2006 para mais detalhes). Isto ajuda a identificar alelos que se desviam da periodicidade de repetição esperada. Alguns programas têm sido disponibilizados para esta proposta, como por exemplo, ALLELOBIN, FLEXIBIN, TANDEM, AUTOBIN (Idury e Cardon, 1997; Amos et al., 2007; Matschiner e Salzburger, 2009; Guichoux et al., 2011). Neste último, o número de amostras e locos são automaticamente detectados, o tamanho do alelo é classificado e plotado para detectar lacunas relevantes de tamanho. Muitos dados acabam sendo perdidos em função do esforço de fazer a congruência entre os genótipos de diferentes conjuntos de dados. Para isto, programas específicos (ex: ALLELOGRAM e MICROMERGE) têm sido desenvolvidos com a finalidade de normalizar a congruência dos alelos a partir de múltiplas fontes de dados, utilizando um pequeno conjunto como controle. A congruência dos alelos pode também ser realizada com genótipos de referência. Entretanto, recomenda-se muita cautela e inspeção manual também desses conjuntos de dados para evitar a comparação de amostras sem correspondência dos alelos.

Após a leitura dos alelos de cada genótipo, através de checagem manual, todo o conjunto de dados é inserido em planilhas que poderão ser convertidas nos formatos dos arquivos de entrada para diferentes tipos de análises e programas.

Métodos utilizados para análise de matrizes de dados

Aqui neste tópico, vamos apresentar uma breve compilação de análises que podem ser realizadas utilizando dados de SSR em estudos com diversos propósitos, assim como serão apresentados alguns programas computacionais que podem ser utilizados para obtenção dos resultados. De uma forma geral, após a genotipagem se obtém uma matriz com o genótipo de cada indivíduo para cada loco, onde se pode identificar o homozigoto (por exemplo, alelo 0101 e 0202 ou 200/200 e 202/202 pb) e o heterozigoto (por exemplo, alelos 0102 ou 200/202 pb). Com base nesta matriz de dados são organizados os arquivos de entrada (*input files)*, ou seja, o arquivo inicial com os dados brutos para as análises seguindo o tutorial de cada programa.

Análises de diversidade genética entre populações ou espécies

Também chamada de estatística descritiva, visa inferir sobre os principais parâmetros de diversidade genética observados entre os indivíduos dentro de uma mesma população, entre populações de uma mesma espécie, entre espécies ou grupo de espécies. Parâmetros como número médio de alelos por loco, número de alelos privados e alelos efetivos, riqueza alélica, heterozigosidade observada e esperada, porcentagem de locos polimórficos, coeficiente de endocruzamento, entre outros, são comumente encontrados em estudos de diversidade genética. Esses dados podem ser calculados a partir da matriz dos genótipos de SSR, levando em consideração as estimativas das

frequências alélicas. Inúmeros programas foram desenvolvidos para estimar os índices de diversidade genética e muitos deles são de acesso livre, como por exemplo: ARLEQUIN, FSTAT, GDA, GENEPOP, GENALEX, MSA, TFPGA (Excoffier e Heckel, 2006), pacotes de análise no R como monpop, microbov e genetics, entre outros também são utilizados. Os níveis de diversidade genética podem ser influenciados por inúmeros fatores, como a natureza do marcador molecular utilizado, amostragem populacional e principalmente por hábitos de vida da espécie estudada. Diferentes hábitos apresentam forte influência nos índices de diversidade genética, como por exemplo, a área de distribuição de uma espécie (ampla, restrita, endêmica). Em plantas o tipo de sistema de cruzamento (cruzado, misto ou autofecundação), o modo de dispersão das sementes (por exemplo, anemocórica, hidrocórica, zoocórica, barocórica e endocórica) são alguns dos fatores que influenciam nos índices de diversidade genética observados (Nybom, 2004), assim como o modo de acasalamento (preferencial ou ao acaso) é um dos fatores que tem influência sobre os níveis de diversidade genética em animais. Desta forma, o conhecimento dos hábitos de vida da espécie estudada auxilia na interpretação dos resultados observados, podendo contribuir para estudos de genética da conservação, por exemplo, ou no conhecimento dos recursos genéticos em bancos de germoplasma.

Outro fator que pode influenciar os parâmetros de diversidade, e que também representa uma das limitações no uso de SSR, é a presença de mutações nas regiões flanqueadoras (*primers*), as quais podem ser responsáveis pela ocorrência de alelos nulos (Chapuis e Estoup, 2007), isto é, alelos que não são amplificados na reação de PCR em função de problemas no anelamento do *primer* em virtude da ocorrência de uma mutação nessa região. Consequentemente, indivíduos heterozigotos podem ser erroneamente identificados como homozigotos para esses locos, diminuindo a proporção de indivíduos heterozigotos observados em relação ao esperado em Equilíbrio de *Hardy-Weinberg*, subestimando os níveis de diversidade genética.

Estrutura populacional

Estimativas de estrutura populacional visam inferir sobre a estruturação populacional dentro de uma espécie, podendo estimar os níveis de fluxo gênico e dispersão entre populações, além de fornecer dados sobre a diferenciação genética dessas populações ao longo da distribuição geográfica do organismo, sendo cruciais para entender a conectividade entre elas. Os marcadores de SSR são utilizados rotineiramente para avaliar a estruturação genética de populações naturais, sendo importantes ferramentas em estratégias conservacionistas, assim como sobre a biologia reprodutiva e ecologia das espécies. Para estimar a conectividade e os padrões de fluxo gênico entre as populações, comumente são apresentadas as estimativas da estatística hierárquica "F" (F_{IS}, F_{ST} e F_{IT}) (Weir e Cockerham, 1984). O F_{ST} mede o déficit de heterozigotos devido à subestruturação das populações (efeito *Wahlund*), podendo variar de 0 a 1. Existem outras estimativas que podem ser calculadas com dados de SSR para inferir estruturação populacional, como o G_{ST}, G'_{ST}, D, estatística-Φ e R_{ST}. Esta última foi desenvolvida especificamente para dados de SSR, assumindo o modelo mutacional "*Stepwise Mutation Model* – SMM" (Box 6.1). Para frequências alélicas de genomas haploides (mitocondrial ou plastidial) é utilizado o N_{ST}, o qual leva em consideração as distâncias filogenéticas entre os haplótipos.

Hábitos de vida como sistema de acasalamento, barreiras ambientais, processos históricos, entre outros, podem influenciar na estrutura genética das populações, incluindo também a ação antrópica. Barreiras genéticas podem ser estimadas através dos programas SAMOVA e BARRIER (Dupanloup et al., 2002; Manni et al., 2004). Este

último emprega o algoritmo de Monmonier de máxima diferenciação, identificando descontinuidades genéticas através da utilização de dados de diferenciação populacional como F_{ST} ou seus análogos e utilizando as coordenadas geográficas de cada população através do cálculo de triangulação de Delaunay. Com esses resultados é possível inferir possíveis barreiras geográficas ao fluxo gênico entre as populações. Análises de correlação entre matriz de dados genéticos e matriz de distância geográfica entre as populações também podem ser realizadas com dados de SSR, como por exemplo, o Teste de Mantel, o qual é implementado em programas como ARLEQUIN, FSTAT, GENALEX, GENETIX, GENEPOP e SPAGEDi (Excoffier e Heckel, 2006). Outra análise muito utilizada com dados de microssatélites é a análise de variância molecular (AMOVA), com três níveis hierárquicos, os quais avaliam qual a porcentagem da diversidade genética está dentro das populações estudadas (F_{SC}), o quanto desta diversidade está entre as populações (F_{ST}) e entre os grupos estipulados hierarquicamente (F_{CT}). Tanto a análise de variância molecular quanto a identificação de barreiras genéticas utilizando o programa BARRIER podem ser realizadas com dados de SSR nucleares ou organelares.

Abordagens Bayesianas também tem sido amplamente utilizadas com dados de microssatélites para inferir estruturação populacional. Numerosos programas foram desenvolvidos com esse propósito, como STRUCTURE, BAPS, BAYES, BAYESASS+, GENECLUST, GENELAND, STRUCTURAMA (Excoffier e Heckel, 2006), entre outros, levando em consideração diferentes estratégias e inferências computacionais. Os métodos bayesianos de análise combinam informações *a priori* com a probabilidade da informação observada para calcular uma distribuição *a posteriori.* A distribuição *a posteriori* é calculada, na maioria dos casos, utilizando Cadeias de Markov Monte Carlo (MCMC). Os programas STRUCTURE (Pritchard et al., 2000) e BAPS ("*Bayesian Analysis of Population Structure*"; Corander et al., 2004) têm sido mais amplamente utilizados em estudos de genética de população e filogeografia, os quais permitem inferir sobre a presença ou não de estrutura nas populações, atribuindo indivíduos às populações, identificando migrantes e indivíduos miscigenados. Ambas as análises podem ser realizadas com dados nucleares ou organelares e também é possível incluir as coordenadas geográficas das populações, realizando inferências espaciais.

Identificação de híbridos e estimativas de fluxo gênico

Marcadores SSRs também tem sido utilizados na identificação de híbridos naturais e artificiais. Hibridação é um fenômeno natural bem documentado na literatura, tendo uma função importante na evolução de plantas e animais, sendo considerada uma poderosa força evolutiva que cria oportunidades para a diversificação adaptativa e especiação em populações (Rieseberg e Carney, 1998). Estudos de zonas híbridas avaliando o grau de fluxo gênico interespecífico podem fornecer importantes informações quanto ao tipo e poder do isolamento reprodutivo entre as espécies que estão hibridizando, contribuindo para a compreensão da dinâmica evolutiva envolvida no processo de especiação de linhagens recentes.

Análises Bayesianas implementadas em programas como STRUCTURE, NEWHYBRIDS, GENELAND, HEXT, LEA, HINDEX (Excoffier e Heckel, 2006), pacotes de análise no R como Introgress têm sido utilizadas em estudos de zonas híbridas naturais ou artificiais. Indivíduos com perfil molecular intermediário entre as espécies parentais são identificados como tendo provável origem híbrida. Algumas análises também permitem, por exemplo, a classificação dos indivíduos em distintas categorias híbridas (F1, F2 ou retrocruzamentos com os parentais), o que é especialmente importante para documentar se existe introgressão e em qual direção ela

ocorre (Anderson e Thompson, 2002). O relacionamento haplotípico obtido a partir de SSRs organelares também tem sido utilizado para inferir graus de introgressão e fluxo interespecífico utilizando o programa NETWORK (disponível em: http://www.fluxus-engineering.com).

Estimativa do grau de fluxo gênico e migração entre populações ou espécies que divergiram recentemente, bem como em zonas híbridas também podem ser realizadas com dados de SSR. Além da estimativa de fluxo gênico entre populações também é possível detectar a colonização de novas áreas, e identificar possível origem de uma espécie introduzida. O programa BAYEASS, o qual utiliza uma abordagem Bayesiana, pode ser utilizado para este fim, onde é possível estimar a direção e taxa de migração contemporânea (Wilson e Rannala, 2003). BAYESASS toma relativamente poucas suposições sobre demografia e pode ser aplicado a populações que não estão em equilíbrio de *Hardy-Weinberg* ou mutação-deriva. BAYESASS calcula o nível de fluxo gênico com base na taxa de migração dentro das últimas três gerações, fornecendo estimativas mais confiáveis quando as taxas de migração são baixas e a diferenciação genética entre as populações é alta. Os resultados da análise BAYESASS podem ser considerados confiáveis quando valores de F_{ST} > 0,05 são observados entre as populações ou espécies a serem analisados. Uma abordagem interessante é utilizar combinadamente os resultados obtidos pelo STRUCTURE e BAYESASS, pois fornecem informações complementares sobre o fluxo gênico recente. Enquanto o STRUCTURE utiliza um modelo Bayesiano probabilístico para atribuir os indivíduos a agrupamentos (*clusters*), BAYESASS usa um algoritmo Bayesiano de atribuição para estimar a probabilidade posterior da história de migração individual. Outro programa que pode ser utilizado para estimativas de migração é o MIGRATE-N implementado com algoritmo Bayesiano e baseado na teoria da coalescência (Beerli e Felsenstein, 1999). Com este programa é possível estimar além de taxas de migração por geração, o tamanho efetivo populacional em valores de Theta ($4Ne\mu$ para populações diploides, $2Ne\mu$ para populações haploides e $1Ne\mu$ para populações haploides transmitidas por apenas um dos sexos, onde *Ne* é o tamanho efetivo populacional e μ a taxa de mutação).

Mapeamento genético

Pelo fato de serem amplamente distribuídos no genoma, os SSRs têm se mostrados úteis na construção de mapas genéticos, por permitirem uma ampla cobertura do genoma e a integração dos dados genéticos a mapas físicos de DNA. Através da utilização de metodologia de construção de mapas físicos é possível determinar a posição dos marcadores de SSR nos cromossomos da espécie em estudo, sendo a distância entre os marcadores medida em Centimorgan (cM) ou unidade de mapa (m.u. – do inglês "*map unit*"). Centimorgan é uma unidade de frequência de recombinação utilizada para medir distância genética, sendo utilizada para inferir a distância genética de dois locos em um mesmo cromossomo. O cálculo leva em consideração a taxa de recombinação entre esses locos. Se esses locos forem separados por poucos cMs, por exemplo, significa que ocorre pouca recombinação entre eles, ou seja, baixa taxa de *crossing-over* durante a meiose, e eles estão proximamente distribuídos no genoma. Cabe ressaltar que Centimorgan é uma distância genética e não física, em humanos, por exemplo, 1 cM corresponde em média a cerca de 1 milhão de pares de bases.

Um mapa genético de uma espécie é obtido a partir da análise de uma população, onde é possível avaliar a taxa de recombinação entre os marcadores utilizados. O número necessário de marcadores para construir um mapa genético depende do tamanho e complexidade do genoma em análise, assim como o número de cromossomos e a frequência de recombinação. Um mapa é considerado satisfatório quando se obtém um

número de grupos de ligações semelhante ao número de cromossomos da espécie em estudo e quando os marcadores genéticos estão ligados e próximos, sem grandes espaços entre eles. Quanto maior o número de marcadores mais preciso será o mapa genético, e em consequência disto, estudos de mapeamento costumam ser laboriosos e de alto custo, por ser necessário avaliar um grande número de marcadores.

Mapas genéticos e físicos têm sido muito utilizados em estudos de melhoramento genético de plantas e animais com interesse econômico, sendo denominado mapeamento QTL (*Quantitative Trait Loci*). Mapeamento QTL consiste na detecção, localização e estimativa dos efeitos genéticos de determinada região do genoma sobre o fenótipo das espécies, através de uma análise de correlação, em que se pode estimar se esses efeitos podem ser aditivos, de dominância, epistáticos, entre outros (como descrito no Capítulo 3). Inúmeros programas computacionais são utilizados na construção de mapas de ligação, entre eles: MapDisto, ICIMapping, Carthagene, QTLMap, Cartographer, MapQTL, QGene e R/qtl *package*. Algumas revisões estão disponíveis na literatura, auxiliando na delimitação experimental para obtenção da população em estudo, na determinação do número e tipo de marcadores e análises comparativas entre os principais programas computacionais disponíveis (Young, 1996; Manly e Olson, 1999; Broman, 2001; Collard et al., 2005). De uma forma geral, os microssatélites têm se mostrado muito efetivos na construção dos mapas genéticos em função do seu alto grau de polimorfismo, boa cobertura do genoma, pela sua característica codominante e por apresentarem segregação mendeliana. Trabalhos recentes têm integrado dados massivos de SNPs (*Single Nucleotide Polymorphism* – Capítulo 8) provenientes de sequenciamento de nova geração com dados de microssatélites, na tentativa de obter um mapa mais representativo e auxiliar na identificação dos genes responsáveis pelo fenótipo em estudo.

Análise de parentesco

Para as análises de parentesco é necessário que se utilize um número razoável de locos não ligados altamente polimórficos, caso a espécie em estudo tenha baixa variabilidade é necessário aumentar o número de locos analisados. Para determinar o grau de parentesco também é importante ter o conhecimento das frequências alélicas populacionais dos marcadores utilizados, uma vez que pode haver variações entre grupos populacionais. Análises de parentesco têm sido utilizadas em diversos estudos ecológicos e evolutivos como a seleção sexual, padrões de dispersão e recrutamento de sementes, estimativa de parâmetros genéticos quantitativos e biologia da conservação, estimativas de sistemas de cruzamentos em plantas, assim como no melhoramento genético e determinação de *pedigree*. Os marcadores SSRs têm se mostrado muito efetivos na determinação de parentesco, para isso é necessário selecionar locos altamente polimórficos, ou seja, que apresentem um alto Conteúdo de Informação Polimórfica (PIC). Quanto maior for o PIC mais eficiente será o marcador SSR (PIC $\geq$ 0,5), e um número menor de locos poderão ser analisados. O princípio geral da análise de parentesco é determinar o grau de relacionamento genético (parentesco) entre indivíduos. Muitas abordagens, baseadas nas estimativas de máxima verossimilhança e/ou inferência Bayesiana têm sido desenvolvidas recentemente, como, por exemplo, as implementadas nos programas Cervus, Colony, MLTR, PARENTER, MER, FazMoz, Relatedness, PRDM.

Análises forenses e perfil molecular

Em 7 de março de 1985, em uma edição da revista *Nature*, Alec Jeffreys e colaboradores descreveram pela primeira vez um teste baseado em sondas multilócus

capazes de detectar variabilidade em diversos locos no genoma humano, batizando o método de "*fingerprinting* de DNA" (impressão digital de DNA ou perfil molecular) (Jeffrey et al., 1985). Foram usadas para o desenvolvimento desta técnica, sondas que hibridizavam com regiões conhecidas do genoma por apresentarem grande variabilidade no DNA humano, as repetições em tandem conhecidas como minissatélites ou VNTRs (*Variable Number of Tandem Repeats*). Este foi um importante marco na análise de DNA, a origem dos métodos de identificação humana, com finalidades na genética forense, sendo uma importante ferramenta em investigações criminais e em testes de paternidade (Gill et al., 1985). A partir de 1987 os resultados de análise de DNA foram admitidos como provas em julgamentos de cortes americanas e inglesas. Avanços nas técnicas de *fingerprinting* de DNA deram origem aos métodos que usam regiões SSRs. Análises forenses utilizando dados genéticos são aplicadas mundialmente em humanos desde então, mais recentemente outros organismos têm sido utilizados como recursos de provas de crimes, incluindo a identificação de espécies de animais silvestres que sofrem com caça predatória. Em humanos, o primeiro banco de dados de perfis genéticos de criminosos foi criado na Inglaterra. Atualmente o banco mais importante, pertence ao FBI nos Estados Unidos, sendo denominado de Sistema de índice de DNA combinado (CODIS – *Combined DNA Index System*). Os perfis genéticos humanos são baseados na análise de locos SSRs multiplex, buscando analisar os mesmos locos nos diferentes laboratórios, permitindo a posterior troca de informações entre eles. O CODIS utiliza atualmente 20 locos SSRs altamente polimórficos, os quais são reconhecidos internacionalmente como referência para a determinação do perfil molecular em humanos. O CODIS também possui um robusto programa de análise e pesquisa com o mesmo nome. O Brasil faz uso deste sistema desenvolvido pelo FBI. Caso obtenham-se amostras de DNA de baixa qualidade, alternativamente pode-se optar por sequenciar regiões hipervariáveis do DNA mitocondrial.

Os SSRs representam aproximadamente 3% do genoma humano, sendo que o cromossomo 19 apresenta o maior número deste tipo de sequência. Normalmente causadas pelo escorregamento da DNA polimerase na replicação, mutações nessas regiões podem ocorrer em uma frequência de 1×10^{-4} a 1×10^{-3} mutações por loco por geração em humanos, taxa maior que a de substituição de nucleotídeos, a qual é de 1×10^{-8}. A maioria consiste em repetições de 3 e 6 nucleotídeos compreendendo de 500 a 1000 pb/Mb e repetições de 2, 4 e 6 nucleotídeos compreendendo de 2.000 a 3.000pb/Mb. Tri e hexa-nucleotídeos tendem a ocorrer mais nos éxons, enquanto que as demais variações ocorrem principalmente em regiões não codificantes. Este tipo de variante acaba sendo mais informativo que os SNPs, uma vez que podem apresentar uma ampla gama de variações no número de repetições e os SNPs em geral são bialélicos (Subramanian et al., 2003; Ellegren, 2004).

O genoma humano é diploide, sendo que herdamos uma cópia do material cromossômico do pai e uma cópia da mãe. A análise de microssatélites consiste em identificar o número de repetições das regiões alvo nos dois alelos de um indivíduo. Os testes de paternidade consistem em identificar o número de repetições de cada um dos possíveis genitores, e do filho, e cada uma das cópias dos locos analisados no filho deve ter o mesmo número de repetições em um dos genitores. A análise do trio sempre garantirá maior fidelidade ao resultado.

Testes de identificação humana podem ser obtidos na forma de *kits* comerciais ou amplificação por PCR das regiões escolhidas seguida de eletroforese para estimativa do número de repetições. Em um exemplo de *kit* comercial são amplificados 15 locos SSRs, incluindo 13 locos do CODIS, mais o gene da amelogenina, para determinação do gênero do indivíduo investigado, em uma única reação. O que permite a

amplificação de todos os alvos em uma única reação é a marcação de *primers* com diferentes fluoróforos para os determinados locos. Para o *kit* em questão, os locos e fluoróforos utilizados são: Penta E, D18S51, D21S11, TH01 e D3S1358 com um dos primers marcado com fluoresceina (FL); FGA, TPOX, D8S1179, vWA e Amelogenina com um dos *primers* marcados com carboxitetrametilrodamina (TMR); e Penta D, CSF1PO, D16S539, D7S820, D13S317 e D5S818 com um dos primers marcado com 6-carboxy-4´,5´-dichloro-2´,7´-dimethoxy-fluorescein (JOE). Outros *kits* disponíveis usam os mesmos alvos, mas com fluoróforos diferentes. As amostras amplificadas através do *kit* são migradas em eletroforese capilar, em sequenciador automatizado, juntamente com um padrão de peso molecular (ISS) e sendo então interpretadas por *softwares* como, por exemplo, o GeneMapper. Os softwares de análise são capazes de interpretar o número de repetições de cada alelo baseado nos picos/bandas apresentados pelo padrão de peso molecular.

Cada loco será utilizado para calcular o índice de paternidade (IP) através da fórmula IP=X/Y, onde X é a probabilidade do possível pai transmitir o alelo à criança e Y a probabilidade de um indivíduo ao acaso na mesma população ter transmitido. É então calculado o índice de paternidade combinado (IPC), que é o produto dos IPs de cada SSR. O cálculo de probabilidade de paternidade (W), é calculado com a fórmula W=IPC/(IPC+1)x100, para um resultado em porcentagem. A interpretação do resultado de W pode variar, onde a paternidade é confirmada em W≥99,8%, improvável quando W<10,0% (Pena et al., 1994).

Este tipo de teste de *fingerprinting* também é utilizado para identificação de diferentes tipos celulares, por exemplo, na distinção de células em cultura. Muitas das culturas celulares utilizadas em pesquisa são derivadas de diferentes tipos de tumores e cânceres, e vale ressaltar que a incidência de instabilidade genética em SSRs nessas células não é incomum, diferindo geneticamente de células normais. Além disso, células em cultura por si só são capazes de acumular mutações. Outra particularidade da análise de células é um desbalanço nos picos analisados, que pode decorrer de duplicação de genes, mutações somáticas, trissomias, aneuploidias, ou populações de células quiméricas.

A técnica de perfil molecular também é aplicada em estudos agronômicos. Os microssatélites têm sido utilizados na identificação de cultivares (proteção varietal). Esses dados são agregados em processos de patentes de novas cultivares junto a União Internacional para a Proteção das Obtenções Vegetais (UPOV). Comumente são observadas as seguintes análises para determinar o perfil molecular: heterozigosidade esperada (H_E); PIC, probabilidade de identidade genética (*I*), a qual corresponde à probabilidade de dois indivíduos aleatórios apresentarem o mesmo genótipo e probabilidade de exclusão de paternidade (*Q*), as quais podem ser calculadas em programas como GIMLET, GENECAP, CERVUS, IDENTITY entre outros.

Aplicações

Os marcadores microssatélites são altamente polimórficos devido à variação no número de repetições. Este tipo de marcador molecular pode ser facilmente detectado por PCR com o uso de duas regiões que flanqueiam os locos de SSR. Os locos de microssatélites têm sido muito utilizados devido a sua herança codominante, ampla cobertura e relativa abundância no genoma, alta reprodutibilidade, além do alto conteúdo de informação (multialélico) e facilidade de genotipagem. A capacidade de distinção entre indivíduos proximamente relacionados é também uma característica que faz deste tipo de marcador um dos mais utilizados em várias áreas da biologia

molecular. As principais aplicações dos microssatélites são: 1) genética da conservação, por exemplo, a avaliação da estrutura e diversidade genética de populações naturais, bancos germoplasma *in situ* e *ex situ*, e populações de cativeiro; ecologia molecular e evolução; 2) aplicações forenses, como a identificação de indivíduos e teste de paternidade; 3) mapeamento genético, epidemiologia molecular e patologia.

Os microssatélites têm sido amplamente utilizados na área da conservação biológica devido a algumas características que fazem destes marcadores especialmente apropriados para estes estudos. Por exemplo, os microssatélites são facilmente obtidos em espécies não modelo, seja através do isolamento de locos espécie-específicos, ou pela transferência de marcadores isolados para espécies próximas evolutivamente (Barbara et al., 2007). Além disso, pelo fato dos microssatélites serem amplificados por PCR eles podem utilizar materiais amostrados de forma não invasiva (por exemplo, pêlos, fezes, saliva, etc), ou seja, sem o contato direto com o animal em estudo.

Na área de genética da conservação e evolução os SSRs têm sido utilizados principalmente em estudo de hibridação (fluxo gênico interespecífico), delimitação de espécies, estrutura e diversidade genética populacional, fluxo gênico histórico e contemporâneo, filogeografia e descrição de diversidade em espécies rara e endêmicas. Estudos com estes marcadores permitem acessar informações quanto a eventos demográficos populacionais, comportamento reprodutivo, estrutura social, especialização ecológica, modo de dispersão e capacidade de colonização de populações naturais e sistema reprodutivo em plantas (autógamo ou alógamo).

Em estudos de hibridação de populações naturais, a utilização dos microssatélites tem permitido a detecção dos níveis de fluxo gênico entre espécies puras (introgressão), demonstrando que em muitos casos o isolamento reprodutivo entre as espécies pode não ser completo para que o processo de especiação ocorra (Wu et al., 2001). Em um estudo com duas espécies de bromélias que ocorrem em simpatria em afloramentos rochosos da cidade do Rio de Janeiro, a partir de resultados obtidos com marcadores SSRs, foi observado que estas espécies produziam híbridos férteis e os eventos de retrocruzamento entre os híbridos férteis e as espécies parentais promovia o fluxo de genes entre as espécies. Apesar da ocorrência de hibridação e introgressão, a integridade genética e morfológica de cada uma das espécies ainda assim era mantida (Palma-Silva et al., 2011). Outro estudo utilizou marcadores SSRs com foco em hibridação em onça parda: *Leopardus geoffroyi* e *L. tigrinus* mostrando que as espécies divergiram a cerca de 1 milhão de anos atrás. Estas duas espécies ocorrem em alopatria em quase toda a América do Sul e possuem uma estreita zona de contato no sul do Brasil, onde foram identificados, através dos microssatélites, indivíduos híbridos com sinais de introgressão entre as espécies. A estrutura genética das populações brasileiras de *L. tigrinus*, parece ter sido afetada pelo processo de introgressão, mostrando um gradiente de diferenciação de *L. geoffroyi* correlacionado com a distância da zona de contato (Trigo et al., 2008).

A delimitação de espécies tem se tornado um tópico bastante importante na conservação e evolução da biodiversidade. Grande ênfase na delimitação de espécies tem surgido, principalmente pela crescente preocupação da conservação da biodiversidade, e assim poder descrever o maior número de espécies possíveis, o mais rápido e preciso possível, antes que elas sejam extintas (Wiens, 2007). O uso de marcadores moleculares SSR para análises de delimitação de espécies tem sido bastante explorado recentemente. Por exemplo, a delimitação de espécies dentro do Clado Atlântico do gênero de orquídeas *Epidendrum* (Subg. Amplhyglottium) foi investigado através do uso de SSR e análises morfométricas. Os autores puderam delimitar claramente quatro espécies já reconhecidas para o Clado Atlântico e identificaram uma

nova espécie ainda não descrita (Pessoa et al., 2012). Marcadores SSR também foram utilizados para descrição da diversidade genética em uma espécie de planta rara e restrita a um específico habitat no Bioma Pampa. Neste estudo os autores identificaram que a raridade da espécie está associada a sua história evolutiva, bem como foi identificado dois *pools* genéticos associados a ambientes diferentes (Turchetto et al., 2016).

Em estudos filogeográficos, os microssatélites têm sido utilizados, tanto em animais como em plantas, combinados a outros marcadores mitocondriais ou plastidiais (Turchetto-Zolet et al., 2013). Em um estudo com duas espécies de sapos endêmicos da Caatinga, após as análises exploratórias de estrutura genética dentro e entre as espécies, os dados de microssatélites foram utilizados para testar hipóteses biogeográficas. Este trabalho demonstrou que as quebras filogeográficas foram geograficamente coincidentes entre as espécies e que expansões da floresta Amazônica e Atlântica sobre o que hoje é o Bioma Caatinga moldaram a estrutura genética destas espécies de sapos através de eventos de vicariância (Thomé et al., 2016).

Os microssatélites são os marcadores mais utilizados em análises de sucesso reprodutivo, estrutura social e parentesco. Nos estudos de evolução do comportamento social os microssatélites permitem inferências nunca antes acessadas. De particular importância para a conservação da biodiversidade é conhecer como as populações são estruturadas socialmente, e como o endocruzamento pode afetar o potencial evolutivo das populações. Neste contexto, vários estudos têm conectado a estrutura social e genética de populações naturais ameaçadas. Por exemplo, em um estudo realizado na Islândia, foi investigado o fenômeno no qual as fêmeas de patos selvagens aumentam o seu sucesso reprodutivo colocando ovos nos ninhos de outras fêmeas da mesma espécie, que chocam esses ovos parasitas em seus próprios ninhos. Nas análises de parentesco baseado em sete locos de microssatélites o estudo demonstrou que as fêmeas parasitadas eram geneticamente próximas das fêmeas parasitas. Os resultados também indicaram que as fêmeas parasitadas eram mais velhas do que as fêmeas parasitas, que a porcentagem de ninhos com ovos parasitas aumentava com a idade das fêmeas e que o número de ovos da própria fêmea no ninho diminuía com a idade. Assim, as fêmeas mais velhas, que tem menos capacidade de colocar muitos ovos teriam um ganho evolutivo indireto ao chocar os ovos de suas parentas mais novas. Este exemplo constitui a cooperação entre gerações de fêmeas proximamente relacionadas nesta espécie de patos selvagens (Tiedemann et al., 2011).

Os microssatélites também são muito úteis para avaliar interações ecológicas através dos níveis de fluxo gênico contemporâneo e a variação no sistema reprodutivo de populações naturais. Por exemplo, estudo com progênies de uma espécie de bromélia, *Vriesea gigantea*, de ampla distribuição no sul e sudeste brasileiro observou-se que a espécie é autocompatível, porém apresenta uma variação nas taxas de fecundação cruzada e autofecundação entre as populações estudadas, estando a variação relacionada a diferentes interações ecológicas desta espécie ao longo de sua distribuição geográfica. Mais ao sul da distribuição, onde as plantas eram polinizadas principalmente por beija-flores as taxas de autofecundação eram mais altas do que ao norte da distribuição, onde as populações eram polinizadas também por morcegos. As análises com microssatélites também demonstraram uma limitada dispersão dos grãos de pólen, que dispersam somente cerca de 170 m^2 ao redor da planta mãe, o que faz com que as populações tenham uma alta estruturação genética e variados níveis de endogamia (Paggi et al., 2015).

Os microssatélites são muito usados em identificação forense de indivíduos ou partes deles e testes de paternidade em humanos, mas também em animais e plantas

com o objetivo de conservação. Como comentado anteriormente, várias características fazem dos microssatélites marcadores excelentes para aplicações forenses: como a facilidade e precisão de genotipagem, alto polimorfismo que permite que cada indivíduo tenha um próprio perfil genético e a necessidade de uma quantidade muito pequena de amostra (DNA) para ser analisada, além de sua alta reprodutibilidade, permitindo interações de resultados entre diferentes laboratórios de análise. Em humanos o objetivo é geralmente, conectar um suspeito a uma amostra (sangue, saliva, esperma, etc) retirado da cena de um crime e que supostamente seja do culpado. Os testes de paternidade são também utilizados para estabelecer a paternidade (ou maternidade) entre indivíduos, em casos: de estupro ou incesto, de desaparecimento de pessoas, de reunião de familiares, ou de requerimento de direitos legais (Balding, 1999).

Regiões SSRs foram extremamente importantes na construção de mapas de ligação onde genes envolvidos em importantes doenças genéticas humanas, no qual foi possível identificar gene como *CARD15*/*NOD2*, relacionado a doença de Crohn, *PTPN22* na diabetes tipo 1, *TCF7L2* na diabetes tipo 2 e *STAT4* na artrite reumatoide e no lúpus eritematoso sistêmico (Ku et al., 2010).

SSRs podem ser importantes marcadores para o diagnóstico de doenças. Entre as doenças que podem ser rastreadas pela análise de SSRs está uma síndrome causadora de câncer, a Síndrome de Lynch, também conhecida como câncer colorretal hereditário não poliposo (HNPCC). Esta é uma síndrome de câncer hereditário associada ao câncer colorretal e outros órgãos pélvicos e abdominais. A condição é causada por alterações em cinco genes envolvidos no reparo por mal pareamento (via MMR – *DNA mismach repair*) durante a replicação do DNA, entretanto, ao invés de investigar os cinco genes para o diagnóstico molecular, que levaria tempo e seria oneroso, um primeiro passo de rastreamento de instabilidade de microssatélites (MSI) é usualmente realizado. O aumento da instabilidade de microssatélites está associado ao maior risco do paciente possuir a síndrome, uma vez que a instabilidade é encontrada no tecido tumoral de 90% dos pacientes com HNPCC, e em 15% dos pacientes que têm câncer colorretal esporádico, para então se partir para a pesquisa dos genes envolvidos nos casos rastreados (Parsons et al., 1995).

Diferentes doenças humanas envolvem diretamente a expansão de regiões SSRs. Entre as mais conhecidas estão a doença de Huntington, Síndrome do X-Frágil, as ataxias espinocerebelar e distrofia miotônica. A maior parte das doenças de expansão de nucleotídeos é de herança autossômica dominante. Quanto maior for a expansão, maior a chance de ocorrer aumento das expansões nas próximas gerações, e o tamanho da expansão, também conhecida como mutação dinâmica, apresenta uma correlação positiva com a gravidade dos sintomas e idade de início, fenômeno conhecido como antecipação. Neste contexto, é introduzido o conceito de pré-mutação, que é a observação do aumento do número de repetições, mostrando instabilidade do alelo, mas em um número onde não há presença de sintomas, sendo uma situação de atenção para a expansão nas próximas gerações levando à apresentação de sintomas.

A doença de Huntington e a Síndrome do X-Frágil estão entre as doenças mais frequentes de expansão de regiões microssatélites. A doença de Huntington tem uma frequência de 1/20.000 caucasianos, mas é rara entre orientais e negros. A expansão de repetições do trinucleotídeo CAG no gene *IT15*, codificador da huntingtina é o que causa a doença. Pessoas não afetadas apresentam de 11 a 35 repetições, e pessoas com a doença apresentam de 36 a 100 repetições do trinucleotídeo. O acúmulo de poliglutamina forma agregados tóxicos especialmente para o sistema nervoso central, acarretando em morte neuronal precoce. Entre os principais sintomas estão a perda substancial de neurônios, perda progressiva do controle motor (coreia), distúrbios

psiquiátricos (demência e distúrbios afetivos), com curso progressivo, sendo em torno de 15 anos o tempo entre o diagnóstico e a morte do paciente. A morte normalmente decorre de dificuldade de deglutição seguida de pneumonia de aspiração, hematoma subdural e suicídio (Walker, 2007).

A Síndrome do X-frágil tem uma frequência de 1:1.250 entre homens e 1:2.500 entre mulheres, sendo a causa mais comum de retardo mental isolado herdado, representando cerca de 40% dos casos. O gene afetado é o *FMR1*, localizado na posição Xq28 do cromossomo X, e a expansão de nucleotídeos no promotor do gene acarreta na sua metilação anormal e ausência do produto gênico. Na situação normal, o promotor do gene contém de 6 a 50 cópias da repetição CGG, no estado de pré-mutação contém de 50 a 230 repetições, e os indivíduos afetados apresentam de 230 a 1.000 cópias da repetição. A proteína gerada por esse gene em sua função normal se liga a RNA, sendo envolvido no tráfego de mRNA do núcleo para o citoplasma e é associado a polissomos. A doença tem penetrância de 80% entre homens e 30% entre mulheres, apresentando como sintomas clínicos entre os homens a face alongada, orelhas e mandíbula proeminente, hipermobilidade articular, macro-orquidismo na puberdade e retardo mental, e nas mulheres um fenótipo parcial que inclui menor grau de comprometimento cognitivo e neurocomportamental (Mandel et al., 2004).

Em animais e plantas, os testes de identificação através de métodos forenses moleculares estão sendo cada vez mais utilizados para auxiliar a criação e o cumprimento de leis de conservação e assim processar caçadores e coletores ilegais, na tentativa de minimizar a exploração ilegal da biodiversidade. Por exemplo, a origem de um indivíduo pode ser identificada como sendo de uma população protegida ou de um criador registrado e certificado. Em um estudo baseados em microssatélites e estrutura genética de uma espécie de veado europeu (*Cervus elaphus*), foi possível demonstrar a translocação ilegal de pelo menos quatro animais que tinham sido ilegalmente introduzidos a grupo de veados de uma área de caça na região de Luxemburgo (Frantz et al., 2006).

O mapeamento genético representa outra importante área da pesquisa onde os marcadores microssatélites têm sido aplicados. O mapeamento genético é feito através de análise de ligação entre um marcador e o loco de interesse. Esta análise é feita com base nos padrões de segregação entre os locos através das gerações. Os estudos de associação baseados em desequilíbrio de ligação (DL) têm sido utilizados como estratégia de mapeamento, para a investigação da associação entre o loco de certa característica ou doença e o loco do marcador molecular (Schlötterer, 2004). Os microssatélites têm sido muito aplicados na construção de mapas de ligação, mapeamento de características quantitativas de interesse e identificação de genes e regiões genômicas responsáveis por características fenotípicas (Kalia et al., 2011). Os estudos de mapeamento genético requerem um grande número de marcadores distribuídos ao longo de todo o genoma, o que faz dos microssatélites bons marcadores moleculares para essa aplicação.

O enriquecimento dos mapas genéticos com marcadores seletivamente neutros beneficia o mapeamento e a caracterização de genes responsáveis por importantes características médicas, agrícolas e evolutivas. Isso também disponibiliza a oportunidade da seleção assistida por marcadores em espécies de importância comercial. Desta forma, os lócus de microssatélites que ocorrem em regiões não codificantes do genoma são muito úteis para a construção de uma primeira estrutura de mapas genéticos de alta resolução, enquanto que os lócus de microssatélites que ocorrem em regiões codificadoras podem ser incorporados a esses mapas e assim

permitir a identificação de características fenotípicas de interesse, uma vez que estes marcadores podem mostrar a localização dos genes de interesse no mapa de ligação.

Apesar dos microssatélites serem considerados seletivamente neutros, eles podem também representar polimorfismos funcionais relevantes, uma vez que podem estar localizados em regiões codificantes e expressas do genoma. Os SSRs localizados em regiões expressas do genoma são conhecidos como SSR - EST (*Expressed sequence tag*). Estes marcadores são uteis porque representam genes transcritos e sua função putativa pode facilmente ser deduzida por buscas de homologia.

Considerações Finais

Os microssatélites são uma das classes mais polimórfica de marcadores moleculares presentes nos genomas. A utilização de SSRs tornou-se uma importante ferramenta de análise genética a partir do momento em que sua variabilidade individual foi caracterizada entre as mais diferentes espécies, permitindo que desde microrganismos até plantas, animais e seres humanos sejam diferenciados por métodos de biologia molecular. Além do mais, SSRs apresentam ampla aplicabilidade para responder diversas questões biológicas. Nos últimos anos foram realizados consideráveis progressos no desenvolvimento da metodologia de identificação e genotipagem de marcadores SSRs, especialmente em combinação com a publicação dos genomas completos e com os avanços na área de bioinformática. O uso de NGS para obter dados de sequência e identificar SSRs em espécies não modelo é extremamente promissora e tem permitido uma diminuição considerável no tempo e esforço investidos neste tipo de análise. Mapas genéticos puderam ser construídos de forma mais acurada na integração de resultados obtidos por outros métodos. O conhecimento das regiões flanqueadoras dessas repetições facilitaram o desenvolvimento dos métodos para amplificação das regiões que envolvem os SSRs. A utilização de marcadores SSR também foi um importante marco nas análises genéticas envolvendo seres humanos, permitindo a solução de crimes, provando paternidades e diagnosticando doenças.

As altas taxas de mutação desses marcadores permitem que avaliação de risco ambiental seja estimada, como discutido ao longo do capítulo, por exemplo, na identificação de animais em situação de risco, exposição de organismos a agentes tóxicos, e perda de variabilidade genética entre espécies, em tempos onde a preservação do meio ambiente se mostra tão importante.

Análises de SSRs estão presentes em diferentes áreas da genética molecular entre os mais diferentes organismos, e as perspectivas são de que a variação altamente polimórfica desses sítios continuará por muito tempo sendo utilizadas como uma importante ferramenta investigativa.

Referências Bibliográficas

Amos W, Hoffman JI, Frodsham A, et al. (2007) Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes* 7: 10–14.

Armour JA, Neumann R, Gobert S, Jeffreys AJ (1994) Isolation of human simple repeat loci by hybridization selection. *Human Molecular Genetics* 3: 599—565.

van Asch B, Pinheiro R, Pereira R, et al. (2010) A framework for the development of STR genotyping in domestic animal species: characterization and population study of 12 canine X-chromosome loci. *Electrophoresis* 31: 303–308.

Blanca J, Canizares J, Roig G, Ziarsolo P, Nuze F, Pico B (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (Cucurbitaceae). *BMC Genomics* 12: e104 – 118.

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays* 28: 1040–1050.

Butler JM, Ruitberg CM, Vallone PM (2001) Capillary electrophoresis as a tool for optimization of multiplex PCR reactions. *Fresenius Journal of Analytical Chemistry* 369: 200–205.

Butler JM, Buel E, Crivellente F, McCord BR (2004) Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis* 25 1397–1412.

Butler JM (2005) Constructing STR multiplex assays. *Methods in Molecular Biology* 297: 53–65.

Cavagnaro PF, Senalik DA, Yang L, et al. (2010) Genome-wide characterization of simple sequence repeats in cucumber (Cucumis sativus L.). *BMC Genomics* 11: 569 – 586.

Chamberlain JS, Gibbs RA, Rainer JE, Nguyen PN, Thomas C (1988) Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Research* 16: 11141–11156.

Chambers GK, MacAvoy ES (2000) Microsatellites: consensus and controversy. Comparative Biochemistry and Physiology—Part B: *Biochemistry and Molecular Biology* 126: 455–476.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* 24: 621-631.

Chistiakov DA, Hellemans B, Volckaert F A (2006). Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. *Aquaculture* 255(1): 1-29.

Cifarelli RA, Gallitelli M, Cellini F (1995) Random amplified hybridization microsatellites (RAHM): isolation of a new class of microsatellite-containing DNA clones. *Nucleic Acids Research* 23: 3802—3803.

Collins HE, Li H, Inda SE, et al. (2000) A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Human Genetics* 106: 218–226.

Cryer N, Butler D, Wilkinson M (2005) High throughput, high resolution selection of polymorphic microsatellite loci for multiplex analysis. *Plant Methods* 1: 3.

Dieringer D, Schlötterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Research* 13: 2242–2251.

Edwards A, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am J Hum Genet* 49:746–756.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5: 435–445.

Estoup A, Wilson IJ, Sullivan C, Cornuet J-M, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. *Genetics* 159: 1671–1687.

Frantz AC, Pourtois JT, Heuertz M, et al. (2006). Genetic structure and assignment tests demonstrate illegal translocation of red deer (Cervus elaphus) into a continuous population. *Molecular Ecology* 15(11): 3191-3203.

Gill P, Jeffreys A J, Werrett DJ (1985) Forensic application of DNA 'fingerprints'. *Nature* 318(6046): 577-579.

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST–SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol Genet Genomics* 270: 315–323.

Gusmão L, Butler JM, Carracedo A, et al. (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *International Journal of Legal Medicine* 120: 191–200.

Hamilton MB, Pincus EL, Di-Fiore A, Fleischer RC (1999) Universal linker and ligation procedures for construction of genomic DNA libraries enriched for microsatellites. *Biotechniques* 27: 500—507.

Hill CR, Butler JM, Vallone PM (2009) A 26plex autosomal STR assay to aid human identity testing. *Journal of Forensic Sciences* 54: 1008–1015.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* 14: 599–612.

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques* 46: 511–517.

Jacob HJ, Lindpaintnesr K, Kusumir EL, et al. (1991) Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. *Cell* 67: 213–224.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* 11: 424–429.

Jayashree B, Reddy PT, Leeladevi Y, et al. (2006) Laboratory Information Management Software for genotyping workflows: applications in high throughput crop genotyping. *BMC Bioinformatics* 7: 383.

Jeffreys A J, Wilson V, Thein S L (1985) Hypervariable'minisatellite' regions in human DNA. *Nature* 314(6006): 67-73.

Karagyozov L, Kalcheva ID, Chapman VM (1993) Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. *Nucleic Acids Research* 21: 3911—3912.

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011). Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177(3): 309-334.

Kaplinski L, Andreson R, Puurand T, Remm M (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics* 21: 1701–1702.

Kelkar YD, Strubczewski N, Hile SE, et al. (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A / T and GT / AC repeats. *Genome Biology and Evolution* 2: 620–635.

Kijas JM, Fowler JC, Garbett CA, Thomas MR (1994) Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *Biotechniques* 16: 656—662.

Kim TS, Booth J, Gauch H, et al. (2008) Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference. *BMC Genomics* 9: 31.

Kolpakov R, Bana G, Kucherov G (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* 31: 3672 – 3678.

Kraemer L, Beszteri B, Gabler-Schwarz S, et al. (2009) STAMP: extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. *BMC Bioinformatics* 10: 41.

Ku C S, Loy E Y, Salim A, Pawitan Y, Chia K S (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of human genetics* 55(7): 403-415.

Kumar M, Choi JY, Kumari N, Pareek A, Kim SR (2015) Molecular breeding in Brassica for salt tolerance: importance of microsatellite (SSR) markers for molecular breeding in Brassica. *Frontier in Plant Science* 6: 688.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203–221.

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 44: 397–401.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* 11: 2453–2465.

Lunt DH, Hutchinson WF, Carvalho GR (1999) An efficient method for PCR-based identification of microsatellite arrays (PIMA). *Molecular Ecology* 8: 893—894.

Maia LC da, Palmieri DA, DE Souza VQ, et al. (2008) SSR locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* 41: 2696.
Mandel JL, Biancalana V (2004) Fragile X mental retardation syndrome: from pathogenesis to diagnostic issues. *Growth hormone & IGF research* 14: 158-165.
Matschiner M, Salzburger W (2009) TANDEM: integrating automated allele binning into genetics and genomics workflows. *Bioinformatics* 25: 1982–1983.
Martins WS, Lucas DC, Neves KF, Bertioli DJ (2009) WebSat — A web software for microsatellite marker development. *Bioinformation* 3: 282 – 283.
Meglécz E, Costedoat C, Dubut V, et al. (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* 26: 403–404.
Meldgaard M, Morling N (1997) Detection and quantitative characterization of artificial extra peaks following polymerase chain reaction amplification of 14 short tandem repeat systems used in forensic investigations. *Electrophoresis* 18: 1928–1935.
Merritt BJ, Culley TM, Avanesyan A, Stokes R, Brzyski J (2015) An empirical review: characteristics of plant microsatellite markers that confer higher levels of genetic variation. *Applications in Plant Sciences* 3(8): 1500025.
Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology* 29: 294–307.
Ostrander EA, Jong PM, Rine J, Duyk G (1992) Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. *Proceedings of the National Academy of Sciences of the USA* 89: 3419—3423.
Paetkau D (1999) Microsatellites obtained using strand extension: An enrichment protocol. *Biotechniques* 26: 690—697.
Parsons RE, Jen J, Papadopolous N, Peltomäki P (1995) Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nature genetics* 9.
Pena SD, Chakraborty R (1994). Paternity testing in the DNA era. *Trends in Genetics* 10(6): 204-209.
Pickett BD, Karlinsey SM, Penrod CE, et al. (2016) SA-SSR: A Suffix Array-Based Algorithm for Exhaustive and Efficient SSR Discovery in Large Genetic Sequences. *Bioinformatics Advance Access*.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
Rachlin J, Ding C, Cantor C, Kasif S (2005) MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Research* 33: 544–547.
Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucleic acids research* 20: 211-215.
Shen Z, Qu W, Wang W, et al. (2010) MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics* 11: 143.
Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18: 233–234.
Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecology Letters* 9: 615 – 629.
Subirana JA, Messeguer X (2008) Structural families of genomic microsatellites. *Gene* 408: 124–132.
Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biology* 4(2): R13.
Sun X, Liu Y, Lutterbaugh J, et al. (2006) Detection of mononucleotide repeat sequence alterations in a large background of normal DNA for screening high-frequency microsatellite instability cancers. *Clinical Cancer Research* 12: 454–459.
Sun JX, Mullikin JC, Patterson N, Reich DE (2009) Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution* 26: 1017–1027.

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acid Res* 17: 6463–6471.

Tautz D, Schlötterer C (1994) Simple sequences. *Current Opinion in Genetics & Development* 4: 832–837.

Tiedemann R, Paulus KB, Havenstein K, et al. (2011). Alien eggs in duck nests: brood parasitism or a help from Grandma? *Molecular Ecology* 20(15): 3237-3250.

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L). *Theoretical and Applied Genetics* 106: 411 – 422.

Turchetto-Zolet AC, Pinheiro F, Salgueiro F, Palma-Silva C. 2013. Phylogeographical patterns shed light on evolutionary process in South America. *Molecular Ecology* 22: 1193-1213.

Turchetto C, Segatto ALA, Mäder G, Rodrigues DM, Bonatto SL, Freitas LB (2016) High levels of genetic diversity and population structure in a endemic and rare species: implication for conservation. *AoB Plants* 8: plw002.

Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques* 37: 226–231.

Varshney RK, Thudi M, Aggarwal R, Börner A (2007) *Genic molecular markers in plants: development and applications*. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement: genomics approaches and platforms, vol 1. Springer, Dordrecht, pp 13–29.

Victoria FV, Maia LC, Oliveira AC (2011) In silico comparative analysis of SSR markers in plants. *BMC Plant Biology* 11: 15.

Xia EH, Yao QY, Zhang HB, Jiang JJ, Zhang LP, Gao LZ (2016) CandiSSR: An Efficient Pipeline used for Identifying Candidate Polymorphic SSRs Based on Multiple Assembled Sequences. *Frontiers in Plant Sciences* 6: 1171.

Walker FO (2007). Huntington's disease. *The Lancet* 369(9557): 218-228.

Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE (2014) A Review of the prevalence, utility, and caveats of using chloroplast Simple Sequence Repeats for studies of plant biology. *Applications in Plant Sciences* 2 (12): 1400059.

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* 18: 6531-6535.

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

Zalapa JE, Cuevas H, Zhu H, et al. (2012) Using next generation sequencing approaches to isolate Simple Sequence Repeats (SSR) loci in the plant sciences. *American Journal of Botany* 99(2): 193–208.

Zane L, Bargelloni L, Artanello TP (2002) Strategies for microsatellite isolation: A review. *Molecular Ecology* 11: 1-16.

Capítulo 7

DArT: marcadores baseados em Diversity Arrays Technology

Dra. Franceli Rodrigues Kulcheski

Considerações gerais

Diversidade genética é geralmente definida como a quantidade da variabilidade genotípica (em nível de DNA) presente em um grupo de indivíduos. Esta diversidade é extremamente importante, pois fornece às espécies a habilidade de se adaptar às mudanças ambientais, como, adaptação a novas condições climáticas (por exemplo, plantas que sobrevivem em regiões secas) ou adaptação a novas doenças ou predadores (casos de organismos que resistem a doenças que antes seriam letais). Um considerável número de métodos para análise da diversidade genética entre diferentes germoplasmas tem sido desenvolvido nos últimos anos. Muitos destes métodos são baseados na geração de marcadores moleculares (Avise, 2004). Um marcador molecular representa a variação de um sítio em particular do genoma, o qual é herdado de uma maneira Mendeliana, fácil de ser testado e pode ser mantido através das gerações. Adicionalmente, eles podem ser ordenados nos cromossomos e neste contexto, o genoma inteiro de um indivíduo pode ser estimado via mapas genéticos.

O desenvolvimento de marcadores para a detecção de diversidade genética entre distintos organismos tem sido um dos recursos mais importantes na área da genética molecular e biotecnologia. Atualmente já foram descritos vários tipos de marcadores, os quais podem ser originados de processos de restrição enzimática, ou de amplificações pela reação em cadeia da Polimerase (PCR, de *Polimerase Chain Reaction*) e ainda aqueles detectados via sequenciamento. Nos últimos anos, algumas técnicas foram desenvolvidas abordando ensaios baseados em hibridização, dentre as quais se podem destacar a *Diversity Arrays Technology* (DArT). Esta tecnologia foi originalmente desenvolvida no *Centre for the Application of Molecular Biology to International Agriculture* (CAMBIA) na Austrália, e em 2001 estabelecida na empresa que leva o nome da mesma (Diversity Arrays Technology Pty. Ltd.), com o objetivo de resolver limitações de outras técnicas existentes até aquele momento (Jaccoud et al., 2001). Inicialmente desenvolvidos em arroz e posteriormente validados em outras culturas, os marcadores DArT são baseados na hibridização de microarranjos a fim de detectar polimorfismos (presença ou ausência de determinada sequência nucleotídica) ao longo de um genoma. Os inúmeros trabalhos realizados até o momento demonstraram a eficiência desta técnica na obtenção de centenas de *loci* polimórficos espalhados pelo genoma. Desta forma os marcadores DArT têm apresentado um grande potencial nos estudos de diversidade genética e de mapeamento, demonstrando alta informação de polimorfismos e revelando relações genéticas consistentes entre amostras (Wenzl et al., 2004; 2006).

Na tecnologia DArT, amostras de DNA a serem investigadas são transformadas em uma "representação genômica" através do uso de enzimas de restrição. Inicialmente o DNA de interesse é digerido e aos fragmentos gerados são adicionados adaptadores

para os quais existem *primers* complementares. A amplificação exclusiva dos fragmentos gerados pela digestão ocasiona uma redução da complexidade do genoma. Os fragmentos amplificados são clonados, novamente amplificados, purificados e arranjados em um suporte sólido (microarranjo), resultando em um arranjo de descoberta ou do inglês "*discovery array*". Estes arranjos de descoberta servem como plataformas para posteriores hibridizações com outros genomas individuais a fim de identificar polimorfismos. Assim, os marcadores DArT detectam polimorfismos através da intensidade de sinal proveniente de hibridizações variáveis entre diferentes genótipos.

Devido às características da tecnologia DArT, atualmente este sistema tem sido amplamente utilizado em espécies com genomas poliplóides altamente complexos (característica da maioria das plantas cultivadas), bem como em organismos com pouca informação do genoma, resultando em uma ferramenta de amplo espectro de aplicação (Grzebelus, 2015). Nos tópicos seguintes serão discutidas as características, desenvolvimento, vantagens, limitações e aplicações dos marcadores DArT.

Características da tecnologia DArT

Princípios básicos

O princípio da DArT é baseado na investigação de polimorfismos no DNA genômico para a presença ou ausência de fragmentos individuais (Figura 7.1). Este DNA genômico pode ser composto tanto por um grupo de indivíduos que representem um germoplasma de interesse, como pode corresponder a um "*pool*" ou conjunto de genomas de espécies cultivadas, estar restrito aos genomas parentais de um cruzamento (se o propósito for uma rápida criação de um mapa de ligação) ou ainda expandido a conjuntos de genes secundários ou terciários (Kilian et al., 2012). Todos os arranjos de DArT são desenvolvidos com o principal foco na captura da diversidade alélica do organismo de interesse a fim de limitar o viés de certificação ou averiguação (viés introduzido quando se utiliza marcadores desenvolvidos a partir de um grupo amostral pequeno, e consequentemente, restrito quanto ao número de genótipos). Estudos filogenéticos em populações de Eucaliptos demonstraram que os marcadores DArT estavam livres de viés de averiguação (Sansaloni et al., 2010; Steane et al., 2011).

A coleção de DNA representando o *pool* gênico de interesse é processada utilizando um método de redução da complexidade, um processo que seleciona reprodutivelmente (isto é, livre de efeitos aleatórios) uma fração definida de fragmentos genômicos. O alto nível de precisão que caracteriza este processo é mediado pela ação das enzimas de restrição específicas. Assim a coleção resultante desta fragmentação, chamada de representação, é então utilizada para construir uma biblioteca de clones em *Escherichia coli*. Este processo de clonagem em bactérias garante uma individualização dos fragmentos para a representação (Kilian et al., 2012). Os insertos clonados serão amplificados e utilizados como sondas moleculares nos arranjos de DArT. É importante salientar, que o processo de individualização dos fragmentos é um dos pontos que difere a tecnologia DArT de outros métodos como o polimorfismo baseado no comprimento de fragmentos amplificados (AFLP, de *Amplified Fragment Lenght Polymorphism*), o qual permite a caracterização dos marcadores apenas via tamanho dos fragmentos.

A tecnologia DArT não requer altas quantidades de DNA genômico, geralmente 100ng de DNA são suficientes para genotipar simultaneamente mais de 7000 *loci* em uma única reação. Estes marcadores são estritamente bi-alélicos e são classificados em variantes de presença versus ausência, onde o estado de presença é dominante sobre o

estado de ausência. Sendo assim caracterizados como marcadores dominantes. Entretanto, algumas vezes, pesquisadores declaram que os mesmos podem ser empregados como co-dominantes, pois avaliam os polimorfismos levando em consideração a intensidade do sinal detectado como um reflexo do efeito de dose, isto é, dose dupla, dose única ou ausente (Grzebelus, 2015). O polimorfismo observado geralmente resulta de substituições nucleotídicas únicas dentro de sítios de restrição, ou de inserções e deleções (InDels) incluindo alterações nos próprios sítios de restrição. Outra variante pode ainda ocorrer no padrão de metilação destes sítios, entretanto, o polimorfismo estrutural colabora com mais de 90% da variabilidade identificada (Wittenberg et al., 2005).

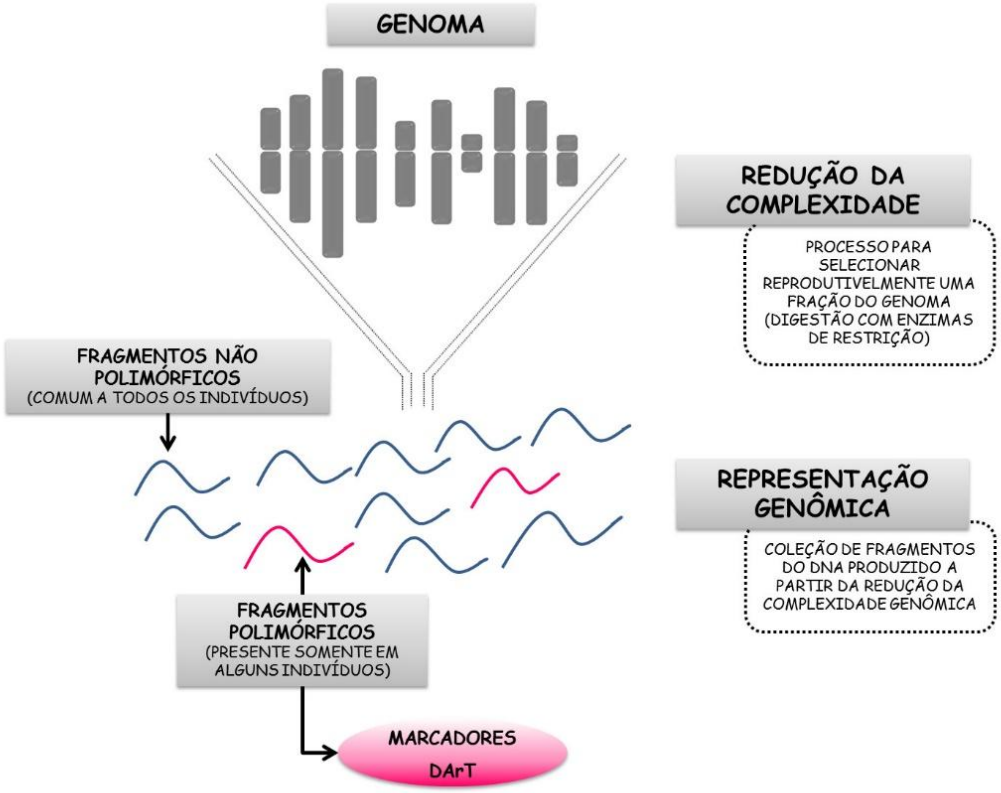

Figura 7.1 - **Princípio dos marcadores DArT**. Inicialmente é realizada uma redução da complexidade genômica do pool gênico de interesse, seguido da construção de uma biblioteca ou coleção dos fragmentos deste *input* de DNA. Ao final, a genotipagem de um indivíduo é determinada pela presença ou ausência de fragmentos polimórficos identificados na representação genômica.

Comparando DArT com outras tecnologias

A tecnologia DArT tanto compartilha algumas características com outras técnicas, como apresenta aspectos únicos. No seu princípio, DArT é mais similar à técnica de polimorfismo de comprimento de fragmento de restrição (RFLP, de *Restriction Fragment Length Polimorphism*) que consiste na fragmentação do DNA através de enzimas de restrição com posterior hibridização destes fragmentos a sequências homólogas de DNA marcadas com radioatividade ou compostos que desencadeiam uma reação de luminescência. Entretanto, DArT é realizada em reverso, isto é, as sondas são

fixadas em um suporte sólido também chamado de *slide* de uma maneira altamente paralela, isto é, em uma única análise pode-se investigar centenas a miliares de fragmentos.

DArT é superior na geração de fragmentos polimórficos quando comparado à AFLP. Embora esta última também utilize enzimas de restrição e amplificação de fragmentos através de *primers* com sequências complementares aos adaptadores adicionados às extremidades geradas, a cobertura de um genoma com um único ensaio de AFLP é cerca de duas vezes menor que o ensaio com DArT (Kilian et al., 2012).

Com relação a metodologias que detectam SNPs (ver descrição detalhada no Capítulo 8), DArT apresenta vantagens àquelas que fazem uso de *primers* quando aplicada a estudos de espécies poliplóides. O bom desempenho dos marcadores DArT em espécies de qualquer nível de poliploidia (como a cana-de-açúcar) se deve ao fato de que a detecção dos SNPs é baseada na alta fidelidade das enzimas de restrição ao invés de anelamento de *primers*. Este é um aspecto bastante importante considerando que, um aumento no nível de poliploidia ocasiona maior competição entre os sítios alvos para o anelamento dos iniciadores.

Além disto, é interessante salientar que algumas novas técnicas foram desenvolvidas baseadas no princípio da DArT. Este é o caso da técnica *Restriction Site Tagged Microarrays* (RST), que faz algumas modificações na forma como reduz a complexidade genômica inicial (Zabarovsky et al., 2003). Uma característica de RST é que os microarranjos são gerados a partir das sequências flanqueadoras de sítios de restrição específicos (por exemplo, sequências ligadas exclusivamente a sítios de *NotI*). Outra variação da DArT é a *Restriction site-associated DNA* (RAD) a qual também varia na etapa de redução da complexidade, bastante similar à descrita anteriormente. RST tem sido relatada em estudos de composição de populações microbianas (Zabarovsky et al., 2003) e RAD, em genotipagem de *Drosophila melanogaster* (Miller et al., 2007). Ambas as técnicas desenvolvidas a partir de DArT também detectam polimorfismos baseados na variação existente nos sítios das enzimas de restrição utilizadas.

Desenvolvendo marcadores do tipo DArT

O desenvolvimento de marcadores DArT pode ser concentrado em três principais etapas: (1) desenvolvimento dos painéis de diversidade, também descritos como construção das bibliotecas, (2) genotipagem das amostras, e (3) análises dos dados. Todas as etapas serão descritas detalhadamente nas próximas sessões e estão ilustradas na Figura 7.2.

Desenvolvimento dos painéis de diversidade

Nesta primeira etapa desenvolve-se um painel de diversidade (*array*) baseado na construção de uma biblioteca que represente um genoma de interesse. Esta biblioteca pode ser proveniente de um grupo (*pool*) de DNA genômico de vários indivíduos que representem este genoma. As representações genômicas são geradas através da redução da complexidade do DNA. Para isto é realizado o tratamento das amostras com enzimas de restrição específicas, ou seja, utilizando a ação combinada de duas enzimas, uma de corte raro e outra de corte frequente. Exemplos de enzimas utilizadas no DArT são: *PstI* (para corte raro) *e TaqI*, *MseI*, *MspI* entre outras (para corte frequente). A enzima de corte raro *PstI* é sensível à metilação CpG, isto significa que ela corta preferencialmente em regiões pouco metiladas ou hipometiladas do DNA, o que geralmente são regiões

ricas em sequências gênicas e de alta complexidade. Já as enzimas de corte frequente são utilizadas para remover fragmentos *PstI*/*PstI* longos, enriquecendo assim fragmentos curtos que serão mais adequados para os passos seguintes: ligação de adaptadores e amplificação por PCR. A ligação de adaptadores específicos aos sítios de restrição produzidos pelas enzimas de corte raro é realizada logo após a digestão. Desta forma, *primers* contendo sequências complementares aos adaptadores amplificarão somente fragmentos contendo sítios *PstI*/*PstI*, produzindo assim uma representação genômica com reduzida complexidade. Geralmente, após a amplificação, parte da reação é verificada em gel de agarose a fim de confirmar a combinação das enzimas de restrição, que deverá ser aquela que apresente um rastro homogêneo de fragmentos. As reações que possuem padrões de bandas detectáveis não são de interesse, pois elas revelam a presença de fragmentos repetitivos ou multicópias, os quais serão redundantes na biblioteca e na subsequente descoberta de marcadores (Kilian et al., 2012; Sansaloni, 2012). O passo seguinte é a clonagem destes fragmentos em um vetor de entrada, com posterior amplificação dos fragmentos inseridos via PCR. Como descrito anteriormente (seção Princípios da técnica), o processo de clonagem tem como função a individualização dos fragmentos. Os produtos desta PCR serão tratados e então impressos em lâminas de vidro (também denominadas *slides*). Desta forma, a biblioteca de referência (também chamada microarranjo - *microarray*), está pronta para as análises subsequentes.

Genotipagem das amostras alvo ou targets

A segunda etapa de obtenção dos marcadores DArT consiste na genotipagem utilizando o painel de diversidade (microarranjo) gerado na etapa anterior. Desta forma, são preparadas as amostras a serem testadas ou genotipadas, designadas de DNA-alvo ou *target*. Uma vez que estiverem prontos, estes alvos são hibridizados ao microarranjo. Inicialmente o DNA-alvo é processado como descrito para a construção das bibliotecas, mas após a amplificação o produto da PCR é marcado com fluoróforos, como Cy3 (verde) ou Cy5 (vermelho). A região de multiclonagem (*polylinker*) do vetor de clonagem é utilizada como referência de qualidade, isto é, a intensidade do sinal de hibridização deste fragmento é analisada pelo *software*, determinando assim, para cada clone a quantidade de DNA impresso no arranjo (Sansaloni, 2012). Por isto, esta referência é marcada com um fluoróforo de cor diferente do utilizado para marcar as amostras. Em seguida promove-se a hibridização do DNA-alvo com a biblioteca de referência impressa nos slides. Esta hibridização ocorre em câmaras especias com capacidade para 8 *slides*, fator que delimitará o número de análises diárias. Após a hibridização, estes *slides* são lavados e secos para seguir para o escaneamento. Um escâner para microarranjos (com laser confocal) realiza a leitura da intensidade da fluorescência. As imagens detectadas serão processadas com o uso do *software* DArTSoft (DArT Pty/Ltd).

Análises de DArT

As imagens geradas são analisadas em um *software* desenvolvido pela própria companhia denominado de DArTsoft. O DArTsoft é utilizado para analisar a intensidade das hibridizações, ou seja, após o escaneamento dos microarranjos as imagens são analisadas por este *software* que localiza automaticamente os *spots* (ou pontos) individuais sobre o microarranjo, considerando o diâmetro dos *spots* (em pixels) e a resolução com a qual os *slides* foram digitalizados (em microns). A análise da imagem é sempre feita em par, isto é, uma imagem detectando a referência (DNA do vetor) e o alvo (DNA da amostra). As imagens são transformadas em matrizes de

números com valores variando de "0" (correspondendo à ausência de sinal) a "65553" (um valor extremamente alto que corresponde à saturação). O *software* localiza automaticamente os *spots*, e após esta localização extrai as informações da imagem. Para cada um dos *spots*, o DArTsoft calcula uma gama de parâmetros, dentre os quais destaca-se a avaliação do sinal de intensidade detectado da referência, o qual servirá como controle de qualidade para cada elemento do arranjo, sendo também utilizado como um normalizador do sinal de hibridização proveniente dos alvos. Este parâmetro calculado por log[alvo/referência] reduz o ruído gerado pela diferença na quantidade de DNA aplicado no arranjo (Kilian et al., 2012). Os clones polimórficos são identificados como os *spots* que possuem diferenças significantes na intensidade dos sinais de hibridização entre as amostras testadas. Assim, dentro do DArTsoft, análises estatísticas são utilizadas para converter a intensidade dos sinais de hibridização dos clones polimórficos em valores, e assim a genotipagem é classificada de uma forma binária como "0" ou "1".

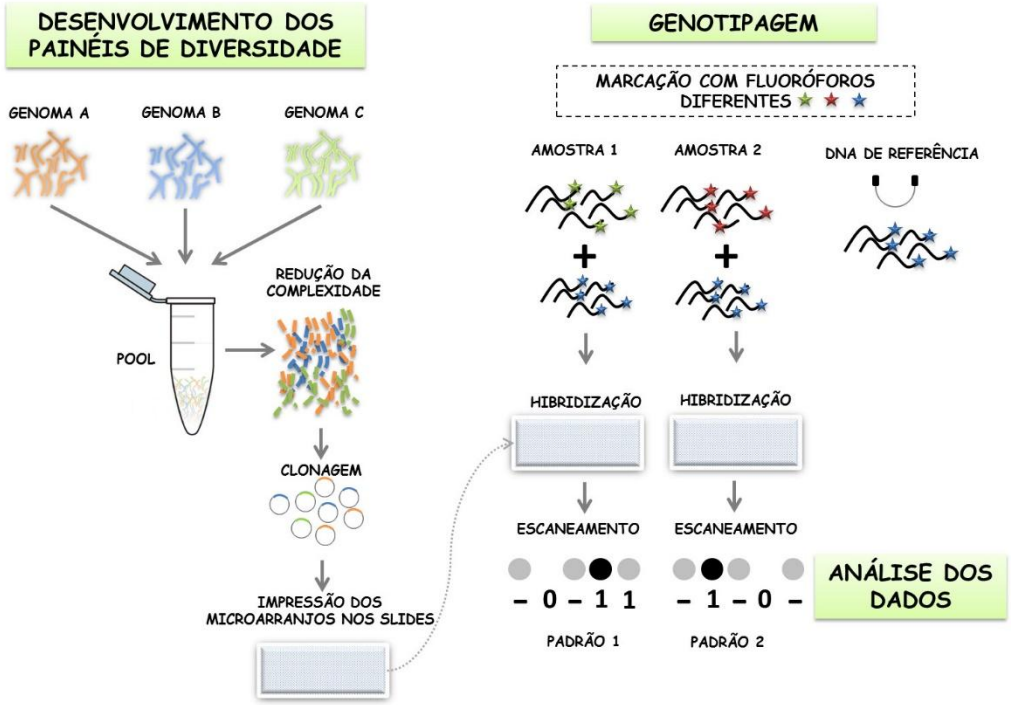

Figura 7.2 - **Etapas do desenvolvimento dos marcadores DArT**. Inicialmente o desenvolvimento dos painéis de diversidade é caracterizado pela mistura de um conjunto de DNA que pode ser proveniente de diferentes espécies, de uma população ou de genótipos parentais, seguido da redução da complexidade via atividade de enzimas de restrição, com posterior clonagem, amplificação e impressão dos fragmentos sobre um *slide* de vidro (microarranjo). A genotipagem é realizada com a preparação das amostras a serem testadas, isto é, redução da complexidade, marcação com distintos fluoróforos e hibridização com o arranjo anteriormente preparado. Utiliza-se um DNA de referência, geralmente proveniente da região de multiclonagem do vetor, que servirá como um normalizador nas análises de imagem. Os arranjos hibridizados são escaneados e as imagens são convertidas em uma matriz binária (0 ou 1) para posteriores análises (adaptado de Jaccoud et al., 2001).

Vantagens da tecnologia DArT

Os marcadores DArT são caracterizados pela sua alta robustez, baixo custo e alta confiabilidade com que são gerados os dados. Ao longo do seu desenvolvimento, e com o decorrer das experiências dos grupos de pesquisas que fazem uso destes marcadores, pode se apontar alguns dos principais atrativos desta técnica:

Alta reprodutibilidade: os marcadores DArT são altamente reprodutíveis. Em estudos utilizando *Arabidopsis thaliana*, os autores encontraram uma reprodutibilidade de 99,8%, isto é, uma acurácia de 1 erro para cada 500 genotipagens (Wittenberg et al., 2005), corroborando com os achados em cevada (Wenzl et al., 2004).

Não necessita prévio conhecimento do genoma: este marcador pode ser desenvolvido sem nenhuma informação prévia do genoma da espécie em questão. Assim, DArT independe de investimentos em sequenciamento, sendo um marcador de especial interesse para espécies as quais há uma limitada ou ausente fonte de informações genéticas, e adicionalmente facilita o trabalho com espécies poliploides onde o sequenciamento é bastante trabalhoso.

Análises em paralelo: alguns marcadores moleculares ainda são dependentes de géis de eletroforese para posterior identificação e desenvolvimento dos mesmos. No caso dos marcadores DArT, a independência com relação a géis de eletroforese facilita e agiliza suas análises. Assim, a mesma plataforma é utilizada tanto para descobrir novos marcadores quanto para a avaliação dos mesmos nas amostras investigadas. Desta forma nenhum ensaio específico precisa ser realizado após a descoberta do marcador, exceto em estudos de metagenômica, nos quais é necessária a montagem de um arranjo inicial composto por todos os marcadores polimórficos detectados. Nestes casos o arranjo de genotipagem contendo apenas marcadores polimórficos será rotineiramente utilizado para genotipagem (Huttner et al., 2005).

Alta acurácia: o *software* DArTsoft, especificamente desenvolvido para as análises dos marcadores DArT, analisa um grande número de dados gerados em cada experimento. O *software* analisa as imagens de microarranjo e subsequentemente classifica os marcadores como descrito na seção 7.2.3. Este programa permite calcular uma gama de parâmetros qualitativos para cada marcador. Os *thresholds* (ou pontos de corte) para estes parâmetros qualitativos podem ser estipulados pelo usuário, podendo assim aumentar os critérios restringentes e selecionar um conjunto de marcadores com alta qualidade e reprodutibilidade.

Flexibilidade na aplicação: as bibliotecas podem ser geradas desde genomas individuais a genomas complexados em um *pool* (metagenoma), dependendo da aplicação desejada. Para os estudos de mapeamento, podem-se utilizar os indivíduos parentais de uma população segregante, enquanto que, em estudos de diversidade genética o DNA pode ser proveniente desde variedades cultivadas a espécies aparentadas selvagens. Além disto, a plataforma de microarranjos também pode ser alterada, isto é, se marcadores forem identificados no painel de diversidade em um experimento inicial, estes podem então ser rearranjados em novos slides servindo como um novo arranjo para genotipagem.

Aplicações dos marcadores DArT

Até o momento inúmeros estudos (dos quais alguns estão citados na Tabela 7.1) já desenvolveram marcadores do tipo DArT. Sobretudo, destacam-se dois grandes focos na utilização destes marcadores: os estudos de diversidade genética e o melhoramento genético de plantas, sendo que em ambos os casos a grande maioria se concentra em plantas de interesse agronômico.

Estudos de diversidade genética

O grande número de marcadores que são simultaneamente testados por DArT fornece uma resolução bastante alta em estudos de diversidade genética. Além disto, as estimativas da distância genética obtida por marcadores DArT possuem maior probabilidade de serem mais precisas devido à natureza aleatória deste tipo de marcador (Wenzl et al., 2008). No trabalho onde foi descrito o desenvolvimento da tecnologia DArT (utilizando-se o cereal modelo *Oryza sativa* - arroz) também foi demonstrada a eficiência destes marcadores na detecção da diversidade genética de diferentes cultivares (Jaccoud et al., 2001). Resultado este corroborado por um segundo estudo utilizando diferentes populações de arroz (Xie et al., 2006). Os estudos de diversidade genética envolvendo marcadores DArT também têm sido amplamente utilizados em outros cereais como trigo (Akbari et al., 2006), cevada (Ovesná et al., 2013) e aveia (Oliver et al., 2011). Outra aplicação bastante utilizada de DArT tem sido a identificação genética ou DNA *fingerprinting* de várias espécies cultivadas. Por exemplo, marcadores DArT foram eficientes na investigação genética de 38 acessos de mandioca (*Manihota esculenta*), obtendo sucesso na segregação entre genótipos selvagens dos cultivados (Xia et al., 2005). Além disto, este mesmo painel foi utilizado em um segundo estudo, que a partir de 435 marcadores DArT, conseguiu separar populações de mandiocas provenientes da África e da América Latina (Hurtado et al., 2008). Análises de identidade genética também foram realizadas para espécies arbóreas (Lezar et al., 2004; Sansaloni et al., 2010); leguminosas (Briñez et al., 2012; Hang Vu et al., 2012); frutíferas (Amorim et al., 2009; Risterucci et al., 2009) e oleaginosas (Raman et al., 2011; Atienza et al., 2013). DArT também tem sido aplicado em estudos de diversidade gênica em outros organismos além de plantas. Com o objetivo de identificar genes do mosquito *Aedes aegypti* associados com a resistência a proteínas inseticidas de *Bacillus thuringiensis* var *israelensis* (*Bti*), populações de *A. aegypti* foram exploradas em nível genômico a fim de identificar loci candidatos (Bonin, 2008).

Melhoramento genético vegetal

Dentro do melhoramento genético os marcadores DArT têm sido aplicados com vários enfoques, sendo propícios para a geração de mapas genéticos, análises de *loci* responsáveis por características quantitativas (QTLs, de *quantitative trait loci*), análises de agrupamentos segregantes (BSA, de *Bulked Segregant Analysis*), bem como, demonstram um potencial na seleção assistida por marcadores (MAS, de *Marker-assisted selection*). A construção de mapas genéticos baseados em marcadores moleculares é um pré-requisito para vários fatores como (i) definição da base genética de características qualitativas e quantitativas de importância agronômica, (ii) descoberta de novos genes envolvidos no controle de variações fenotípicas e (iii) identificação de marcadores moleculares utilizados em MAS (Marone et al., 2012). Devido à geração de um alto número de marcadores em um único ensaio, DArT é altamente utilizado para produzir mapas genéticos, sendo já desenvolvidos em diversas culturas como: aveia (Oliver et al., 2011), oliveira (Domínguez-García et al., 2012), macieira (Schouten et al.,

2012) entre outros. Além disto, DArT têm sido utilizados em integração com outros marcadores, os quais combinados incrementam mapas genéticos. DArT já foi agregado com marcadores *Simple Sequence Repeat* (SSR), RFLP e *Sequence Tagged Site* (STS) em análises de cevada (Wenzl et al., 2006). Em canola, grupos de ligações foram estabelecidos integrando DArT com SSR, *intron polymorphism* (IP) e marcadores baseados em genes (Raman et al., 2011). Outro exemplo de associação de DArT com marcadores genes específicos foi realizada em estudos com grão-de-bico (King et al., 2013). Os marcadores DArT também são utilizados no mapeamento de QTLs que são *loci* responsáveis por características de natureza quantitativa, isto é, aqueles associados a traços fenotípicos que apresentam uma variação contínua, como por exemplo, altura, peso ou rendimento de uma planta. Diversos *loci* já foram detectados empregando DArT, como por exemplo, regiões gênicas associadas ao rendimento de grãos (Cui et al., 2014) e cor do endosperma (Pozniak et al., 2007) do trigo, tolerância a alagamento em cevada (Li et al., 2008), conteúdo de carotenóides em batata (Campbell et al., 2014) e resistência à patógenos em cevada (Ziems et al., 2014). DArT ainda pode auxiliar no melhoramento genético quando aplicado à BSA. Análises de BSA identificam marcadores associados a um fenótipo através de um *screening* ou triagem de dois *pools* de DNA proveniente de plantas fenotipicamente distintas (Wenzl et al., 2007). Embora os marcadores DArT sejam tipicamente classificados como dominantes, as intensidades geradas nos ensaios de hibridização possuem uma natureza quantitativa, pois refletem a abundância dos fragmentos de DNA individuais nas representações genômicas. Associada a grande reprodutibilidade da técnica, as intensidades de hibridização podem ser utilizadas para medir a frequência alélica nos *pools* de DNA, como foi observado em análises de BSA entre *bulks* de cevada contrastantes quanto às folhas pubescentes. O alelo responsável pela presença de pêlos na superfície foliar foi identificado diferencialmente nos *pools* através do uso de marcadores DArT (Wenzl et al., 2007). Uma vez que a ligação entre uma característica de interesse e um marcador molecular seja estabelecida, estes podem auxiliar na MAS. Assim, conjuntos de marcadores DArT têm sido testados quanto a sua ligação a *locus* de interesse em plantas agronômicas (Chen et al., 2016).

Tabela 7.1 - Exemplos de organismos para os quais já foram desenvolvidos marcadores DArT e sua aplicações.

Espécie	Aplicação	Referência
Arabidopsis thaliana	Validação da técnica	(Wittenberg et al., 2005)
Aedes aegypti	Diversidade genética	(Bonin et al., 2008)
Avena sativa	Mapeamento genético	(Oliver et al., 2011)
Brassica napus	Mapeamento genético	(Raman et al., 2011)
Cajamus cajan	Diversidade genética	(Yang et al., 2006)
Cicer arietinum	Mapeamento genético	(Thudi et al., 2011)
Cucumis sativus	Sequenciamento do genoma	(Huang et al., 2009)

Daucus carota	Diversidade e mapeamento genético, genotipagem	(Macko-Podgorni et al., 2014)
Eucalyptus grandis	Identidade genética, Genotipagem em larga escala	(Lezar et al., 2004; Sansaloni et al., 2010)
Fragaria ananassa	Mapeamento e diversidade genética, Estruturação populacional	(Sanchez-Sevilla et al., 2015)
Fusarium oxysporum	Diversidade genética	(Sharma et al., 2004)
Hordeum vulgare	Mapeamento genético BSA	(Wenzl et al., 2006; Wenzl et al., 2007)
Humulus lupulus	Identificação de QTLs	(McAdam et al., 2013)
Lupinus angustifolia	Diversidade genética, MAS	(Chen et al., 2016)
Malus domestica	Mapeamento genético	(Schouten et al., 2012)
Manihot esculenta	Genotipagem	(Xia et al., 2005)
Musa sp.	Análises de germoplasma	(Risterucci et al., 2009)
Mycophaerella gramicola	Caracterização genética	(Wittenberg et al., 2009)
Nicotiana tabacum	Diversidade genética	(Lu et al., 2013)
Olea europaea	Mapeamento genético	(Domínguez-García et al., 2012)
Oryza sativa	Desenvolvimento da tecnologia, Genotipagem em larga escala	(Jaccoud et al., 2001; Xie et al., 2006)
Phaseolus vulgaris	Identificação de QTLs	(Oblessuc et al., 2013)
Saccharum officinarum	Mapeamento e genotipagem	(Wei et al., 2010)
Solanum michoacanum	Mapeamento genético	(Sliwka et al., 2012)
Sorgum bicolor	Diversidade e mapeamento genético	(Mace et al., 2008)
Triticum aestivum	Identidade genética, Mapeamento genético	(Akbari et al., 2006; Paux et al., 2008; Tsilo et al., 2010)
Triticum turgidum	Identificação de QTLs	(Pozniak et al., 2007)

Considerações finais

O aumento no número de estudos utilizando a tecnologia DArT tem comprovado o sucesso deste marcador entre diferentes grupos de pesquisa ao redor do mundo. Um dos principais fatores para a ampla utilização destes marcadores se deve ao fato de que

um arranjo de DArT pode ser desenvolvido em um curto período de tempo, mesmo nos casos de espécies sem prévio conhecimento do seu genoma. Além disto, uma série de outras vantagens é associada a estes marcadores, como: baixo custo de genotipagem por *locus*, procedimento rápido e robusto, além da geração automatizada de dados, pois independe de análises baseadas em géis de eletroforese.

Os marcadores DArT se popularizaram em estudos de genomas vegetais, pois devido sua ampla cobertura, viabilizaram estudos que pareciam limitados por características complexas. Um exemplo é a aplicação em espécies com o alto grau de poliploidia, característica comum de plantas cultivadas. Entretanto, devido a plataforma DArT ser uma tecnologia aberta, estes marcadores podem se difundir entre outros organismos (como fungos, bactérias e animais) e assim incrementar pesquisas nos mais diversos campos da genética e da biotecnologia.

Box 7.1

Box 7.1 - DArTseq: Marcadores DArT associados a NGS

Nos últimos anos grandes avanços foram realizados na área de sequenciamento de ácidos nucleicos, gerando plataformas conhecidas como *next-generation sequencing* (NGS) ou sequenciamento de última geração. Estas plataformas são bastante competitivas, pois geram uma quantidade massiva de dados barateando o custo por análise de amostra. Desta forma, algumas modificações realizadas na metodologia DArT, permitiram implementar esta tecnologia com NGS, modificando a etapa de hibridização dos microarranjos para detecção do polimorfismo, sendo assim denominada de DArTseq. Esta nova abordagem combina o protocolo de redução da complexidade genética empregando o sistema tradicional de DArT seguido pela genotipagem por sequenciamento baseado na plataforma Illumina (*short read sequencing*). Como resultado duas tabelas de classificação são geradas, combinando polimorfismos gerados por DArT bem como variações de SNPs. Esta metodologia já demonstrou ser bastante eficiente em dois estudos, um reportado por Cruz et al. (2013) envolvendo o gênero *Physaria spp*., o qual gerou mais de 27.000 marcadores polimórficos, bem como, em outro trabalho que revelou 17.000 marcadores polimórficos uniformemente distribuídos no genoma de arroz (Courtois et al., 2013). A vantagem da metodologia DArTseq comparada ao DArT tradicional é relacionada a aplicações que requerem a detecção de um número massivo de marcadores (algo como milhares de marcadores). Isto é ocasionado pela ausência da etapa de hibridização dos alvos nos *slides* e consequentemente não tem a limitação do número de genotipagens devido o espaço restrito nas câmaras de hibridização. Desta forma, DArTseq apresenta o perfil de sequenciamento de larga escala, sendo propícia para a geração de mapas de alta resolução e na exploração de detalhes genéticos de características de interesse.

Referências Bibliográficas

Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, et al. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor Appl Genet* 113: 1409-1420.

Amorim EP, Vilarinhos AD, Cohen KO, et al. (2009). Genetic diversity of carotenoid-rich bananas evaluated by Diversity Arrays Technology (DArT). *Genet Mol Biol* 32: 96-103.

Atienza SG, De La Rosa R, Domínguez-García MC, Martín A, Kilian A, Belaj A (2013). Use of DArT markers as a means of better management of the diversity of olive cultivars. *Food Res Int* 54**:** 8.

Avise JC (2004). *Molecular Markers, Natural History, and Evolution.* Sunderland, MA: Sinauer.

Bonin A, Paris M, Després L, Tetreau G, David JP, Kilian A (2008) A MITE-based genotyping method to reveal hundreds of DNA polymorphisms in an animal genome after a few generations of artificial selection. *BMC Genomics* 9: 459.

Briñez B, Blair M, Kilian A, Carbonell S, Chiorato A, Rubiano L (2012) A whole genome DArT assay to access germplasm collection diversity in common beans. *Mol Breed* 30: 13.

Campbell R, Pont SD, Morris JA, et al. (2014) Genome-wide QTL and bulked transcriptomic analysis reveals new candidate genes for the control of tuber carotenoid content in potato (Solanum tuberosum L.). *Theor Appl Genet* 127: 1917-1933.

Chen Y, Shan F, Nelson MN, Siddique KH, Rengel Z (2016) Root trait diversity, molecular marker diversity, and trait-marker associations in a core collection of Lupinus angustifolius. *J Exp Bot* 67: 3683-3697.

Courtois B, Audebert A, Dardou A, et al. (2013) Genome-wide association mapping of root traits in a japonica rice panel. *PLoS One* 8: e78037.

Cruz VM, Kilian A, Dierig DA (2013) Development of DArT marker platforms and genetic diversity assessment of the U.S. collection of the new oilseed crop lesquerella and related species. *PLoS One* 8: e64062.

Cui F, Zhao C, Ding A, et al. (2014) Construction of an integrative linkage map and QTL mapping of grain yield-related traits using three related wheat RIL populations. *Theor Appl Genet* 127: 659-675.

Domínguez-García MC, Belaj A, De La Rosa R, et al. (2012) Development of DArT markers in olive (Olea europaea L.) and usefulness in variability studies and genome mapping. *Scientia Horticulturae* 136: 50 - 60.

Grzebelus D (2015) "Diversity Arrays Technology (DArT) Markers for Genetic Diversity," in *Genetic Diversity and Erosion in Plants,* eds. M.R. Ahuja & S. Mohan Jain. Springer International Publishing), 295-309.

Hang Vu TT, Lawn RJ, Bielig LM, Molnar SJ, Xia L, Kilian A (2012) Development and initial evaluation of diversity array technology for soybean and mungbean. *Euphytica* 186: 14.

Huang S, Li R, Zhang Z, et al. (2009) The genome of the cucumber, Cucumis sativus L. *Nat Genet* 41: 1275-1281

Hurtado P, Olsen KM, Buitrago C, et al. (2008) Comparison of simple sequence repeat (SSR) and diversity array technology (DArT) markers for assessing genetic diversity in cassava (Manihot esculenta Crantz). *Plant Genetic Resources: Characterization and Utilization* 6: 208-214

Huttner E, Wenzl P, Akbari M, et al. (2005) "Diversity Arrays Technology: A Novel Tool for Harnessing the Genetic Potential of Orphan Crops," in *Discovery to Delivery: BioVision, Proceedings of the 2004 Conference of The World Biological Forum,* eds. I. Serageldin & G.J. Persley. (Alexandria 2004: CABI Publishing), 145-155.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res* 29: E25.

Kilian A, Wenzl P, Huttner E, et al. (2012) Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol Biol* 888: 67-89.

King J, Thomas A, James C, King I, Armstead I (2013) A DArT marker genetic map of perennial ryegrass (Lolium perenne L.) integrated with detailed comparative mapping information; comparison with existing DArT marker genetic maps of Lolium perenne, L. multiflorum and Festuca pratensis. *BMC Genomics* 14: 437.

Lezar S, Myburg AA, Berger DK, Wingfield MJ, Wingfield BD (2004) Development and assessment of microarray-based DNA fingerprinting in Eucalyptus grandis. *Theor Appl Genet* 109: 1329-1336.

Li H, Vaillancourt R, Mendham N, Zhou M (2008) Comparative mapping of quantitative trait *loci* associated with waterlogging tolerance in barley (Hordeum vulgare L.). *BMC Genomics* 9: 401.

Lu, XP, Xiao BG, Li YP, Gui YJ, Wang Y, Fan LJ (2013) Diversity arrays technology (DArT) for studying the genetic polymorphism of flue-cured tobacco (Nicotiana tabacum). *J Zhejiang Univ Sci B* 14: 570-577.

Mace ES, Xia L, Jordan DR, Halloran K, Parh DK, Huttner E, et al. (2008) DArT markers: diversity analyses and mapping in Sorghum bicolor. *BMC Genomics* 9: 26.

Macko-Podgorni A, Iorizzo M, Smolka K, Simon PW, Grzebelus D (2014) Conversion of a diversity arrays technology marker differentiating wild and cultivated carrots to a co-dominant cleaved amplified polymorphic site marker. *Acta Biochim Pol* 61: 19-22.

Marone D, Panio G, Ficco DB, et al. (2012) Characterization of wheat DArT markers: genetic and functional features. *Mol Genet Genomics* 287: 741-753.

Mcadam EL, Freeman JS, Whittock SP, et al. (2013) Quantitative trait *loci* in hop (Humulus lupulus L.) reveal complex genetic architecture underlying variation in sex, yield and cone chemistry. *BMC Genomics* 14: 360.

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.

Oblessuc PR, Cardoso Perseguini JM, et al. (2013) Increasing the density of markers around a major QTL controlling resistance to angular leaf spot in common bean. *Theor Appl Genet* 126: 2451-2465.

Oliver RE, Jellen EN, Ladizinsky G, et al. (2011) New Diversity Arrays Technology (DArT) markers for tetraploid oat (Avena magna Murphy et Terrell) provide the first complete oat linkage map and markers linked to domestication genes from hexaploid A. sativa L. *Theor Appl Genet* 123: 1159-1171.

Ovesná J, Kučera L, Vaculová K, et al. (2013) Analysis of the Genetic Structure of a Barley Collection Using DNA Diversity Array Technology (DArT). *Plant Molecular Biology Reporter* 31: 9.

Paux E, Sourdille P, Salse J, et al. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322: 101-104.

Pozniak CJ, Knox RE, Clarke FR, Clarke JM (2007) Identification of QTL and association of a phytoene synthase gene with endosperm colour in durum wheat. *Theor Appl Genet* 114: 525-537.

Raman H, Raman R, Nelson MN, et al. (2011) Diversity array technology markers: genetic diversity analyses and linkage map construction in rapeseed (Brassica napus L.). *DNA Res* 19: 51-65.

Risterucci AM, Hippolyte I, Perrier X, et al. (2009) Development and assessment of Diversity Arrays Technology for high-throughput DNA analyses in Musa. *Theor Appl Genet* 119: 1093-1103.

Sanchez-Sevilla JF, Horvath A, Botella MA, et al. (2015) Diversity Arrays Technology (DArT) Marker Platforms for Diversity Analysis and Linkage Mapping in a Complex Crop, the Octoploid Cultivated Strawberry (Fragaria x ananassa). *PLoS One* 10: e0144960.

Sansaloni CP (2012) *Desenvolvimento e aplicações de DArT (Diversity Arrays Technology) e genotipagem por sequenciamento (Genotyping-bySequencing) para análise genética em Eucalyptus.* PhD, Universidade de Brasília.

Sansaloni CP, Petroli CD, Carling J, et al. (2010) A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. *Plant Methods* 6: 16.

Schouten HJ, Van De Weg WE, Carling J, et al. (2012) Diversity arrays technology (DArT) markers in apple for genetic linkage maps. *Mol Breed* 29: 645-660.

Sharma M, Nagavardhini A, Thudi M, Ghosh R, Pande S, Varshney RK (2014) Development of DArT markers and assessment of diversity in *Fusarium oxysporum* f. sp. *ciceris*, wilt pathogen of chickpea (*Cicer arietinum* L.). *BMC Genomics* 15: 454.

Sliwka J, Jakuczun H, Chmielarz M, et al. (2012) A resistance gene against potato late blight originating from Solanum x michoacanum maps to potato chromosome VII. *Theor Appl Genet* 124: 397-406.

Steane DA, Nicolle D, Sansaloni CP, et al. (2011) Population genetic analysis and phylogeny reconstruction in Eucalyptus (Myrtaceae) using high-throughput, genome-wide genotyping. *Mol Phylogenet Evol* 59: 206-224.

Thudi M, Bohra A, Nayak SN, et al. (2011) Novel SSR markers from BAC-end sequences, DArT arrays and a comprehensive genetic map with 1,291 marker *loci* for chickpea (Cicer arietinum L.). *PLoS One* 6: e27275.

Tsilo TJ, Hareland GA, Simsek S, Chao S, Anderson JA (2010) Genome mapping of kernel characteristics in hard red spring wheat breeding lines. *Theor Appl Genet* 121: 717-730.

Wei X, Jackson PA, Hermann S, Kilian A, Heller-Uszynska K, Deomano E (2010) Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker-trait associations in sugarcane. *Genome* 53: 973-981.

Wenzl P, Carling J, Kudrna D, et al. (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proc Natl Acad Sci U S A* 101: 9915-9920.

Wenzl P, Huttner E, Carling J, et al. (2008) "Diversity Arrays Technology (DArT): A generic high-density genotyping platform", in: *7th Internatitional Safflower Conference.* (eds.) S.E. Knights & T.D. Potter. (Wagga Wagga, Australia).

Wenzl P, Li H, Carling J, et al. (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS *loci* and agricultural traits. *BMC Genomics* 7: 206

Wenzl P, Raman H, Wang J, Zhou M, Huttner E, Kilian A (2007) A DArT platform for quantitative bulked segregant analysis. *BMC Genomics* 8: 196.

Wittenberg AH, Van Der Lee T, Cayla C, Kilian A, Visser RG, Schouten HJ (2005) Validation of the high-throughput marker technology DArT using the model plant Arabidopsis thaliana. *Mol Genet Genomics* 274: 30-39.

Wittenberg AH, Van Der Lee TA, Ben M'barek S, et al. (2009) Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen Mycosphaerella graminicola. *PLoS One* 4: e5863.

Xia L, Peng K, Yang S, et al. (2005) DArT for high-throughput genotyping of Cassava (Manihot esculenta) and its wild relatives. *Theor Appl Genet* 110: 1092-1098.

Xie Y, Mcnally K, Li CY, Leung H, Zhu YY (2006) A high-throughput genomic tool: Diversity Arrays Technology complementary for rice genotyping. *Journal of Integrative Plant Biology* 2.

Yang S, Pang W, Ash G, et al. (2006) Low level of genetic diversity in cultivated Pigeonpea compared to its wild relatives is revealed by diversity arrays technology. *Theor Appl Genet* 113: 585-595.

Zabarovsky ER, Petrenko L, Protopopov A, et al. (2003) Restriction site tagged (RST) microarrays: a novel technique to study the species composition of complex microbial systems. *Nucleic Acids Res* 31: e95.

Ziems LA, Hickey LT, Hunt CH, et al. (2014) Association mapping of resistance to Puccinia hordei in Australian barley breeding germplasm. *Theor Appl Genet* 127: 1199-1212.

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).

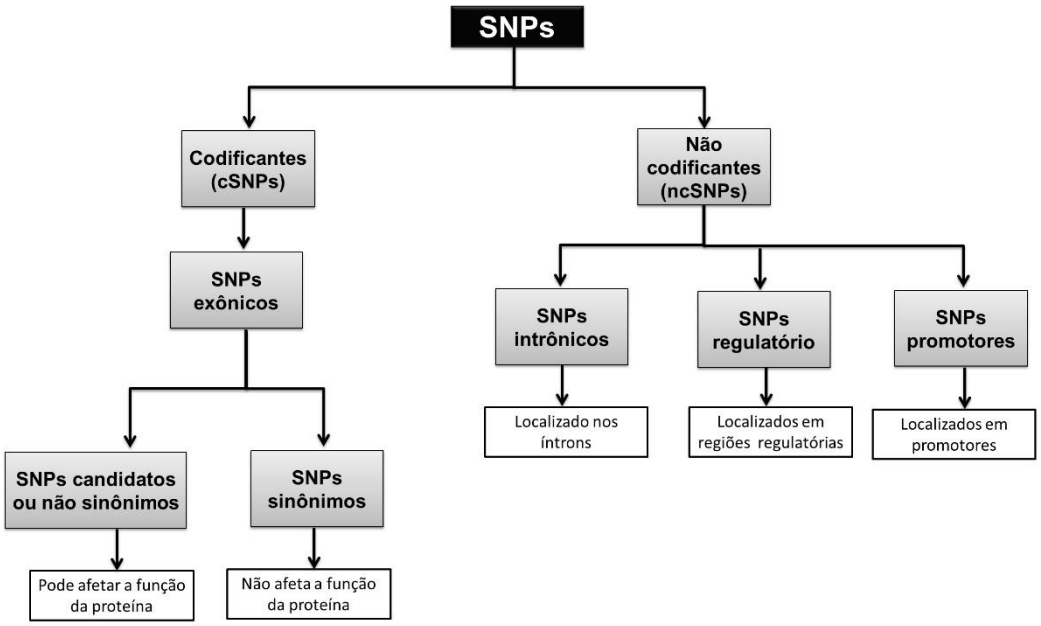

Figura 8.1 - Classificação dos SNPs quanto à localização no genoma e quanto ao impacto causado na proteína ou fenótipo.

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.

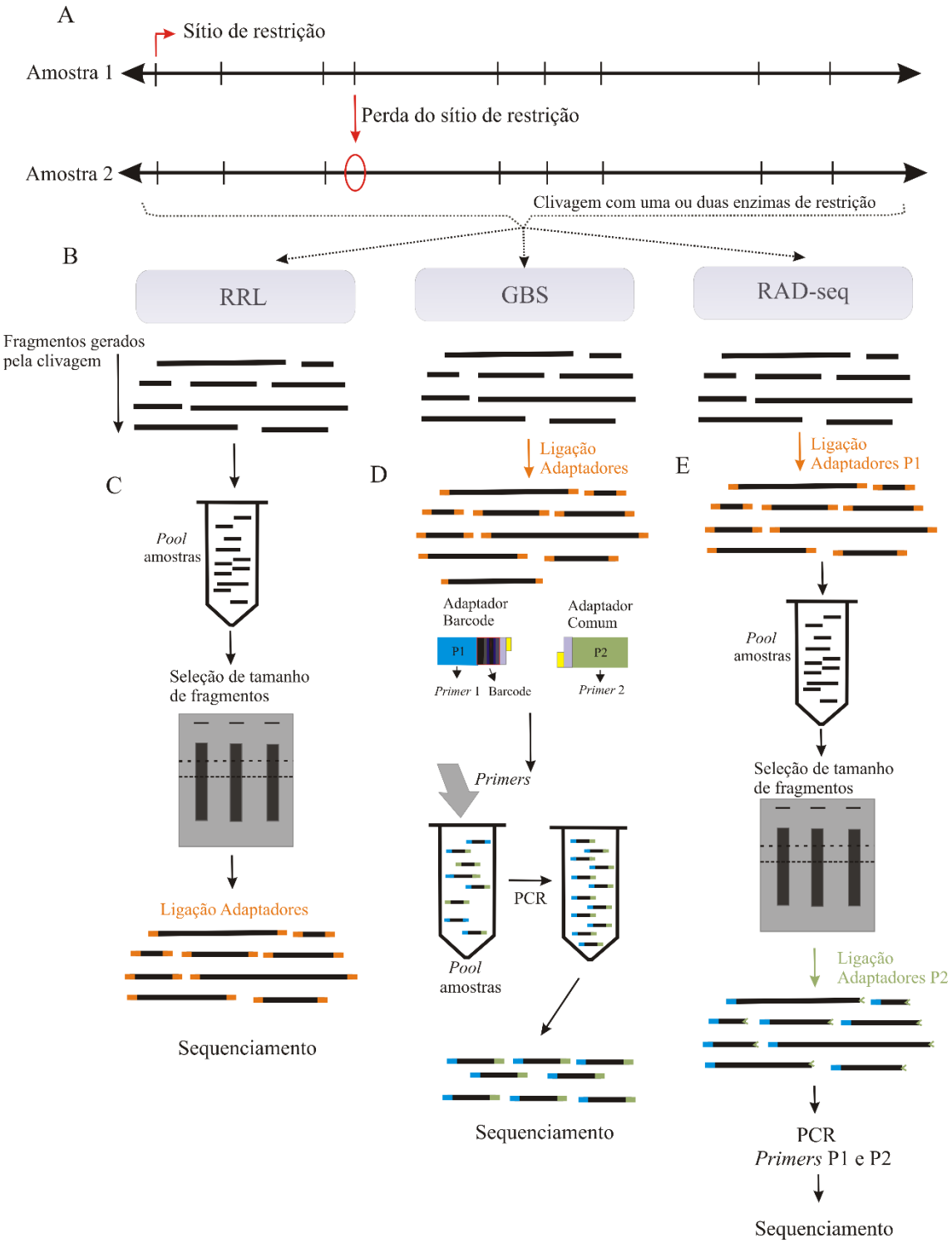

Figura 8.2 - (A, B) Visão geral do processo de clivagem por enzimas de restrição em uma região genômica de dois organismos; etapa comum aos três métodos (RRL, GBS e Rad-seq). A perda de sítios de restrição resulta na variação do número e tamanho dos fragmentos gerados com a clivagem. (C) Medotologia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

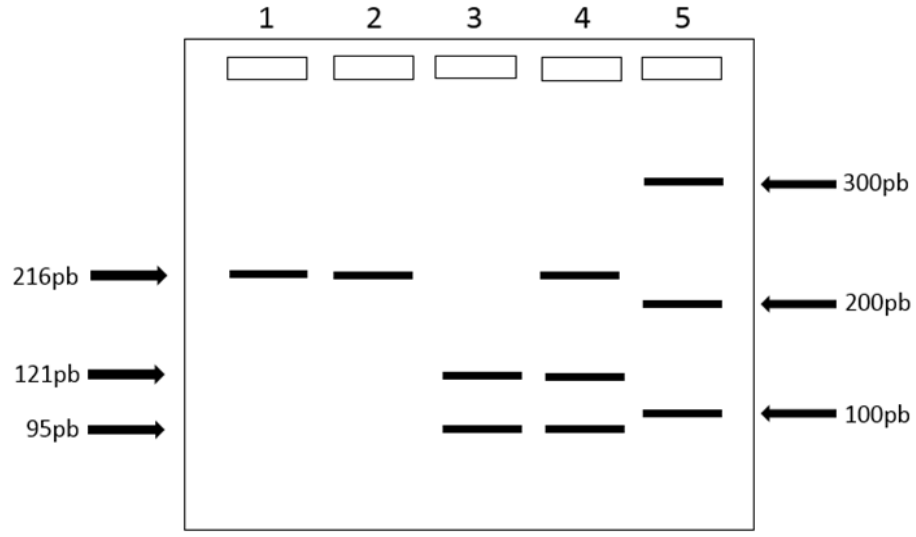

Figura 8.3 - Diagnóstico molecular de hipolactasia: eletroforese em gel de agarose 2,5% do produto de PCR da região à montante do gene LCT e a respectiva clivagem da posição -13910 com a enzima BsmFI. Produto não clivado (1), indivíduo homozigoto para o alelo C (2), indivíduo homozigoto para o alelo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