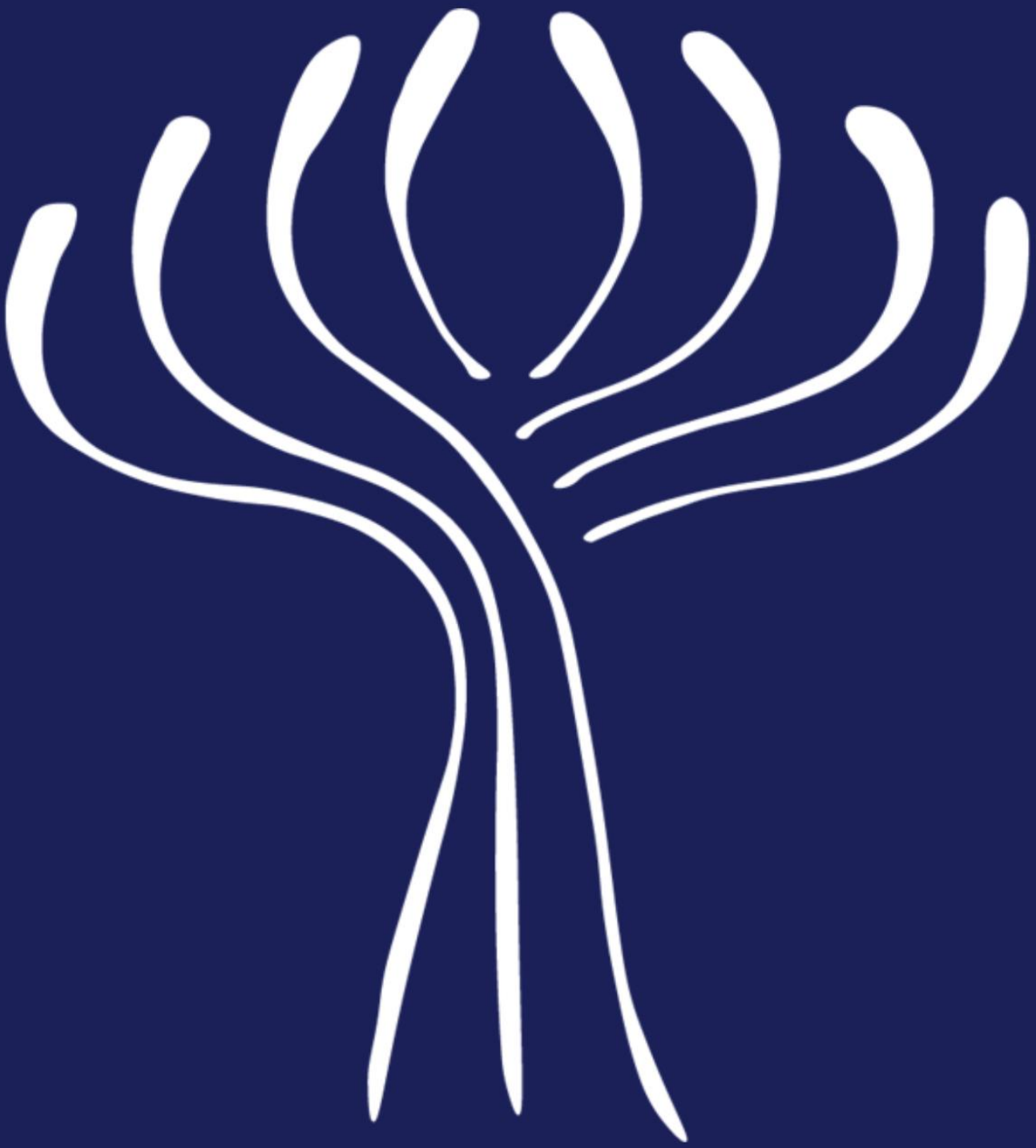


XIII BOTÂNICA NO INVERNO

2024

Universidade de São Paulo
Instituto de Biociências
Departamento de Botânica
Apostila
XIII Botânica no Inverno 2024

Organizadores

Gabriela Naomi Haseyama dos Santos
gabriela.naomi.santos@usp.br
Ítalo Vinicius Cantanhede Santos
italosantos@usp.br
José Laurindo dos Santos Júnior
juniorsantos.laurindo@gmail.com
Joyce Gomes Falcão
joyce.gomes@usp.br
Marcos Lorenzi Martins
marcos.lorenzi@usp.br
Raquel Tsu Ay Wu
raquelwu@usp.br

Professora responsável

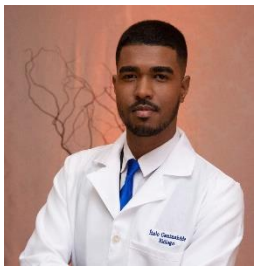
Cláudia Maria Furlan - furlancm@ib.usp.br

Autores dos capítulos

Adriana dos Santos Lopes - adriana.lopes@usp.br
Alan Novaes dos Santos - alannssantos@hotmail.com
Aline Possamai Della - alinedella@usp.br
Antônio Azeredo Coutinho Neto - antonioacneto@biologo.bio.br
Arthur Kim Chan - arthur.chan@usp.br
Bárbara Sousa dos Santos - barbarabertagia@gmail.com
Bruno Edson Chaves - brunoedch02@gmail.com
Carlos Eduardo Valério Raymundo - carlos.raymundo@usp.br
Douglas Santos Oliveira - dougs1935@gmail.com
Edson Moura dos Santos - edson_moura01@outlook.com
Guilherme de Ornellas Paschoalini - guilherme.paschoalini@usp.br
Haissa de Abreu Caitano - haissa.caitano@gmail.com
Ivan Hurtado Caceres - ivanhc@ib.usp.br
Jessica Soares de Lima - jessicadelimaa@gmail.com
Karla Menezes e Vasconcelos - karlamenezes.vasconcelos@gmail.com
Luana Sauthier - sauthier@ib.usp.br
Marcelo Fernando Devecchi - mfdevecchi@gmail.com
Marcos Marchesi da Silva - marcoschesi.silva@gmail.com
Maria Fernanda da Costa Oliveira - mfdacostaoliveira@gmail.com
Marilia de Freitas Silva - marilia.freitas.silva@usp.br
Marina Fonseca Lima -
Matheus Januario - januarioml.eco@gmail.com
Ricardo da Silva Ribeiro - ricardo.silva@unemat.br
Sandra Reinales - spreinalesl@ib.usp.br
Thayná Juliane Guerra da Silva - thaynajuliane95@gmail.com

Sobre os organizadores

Gabriela N. H. dos Santos é Bacharel e Licenciada pela Universidade de São Paulo (IB-USP). Realizei minha IC no Laboratório de Anatomia Vegetal do IB-USP com a Profa. Dra. Gladys Flávia de Albuquerque Melo-de-Pinna. Atualmente mestranda na área de Anatomia Vegetal, no mesmo laboratório e sob orientação da mesma professora. No âmbito do projeto de pesquisa, estamos estudando o desenvolvimento foliar de 3 espécies de *Trichilia* (Meliaceae).



Ítalo V. C. Santos é Bacharel em Ciências Biológicas pela Universidade Federal do Maranhão (UFMA) e Mestrando em Botânica pela Universidade de São Paulo. Têm experiência na área de botânica, com ênfase em fisiologia vegetal. Durante a graduação trabalhou com sistemática, ecologia e cultivo in vitro de espécies de orquídeas nativas da Amazônia e cerrado no Laboratório de Estudos sobre Orquídeas da UFMA. Atualmente, desenvolve sua pesquisa no Laboratório de Fisiologia do Desenvolvimento Vegetal sob orientação da Profª Drª Helenice Mercier, onde estuda o metabolismo do nitrogênio em plantas de arroz transgênicas expressando gene de uma bromélia epífita nativa da Mata Atlântica, visando a produção de linhagens de arroz com maior eficiência do uso desse macronutriente.

José Laurindo d. S. Jr possui graduação em Ciências Biológicas (licenciatura plena) pela Universidade Federal de Sergipe. Tem experiência em Fisiologia Vegetal, principalmente, fisiologia e bioquímica de espécies cultivadas, florestais e de ambientes semiáridos sob estresse abióticos, bem como fisiologia molecular e fisiologia de sementes. É doutorando em Botânica, sob orientação do Professor Dr. Luciano Freschi no laboratório de Fisiologia do Desenvolvimento Vegetal. Ele investiga como a alta temperatura influencia um componente de percepção luminosa e suas rotas de sinalização, e os impactos no desenvolvimento reprodutivo de tomateiro.

Joyce G. Falcão é Graduada em Ciências Biológicas (Bacharelado) pela Universidade Federal do Maranhão. Realizei iniciação científica no laboratório de Fisiologia e Anatomia Vegetal. Atualmente mestranda no Laboratório de Fisiologia Vegetal e o projeto de mestrado tem como tema: Desenvolvimento radicular em linhagens de arroz transgênicos, expressando constitutivamente o gene de aquaporina VgPIP1;2 de bromélia, cultivadas em diferentes concentrações de amônio.

Marcos L. Martins é Graduado em Ciências Biológicas (Bacharel e Licenciatura) pela Universidade São Judas Tadeu (2018). Especialista em Toxinas de Interesse em Saúde pelo Instituto Butantan (CEFOR-SUS). Mestre em fitoquímica (IB-USP). Atualmente é doutorando no programa de pós-graduação em Botânica da USP na linha de pesquisa em Recursos Econômicos Vegetais, desenvolvendo um projeto que visa o estudo fitoquímico e análise de bioatividades de espécies de Asteraceae endêmicas aos campos rupestres, sob a orientação do Prof. Dr. Marcelo J. Pena Ferreira.

Raquel T. A. Wu é graduada em Ciências Biológicas pela Universidade de São Paulo. Atualmente é doutoranda no Laboratório de Genética Molecular de Plantas no Instituto de Biociências da USP, seu projeto consiste na caracterização funcional de genes de tomateiro que codificam proteínas da família B-box e seus interatores. O objetivo deste projeto é caracterizar fenotipicamente mutantes para os genes alvos, de forma a compreender seu papel em tomateiro.

Cláudia Maria Furlan é formada em Biologia, possui mestrado (1995) e doutorado (2004) em Ecologia pela Universidade de São Paulo, e pós-doutorado pela Universidade de São Paulo (2008) e pela Universidade de Turku, Finlândia (2016). Atua na graduação, pós-graduação e em atividades de extensão universitária desde a contratação como docente do IBUSP em 2010. Minhas pesquisas têm como foco a identificação e quantificação de produtos naturais, em especial polifenóis em duas grandes abordagens: 1. Como polifenóis estão envolvidos na resposta de defesa vegetal a fatores de estresse abiótico (ozônio, elevadas concentrações de CO2) e bióticos (parasitismo entre plantas, herbivoria); 2. Caracterização química de Angiospermas nativas buscando por metabólitos bioativos

Ficha catalográfica elaborada pelo Serviço de Biblioteca do Instituto de Biociências da USP com os dados fornecidos pelos autores

B748 XIII Botânica no Inverno 2024 : apostila / organização Gabriela Naomi Haseyama dos Santos, Ítalo Vinicius Cantanhede Santos, José Laurindo dos Santos Júnior, Joyce Gomes Falcão, Marcos Lorenzi Martins, Raquel Tsu Ay Wu, Cláudia Maria Furlan. – São Paulo : Instituto de Biociências, Universidade de São Paulo, 2024.
272 p. : il.

ISBN: 978-65-88234-18-1

1. Botânica. 2. Biologia Vegetal (experimentação e investigação) 3. Plantas. I. Santos, Gabriela Naomi Haseyama dos (org.). II. Santos, Ítalo Vinicius Cantanhede (org.). III. Santos Júnior, José Laurindo dos (org.). IV. Falcão, Joyce Gomes (org.). V. Martins, Marcos Lorenzi (org.). VI. Wu, Raquel Tsu Ay (org.). VII. Furlan, Cláudia Maria. (org.).

LC: QK45

Ficha Catalográfica elaborada por Elisabete da Cruz Neves. CRB-8/6228

SUMÁRIO

O conteúdo dos capítulos é de responsabilidade dos respectivos autoras e autores.

PREFÁCIO

Fundado em 1934 pelo professor Felix Kurt Rawitscher (1890-1957), o Departamento de Botânica do Instituto de Biociências da Universidade de São Paulo (IB-USP) é uma referência internacional em pesquisa e ensino, O Programa de Pós-Graduação em Botânica começou em 1970, oferecendo os cursos de Mestrado e Doutorado. Atualmente, o programa de pós-graduação em Botânica conta com 29 docentes mais três seniores e 76 pós-graduandos, distribuídos em oito grandes linhas de pesquisa nas principais subáreas da Botânica.

O Departamento de Botânica apresenta como infraestrutura 11 laboratórios, um herbário com coleção de plantas, algas e madeiras estimado em 300.000 espécimes, um fitotério com coleção de plantas vivas para uso didático, estufas e casas de vegetação. Somando-se ao grande número de pós-graduandos (dentre esses, estrangeiros) e a alta atividade científica dessa comunidade, a Pós-Graduação de Botânica possui conceito CAPES 7, o conceito máximo em avaliação de pós-graduação em Botânica do Brasil.

Realizado desde o ano 2011, o curso de Botânica no Inverno é uma iniciativa de pós-graduandos para divulgar o trabalho desenvolvido no Departamento de Botânica e que possibilita o acolhimento de potenciais alunos e pesquisadores. O Curso de Botânica no Inverno pretende, com os alunos de graduação e recém-formados, revisar e atualizar conceitos fundamentais de diversas subáreas da Botânica, além de apresentar as atividades realizadas em nossos laboratórios, com o objetivo incentivar futuros acadêmicos/pesquisadores a se engajarem em programas de pós-graduação nas diferentes áreas da Botânica, principalmente no Departamento de Botânica do IB-USP.

Para a realização do XIII Botânica no Inverno, agradecemos à Universidade de São Paulo, à direção do Instituto de Biociências, a chefia do Departamento de Botânica, à Comissão Coordenadora do Programa de Pós-Graduação em Botânica, às agências de fomento, CAPES, CNPq e FAPESP e os patrocinadores, Sociedade Botânica de São Paulo, Ciencor, Exxtend - Solução em Oligos e o Programa de Pós-graduação em Botânica do Instituto de Biociências da Universidade de São Paulo.

O conteúdo dos capítulos é de total responsabilidade dos respectivos autores.

Desejamos a todos um bom curso.

Comissão Organizadora do XIII Botânica no Inverno

CAPÍTULO V

Introdução ao uso de dados genômicos em Sistemática

Sandra Reinales (*Universidade de São Paulo*)
Luana J. Sauthier (*Universidade de São Paulo*)

A primeira metodologia de sequenciamento de DNA, ou sequenciamento de primeira geração, surgiu no final da década de 70 devido aos esforços do bioquímico Frederick Sanger e seu grupo (Sanger et al. 1977), possibilitando a obtenção de sequências curtas após intensivo e custoso trabalho de bancada. Uma década depois, a descoberta da técnica de amplificação de sequências de interesse curtas por meio da técnica de PCR (*Polymerase Chain Reaction*, ou reação em cadeia da polimerase) aliada à comercialização da técnica de sequenciamento de DNA que ficou conhecida como Sanger, revolucionaram a análise molecular, pois tornou possível a obtenção de várias cópias de fragmentos de interesse em questão de horas. Essa capacidade de amplificação foi um dos fatores que permitiu, a partir de 1990, o início de projetos como o sequenciamento do Genoma Humano, finalizado em 2000. Na botânica sistemática, as primeiras filogenias moleculares também surgiram nos anos 90 e inicialmente contavam com sequências de apenas um gene (Chase et al. 1993; Soltis et al. 1997).

A partir de 2010, essas tecnologias foram superadas pelas tecnologias de sequenciamento de nova geração (*Next generation sequencing* - NGS), como as plataformas Illumina ou Roche 454. Essa nova tecnologia permite sequenciar um volume de dados muito superior em comparação à tecnologia anterior. Desde então, passaram-se mais de 10 anos, e o termo "sequenciamento de nova geração" já não é mais adequado. Tecnologias superiores em relação ao volume de dados e tamanho dos fragmentos gerados, como PacBio e Nanopore, já estão disponíveis no mercado (sequenciamento de terceira geração), no entanto seus custos ainda são elevados. Assim, passou-se a usar o termo "sequenciamento de alto rendimento" ou sequenciamento de larga escala (*High Throughput Sequencing* - HTS), fazendo juz ao volume de dados obtidos em curtos períodos de tempo.

Esses avanços refletem em como entendemos as relações entre linhagens, pois cada vez mais dados são incluídos nas análises. Um exemplo clássico é o APG (*Angiosperm Phylogeny Group*). À medida que novos marcadores do cloroplasto (mas também alguns nucleares e mitocondriais) foram sendo incluídos nas inferências, diferentes relações filogenéticas foram sendo propostas, e novas edições publicadas (APG I - 1998, II - 2003, III - 2009, IV - 2016). Outro exemplo é o projeto *Tree of Life*, liderado pelo *Royal Botanic Gardens*, Kew (Reino Unido), que está inferindo uma árvore da

vida incluindo todas as plantas com flores, baseada em sequenciamento de alto rendimento de genes nucleares (Baker et al. 2022; Zuntini et al. 2024).

Podemos encontrar vários outros exemplos de inferências filogenéticas baseadas em sequenciamento de primeira geração e sequenciamento de alto rendimento em repositórios de artigos científicos, teses e dissertações. Neste capítulo, faremos uma introdução ao uso de dados genômicos em sistemática, comparando o sequenciamento de primeira geração com o sequenciamento de alto rendimento. Em seguida, abordaremos brevemente os princípios das técnicas mais aplicadas na área da sistemática recentemente que usam NGS na geração dos dados.

Sequenciamento Sanger e Sequenciamento de alto rendimento (HTS)

O sequenciamento genômico permite identificar a sequência de nucleotídeos de uma molécula de DNA ou RNA (Fietto & Maciel 2015). Ter acesso a essa sequência nucleotídica é o primeiro passo para a inferência filogenética baseada em dados moleculares, pois após alinhadas as sequências, a posição de cada um desses nucleotídeos constitui um carácter, e suas variantes (A, T, C, G) vão informar as relações evolutivas entre os táxons de interesse. Muitos métodos de sequenciamento e técnicas moleculares para obter sequências de diversas partes dos genomas estão disponíveis atualmente. Aqui, abordaremos dois dos métodos mais empregados em estudos de Sistemática: o sequenciamento Sanger e o sequenciamento de alto rendimento (HTS).

Sequenciamento Sanger

A técnica de sequenciamento Sanger consiste no sequenciamento de um único fragmento de interesse do genoma total por vez. Para que a síntese dos fragmentos de interesse aconteça, é necessário primeiro desenhar *primers* específicos para cada uma dessas regiões (marcadores). Esses *primers* se ligam à fita de DNA por similaridade de sequência, indicando à DNA-polimerase onde ela deverá iniciar a síntese da nova fita de DNA. Para cada região de interesse, os *primers* precisam ser desenhados tanto no sentido 5'-3' (*forward*, para frente) quanto 3'-5' (*reverse*, sentido reverso).

Assim, se temos quatro regiões de interesse, precisaremos desenhar quatro pares de *primers*. Geralmente os marcadores de interesse são selecionados com base na literatura, e a síntese dos *primers* é realizada por empresas especializadas.

Com os *primers* em mãos, podemos seguir com a preparação das reações de PCR. Na primeira reação, as fitas de DNA são desnaturadas e os fragmentos de interesse selecionados, de acordo com os *primers,* e amplificados para seu posterior sequenciamento. O DNA alvo (Fig. 1A) é adicionado em um tubo contendo DNA-polimerase e um par de fragmentos de DNA-iniciador (*primers*). Esse mix é levado ao termociclador para a Reação em Cadeia da Polimerase - PCR onde vão acontecer três

etapas ou ciclos. A primeira é a desnaturação da dupla fita do DNA alvo (Fig. 1B), originando duas fitas de DNA molde, que são então usadas como modelo na síntese das novas fitas. O segundo ciclo, é o anelamento dos *primers* ao DNA molde (Fig. 1C). Em seguida ocorre o terceiro ciclo ou amplificação do DNA onde a DNA-polimerase utiliza os desoxinucleotídeos na síntese de novas fitas (Fig. 1D). Como resultado desta primeira reação de PCR, as regiões de interesse foram amplificadas, podendo ser visualizadas em gel de agarose, e serão usadas na segunda etapa de PCR.

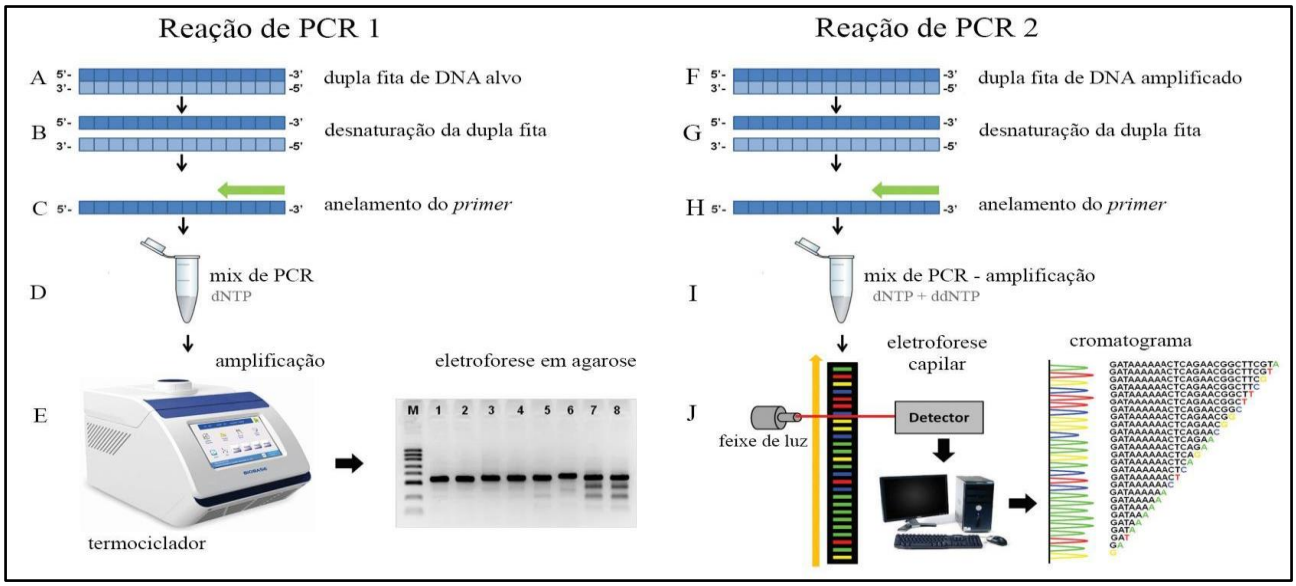


Figura 1: Sequenciamento Sanger diferenciando as etapas um (A-E) e dois (F-J) de reações de PCR. (A-B; F-J) desnaturação da dupla fita do DNA alvo; (C; G) anelamento dos *primers;* (D-E; I-J) transcrição da nova fita; (J) sequenciamento. Fonte: adaptado de Fietto & Maciel (2015).

Na segunda reação, um novo mix de PCR é preparado, incluindo os seguintes reagentes: o DNA amplificado (Fig 1F), uma enzima DNA-polimerase, o *primer*, uma solução tampão necessária ao funcionamento da DNA-polimerase, e os quatro nucleotídeos necessários para sintetizar as novas fitas de DNA; esses nucleotídeos são de dois tipos, desoxinucleotídeos (dATP, dTTP, dCTP, dGTP) e dideoxinucleotídeos ou terminadores de cadeia (ddATP, ddTTP, ddCTP, ddGTP), em menor concentração. Os dideoxinucleotídeos estão marcados com substâncias fluorescentes (fluoróforos) diferentes para permitir o processo de sequenciamento. O mix de reagentes é levado ao termociclador, e os ciclos são repetidos: desnaturação da dupla fita do DNA alvo (Fig. 1G), anelamento dos *primers* ao DNA molde (Fig. 1H), e amplificação (Fig. 1I).

No terceiro ciclo a DNA-polimerase utiliza desoxinucleotídeos e didesoxinucleotídeos na síntese da nova fita, sem distingui-los (Fig. 1D). Se um didesoxinucleotídeo é incorporado na cadeia recém sintetizada, o processo é interrompido. Após determinado tempo de reação, vários conjuntos de fragmentos de DNA sintetizados com tamanhos variados e terminando em um didesoxinucleotídeo corado são produzidos. Esses fragmentos são separados por eletroforese capilar no sequenciador. À

medida que os fragmentos viajam pelo capilar, um feixe de luz excita o corante ligado aos didesoxinucleotídeos. Um detector registra a cor fluorescente emitida por cada fragmento, e transmite essa informação para um software (Fig. 1J). Com isso, o software é capaz de determinar a ordem dos nucleotídeos na sequência de DNA original. No final do processo, o software gera um cromatograma correspondente à sequência de DNA complementar ao DNA molde utilizado (Fietto & Maciel 2015).

Sequenciamento de alto rendimento

Através do sequenciamento Sanger, os trabalhos em sistemática sempre focaram em poucos marcadores de evolução lenta para que *primers* conservados pudessem ser desenhados (Phillips 2001). Esse foco mudou com o advento das tecnologias de sequenciamento de alto rendimento, onde passou-se a utilizar grandes volumes de dados genômicos. O primeiro sequenciador de alto rendimento foi lançado em 2005 (Lemmon & Lemmon 2013), e também operava com base em coloração fosforescente. A partir daí outras metodologias foram desenvolvidas, como por exemplo, a plataforma Illumina (Fietto & Maciel 2015). O princípio do sequenciamento por esta plataforma é similar ao método proposto por Sanger, pois se baseia em bloquear a ação da DNA-polimerase associada com fluoróforo. Mas enquanto no método Sanger se usa didesoxinucleotídeos, o método Illumina usa bloqueadores da hidroxila do carbono 3 .

A diferença crítica entre os dois é o volume do sequenciamento. Enquanto Sanger sequencia apenas um único fragmento de DNA por vez, HTS sequencia milhões de fragmentos simultaneamente em cada corrida. HTS também oferece maior poder de descoberta para detectar variantes novas ou raras com sequenciamento profundo. Além disso, HTS corta o DNA e sequencia os fragmentos aleatoriamente sem focar em regiões alvo como no método Sanger. Outra questão apontada é o custo por nucleotídeo sequenciado. O custo estimado é de $500 por Mb para Sanger, e $0,10 a $6,50 por Mb para Illumina, dependendo do sequenciador (Frank et al. 2013).

Para sequenciar fragmentos de DNA com um método HTS, após a extração do DNA genômico, o primeiro passo consiste em preparar as bibliotecas a serem sequenciadas. Estas bibliotecas podem ser de diferentes tipos, e o protocolo para sua preparação vai depender tanto do tipo de dados de interesse (e.g. regiões codificantes do genoma nuclear) quanto da plataforma de sequenciamento a ser usada (e.g. Illumina). Usualmente o protocolo consiste em fragmentar o DNA, ligar os fragmentos de DNA a adaptadores em ambas as extremidades 5’ e 3’ (Fig. 2A), e amplificar esse conjunto (explicação detalhada na próxima seção).

Na plataforma Illumina, especificamente, o sequenciamento se inicia quando os adaptadores fixam as bibliotecas genômicas à placa de vidro do sequenciador (Fig. 2B), onde acontecerá todo o processo de amplificação. Nesta etapa são fornecidos apenas os desoxinucleotídeos (não marcados)

(Fig. 2C). Durante a amplificação, a extremidade da fita recém-sintetizada se anela ao *primer* na extremidade da fita molde (Fig. 2D) (Fietto & Maciel 2015). Posteriormente, ocorre uma elevação de temperatura desnaturando as duas fitas (Fig. 2E). Essas fitas encontram outros dois *primers,* reiniciando o processo (segundo ciclo) (Fig. 2F). A clonagem *in vitro* é finalizada após 35 ciclos, resultando em milhares de *clusters* (Fig. 2G-H), cada um representando um fragmento a ser sequenciado.

Biblioteca: coleção de fragmentos de DNA ou RNA que foram preparados e marcados para o sequenciamento.

Reads: sequências individuais geradas durante o sequenciamento. Cada *read* contém uma sequência de nucleotídeos que corresponde a uma parte do fragmento original.

Cluster: coleção de sequências de DNA amplificadas.

O sequenciamento pode gerar *paired-end reads*, com sequenciamento nas duas extremidades do fragmento de DNA (*forward* e *reverse*), ou *single-end reads*, com sequenciamento de apenas uma das extremidades do fragmento. O tamanho das sequências geradas depende do sequenciador utilizado, e pode variar entre 150 e 300 nucleotídeos (Rydmark et al. 2022).

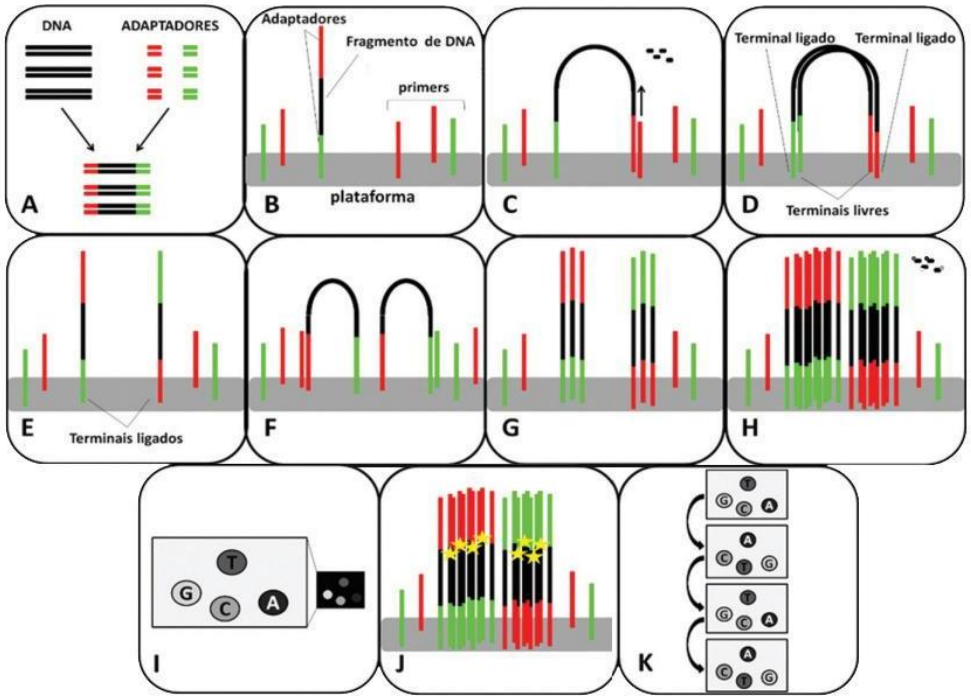


Figura 2: Sequenciamento de alto rendimento pela plataforma Illumina. (A) ligação dos adaptadores em ambas as extremidades dos fragmentos de DNA durante a preparação das bibliotecas. (B-K) Etapas do sequenciamento; (B) fixação dos adaptadores das bibliotecas à placa de sequenciamento; (C-D) amplificação (E) desnaturação das fitas; (F) fixação dos adaptadores livres aos adaptadores complementares na placa, iniciando um novo ciclo; (G) formação do *cluster*, contendo mais de um milhão de cópias do mesmo fragmento; (H) incorporação dos desoxinucleotídeos marcados e bloqueados; (I) captura da imagem e etapa de lavagem para remoção do grupo bloqueador presente na extremidade 3' junto com o fluoróforo; (J) repetição do ciclo; (K) as imagens registradas em cada ciclo são decodificadas para determinar a sequência de bases de cada *cluster* na placa. Fonte: adaptado de Fietto & Maciel (2015).

Sequenciamento: após os 35 ciclos de sequenciamento, é adicionada uma solução contendo os quatro tipos de desoxinucleotídeos marcados com fluorescência e a DNA-polimerase, que fará a incorporação dos desoxinucleotídeos (Fig. 2H) (Fietto & Maciel 2015). Raios laser excitam os fluoróforos ligados aos desoxinucleotídeos, emitindo uma fluorescência com intensidade proporcional ao número de fragmentos dos *clusters*. Uma imagem contendo a cor da fluorescência é capturada para cada posição dos *clusters* (Fig. 2I). Em seguida, a extremidade 3' é desbloqueada com consequente remoção dos reagentes em excesso e do fluoróforo do nucleotídeo incorporado no ciclo anterior, permitindo o início de um novo ciclo (Fig. 2J). Este processo se repete até que todas as bases de um determinado fragmento sejam determinadas. Por fim, as imagens das cores fluorescentes registradas em cada ciclo são decodificadas, determinando assim, a sequência de bases de cada cluster na placa (Fig. 2K) (Fietto & Maciel 2015).

Principais técnicas usadas para a obtenção de sequências de DNA em filogenômica

As tecnologias modernas de sequenciamento, como HTS, tornaram possível a geração de milhares de dados genômicos em larga escala. Essa mudança drástica na disponibilidade de dados, conjuntamente com os avanços nas ferramentas de bioinformática e a diminuição nos custos de sequenciamento, incrementaram, nas últimas décadas, o número de projetos colaborativos visando gerar filogenias robustas para diferentes grupos. Exemplos dessas colaborações incluem uma filogenia para todas as plantas e fungos como *the Plant and Fungal Tree of Life* (PAFTOL; Royal Botanic Gardens Kew, UK), a filogenia para todas as plantas flageladas (briófitas, licófitas, samambaias e gimnospermas) nominado *the Genealogy of Flagellate Plants* (GoFlag; University of Florida, USA), até projetos para produzir genomas de referência como *the 10 000 Plants Genomes Project* (10KP; China National GeneBank, China) e o Darwin Tree of Life (Wellcome Sanger Institute, Reino Unido), que objetiva sequenciar o genoma de 70 mil organismos eucariontes da Irlanda e Grã-Bretanha.

Esses projetos são possíveis graças ao desenvolvimento de diversas técnicas que permitem o sequenciamento de centenas de *loci* nucleares independentes, bem como sequências dos genomas plastidiais e mitocondriais (Pezzini et al. 2023; Mckain et al. 2018). Tais técnicas incluem, por exemplo, sequenciamento de fragmentos de DNA associados a sítios de restrição (RAD--seq), transcriptômica, *genome skimming* e *target capture*. Entre estas, *genome skimming* e *target capture* são comumente usadas na sistemática, filogenômica e biogeografia, enquanto RAD-seq e suas variantes são mais usadas para estudos de genética de populações, delimitação de espécies e análises de introgressão (Mckain et al. 2018). A seguir são abordados os objetivos, tipos de dados, vantagens

e limitações de cada uma dessas técnicas baseadas em sequências curtas (*short-read sequencing*), focando em *target capture* e sua utilidade na sistemática.

Sequenciamento de fragmentos de DNA associados a sítios de restrição (RAD--seq)

O nome RAD-seq reúne um conjunto de técnicas de sequenciamento de alto rendimento que utiliza enzimas de restrição (RE) para fragmentar o DNA genômico durante a preparação das bibliotecas. Esta técnica produz sequências para centenas ou milhares de regiões (codificantes ou não codificantes) ao longo do genoma associadas aos sítios de corte das RE (Fig. 3A), sem a necessidade de um genoma de referência (McKain et al. 2018). As *reads* resultantes, geralmente de 75–250 pb, formam pilhas (do inglês *scaffolds*) associadas aos sítios de restrição (Fig. 3A). O número de *reads* por pilha (cobertura) e a quantidade de *loci* sequenciados (representação) são determinados pelo tamanho do genoma do grupo de interesse, pela frequência dos sítios de restrição selecionados no genoma, e pelos parâmetros usados para a seleção por tamanho dos fragmentos (Peterson et al. 2012; Hühn et al. 2022).

Este tipo de sequenciamento depende do grau de conservação dos sítios de reconhecimento das RE para isolar fragmentos de DNA homólogos. Mutações sobre esses sítios levam à perda de informação (*missing data*). Consequentemente, amostras mais distantes filogeneticamente recuperam menos regiões compartilhadas (Peterson et al. 2012; Hühn et al. 2022). Outros fatores como a quantidade/qualidade do DNA inicial, o tamanho reduzido das *reads*, vieses no processo de amplificação e baixa qualidade do sequenciamento, também contribuem para a diminuição da quantidade média de sítios informativos por *loci,* além da uniformidade com que essa informação está distribuída entre as amostras de interesse (Eaton et al. 2017).

Locus (plural loci): posição específica em um cromossomo, onde está localizado um gene ou marcador genético.

Enzima de restrição: enzimas que reconhecem sequências nucleotídicas específicas e cortam a molécula de DNA nessas regiões.

Homólogo: compartilhado entre diferentes amostras/táxons por ancestralidade comum.

Pseudo-coalescência: quando sequências nucleotídicas de diferentes linhagens compartilham um ancestral comum mais recente. Indica ancestralidade entre linhagens.

SNP: variação na sequência do DNA que afeta um único nucleotídeo em uma posição específica no genoma.

Tais questões limitam a aplicabilidade dos dados gerados com RAD-seq na inferência de árvores de espécies a partir de árvores de genes. Isso acontece devido à pouca informação presente nos *loci* individuais e a disparidade nas amostras representadas (Guo et al. 2022; Hühn et al. 2022). Por esse motivo, RAD-seq tem sido amplamente utilizado em métodos de genética evolutiva baseados na identificação de Polimorfismos de Nucleotídeo Único (*single-nucleotide polymorphism* - SNP)

para o estudo de radiações recentes e delimitação de espécies, nos quais espera-se encontrar maior variação das regiões de DNA de interesse. Outra possibilidade são as análises de introgressão e reticulação, nas quais uma amostragem de milhares de genes ao longo do genoma é requerida para fornecer poder estatístico (McKain et al. 2018).

Avanços metodológicos buscando o aproveitamento máximo da quantidade de dados gerados a um baixo custo usando RAD-seq são cada vez mais frequentes. Por exemplo, aumentando a amostragem e profundidade do sequenciamento, pode-se aumentar a utilidade filogenética dos conjuntos de dados gerados com RAD-seq (Eaton et al. 2017). Outro exemplo é a combinação das técnicas RAD-Seq + *target capture* (Lang et al. 2020). Com algumas modificações na preparação das bibliotecas usando uma digestão dupla (ddRADseq; Fig. 3B) juntamente com ajustes nos parâmetros de seleção por tamanho dos fragmentos, é possível sequenciar um menor número de *loci* com uma maior cobertura e menor número de dados faltantes (Hühn et al. 2022).

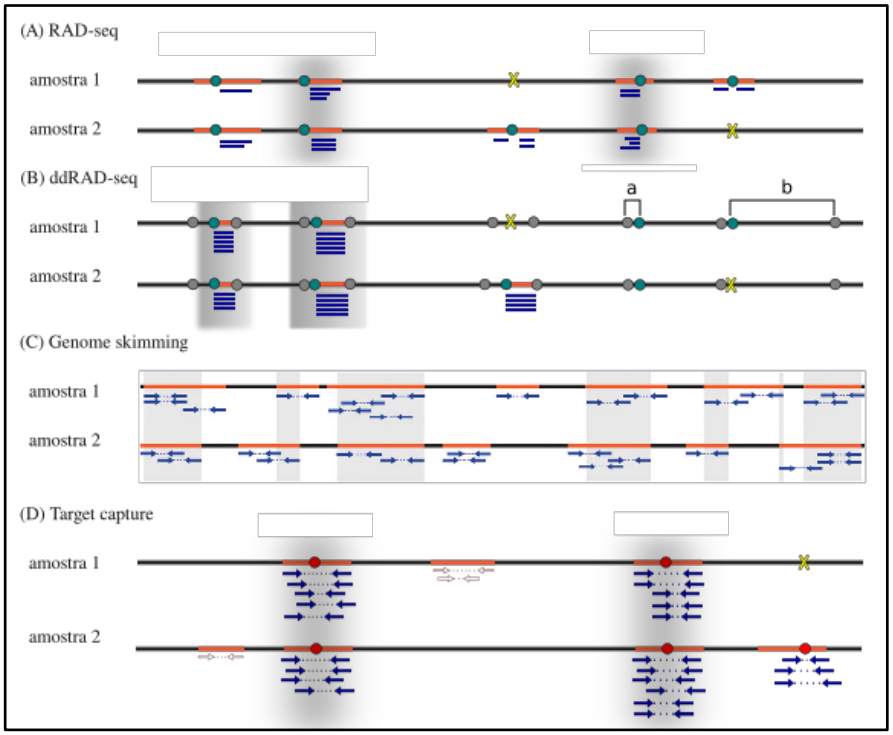


Figura 3: Comparação da representação e cobertura das diferentes técnicas de obtenção de dados genômicos usados na sistemática. (A) RAD-seq gera *reads* (linhas azuis) dos fragmentos do genoma (regiões laranjas) associados aos sítios de reconhecimento das RE (pontos verdes). Esses fragmentos foram previamente selecionados por tamanho, gerando uma representação reduzida (alguns *loci*) com alta cobertura do genoma total das amostras de interesse. Todas as *reads* começam nos locais de corte da RE e se estendem até atingir 75–300 pb. (B) O sequenciamento de digestão dupla ddRAD-seq usa duas RE para fragmentar o DNA, uma que reconhece regiões de corte comuns (pontos verdes) e outra que reconhece regiões de corte raros no genoma (pontos cinzas). Uma seleção precisa de tamanho para excluir regiões flanqueadas por sítios de reconhecimento muito próximos [a] ou muito distantes [b], gera bibliotecas compostas apenas por fragmentos do tamanho desejado (regiões laranjas). Assim, a cobertura é mais constante entre as amostras (linhas azuis), embora a representação ao longo do genoma seja menor. (A-B) Os sítios de reconhecimento da RE podem sofrer mutação em algumas amostras, impedindo que os fragmentos de DNA homólogos sejam sequenciados (xis amarelo). (C) *Genome skimming* gera dados com uma representação fragmentada do genoma completo com uma baixa cobertura (setas

azuis). Regiões homólogas são estabelecidas aleatoriamente por sobreposição das *reads* (caixas cinza). (D) As *reads* produzidas por *target capture* estão concentradas (setas azuis) em regiões homólogas ao longo do genoma nuclear (região cinza), estabelecidas nos locais onde as *baits* hibridizam (pontos vermelhos) com os fragmentos de DNA de interesse (regiões laranja). Consequentemente, a representação é proporcional à quantidade de regiões alvo incluídas nas *baits*, e a cobertura é alta. Fragmentos de DNA fora das regiões de interesse podem ser sequenciados gerando *reads off-target* (setas vermelhas). Em algumas amostras, a hibridização das *baits* com o genoma pode não funcionar, portanto, essas regiões de interesse não são representadas nos dados (xis amarelo). Fonte: adaptado de Peterson et al. (2012) e Hollingsworth et al. (2016).

Genome skimming

No *genome skimming* o genoma total de um organismo é sequenciado com uma baixa cobertura (e.g. 0.05X cobertura), isto significa que cada uma das regiões sequenciadas está representada por poucas *reads* (Fig. 3C). Como consequência, o DNA dos genomas que naturalmente apresentam múltiplas cópias, como o mitocondrial, plastidial, DNA ribossômico nuclear (nrDNA) e elementos altamente repetitivos (transposons), ficam muito melhor representados do que o DNA nuclear. DNA de contaminantes, patógenos, microbiomas e simbiontes também podem estar presentes (Mckain et al. 2018). Tanto o genoma da mitocôndria quanto o do cloroplasto podem ser montados total ou parcialmente a partir de dados gerados com esta técnica. No entanto, a qualidade e concentração das bibliotecas, a fonte do tecido (e.g. tecido fresco ou material de herbário) e a proporção entre os tamanhos dos genomas plastidiais e nuclear afetam o sucesso da montagem (Pezzini et al. 2023; Mckain et al. 2018). Plantas com genomas nucleares maiores geralmente requerem uma maior cobertura para aumentar a probabilidade de recuperação de plastomas ou mitogenomas completos usando esta técnica (Pezzini et al. 2023).

A identificação de genes nucleares a partir de dados de *genome skimming* usualmente é difícil devido à baixa cobertura oferecida pela técnica. Uma cobertura do genoma nuclear tão baixa quanto 0.01X pode ser suficiente para produzir dados úteis para estudos evolutivos, mas depende da complexidade e tamanho do genoma e do número de plastídios presentes no tecido do qual o DNA foi extraído (Straub et al. 2012). Recentemente alguns protocolos foram propostos para recuperar marcadores nucleares de baixa cópia (*Single-to-low-copy nuclear* SLCN) a partir de dados gerados com *genome skimming* (e.g. Reginato 2022).

Target capture

Embora a obtenção de genomas inteiros dos cloroplastos graças à técnica *genome skimming* revolucionou o código de barras do DNA, as regiões plastidiais nem sempre têm a variabilidade suficiente para resolver as relações filogenéticas, principalmente quando os táxons divergiram recentemente (Woudstra et al. 2022). Os genes nucleares são atualmente favorecidos na filogenômica por vários motivos: i) maior taxa de evolução em comparação com os genes plastidiais (Woudstra et

al. 2022) dada, entre outros processos, pela recombinação; ii) evolução independente (i.e., genes não ligados), que faz com que cada um dos genes sequenciados represente uma história evolutiva diferente dentro do genoma, os tornando ideais não só para fazer inferência filogenética (Pezzini et al. 2023), mas também para testar diferentes processos de evolução molecular no grupo de interesse como eventos de hibridização, duplicação, retenção de polimorfismo ancestral (*incomplete lineage sorting* - ILS), entre outros.

Diversos métodos chamados coletivamente de "*target capture*" ou "*target enrichment*" surgiram com o intuito de sequenciar com uma alta cobertura, múltiplos genes selecionados ao longo do genoma nuclear. Esses métodos incluem abordagens como *ultraconserved elements* - UCEs, *anchored phylogenomics*, *exon capture* e *Hyb-Seq*. Eles diferem nas regiões de DNA que visam capturar, sendo regiões genômicas de evolução lenta associadas a regiões flanqueantes variáveis como os UCEs, ou genes codificadores de proteínas como no caso do *exon capture*. De modo geral, esses métodos usam uma biblioteca de DNA genômico a partir da qual capturam e amplificam fragmentos dos genes de interesse usando um conjunto de *baits* pré-selecionadas (Mckain et al. 2018; Woudstra et al. 2022). Esse processo é conhecido como enriquecimento das bibliotecas. Os fragmentos de DNA selecionados são posteriormente sequenciados, enquanto o restante da biblioteca é descartado. Usualmente fragmentos de DNA diferentes aos genes de interesse são também sequenciados (sequências *off-target*), incluindo fragmentos dos genomas do cloroplasto e da mitocôndria (Fig. 3D), sendo possível montar parcialmente esses genomas a partir das *reads off-target* (e.g. *Hyb-seq*; Weitemier et al. 2014). Para aumentar a cobertura das regiões plastidiais, é possível incorporar um volume das bibliotecas descartadas nas bibliotecas enriquecidas antes do sequenciamento (*genome spiking*).

<u>*Cobertura*</u>: número médio de *reads* únicas que incluem um determinado nucleotídeo dentro do genoma.

<u>*Incomplete lineage sorting - ILS*</u>: cópias de genes ancestrais falham em coalescer em uma cópia ancestral comum até um ponto mais profundo do que eventos anteriores de especiação. Assim, uma árvore produzida por um único gene difere da árvore em nível de população ou espécie, produzindo uma árvore discordante.

<u>*Bait* ou *probe:*</u> fragmento de oligonucleotídeos de RNA ou DNA curtas (80–120 pb) desenhados para capturar fragmentos alvo numa biblioteca de DNA genômico.

Uma das principais vantagens da técnica *target capture* é sua aplicabilidade em espécimes de herbário, que usualmente contém o DNA muito degradado (Brewer et al. 2019). No entanto, um dos grandes desafios dessa técnica é a identificação de genes de baixa cópia - SLCN com variabilidade suficiente para informar a história evolutiva de táxons em diferentes níveis taxonômicos. Eventos de duplicação de genoma completo são muito comuns nos genomas nucleares das plantas. Só para as Angiospermas, cerca de 100 eventos já foram descritos (Landis et al. 2018). Além disso, o genoma

das plantas contém muitas sequências repetitivas e transposons, podendo constituir mais do 80% do genoma (Novák et al. 2020; Woudstra et al. 2022).

A disponibilidade de dados públicos do transcriptoma de mais de 1.400 espécies de plantas representativas dos grandes clados no projeto OneKP (Matasci et al. 2014) facilitou a identificação desses SLCN, simplificando o processo de desenho das *baits* usadas para capturá-los (Mckain et al. 2018). Diversos conjuntos de *baits* estão disponíveis no mercado atualmente, sendo específicos para capturar sequências de um grupo taxonômico particular como o *Compositae1061* ou COS, específico para Asteraceae, ou universais como o *Angiosperms353*, que inclui *baits* para capturar 353 *loci* em qualquer espécie dentro das angiospermas, ou o GoFlag que contém baits para 248 loci de briófitos, samambaias e coníferas.

Diversos estudos têm comparado o rendimento de *baits* específicas e universais no mesmo conjunto de amostras (e.g. Chau et al. 2018; Siniscalchi et al. 2021; Yardeni et al. 2022; Fonseca et al. 2023). Dado que os conjuntos de *baits* universais são desenhados para capturar sequências de grupos taxonomicamente afastados, eles favorecem regiões mais conservadas do genoma e, portanto, fornecem uma menor resolução em escalas de tempo microevolutivas, como espécies irmãs de divergência recente e populações (Pezzini et al. 2023). Apesar de terem potencialmente menor resolução, os kits universais permitem combinar dados gerados para diferentes grupos em diferentes projetos, possibilitando a obtenção de filogenias abrangentes e bem amostradas. Além disso, o desenvolvimento de kits específicos é geralmente mais custoso. Existem algumas alternativas para combinar os dados gerados com *baits* específicas e universais, aproveitando o melhor dos dois conjuntos de sondas. Uma alternativa é gerar *reads* com cada um dos conjuntos de *baits* independentemente, e depois analisar todos os dados conjuntamente (e.g. Shah et al. 2021). Nesses casos, o desafio é lidar com os dados faltantes devido aos táxons não compartilhados entre os conjuntos de dados. Mesmo assim é possível salvar tempo de laboratório e recursos, combinando as *baits* específicas e universais numa mesma reação de hibridização (Hendriks et al. 2021). Desta forma, os fragmentos de DNA capturadas com cada um dos conjuntos de *baits* são sequenciados ao mesmo tempo. Essa alternativa permite combinar novos dados com dados preexistentes gerados a partir de cada um dos conjuntos de *baits* independentemente, aproveitando todos os dados genômicos existentes no grupo de interesse para gerar filogenias melhor amostradas.

Protocolo de preparação das bibliotecas e sequenciamento usando *target capture* e Illumina

O protocolo padrão usado em estudos de sistemática para obtenção de sequências com *target capture* em sequenciador Illumina consiste nas seguintes etapas: i) extração de DNA genômico, ii)

preparação das bibliotecas, iii) captura dos fragmentos de DNA de interesse mediante hibridização, iv) sequenciamento.

Extração de DNA genômico

A extração de DNA pode ser feita a partir de tecido fresco, preservado em sílica-gel ou de herbário (Fig. 4A). Métodos como CTAB (brometo de cetiltrimetilamônio, Doyle & Doyle 1987) geram concentrações adequadas de DNA com um baixo custo. Para espécies com concentrações altas de metabólitos secundários e polissacarídeos, existem modificações do protocolo original, como CTAB + Sorbitol (Inglis et al. 2016), ou ainda protocolos específicos para plantas do cerrado, como Souza & Teixeira (2019). Após a extração, as amostras devem ser quantificadas com um método fluorimétrico que detecte especificamente as moléculas de DNA de dupla fita, como o Qubit, e também analisadas com um gel de agarose para determinar o tamanho dos fragmentos (Fig. 4B). Os protocolos de preparação das bibliotecas usualmente exigem fragmentos de DNA inicial de 50–600 pb a partir de apenas 120 ng de DNA genômico (Rydmark et al. 2022; Pezzini et al. 2023).

Preparação das bibliotecas

O protocolo de preparação das bibliotecas depende da plataforma de sequenciamento a ser usada e da técnica de interesse (e.g. RAD-seq, *target capture*). Para as plataformas Illumina, primeiro o DNA é fragmentado mecânica ou químicamente, gerando fragmentos de ~350 bp (Fig. 4B). O DNA proveniente de material de herbário geralmente está altamente degradado e não precisa de fragmentação. Cada fragmento de DNA genômico é ligado com três tipos diferentes de moléculas: i) um fragmento de DNA de ~80 pb chamado adaptador, onde os *primers* de amplificação vão se ligar, ii) duas cadeias de oligonucleotídeos chamados P5 e P7, que permitem a ligação dos fragmentos de DNA na superfície da placa do sequenciador, e iii) duas cadeias de nucleotídeos chamados *index* ou *barcodes* que, em conjunto, servem como um identificador único para cada uma das amostras (Rydmark et al. 2022). Esse conjunto de fragmentos de DNA genômico de uma amostra ligados com seus adaptadores e *barcodes* constituem uma biblioteca (Fig. 4C).

Existem diferentes formas de preparação dessas bibliotecas, seja usando kits disponíveis no mercado, como TruSeq (Illumina) e NEBNext Ultra II (BioLabs) ou protocolos não baseados em kits como o protocolo para DNA degradado (Troll et al. 2019). É possível usar a metade do volume sugerido por esses kits, reduzindo o custo por amostra sem perda significativa do rendimento (Rydmark et al. 2022; Shah et al. 2021). Para um entendimento detalhado do processo de preparação das bibliotecas a serem sequenciadas com a plataforma Illumina: https://www.illumina.com/techniques/sequencing/ngs-library-prep.html.

Captura dos fragmentos de DNA de interesse mediante hibridização com as baits selecionadas

As bibliotecas são hibridizadas junto com as *baits* pré-selecionadas por 16–48h a uma temperatura constante entre 60–65ºC, dependendo da especificidade do conjunto de *baits* e a complexidade das bibliotecas (https://arborbiosci.com/wp-content/uploads/2020/08/myBaits_v5.0_Manual.pdf; Woudstra et al. 2022). Os fragmentos de DNA de interesse unidos com as *baits* são capturadas por esferas magnéticas e amplificados por PCR. Fragmentos de DNA não hibridizados são geralmente descartados (Fig. 4E). Nessa etapa, determinar o número de ciclos de PCR usados para amplificar as bibliotecas hibridizadas é extremamente importante. Usualmente 8–14 ciclos são necessários para obter bibliotecas hibridizadas com uma concentração suficiente, geralmente ≥ 3 nM, para serem sequenciadas (Woudstra et al. 2022). Muitos ciclos de PCR podem gerar clones ou duplicatas desnecessárias e reduzir a diversidade das bibliotecas. Além disso, podem aumentar a chance de formação de dímeros de adaptadores.

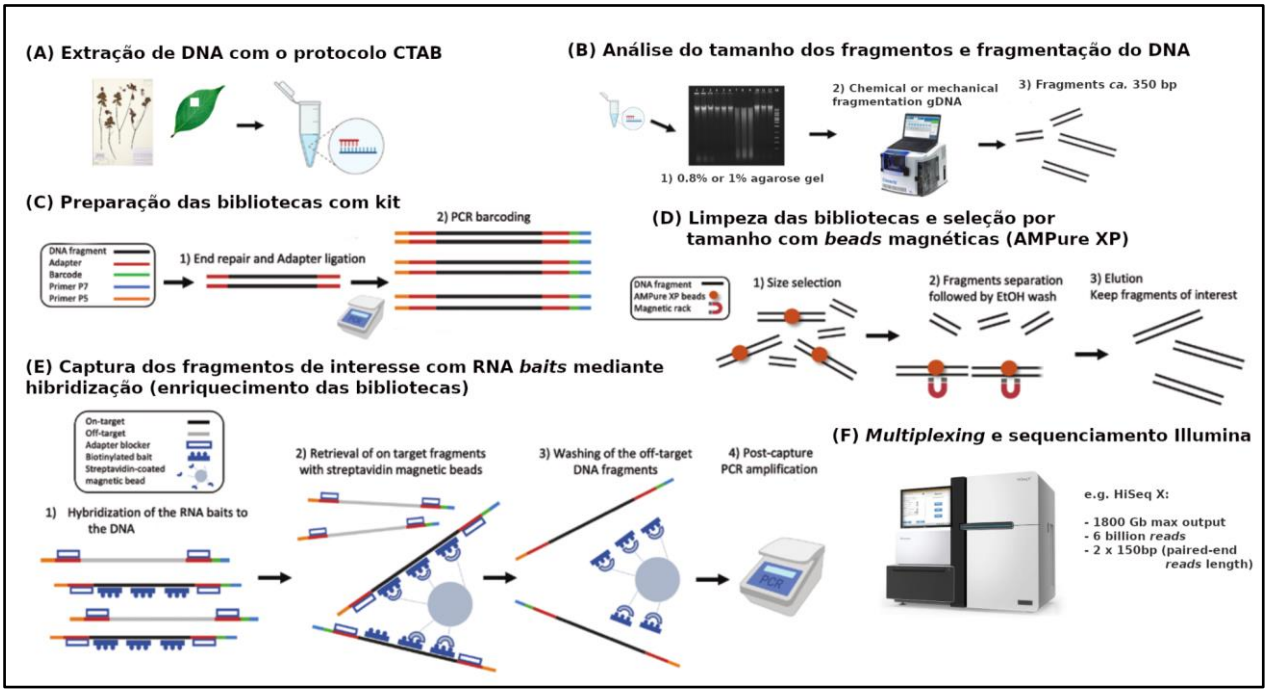


Figura 4. Protocolo de preparação de bibliotecas genômicas usando *target capture* e sequenciamento Illumina. (A) Extração de DNA. (B) Fragmentação do DNA e seleção por tamanho. (C) Ligação dos fragmentos de DNA com os adaptadores e os *barcodes*. (D) Seleção por tamanho das bibliotecas amplificadas. (E) Hibridização das bibliotecas com as *baits* pré-selecionadas, para capturar os fragmentos de DNA de interesse. (F) Sequenciamento. Fonte: Adaptado de Quatela et al. (2023).

Com o intuito de reduzir o número de reações de hibridização, e consequentemente o volume de *baits* usadas, é possível combinar bibliotecas de diferentes amostras num mesmo tubo para fazer uma hibridização simultânea (*pooling*). As amostras serão identificadas e separadas posteriormente utilizando os *barcodes* que foram ligados aos fragmentos de DNA previamente. Estudos focados em sistemática geralmente usam *pools* de 6, 8, 12, 24 ou até 48 amostras/bibliotecas por *pool*. O tamanho

dos *pools* depende da qualidade das bibliotecas, i.e. concentração, tamanho dos fragmentos e fonte do tecido (e.g fresco ou herbário). Embora o *pooling* seja uma estratégia eficiente para reduzir custos e tempo, alguns aspectos devem ser levados em consideração para não prejudicar o desempenho da captura. Bibliotecas com tamanhos de fragmento muito diferentes não devem ser combinadas, pois fragmentos menores tendem a ser hibridizados e sequenciados preferencialmente ficando sobre-representados. Do mesmo jeito, bibliotecas com uma alta concentração competirão pelas *baits* na reação de hibridização, desfavorecendo as bibliotecas menos concentradas. Por esse motivo, o número de fragmentos de DNA das bibliotecas a serem combinadas no mesmo *pool* deve ser igual (quantidades equimolares). Uma consequência de estratégias de *pooling* subótimas é a diminuição drástica da diversidade de *reads* após sequenciamento, deixando poucas amostras com muitas *reads* e algumas amostras não representadas (Rydmark et al. 2022; Woudstra et al. 2022).

Sequenciamento de alto rendimento com Illumina

Dependendo da plataforma, o sequenciamento pode gerar bilhões de *paired-end reads* de 150–300 pb (Rydmark et al. 2022). Uma forma eficiente de otimizar o resultado de uma corrida de sequenciamento e reduzir os custos é combinar diferentes bibliotecas previamente indexadas numa mesma corrida (*multiplexing*). A quantidade de bibliotecas por corrida depende da plataforma de sequenciamento a ser usada e da quantidade esperada de *reads* por amostra. Estratégias 96-plexes funcionam bem na maioria dos casos (Woudstra et al. 2022), podendo chegar até mais de 300 amostras. Empresas especializadas prestam o serviço de sequenciamento e enviam como resultado as *reads* geradas por amostra, o que significa que a empresa faz o *demultiplexing* usando a informação dos identificadores que foram ligados aos fragmentos de DNA na preparação das bibliotecas. As *reads* são armazenadas num formato especial chamado FastQ. O número de arquivos recebidos depende da estratégia de sequenciamento selecionada. No caso de sequenciamento *paired-end*, o usuário recebe dois arquivos por amostra, correspondentes com as *reads forward* e *reverse*. No caso de sequenciamento *single-paired*, somente um arquivo por amostra é gerado.

Hibridização: técnica usada na biología molecular para identificar e capturar sequências específicas de DNA ou RNA, mediante ligações de hidrogênio entre nucleotídeos complementares.

Dímero: molécula resultante da interação intermolecular de duas moléculas idênticas, i.e. que apresentam nucleotídeos complementares, que ficaram próximas por acaso numa reação.

Fluxo de trabalho para obtenção de sequências de DNA após sequenciamento usando *target capture*

Formato FastQ

As *reads* resultantes do HTS com Illumina são armazenadas em arquivos de texto especiais chamados *FastQ*, desenhados para conter a sequência de nucleotídeos e informação da qualidade com que cada um desses nucleotídeos foi sequenciado. Essa informação de qualidade é determinada com base na forma, resolução e outras propriedades dos picos obtidos do sequenciador (e.g. Fig. 1E), e transformada em um índice de qualidade chamado *Phred quality score* ou *Q score*. Tanto as letras correspondentes aos nucleotídeos quanto o índice de qualidade são codificados usando um único caractere do *American Standard Code for Information Interchange* - ASCII. Neste sistema as letras, algarismos, sinais de pontuação e de controle são representados usando somente um *byte* de informação (Fig. 5), o tornando extremamente compactado.

O índice de qualidade Q é uma expressão da probabilidade de que um nucleotídeo específico tenha sido sequenciado incorretamente (P_{error}). O valor de qualidade Q mínimo desejado está entre $Q = 30$, correspondente a uma probabilidade de erro $p = 0{,}001$ ou uma precisão de sequenciamento para esse nucleotídeo de 99,9%, e $Q = 20$ correspondente a uma probabilidade de $p = 0{,}01$ ou precisão de 99%. Dependendo do uso posterior dos dados e do nível de precisão desejado para a montagem das sequências, é possível utilizar esse índice de qualidade para filtrar/escolher entre as *reads* sequenciadas.

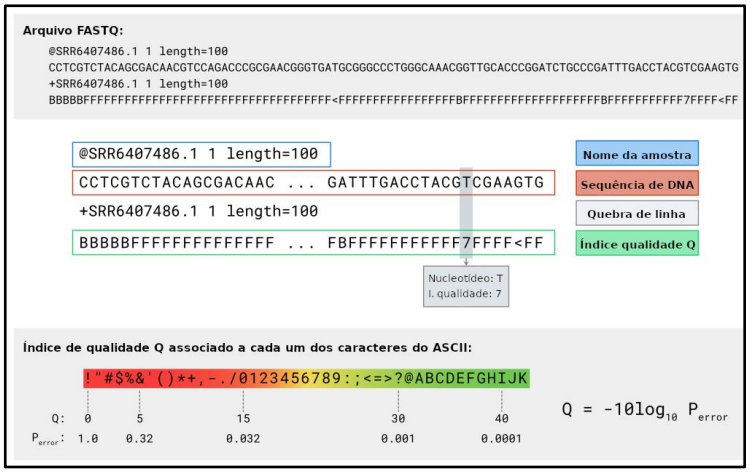


Figura 5. Estrutura do formato *FastQ*. No exemplo, o nucleotídeo T tem um índice de qualidade ou *Phred score* de $Q = 22$ representado com o símbolo 7 do sistema ASCII. Substituindo na equação, a probabilidade de que esse nucleotídeo T tenha sido sequenciado incorretamente é de $p = 0{,}006$. Para a tabela oficial de caracteres ASCII, acesse https://www.ascii-code.com/. Fonte: adaptado de https://gencoded.com/

Controle de qualidade e limpeza das reads

A análise de qualidade das *reads* geradas pelo sequenciador é importante para visualizar a precisão e o rendimento geral, tanto da preparação das bibliotecas quanto do sequenciamento. As métricas de qualidade fornecem informação sobre características pontuais das *reads*, como comprimento, composição de nucleotídeos, presença de sequências dos adaptadores, possíveis contaminantes, entre outros. Isso permite decidir entre os parâmetros a serem usados nas análises subsequentes (e.g. limpeza das *reads*), assim como descartar amostras que não atingiram os padrões de qualidade desejados. FastQC é um dos programas que permite visualizar a informação sobre a qualidade das *reads*. Embora os relatórios individuais sejam úteis, gerar estatísticas agregadas para todas as amostras é mais informativo. Uma forma de analisar os resultados gerados pelo FastQC é por intermédio do programa MultiQC, que produz um sumário comparativo e fácil de interpretar.

Sequências dos adaptadores podem estar presentes nas *reads* quando o fragmento de DNA sequenciado é muito curto. Isso acontece pois, na biblioteca, o fragmento de DNA de interesse vem logo depois do *primer* (Fig. 6), e caso o fragmento de DNA for menor que o número de bases sequenciadas pelo sequenciador (e.g. 250 nt), o sequenciamento vai continuar até atingir parte do adaptador que está na posição 3' (Fig. 6). Além disso, os adaptadores podem formar dímeros que são também sequenciados. A contaminação por adaptadores leva a erros na montagem das sequências e seu posterior alinhamento, motivo pelo qual devem ser detectados e removidos das *reads* para as análises subsequentes.

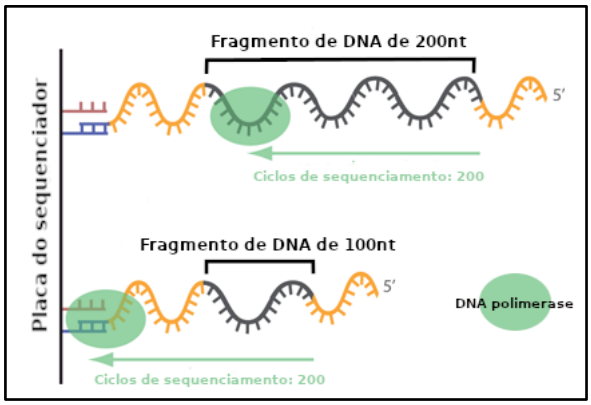


Figura 6. Representação das bibliotecas fixadas na placa do sequenciador Illumina. Acima, o tamanho do fragmento de DNA de interesse igual ou maior que o número de nucleotídeos sequenciados pelo sequenciador (250 nt). Abaixo, o fragmento de DNA de interesse menor (100 nt).

Outro fator que afeta a qualidade das *reads* é a alta quantidade de sequências duplicadas. Elas podem ser produto de um viés no processo de enriquecimento das bibliotecas, ou erros no processo de sequenciamento.

Programas como Trimmomatic ou Fastp permitem detectar tanto sequências de adaptadores quanto sequências duplicadas, e removê-las do conjunto de *reads*. Além disso, Trimmomatic permite

usar um valor de índice de qualidade Q e um valor de comprimento mínimo para filtrar as *reads* antes de continuar com a montagem das sequências. No caso de *reads* muito curtas, i.e. menores do que o comprimento de sequência próprio do sequenciador (Fig. 6B), as *reads* conterão sequências dos adaptadores que são chamadas de *palindrome*. Embora uma sequência completa de adaptador possa ser identificada com relativa facilidade, a identificação confiável de uma sequência parcial é difícil. No entanto, usar o modo *palindome* de Trimmomatic resulta numa deteção e remoção mais eficiente das sequências dos adaptadores. Exemplos do código necessário para rodar Trimmomatic podem ser consultados no manual do programa: *http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf*.

Montagem das reads

Após a etapa de limpeza, as *reads* que atingiram o padrão de qualidade e tamanho desejados são usadas para gerar sequências consenso dos *loci* de interesse para cada uma das amostras. Existem diversos *pipelines* como o HybPiper (Johnson et al. 2016) ou Captus (Ortiz et al. 2023) que usam uma coleção de programas e scripts (linhas de código) para gerar sequências a partir das *reads* (Fig. 7). No caso do Hybpiper, o primeiro passo consiste em alinhar as *reads* contra o conjunto de sequências de referência que foram usadas para desenhar as *baits* (*target file*), estabelecendo uma correspondência entre as *reads* e cada um dos *loci* de interesse, esse processo é conhecido como *mapping* (Johnson et al. 2016; Woudstra et al. 2022). Posteriormente, as *reads* de cada *locus* são comparadas e montadas num processo chamado (*assembly*), gerando uma sequência consenso (*contig*). Alguns programas fazem uma montagem das *reads* sem precisar das sequências de referência (*target file*), e posteriormente mapeiam os *contigs* contra as sequências de referência dos *loci* de interesse (e.g. HybPhyloMaker e Captus); esse processo é chamado *de novo assembly*. Tanto o número de *reads* mapeadas para cada uma das amostras com o comprimento das sequências geradas para cada um dos *loci* de interesse são usados para determinar a proporção de *reads on-target* e a porcentagem dos *loci* que foram sequenciados. Essas estatísticas são uma medida da eficiência do enriquecimento usando as *baits* selecionadas.

Pipeline: conjunto de ferramentas (programas, código) de processamento de dados conectados em série, onde a saída de um elemento é a entrada do próximo.

Transcriptoma: conjunto de sequências do RNA total em uma amostra biológica (RNAs mensageiros, RNAs ribossômicos, RNAs transportadores e os microRNAs).

Diferentes estudos encontraram que *baits* específicas geram maiores valores de *reads on-target* além de comprimentos maiores dos *loci* de interesse em comparação com *baits* universais. Para

um entendimento detalhado do funcionamento do HybPiper e exemplos do código para rodar as análises, consultar o tutorial do *pipeline*: https://github.com/mossmatters/HybPiper/wiki/.

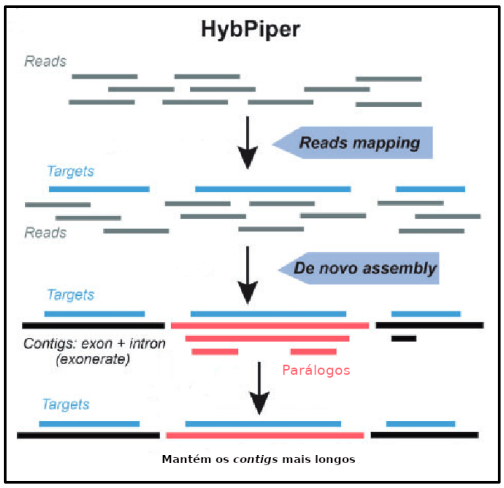


Figura 7. Representação resumida do fluxo de trabalho usado pelo HybPiper para gerar sequências dos genes de interesse a partir das *reads* produzidas pelo sequenciador.

No caso do uso das *baits* universais (*Angiosperms353*), espera-se que 353 *contigs* sejam gerados para cada uma das amostras. No entanto, diversos fatores fazem com que a montagem nem sempre seja completa (e.g. Fig. 3D), ou que mais de um *contig* por *locus* por amostra seja gerado. Quando diferentes *contigs* do mesmo comprimento e a mesma cobertura são gerados, *pipelines* como HybPiper os reconhecem como sequências parálogas (Fig. 7). Essas sequências podem representar variantes alélicas do mesmo gene, podem indicar contaminação entre bibliotecas ou de uma fonte desconhecida, ou podem ainda representar cópias diferentes do mesmo gene na amostra. Nesse último caso, o gene selecionado não é de cópia única, resultando de um processo de duplicação gênica ou um evento de duplicação do genoma completo (Johnson et al. 2016).

Diferenciar entre tipos de parálogos não é fácil, e sua presença na matriz de dados pode ser problemática na inferência filogenética. Por esse motivo, é crucial analisar a distribuição das sequências parálogas nos dados. Uma abordagem amplamente usada nos estudos de sistemática é remover os genes com sinal de paralogia das análises subsequentes. No entanto, essa abordagem é propensa à perda de informação valiosa que poderia ser usada para investigar eventos de duplicação de genoma completo, muito comuns em plantas (e.g. Morales-Briones et al. 2022). Outra alternativa, é analisar os genes com paralogia separadamente (e.g. Frost et al. 2023). Nos últimos anos, diferentes protocolos

Sequências ortólogas: sequências do mesmo gene em diferentes táxons.

Sequências parálogas: diferentes cópias de um gene no mesmo organismo, produto de um processo de duplicação do gene ou do genoma completo na espécie.

têm sido propostos para identificar e analisar sequências parálogas produto do *target capture* incluindo PPD, HybPhaser e Paragone-nf.

Embora a hibridização e o sequenciamento na técnica de *target capture* sejam feitos a partir do DNA genômico das amostras de interesse, a montagem das *reads* para gerar os *contigs* frequentemente foca apenas nas regiões codificantes dos *loci* selecionados. Isso ocorre pois as *baits* são usualmente desenhadas a partir dos transcriptomas das regiões de interesse. No entanto, os *contigs* gerados por HybPiper podem conter, além da região codificante (*exon*), parte das regiões flanqueantes dos exons, que não são codificantes (*introns*; Fig. 7). HybPiper usa Exonerate para extrair os *exons* das sequências montadas para cada um dos *loci* selecionados para cada uma das amostras, que são salvos num arquivo de formato fasta, e alinhado posteriormente para fazer inferência filogenética. As sequências dos *introns* podem ser também recuperadas para serem analisadas independentemente (Johnson et al. 2016). Estudos filogenéticos de espécies proximamente relacionadas ou a nível populacional podem se beneficiar da análise de *exons* e *introns* conjuntamente (*supercontigs*). Isso é possível pois os *introns* contém maior número de sítios variáveis e, portanto, informativos para as análises, enquanto os *exons* são geralmente mais conservados. No entanto, o alinhamento dos *introns* pode ser difícil, resultando em sítios variáveis não homólogos.

Estratégias de filtragem das sequências

Como mencionado anteriormente, nem sempre a captura e montagem das sequências funciona bem para todas as amostras e todas as regiões do genoma de interesse. Os resultados dependem não só do tipo e qualidade do material inicial, mas também da eficiência com que cada uma das etapas do processo foi realizada, desde a montagem das bibliotecas até as análises bioinformáticas. Por isso, após a montagem das sequências é comum que alguns dos genes de interesse (aqueles representados pelas *baits*) não tenham sido recuperados em nenhuma das amostras (usualmente poucos), ou que sequências parciais tenham sido montadas para algumas amostras em alguns genes, gerando matrizes de dados incompletas.

A remoção de genes em função de sua ausência em algumas das amostras parece não ter um efeito positivo na inferência filogenética (Molloy & Warnow 2018). No entanto, alinhamentos de matrizes com poucas amostras de táxons não proximamente relacionados tendem a ser computacionalmente mais difíceis e menos acurados. Além disso, alinhamentos pouco acurados incrementam o erro nas árvores de genes e o grau de conflito entre elas (Mirarab 2019). Não existem níveis ótimos de filtragem das matrizes que possam ser ajustados a qualquer conjunto de dados. Por esse motivo, é importante testar o efeito de diferentes estratégias de filtragem na inferência, tanto das árvores de genes quanto na árvore de espécies. Estudos empíricos em sistemática têm mostrado

melhoras na sustentação dos clados depois de remover amostras com menos de 50% do comprimento total esperado para a matriz de dados completa, no entanto outros estudos usaram um limite de 1⁄3 (e.g. Shah et al. 2021; Fonseca et al. 2023).

As abordagens mencionadas anteriormente podem ser combinadas de diversas formas no desenvolvimento de projetos filogenômicos, aliviando algumas das suas limitações e ampliando o poder de resolução dos dados gerados. Não existe uma resposta correta ou uma abordagem única para a elaboração de um estudo. O conhecimento da biologia do grupo de estudo (e.g. grupo de diversificação recente ou com altas taxas de hibridização), além dos dados genômicos disponíveis em repositórios públicos como EMBL ou NCBI relacionados com a pergunta de interesse, é fundamental para o sucesso do projeto e a optimização dos recursos.

Inferência filogenética usando dados genômicos

Após obter, filtrar e limpar as sequências temos duas opções principais para analisar esses dados, seja pela abordagem de árvores de genes, ou uma abordagem de supermatrix, concatenando todos os marcadores. As árvores de genes refletem o padrão de ancestralidade de um conjunto de cópias de genes homólogos derivadas de um único genoma. Existem diversos programas para inferir as árvores de genes diretamente das sequências alinhadas como RAxML (Stamatakis et al. 2005), IQ-TREE (Nguyen et al. 2015) e MrBayes (Ronquist & Huelsenbeck 2003). RAxML e IQ-TREE utilizam uma abordagem de máxima verossimilhança (*Maximum Likelihood*) onde a incerteza nos parâmetros estimados (e.g. topologia e comprimento dos ramos) é avaliada usando os valores de *bootstrap*. Com essa abordagem geralmente é possível tratar grandes conjuntos de dados contendo milhares de genes e táxons. Já a abordagem Bayesiana, implementada no MrBayes, provê uma estimativa direta de incerteza dos parâmetros, mas só consegue tratar um moderado número de dados, contendo dezenas ou centenas de genes e táxons.

A partir das árvores de genes é possível reconstruir uma árvore de espécies usando um método de sumarização como o ASTRAL (Mirarab et al 2014). Os dados de entrada no ASTRAL são um conjunto de árvores gênicas não enraizadas. A partir delas, o ASTRAL encontra a árvore de espécies que concorda com o maior número de quarteto de árvores induzidas pelo conjunto de árvores gênicas (Mirarab et al 2014). Quando comparado com outros métodos baseados em coalescência de multi-espécie, ASTRAL se mostrou mais estatisticamente consistente. Uma lacuna importante acerca desses métodos baseados em árvores de genes é que eles estimam apenas topologias, e não o comprimento dos ramos, muito importante dependendo da abordagem filogenética (Lemmon & Lemmon 2013). Além disso, esses métodos de sumarização são altamente sensíveis a erros na

estimação das árvores de genes (Molloy & Warnow 2018). Entretanto, os métodos baseados em árvores de genes permitem descrever a incongruência entre topologias independentes e a incerteza das relações em nós pontuais da árvore de espécies, melhorando nosso entendimento sobre os processos de evolução molecular no grupo de estudo como eventos de duplicação gênica, hibridização e retenção de polimorfismo ancestral (*incomplete lineage sorting*, ILS).

Outra possibilidade é a análise de supermatriz. A análise é realizada a partir dos alinhamentos de múltiplas regiões genômicas não relacionadas e concatenadas em uma única matriz, usando um modelo que assume que não há recombinações dentro destas matrizes. Árvores de supermatriz são difíceis de interpretar, pois não há garantia de que elas reflitam uma média ponderada igual para todas as árvores de genes, porque diferentes genes contêm quantidades diferentes de informação filogenética (Lemmon & Lemmon 2013). Contrário aos métodos de sumarização, os métodos de supermatriz geram informação sobre comprimento dos ramos. Entretanto, eles perdem sua acurácia quando os níveis de *incomplete lineage sorting* são muito altos (Molloy & Warnow 2018).

Para saber mais, acesse:

https://github.com/spreinalesl/IntroGenomics_BotInv2024

Referências

Baker, W. J. et al. 2022. A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life, *Systematic Biology* 71(2): 301–319. https://doi.org/10.1093/sysbio/syab035

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., Biggs, N., Cowan, R. S., Davies, N. M. J., Dodsworth, S., Edwards, S. L. et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science* 10:1102. https://doi.org/10.3389/fpls.2019.01102.

Chase, M. W., Soltis, D. E., Olmstead, R. G. et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden* 80: 528-580. https://doi.org/10.2307/2399846

Chau, J. H., Rahfeldt, W. A. & Olmstead, R. G. 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant Sciences* 6: e1032. https://doi.org/10.1002/aps3.1032.

Chen, M. Y., Liang, D., Zhang, P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. *Systematic Biology* 64: 1104–1120. https://doi.org/10.1093/sysbio/syv059.

Doyle, J. J. & Doyle, J. L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 66: 399–412. https://doi.org/10.1093/sysbio/syw092.

Fietto, J. L. R., & Maciel, T. E. F. 2015 Sequenciando genomas. IN: Moreira, L. M. (org.). Ciências genômicas: fundamentos e aplicações. Sociedade Brasileira de Genética, Ribeirão Preto, 403 p.

Fonseca, L. H. M., Asselman, P., Goodrich, K. R., Nge, F. J., Soulé, V., Mercier, K., Couvreur, T. L. P., Chatrou L. W. 2023. Truly the best of both worlds: merging lineage-specific and universal baiting kits to maximize phylogenomic inference. *bioRxiv preprint*. https://doi.org/10.1101/2023.11.16.567445.

Frank, M., Prenzler, A., Eils, R., & and von der Schulenburg, J. G. 2013. Genome sequencing: a systematic review of health economic evidence. *Health Economics Review* 3:29. https://doi.org/10.1186/2191-1991-3-29

Frost, L. Bedoya, A. M. & Lagomarsino, L. 2023. Strong phylogenetic signal despite high phylogenomic complexity in an Andean plant radiation (*Freziera*, Pentaphylacaceae). *bioRxiv preprint.* https://doi.org/10.1101/2021.07.01.450750.

Guo, C., Luo, Y., Gao, L.-M., Yi, T.-S., Li, H.-T., Yang, J.-B., and Li, D.-Z. 2023. Phylogenomics and the flowering plant tree of life. *Journal of Integrative Plant Biology* 65: 299–323. https://doi.org/10.1111/jipb.13415.

Heled, J. & Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27:570–80. https://doi.org/10.1093/molbev/msp274.

Hendriks, K. P., Mandáková, T., Hay, N. M., Ly, E., Hooft van Huysduynen, A., Tamrakar, R., Thomas, S. K. *et al.* 2021. The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences* 9: e11438. https://doi.org/10.1002/aps3.11438.

Hollingsworth, P. M., Li D.-Z., van der Bank, M., Twyford, A. D. 2016. Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions B* 371: 20150338. http://dx.doi.org/10.1098/rstb.2015.0338.

Hühn, P., Dillenberger, M. S., Gerschwitz-Eidt, M., Hörandl, E., Los, J. A., Messerschmid, T. F. E., et al. 2022. How challenging RADseq data turned out to favor coalescent-based species tree inference, a case study in *Aichryson* (Crassulaceae). *Molecular Phylogenetic and Evolution* 167: 107342. https://doi.org/10.1016/j.ympev.2021.107342.

Inglis, P. W., Pappas, M. C. R. & Grattapaglia, D. 2016. Protocolo de Extração de DNA e RNA de Alta Qualidade para Espécies Ricas em Compostos Secundários. Comunicado Técnico 204, Embrapa, Brasília, 5 p.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. C. & Wickett, N. J. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Applications in Plant Sciences* 4: 1600016. https://doi.org/10.3732/apps.1600016.

Landis, J. B., D. E. Soltis, Z. Li, H. E. Marx, M. S. Barker, D. C. Tank, and P. S. Soltis. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* 105: 348–363. https://doi.org/10.1002/ajb2.1060.

Lang, P. L. M., Weiß, C. L., Kersten, S., et al. 2020. Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Molecular Ecology Resources* 20: 1228–1247. https://doi.org/10.1111/1755-0998.13168.

Lemmon, E. M. & Lemmon, A. R. 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 44: 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822.

McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6: e1038. https://doi.org/10.1002/aps3.1038.

Mirarab, S. 2019. Species Tree Estimation Using ASTRAL: Practical Considerations. Quantitative Biology, Populations and Evolution. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.03826.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: i541–i548. https://doi.org/10.1093/bioinformatics/btu462.

Molloy, E. K. & Warnow, T. 2018. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology* 67: 285–303. https://doi.org/10.1093/sysbio/syx077.

Morales-Briones, D. F., Berit Gehrke, Chien-Hsun Huang, Aaron Liston, Hong Ma, Hannah E Marx, David C Tank, Ya Yang. 2022. Analysis of Paralogs in Target Enrichment Data Pinpoints Multiple Ancient Polyploidy Events in *Alchemilla* s.l. (Rosaceae), *Systematic Biology* 71: 190–207. https://doi.org/10.1093/sysbio/syab032.

Nguyen, L.--T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. 2015. IQ--TREE: a fast and effective stochastic algorithm for estimating maximum--likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–-274. https://doi.org/10.1093%2Fmolbev%2Fmsu300.

Ortiz, E. M., Hoewener, A., Shigita, G., Raza, M., Maurin, O., Zuntini, A., Forest, F., Baker, W. J., Schaefer, J. 2023. A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in Cucurbitales. *bioRxiv preprint*. https://doi.org/10.1101/2023.10.27.56436.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., Hoekstra, H. E. 2012. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* 7: e37135. https://doi.org/10.1371/journal.pone.0037135.

Pezzini, F. F. , Ferrari G., Forrest L. L., Hart M. L., Nishii K., and Kidner C. A. 2023. Target capture and genome skimming for plant diversity studies. *Applications in Plant Sciences* 11: e11537. https://doi.org/10.1002/aps3.11537.

Phillips, A. J. 2001. Comparative Phylogenomics: A Strategy for High-throughput Large-scale Sub-genomic Sequencing Projects for Phylogenetics Analysis. IN: DeSalle, R., Giribet, G., & Wheeler, W. (eds.)

Techniques in Molecular Systematics and Evolution. Methods and Tools in Biosciences and Medicine Collection. Springer Basel AG, 407 p. https://link.springer.com/book/10.1007/978-3-0348-8125-8.

Quatela, A. S., Cangren, P., Jafari, F., Michel, T., de Boer, H. J., Oxelman, B. 2023. Retrieval of long DNA reads from herbarium specimens. *AoB Plants* 15: 1–11. https://doi.org/10.1093/aobpla/plad074.

Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19: 1572–74. https://doi.org/10.1093/bioinformatics/btg180.

Rydmark, M. O., Woudstra, Y., de Boer, H. 2022. Chapter 9. Sequencing platforms. In: de Boer, H., Rydmark, M. O., Verstraete, B., Gravendeel, B. (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875.

Shah, T., Schneider, J. V., Zizka, G., Maurin, O., Baker, W., Forest, F., Brewer, G. E., Darbyshire, I. & Larridon, I. 2021. Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits. *American Journal of Botany* 108: 1–16. https://doi.org/10.1002/ajb2.1682.

Siniscalchi, C. M., Hidalgo, O., Palazzesi, L., Pellicer, J., Pokorny, L., Maurin, O., Leitch, I. J., *et al.* 2021. Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences* 9: e11422. https://doi.org/10.1002/aps3.11422.

Souza, D. C. & Teixeira, T. A. 2019. A simple and effective method to obtain high DNA quality and quantity from Cerrado plant species. *Molecular Biology Reports*, 46: 4611–4615. https://doi.org/10.1007/s11033-019-04845-0

Stamatakis, A., Ludwig T. & Meier, H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456–63. https://doi.org/10.1093/bioinformatics/bti191.

Troll, C. J., Kapp, J., Rao, V., Harkins, K. M., Cole, C., Naughton, C., Morgan, J. M., Shapiro, B., Green, R. E. 2019. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics* 20: 1023. https://doi.org/10.1186/s12864-019-6355-0.

Woudstra, Y., Quatela, A. S., Kidner, C., Viruel, J., Zuntini, A., Martin, M. D., Michel T., Grace, O. M. 2022. Chapter 14. Target capture. In: de Boer, H., Rydmark, M. O., Verstraete, B., Gravendeel, B. (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875.

Yardeni, G., J. Viruel, M. Paris, J. Hess, C. Groot Crego, M. de La Harpe, N. Rivera, et al. 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources* 22: 927–945. https://doi.org/10.1111/1755-0998.13523.

Zuntini, A. R., Carruthers, T., Maurin, O. et al. 2024. Phylogenomics and the rise of the angiosperms. *Nature* 629: 843–850. https://doi.org/10.1038/s41586-024-07324-0.