UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-graduação em Bioinformática

Renato Renison Moreira Oliveira

DESENVOLVIMENTO E COMPARAÇÃO DE FERRAMENTAS E PIPELINES PARA ANÁLISES ÔMICAS NOS ESTUDOS DE CONSERVAÇÃO DA DIVERSIDADE VEGETAL

Belo Horizonte
2023

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-graduação em Bioinformática

Renato Renison Moreira Oliveira

DESENVOLVIMENTO E COMPARAÇÃO DE FERRAMENTAS E PIPELINES PARA ANÁLISES ÔMICAS NOS ESTUDOS DE CONSERVAÇÃO DA DIVERSIDADE VEGETAL

Tese apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Doutor em Bioinformática.

Orientador: Guilherme Oliveira

Co-orientador: Thomas Sicheritz-Pontén

Belo Horizonte
2023

043 Oliveira, Renato Renison Moreira.

Desenvolvimento e comparação de ferramentas e pipelines para análises ômicas nos estudos de conservação da diversidade vegetal [manuscrito] / Renato Renison Moreira Oliveira. – 2023.

139 f. : il. ; 29,5 cm.

Orientador: Dr. Guilherme Oliveira. Coorientador: Dr. Thomas Sicheritz-Pontén.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Conservação da Diversidade Biológica. 3. Genoma de Planta. 4. Filogenia. I. Oliveira, Guilherme. II. Sicheritz-Pontén, Thomas. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Jéssica Patrícia Silva de Sá – CRB 6 - 3430

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DE DEFESA DE TESE

RENATO RENISON MOREIRA OLIVEIRA

Às nove horas do dia **16 de agosto de 2023**, reuniu-se, através de videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Desenvolvimento e comparação de ferramentas e pipelines para análises ômicas nos estudos de conservação da diversidade vegetal"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Guilherme Corrêa de Oliveira**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Guilherme Corrêa de Oliveira - Orientador	Instituto Tecnológico Vale	**Aprovado**
Dra. Ana Maria Benko Iseppon	Universidade Federal de Pernambuco	**Aprovado**
Dr. Alessandro de Mello Varani	Universidade Estadual Paulista	**Aprovado**
Dra. Thannya Nascimento Soares	Universidade Federal de Goiás	**Aprovado**
Dr. Henrique Vieira Figueiró	George Mason University	**Aprovado**

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 16 de agosto de 2023.

Documento assinado eletronicamente por **Thannya Nascimento Soares**, **Usuária Externa**, em 16/08/2023, às 11:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Maria Benko Iseppon**, **Usuário Externo**, em 16/08/2023, às 12:46, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alessandro de Mello Varani**, **Usuário Externo**, em 16/08/2023, às 14:11, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Henrique Vieira Figueiró**, **Usuário Externo**, em 17/08/2023, às 08:42, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Guilherme Corrêa de Oliveira**, **Usuário Externo**, em 22/08/2023, às 17:27, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2472466** e o código CRC **1007AA0B**.

Referência: Processo nº 23072.243697/2023-68 SEI nº 2472466

AGRADECIMENTOS

Aos meus amados pais, José Renato Siqueira Oliveira (Dedé, “in memorian”) e Maria Helena Moreira Oliveira, por nunca terem medido esforços para me proporcionarem uma boa educação e ser uma pessoa de bem, sempre me apoiando nas decisões ousadas da vida.

À minha irmã, Maria Oliveira, que me deu apoio e ficou do meu lado em momentos difíceis, e que hoje com sua família feita, continua me trazendo muitas alegrias, como as minhas sobrinhas Clara e Eduarda.

Aos meus tios, tias, primos e primas que se mantiveram juntos a mim e à minha família em momentos difíceis.

Aos amigos que fiz no antigo CEFET, que me fizeram ver que havia um mundo todo além de Icoaraci, e cuja amizade se mantém forte até hoje: Gizele Abdon, Derick Rosa e Felipe Freitas.

Aos amigos que de alguma forma presenciaram minhas preocupações com o doutorado: Derick Rosa (de novo), Marissa Brasil, Michell Cruz, Raíssa Lorena, Helder Paraense.

Ao meu orientador, Prof. Dr. Guilherme Corrêa de Oliveira, por ter me aceitando como aluno de doutorado e ter sido um ótimo orientador, sempre aberto às minhas opiniões e sugestões na pesquisa acadêmica e ter sempre confiado em mim. Ao meu co-orientador Thomas Sicheritz-Pontén e ao colaborador Bent Petersen, por terem me auxiliado em todo o processo de utilização dos clusteres computacionais instalados em Copenhague.

A todos os professores que fizeram parte da minha educação acadêmica e que de certa forma me inspiram a ser um bom pesquisador, em especial ao Professor Miguel Ortega, por ter me recebido muito bem em seu laboratório Biodados, na UFMG-ICB.

Ao Programa Interunidades de Pós-graduação em Bioinformática, em especial aos coordenadores do programa e aos secretários Thiago e Sheyla, sempre bem dispostos a ajudar quando precisamos de qualquer coisa.

Aos amigos feitos no laboratório Biodados, Fenícia Brito, Arthur Fonseca, Lissur Orsine, Beatriz Koury e Fernanda Stussi.

Aos amigos feitos na cidade de Belo Horizonte, Rodrigo Profeta, Amanda Emerich, Caio Herbert, Mônica Rezende, João Almeida, Bruno Marques, por terem feito parte dos meus quatro anos e meio de vivência em BH, tanto para lazer quanto para conversas acadêmicas e filosóficas.

Ao Instituto Tecnológico Vale, pelos anos de vínculo que foram essenciais para a minha formação acadêmica, permitindo conhecer pesquisadores de excelência e aprender bastante com eles.

Aos meus amigos de trabalho que sempre estiveram presentes, dando conselhos não só sobre a pesquisa e ciência, mas também conselhos de vida: Santelmo Vasconcelos, Gisele Nunes, Mariana Costa, Diego Sotero, Guilherme Oliveira, Rafael Valadares e José Bittencourt.

A todos que porventura não estão com o nome aqui, mas que me ajudaram de alguma forma nesta caminhada, meus eternos agradecimentos.

RESUMO

As plantas desempenham um papel fundamental na manutenção da vida no planeta, e diante de ameaças, principalmente antrópicas, é essencial conduzir estudos de conservação para garantir o conhecimento, proteção e perpetuação das espécies. Entre as diversas abordagens nesses estudos, se destacam a Genética e Genômica da Conservação, sendo esta última capaz de fornecer informações que permitam um melhor entendimento de características adaptativas, variabilidade genética e estudos adaptativos sob diferentes estresses ambientais. A Genômica da Conservação ganhou maior relevância com o surgimento das tecnologias de sequenciamento, permitindo um crescimento exponencial na quantidade de genomas depositados nos bancos de dados públicos. Apesar dos custos de sequenciamento terem diminuído consideravelmente e a quantidade de dados gerados ter aumentado, montar genomas grandes de eucariotos, especialmente plantas, ainda é uma atividade complexa e custosa. Características inerentes aos genomas de plantas, como poliploidia, tamanho do genoma e regiões repetitivas ainda causam problemas nos processos de montagem desses genomas. Para solucionar tais problemas, uma grande diversidade de *softwares* montadores foram desenvolvidos, cada um funcionando de uma forma específica. Além do desenvolvimento de diversos *softwares* montadores para superar esses problemas em montagens de genomas grandes e complexos, outras técnicas que vem sendo desenvolvidas e bastante utilizadas nos estudos de biodiversidade são as técnicas de *barcoding* e *metabacording,* que juntas permitem a identificação rápida e precisa de espécies presentes em amostras ambientais, além de atuarem na realização de estudos evolutivos. Este trabalho identificou três necessidades nas áreas de genômica ambiental e, por meio da elaboração de três artigos, procurou suprir tais necessidades. O primeiro artigo apresenta o SPLACE, uma ferramenta que permite o alinhamento e concatenação de genes de forma automatizada, rápida e precisa, tratando os dados faltantes e gerando uma supermatriz que pode ser utilizada para a inferência de árvores filogenômicas. O segundo artigo apresenta o PIMBA, um pipeline para realizar análises de *metabarcoding* de forma rápida e abrangente, permitindo escolher diferentes abordagens de análise e seus diversos parâmetros, além de possibilitar o uso de bancos de dados de referência próprios, além dos bancos tradicionais existentes e do GenBank. Por fim, foi produzido um manuscrito onde buscou-se suprir a necessidade de trabalhos na literatura que fizessem uma comparação justa entre os diversos montadores disponíveis ao se montar os genomas de várias espécies de plantas, além de reunir informações sobre os genomas nucleares completos de plantas que já foram analisados. Estes três trabalhos publicados auxiliaram em diversos estudos importantes no planejamento de estratégias para a conservação da diversidade vegetal, podendo também ser aplicadas a outros grupos de organismos.

PALAVRAS-CHAVE: Conservação, biodiversidade, genoma de plantas, *barcoding*, *,metabarcoding*, supermatriz, filogenômica.

ABSTRACT

Plants play a key role in maintaining life on the planet, and in the face of threats, mainly anthropogenic, it is essential to conduct conservation studies to ensure the knowledge, protection, and perpetuation of species. Among the various approaches in these studies, Conservation Genetics and Genomics stand out, the latter being able to provide information that allows a better understanding of adaptive characteristics, genetic variability and adaptive studies under different environmental stresses. Conservation Genomics gained greater relevance with the emergence of sequencing technologies, allowing an exponential growth in the amount of genomes deposited in public databases. Although the costs of sequencing have decreased considerably and the amount of data generated has increased, assembling large genomes of eukaryotes, especially plants, is still a complex and costly activity. Inherent characteristics of plant genomes, such as polyploidy, genome size, and repetitive regions still cause problems in the assembly processes of these genomes. To solve these problems, a great diversity of assembly software has been developed, each one working in a specific way. Besides the development of several software assemblers to overcome these problems in assembling large and complex genomes, other techniques that have been developed and widely used in biodiversity studies are the barcoding and metabarcoding techniques, which together allow the rapid and accurate identification of species present in environmental samples, as well as to perform evolutionary studies. This work identified three needs in the areas of environmental genomics and, through the elaboration of three articles, sought to supply those needs. The first paper introduces SPLACE, a tool that allows for the alignment and concatenation of genes in an automated, fast, and accurate manner, addressing missing data and generating a supermatrix that can be used for the inference of phylogenomic trees. The second paper presents PIMBA, a pipeline to conduct metabarcoding analyses quickly and comprehensively, allowing users to choose different analysis approaches and their various parameters, and also facilitating the use of one's own reference databases, in addition to the existing traditional databases and GenBank. Lastly, a manuscript was produced aiming to fill a gap in the literature by fairly comparing the various genome assemblers available for several plant species and gathering information about the complete nuclear genomes of plants that have already been analyzed. These three published works assisted in various important studies planning strategies for the conservation of plant diversity and can also be applied to other groups of organisms.

KEYWORDS: Conservation, biodiversity, plant genome, *barcoding*, *,metabarcoding*, supermatrix, phylogenomics.

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ASV	Amplicon Sequence Variant
AT	Adenina/Timina
COI	Citocromo C Oxidase
DNA	Desoxirribonucleic acid
GC	Guanina/Citosina
ITS	Internal Transcribed Spacer
ITV	Instituto Tecnológico Vale
LINE	Long Interspersed Nuclear Element
MPEG	Museu Paraense Emílio Goeldi
NCBI	National Center for Biotechnology Information
OLC	Overlap-Layout-Consensus
OTU	Operational Taxonomic Unit
RAD	Recperação de Áreas Degradadas
RNA	Ribonucleic acid
SINE	Short Interspersed Nuclear Element
SSR	Simple Sequence Repeats
TE	Elementos Transponíveis
TGS	Third-Generation Sequencing

SUMÁRIO

1 INTRODUÇÃO

As plantas possuem um papel vital para o ecossistema e a humanidade, servindo como produtores primários e fonte de alimentos (STUART CHAPIN; MATSON; VITOUSEK, 2012; TILMAN et al., 2011), atuando na produção de oxigênio e no sequestro de carbono (BONAN, 2008), nos serviços ecossistêmicos (COSTANZA et al., 1997), e contribuindo para a manutenção da diversidade animal (PIMM et al., 1995). A diversidade de espécies de plantas é muito alta, constando cerca de 410 mil espécies (https://www.catalogueoflife.org/data/metadata). Diante da importância que as plantas possuem, é essencial que sejam feitos esforços visando sua manutenção e conservação.

Estudos de conservação da biodiversidade consistem na investigação das várias espécies que compõem um ecossistema, assim como suas interações, papéis na manutenção da saúde, entendimento e preservação do ecossistema (CARDINALE et al., 2012). O crescimento nos estudos de conservação de plantas ajudou a entender os papéis que as plantas possuem nas funções e serviços ecossistêmicos. O entendimento das relações entre as espécies de plantas e os papéis que elas cumprem, por sua vez, ajudam a desenvolver estratégias de conservação que ajudem a manter os serviços ecossistêmicos (BALVANERA et al., 2006; BRUMMITT; ARAÚJO; HARRIS, 2021).

Algumas ameaças às espécies de plantas, como a destruição de habitat, surgimento de espécies invasoras e mudanças climáticas, podem ser identificadas por meio dos estudos de conservação (BROOKS et al., 2002). Após identificadas, tais ameaças podem ser estudadas, promovendo o desenvolvimento de estratégias que ajudem a mitigar seus impactos e a priorizar esforços de conservação para espécies que sejam identificadas como mais vulneráveis (MACK et al., 2000; NIC LUGHADHA et al., 2020). Por conta dessas ameaças, se torna imprescindível a preservação de espécies de plantas por meio do conhecimento genético de sua diversidade, algo crucial para a sobrevivência a longo prazo e adaptação das espécies às mudanças do ambiente (JUMP; MARCHANT; PEÑUELAS, 2009), além da identificação genética de espécies e populações de plantas que devem ser foco das atividades de conservação, para que os genes se mantenham disponíveis às próximas gerações (DIRZO; CEBALLOS; EHRLICH, 2022).

Dentre as diferentes abordagens existentes para se realizar estudos de conservação da biodiversidade, duas serão abordadas neste trabalho: Genética da Conservação e Genômica da Conservação. Ambas as abordagens visam preservar a diversidade genética das espécies e das populações, o que se torna essencial para suas sobrevivências, adaptação e potencial genético.

A Genética da Conservação usa técnicas moleculares, como o sequenciamento de DNA e microssatélites, para estudar a diversidade genética, fluxo gênico e a estrutura da população (FRANKHAM; BALLOU; BRISCOE, 2010). É uma abordagem que permite identificar populações geneticamente únicas ou distintas, avaliar os níveis de endogamia e os impactos da fragmentação do habitat no fluxo gênico. Os resultados obtidos por estudos de Genética da Conservação podem ajudar na formação de estratégias como translocação e reintrodução de espécies, além do estabelecimento de corredores de vida selvagem, mantendo a diversidade genética e promovendo a viabilidade da população (ALLENDORF; HOHENLOHE; LUIKART, 2010a).

A Genômica da Conservação é mais recente e surgiu com o aparecimento das tecnologias de sequenciamento de nova geração (NGS), visando realizar o sequenciamento de genomas completos, estudos de transcriptoma e anotação funcional dos genomas sequenciados, e direcionando essas análises e resultados obtidos aos estudos de conservação (SHAFER et al., 2015). Esta abordagem pode fornecer informações sobre a base genética de traços adaptativos, os efeitos da variabilidade genética na adaptação e o papel da adaptação local na conservação de espécies. Pela identificação de genes associados a tipos de adaptações específicas ou estresses ambientais, a Genômica da Conservação pode ajudar a priorizar espécies para esforços de conservação, planejar estratégias de manejo mais eficazes, além de permitir a avaliação do sucesso de intervenções de conservação (FUNK et al., 2012).

As áreas de Genética e Genômica da Conservação contribuem enormemente para os estudos de conservação da biodiversidade, permitindo que se reduza o risco de extinção de espécies, reduzindo incertezas taxonômicas (por meio de técnicas de *barcoding* e *metabarcoding*) e conhecimento dos genes de uma população, por meio da montagem do genoma completo (OUBORG et al., 2010). Muitas das análises feitas nos campos da Genética e Genômica da Conservação utilizam técnicas de bioinformática, que acaba apresentando um papel crucial na análise e interpretação dos dados genéticos e genômicos coletados (ALLENDORF; HOHENLOHE; LUIKART, 2010b).

A importância da bioinformática cresceu com o surgimento das primeiras máquinas de sequenciamento e algoritmos de montagem de genomas (NAGARAJAN; POP, 2013). O avanço das tecnologias de sequenciamento de DNA tem revolucionado o campo da biologia molecular, fornecendo ferramentas poderosas para investigar a diversidade e a evolução das espécies. Essa investigação pode ser feita com diferentes graus de resolução: indo de uma menor resolução (i), como a técnica de *Barcoding*, que permite avaliar regiões específicas a um grupo de organismos, garantindo sua identificação taxonômica de forma rápida e acurada; ao se aumentar a resolução, (ii) a técnica de *Metabarcoding* permite a identificação de diferentes espécies de organismos em uma mesma amostra ambiental; com uma maior resolução, se analisa o genoma completo de um determinado organismo, por meio da (iii) montagem e anotação do seu genoma nuclear, permitindo não só a identificação da espécie, como o conhecimento de todos os genes e da estrutura que esses genomas nucleares apresentam.

O *Barcoding* (também chamado de Código de barras de DNA) é uma técnica que permite identificar e comparar espécies de seres vivos com base em fragmentos específicos do DNA (HEBERT et al., 2003). É uma técnica amplamente utilizada na biologia, ecologia e conservação, ajudando a identificar espécies invasoras, monitorar a biodiversidade e compreender a evolução e relações filogenéticas entre as espécies, se tornando uma ferramenta valiosa para a pesquisa biológica e podendo ajudar a preservar a biodiversidade do planeta (MEIER et al., 2006).

Enquanto o *barcoding* se concentra na identificação de uma espécie a cada amostra, a técnica de *metabarcoding* permite a identificação simultânea de várias espécies a partir de uma única amostra ambiental, ou seja, amostras cujo DNA foi extraído do solo, água, ou qualquer outro ambiente (CREER et al., 2016). A identificação das espécies é feita ao se amplificar e sequenciar os fragmentos específicos de DNA e em seguida realizando a comparação dessas sequências em um banco de dados contendo sequências de DNA específicas para cada espécie estudada (geralmente são os próprios códigos de barra de DNA previamente obtidos) (COISSAC; RIAZ; PUILLANDRE, 2012). Assim como o *barcoding,* o *metabarcoding* também é muito utilizado em estudos de conservação, monitoramento da biodiversidade, e vem sendo também bastante utilizado em inspeção de alimentos (DEINER et al., 2017).

Para se ter uma melhor compreensão de uma espécie a um nível mais profundo, como sua história evolutiva e capacidade funcional, a análise do genoma nuclear completo é indicada. Milhares de genomas de micro-organismos, animais e plantas já foram montados e depositados em bancos de dados públicos, como o NCBI/GenBank (WHEELER et al., 2007).

Com exceção dos vírus de DNA/RNA de até 30 Kb, ainda não existe um método capaz de obter a informação genética completa de uma molécula de DNA de um organismo. Para resolver esse problema, a molécula de DNA pode ser aleatoriamente quebrada em pequenos fragmentos que serão posteriormente analisados pelas máquinas de sequenciamento, gerando as chamadas leituras de sequenciamento (ou como vamos chamar daqui para frente, *reads*). Essas *reads* podem ser usadas como entrada para *softwares* específicos que são capazes de agrupá-las, caso compartilhem sequências de nucleotídeos similares, em estruturas maiores chamadas *contigs*, que por sua vez podem ser agrupadas em estruturas maiores ainda chamadas *scaffolds.* As diferentes formas de se montarem essas *reads* se constitui como o bem conhecido Problema das Montagens de Fragmentos (NAGARAJAN; POP, 2013).

Existem muitos programas e algoritmos capazes de montar um conjunto de *reads* usando métodos diferentes de montagem. A maioria deles é baseada em dois métodos bem conhecidos: a técnica OLC (Overlap-Layout-Consensus) e o Grafo de Bruijn, que usam propriedades dos grafos hamiltonianos e eulerianos, respectivamente (LI et al., 2012).

Os genomas de eucariotos ainda são um desafio para a bioinformática. Propriedades inerentes a esses genomas como tamanho, duplicações e repetição de regiões do genoma dificultam a sua montagem em grandes *contigs* ou *scaffolds* (CLAROS et al., 2012). Como os genomas de plantas são geralmente reportados como sendo maiores em tamanho do que os de animais ou outros eucariotos, muitos esforços foram feitos para superar as dificuldades associadas a estes fatos.

O problema do tamanho do genoma está sendo solucionado com a evolução das máquinas de sequenciamento, que agora podem gerar um alto volume de dados de sequências, podendo cobrir totalmente um genoma grande, por exemplo: Illumina (NextSeq, HiSeq, NovaSeq). Sequências repetitivas em genomas de plantas ainda são um problema para os algoritmos de montagem. A baixa complexidade e o pequeno tamanho dessas regiões dificultam a geração de resultados de montagem bons e confiáveis (BELSER et al., 2018; CLAROS et al., 2012). Além disso, eventos de duplicação do genoma completo (poliploidia) são comumente observados em angiospermas, sendo caracterizado pela presença de pelo menos uma cópia adicional de todo conjunto de cromossomos dentro do núcleo de uma célula gamética (MARKS et al., 2021; MEYERS; LEVIN, 2006), o que pode resultar na má qualidade da montagem de genomas de plantas poliploides. O surgimento das plataformas de sequenciamento de Terceira Geração com a produção de *reads* longas, como as do PacBio e Oxford Nanopore, também podem ajudar na resolução desses problemas das regiões de repetição, além de permitir uma maior contiguidade na montagem dos genomas (DASZKOWSKA-GOLEC; MASCHER; ZHANG, 2023). Muitos *softwares* montadores de genomas foram desenvolvidos para que pudessem lidar com o tamanho grande, repetições e poliploidia dos genomas de plantas. A disponibilidade de tantos *softwares* montadores (a serem discutidos na seção 1.2) começa a levantar uma preocupação no sentido de que até o momento nenhum trabalho foi publicado fazendo uma comparação justa da eficiência e eficácia desses *softwares* ao se analisar genomas de plantas.

Este trabalho pretende reunir informações sobre os genomas nucleares de plantas que estão depositados a nível cromossômico no NCBI, como o tamanho do genoma, número cromossômico, conteúdo GC, e ploidia, além de também extrair informações sobre os sequenciadores e softwares montadores utilizados nas análises desses genomas. Em seguida, é realizada a comparação dos diversos *softwares* montadores de genoma nuclear de plantas que estão disponíveis e sendo utilizados pela comunidade científica. Além disso, também serão mostradas algumas ferramentas de *barcoding* e *metabarcoding* que foram desenvolvidas para auxiliar nos estudos de conservação da diversidade vegetal.

1.1 *DNA Barcoding* e *Metabarcoding* aplicados a estudos de plantas

O *DNA barcoding* (ou Código de Barras de DNA) é uma técnica que permite a identificação rápida e acurada de espécies, por meio do uso de regiões curtas e específicas do DNA de um organismo, podendo ser regiões gênicas ou intergênicas. Essa técnica foi proposta em 2003 (HEBERT et al., 2003) como uma solução para a identificação de espécies em ambientes ricos em biodiversidade, onde muitas espécies de organismos ainda não estão descritas e a dependência dos métodos tradicionais de identificação taxonômica tornam inviáveis o avanço no conhecimento taxonômico da região.

A técnica de *DNA barcoding* foi inicialmente realizada em insetos, no trabalho de Hebert et al., com a utilização do gene mitocondrial Citocromo C Oxidase Subunidade I (*COI*) como marcador para animais, que passou a ser largamente adotado pela comunidade. Para a identificação de animais, o gene *COI* funcionou muito bem, uma vez que a diversidade genética intraespecífica não é grande. Utilizar a técnica de *DNA barcoding* a fim de se realizar a identificação taxonômica em plantas não foi tão simples quanto o realizado para animais, uma vez que a variedade intraespecífica nos genomas e genes de plantas acaba sendo maior, o que dificulta encontrar um gene marcador que possa ser utilizado para se diferenciar os diversos grupos de plantas existentes.

Em 2005 foi publicado o primeiro estudo de *DNA barcoding* aplicado a plantas (KRESS et al., 2005), mostrando que essa técnica poderia ser utilizada para identificação de espécies de plantas, mesmo que pertencentes a gêneros complexos. A diferença de se aplicar essa técnica em plantas é que ao invés de um gene marcador único, poderiam ser utilizados diferentes genes marcadores, cada um apresentando vantagens e desvantagens. Por exemplo, comumente passou-se a utilizar a região nuclear ribossômica *ITS* (*Internal Transcribed Spacer*) para a identificação de espécies, devido ao seu alto nível de variação. Caso se deseje realizar um estudo evolutivo, recomenda-se utilizar regiões com um menor nível de variação, como o gene plastidial *matK* (maturase-K), por exemplo.

Outros genes plastidiais também são muito utilizados como marcadores em estudos de *DNA barcoding*, como os genes codificantes *rbcL* (ribulose-1,5-bisfosfato carboxilase/oxigenasse) e *rpoB* (RNA polimerase subunidade beta). Regiões intergênicas não-codificantes também são bem utilizadas, como as regiões *trnH-psbA* e *psbK-psbI* (CBOL PLANT WORKING GROUP et al., 2009).

A designação taxonômica e os estudos evolutivos que podem ser feitos com essas sequências de *DNA barcode* são realizadas por meio de alinhamento de sequência e filogenia, respectivamente. A inferência filogenética pode ser feita usando apenas uma sequência para cada organismo ou usando múltiplos genes. Para grupos que já apresentam um gene marcador bem definido, como procariotos (*16S*/*18S*) e animais (*COI*), a estratégia de se usar apenas uma sequência para cada organismo pode gerar árvores filogenéticas robustas e confiáveis. Para realizar estudos filogenéticos de plantas, usar apenas uma sequência pode não gerar resultados confiáveis, uma vez que dependendo do gene utilizado, pode haver maior ou menor variação intra- ou interespecífica (BESSE; DA SILVA; GRISONI, 2021). Logo, para se inferir uma filogenia robusta de espécies de plantas, o ideal é usar múltiplos genes.

A inferência de uma árvore de espécies é geralmente feita por meio de duas abordagens: utilizando uma árvore consenso (*supertree*) ou utilizando uma supermatriz. A abordagem de *supertree* envolve gerar uma árvore filogenética para cada sequência de gene, de forma a encontrar congruência entre as diferentes árvores, gerando no final uma árvore consenso (BAUM; SMITH, 2012). Apesar de se utilizar as árvores geradas por todos os múltiplos genes, cada gene pode ter um poder de resolução diferente, gerando árvores que apresentam robustez em níveis diferentes. Isso pode ser um complicador na etapa de gerar a árvore consenso. De forma a contornar este problema, a técnica de supermatriz visa utilizar as sequências alinhadas de todos os genes concatenados em uma única sequência por organismo. Essa estratégia acaba tendo uma acurácia melhor devido a usar um maior número de *loci* do que aconteceria se fosse usado apenas um gene para gerar uma árvore (GADAGKAR; ROSENBERG; KUMAR, 2005).

Além da utilização de *DNA barcoding* para identificação de espécies de plantas e de estudos evolutivos e de conservação (KRESS et al., 2015), a técnica também vem sendo muito utilizada para auxiliar no monitoramento da distribuição e abundância de diferentes espécies de plantas, rastreamento de mudanças em populações de espécies ao longo do tempo em uma determinada região, e detecção de plantas invasoras (KRESS, 2017).

Um estudo recente utilizou a técnica de *DNA barcoding* para analisar a diversidade de plantas presentes em áreas de Canga (solo rico em minério de ferro) na Serra de Carajás, que apresenta várias espécies de plantas que são raras e endêmicas (VASCONCELOS et al., 2021). Como resultado, este trabalho conseguiu gerar 1.130 sequências de *DNA barcode* para 538 espécies de plantas, tendo sido gerado *barcode* pela primeira vez para 344 dessas espécies. As sequências de *barcode* geradas pelo trabalho de Vasconcelos *et al.* passam a compor uma base de dados que futuramente pode permitir a identificação rápida e estudos evolutivos de outras espécies de plantas.

A geração de sequências de *DNA barcode* se tornou mais rápida, barata e efetiva com os avanços das tecnologias de sequenciamento. Isso permitiu a geração de sequências para diversos grupos taxonômicos, que passaram a compor bancos de dados importantes para a identificação de espécies. Alguns exemplos desses importantes bancos de dados são: SILVA (QUAST et al., 2013), RDP (COLE et al., 2014) e Greengenes (DESANTIS et al., 2006), para identificação de procariotos por meio de sequências do gene ribossomal 16S; UNITE (ABARENKOV et al., 2010), para identificação de fungos, por meio da região nuclear intergênica ITS; MIDORI (MACHIDA et al., 2017), para identificação de vertebrados e invertebrados, com o uso do gene mitocondrial COI. Considerando a plasticidade de variação intra- ou interespecífica que um gene marcador de planta pode apresentar em um determinado grupo taxonômico, até o momento não existe nenhum banco oficial para a identificação de plantas. Entretando, o trabalho realizado por Vasconcelos *et al.* permitiu a criação de um banco de dados próprio com sequências de *DNA barcode* para espécies de plantas da biodiversidade amazônica. Todos esses bancos de dados mencionados anteriormente são frutos direto da aplicação das técnicas de DNA *barcoding* e viabilizaram a aplicação da técnica de *metabarcoding*.

Enquanto a técnica de *DNA barcoding* é focada na geração de sequências que permitem a identificação de apenas uma espécie por amostra, no *metabarcoding* ocorre a identificação de múltiplas espécies que possam estar presentes em uma amostra ambiental, como solo, água, ar, e tecido de animais e plantas (CREER et al., 2016). Para que a identificação taxonômica possa ocorrer, primeiramente se extrai o DNA da amostra ambiental que se deseja analisar, e em seguida todo o protocolo laboratorial para a amplificação das regiões de interesse é realizado com a utilização de primers específicos. As regiões de interesse amplificadas (*amplicons*) são usadas nas plataformas de sequenciamento, gerando as *reads* obtidas na amostra. Com o uso de bancos de dados específicos para cada região de interesse e grupo de organismos que se deseja identificar, técnicas de bioinformática são utilizadas, possibilitando assim a identificação das espécies presentes na amostra (COISSAC; RIAZ; PUILLANDRE, 2012).

Em posse das *reads/amplicons* e do banco de dados apropriado, duas abordagens de bioinformática podem ser utilizadas para realizar a identificação das espécies presentes nas amostras ambientais: OTU (*Operational Taxonomic Unit*) ou ASV (*Amplicon Sequence Variant*).

Abordagens com OTU foram as primeiras a serem utilizadas em análise de *metabarcoding*, onde dado um limiar *c*, as reads da amostra são clusterizadas de forma a que leituras que apresentem no mínimo *c%* de similaridade sejam agrupadas em estruturas chamadas OTUs (geralmente $c = 97$). Tais OTUs seriam a representação de um táxon presente na amostra. Por exemplo, se em 1 mil reads sequenciadas em uma amostra ambiental, 20 OTUs forem formadas ao se utilizar um limiar de 97% de similaridade, isso significa que 20 unidades taxonômicas estão presentes nesta amostra (CAPORASO et al., 2010). Na última etapa da abordagem, as OTUs identificadas são comparadas por alinhamento contra o banco de dados específico (por exemplo: SILVA, RDP, UNITE, MIDORI) para o gene marcador e o grupo de organismos a serem analisados. Essa designação taxonômica também pode ser feita com a definição de um limiar *a*, que define que se a sequência de uma OTU encontrada apresentar no mínimo *a%* de similaridade com uma sequência de *DNA barcode* presente no banco de dados, então pode-se inferir que a OTU corresponde ao táxon que a sequência de *DNA barcode* representa.

Enquanto a abordagem com OTU se utiliza de limiar para a formação dos grupos, a abordagem ASV define que apenas reads que difiram de apenas *n* nucleotídeos sejam pertencentes ao mesmo táxon (geralmente $n = 1$). Isso acaba eliminando qualquer ruído oriundo de erros de sequenciamento, mas pode gerar uma super-representação dos taxa presente na amostra (EREN et al., 2013).

Algumas ferramentas de bioinformática foram desenvolvidas para que pudessem realizar análises de *metabarcoding* usando os bancos de dados específicos para cada grupo taxonômico, dentre essas ferramentas, podemos citar: Mothur (SCHLOSS et al., 2009), QIIME (CAPORASO et al., 2010), Obitools (BOYER et al., 2016), mBRAVE (RATNASINGHAM, 2019) e PEMA (ZAFEIROPOULOS et al., 2020). Com exceção da ferramenta PEMA, todas as outras utilizam a abordagem OTU para a identificação taxonômica. Além disso, cada uma dessas ferramentas foi desenvolvida permitindo se utilizar bancos de dados específicos, sem permitir que o usuário possa utilizar bancos de dados próprios ou personalizados. A ferramenta Mothur permite analisar dados de 16S/18S e ITS de fungos, usando os bancos de dados Greengenes e Findley (FINDLEY et al., 2013), respectivamente. QIIME e QIIME2 podem analisar dados de metabarcoding de 16S, 18S e ITS de fungo, usando os bancos Greengens, SILVA e UNITE, respectivamente. A ferramenta Obitools é otimizada para analisar dados de 16S com os bancos SILVA e PR2 (GUILLOU et al., 2013), e permite utilizar o banco GenBank do NCBI, expandindo a possibilidade de se analisar os mais diversos tipos de genes marcadores. O mBRAVE permite utilizar apenas o banco do BOLD (RATNASINGHAM; HEBERT, 2007) como referência. A ferramenta PEMA permite analisar dados de 16S/18S, ITS de fungo e COI animal, usando os bancos SILVA, UNITE e MIDORI, respectivamente.

Alguns estudos de *metabarcoding* já foram feitos em vários tipos de amostras de plantas, incluindo pólen, folhas, solos, raízes e rizosfera (DE MEDEIROS AZEVEDO et al., 2021). Por exemplo, análises de metabarcoding de amostras de pólen podem indicar quais espécies de plantas estão sendo visitadas por espécies polinizadoras (LOWE et al., 2022). Análises de metabarcoding de folhas podem revelar a diversidade de espécies de plantas presentes em um determinado ecossistema (DEINER et al., 2017). Dessa forma, permitir que ferramentas de bioinformática possam usar bancos de dados de referência que permitam a identificação taxonômica de espécies de plantas se torna imprescindível para que mais estudos de metabarcoding em plantas sejam feitos com mais facilidade, acurácia e rapidez.

1.2 Sequenciamento e montagem de genomas de plantas

De acordo com o *Royal Botanic Gardens, Kew,* estima-se que existam mais de 400.000 espécies de plantas conhecidas no mundo (POWO, 2017). Do total de espécies de plantas conhecidas, 217.322 têm pelo menos uma única sequência depositada no GenBank/NCBI (em 13 de abril de 2023). Esse número de sequências depositadas dobrou nos últimos cinco anos, principalmente com o avanço e diminuição de custo das novas tecnologias de sequenciamento (SATAM et al., 2023). Analisar e montar genomas de plantas é extremamente importante, pois permite responder a perguntas biológicas relevantes (MCCOUCH, 2013), conhecer seu conteúdo gênico, a descoberta de novos genes, e realização de estudos evolutivos e adaptativos (HULSE-KEMP et al., 2018).

O primeiro genoma nuclear completo de planta montado (*Arabidopsis thaliana* (L.) Heynh.) foi sequenciado pelo método de Sanger há mais de duas décadas (The Arabidopsis Genome Initiative, 2000). Apesar da alta qualidade das bases sequenciadas por essa tecnologia, ela apresenta nos tempos atuais algumas desvantagens, como baixo rendimento, baixa velocidade de sequenciamento e alto custo (BRÄUTIGAM; GOWIK, 2010). Essas desvantagens foram parcialmente superadas com as plataformas NGS (*Next Generation*

Sequencing) e *Third-Generation Sequencing* (TGS) (Illumina (BENTLEY et al., 2008), Ion Torrent (QUAIL et al., 2012), PacBio (RHOADS; AU, 2015) e MinION Oxford Nanopore (JAIN et al., 2015)), que permitiram a geração de dados com um rendimento muito maior, fornecendo até 1,5 TB de dados, sequenciamento rápido, custos baixos e *reads* mais longas. Inicialmente, os sequenciadores da PacBio e Oxford Nanopore apesar de serem capazes de gerar *reads* mais longas, apresentavam algumas desvantagens em comparação à plataforma Illumina, como baixa cobertura de sequenciamento e baixa qualidade das *reads*. No entanto, novas metodologias vêm sendo desenvolvidas e implementadas nos sequenciadores de *reads* longas, o que tem elevado a qualidade de suas bases sequenciadas. Em linha com esses avanços, a PacBio recentemente lançou novos sequenciadores, como o Sequel II, Sequel IIe e o Revio, projetados para gerar *reads* longas de alta qualidade, com capacidade de gerar uma quantidade de dados maior à um menor custo (WENGER et al., 2019).

As *reads* sequenciadas pelo método Sanger foram montadas pelos primeiros montadores de genoma que foram desenvolvidos: TIGR (SUTTON et al., 1995) e Celera (MYERS, 1995). Esses montadores não eram adequados para lidar com grandes conjuntos de dados, uma vez que dada a época que foram desenvolvidos, os algoritmos utilizados não foram desenvolvidos de forma eficiente a lidar com o uso de muita memória RAM, causando travamentos e longo tempo de espera. À medida que as plataformas NGS e TGS evoluíram e o rendimento dos sequenciamentos aumentou, novos montadores foram desenvolvidos para que pudessem lidar com a grande quantidade de dados gerada. A técnica de montagem de genomas baseadas no Grafo De Bruijn (PEVZNER; TANG; WATERMAN, 2001) teve como objetivo simplificar a representação de milhões de *reads* em estruturas menores chamadas *k-mers*, reduzindo a complexidade computacional das análises para os montadores que usam essa técnica, como o Velvet (ZERBINO; BIRNEY, 2008), ABySS (SIMPSON et al., 2009) e SPAdes (BANKEVICH et al., 2012). Apesar da facilidade de lidar com grandes conjuntos de dados, esses primeiros montadores tiveram dificuldades em lidar com conjuntos de dados originários de genomas mais complexos, como genomas de animais e especialmente de plantas, uma vez que que tais montadores foram inicialmente desenvolvidos para montar genomas de procariotos.

Como mencionado anteriormente, os principais desafios para a montagem do genoma nuclear de plantas são seu tamanho grande do genoma, poliploidia, regiões altamente repetitivas e duplicação do genoma (VALLIYODAN; LEE; NGUYEN, 2017). O menor tamanho de genoma de planta reportado até o momento é da espécie *Genlisea tuberosa* Rivadavia, Gonella & A.Fleischm. (64,9 Mb) e o maior genoma é da *Paris japonica* (Franch. & Sav.) Franch. (149

Gb), o que denota uma plasticidade no tamanho do genoma de plantas de aproximadamente 2.295 vezes. Essa grande variação no tamanho dos genomas de plantas pode acontecer até mesmo dentro da mesma espécie, que podem vir ou não de diferentes populações (LEITCH; LEITCH, 2013). Tal plasticidade no tamanho do genoma nuclear de plantas e sua complexidade pode ser atribuída a diversos fatores, como novos genes que provêm da duplicação de genes e genomas, nível de ploidia das plantas e a propagação de elementos transponíveis (KELLY et al., 2015; VELEBA et al., 2014). À medida que as plataformas NGS e TGS foram sendo desenvolvidas, o tamanho do genoma nuclear das plantas não passou mais a ser visto como um problema, mas sim a complexidade do próprio genoma. Essa complexidade é frequentemente associada às regiões do genoma altamente repetitivas, que são bem difíceis de montar, uma vez que *reads* idênticas ou quase idênticas podem vir de diferentes locais do genoma, gerando *gaps*, ambiguidades e colapsos nos grafos da montagem, enviesando os resultados.

Algumas espécies de plantas chegam a ter 85% do seu genoma preenchido por regiões repetitivas, como os genomas do milho (*Zea may*s ssp. *mays* L.), por exemplo (SCHNABLE et al., 2009a). Os tipos de repetições que são encontrados em genomas de plantas incluem repetições em *tandem* (microssatélites e minissatélites) e os elementos transponíveis (TEs). As repetições em *tandem* são classificados como (i) microssatélites (ou SSR – *Simple Sequence Repeats*), que são repetições curtas de um a seis nucleotídeos variando de 10 a 100 repetições; (ii) minissatétlites, sendo repetições de seis a 100 bp com tamanho total da repetição variando de 0,5 a 30 Kb; e (iii) DNA satélite (satDNA), com repetições de 150 a 400 bp, rico em AT e que comumente formam repetições de até 100 Mb (MEHROTRA; GOYAL, 2014). Os elementos transponíveis (TEs) são segmentos de DNA com a capacidade de se deslocar e se reintegrar em diferentes partes do genoma por meio de dois mecanismos principais: "copia-e-cola" e "recorta-e-cola". Frequentemente apelidados de "genes saltadores", esses elementos estão amplamente distribuídos pelo genoma e têm papéis cruciais na evolução e na diversidade genética. sendo classificadas como: (iv) Classe I, que precisam de uma molécula de RNA intermediária para reinserir a região no DNA (como o retrotransposon de repetição terminal longa (ou LTR-Retrotransposon), SINEs ou LINEs), ou do tipo Classe II, que não precisam de um RNA intermediário, como os transposons "recorta e cola" ou os transposons "rolantes" (WICKER et al., 2007). Contudo, dependendo da extensão da região, o problema de sequências repetitivas no genoma pode ser abordado com o uso intensivo de sequenciadores de *reads* longas, como o PacBio ou Oxford Nanopore, em que uma única sequência pode abranger os dois lados de uma região de repetição (CLAROS et al., 2012).

A poliploidia (ou duplicação do genoma completo) é uma condição em que existe pelo menos uma cópia adicional de todo o conjunto de cromossomos dentro do núcleo de uma célula gamética, sendo uma das principais forças motrizes na evolução, especiação e diversidade de plantas (BENTO et al., 2011; PATERSON et al., 2010). Existem dois tipos principais de poliploidia: (i) Autopoliploidia, quando a duplicação do conjunto de cromossomos acontece nas células do próprio indivíduo ou pelo cruzamento de indivíduos da mesma espécie; (ii) Alopoliploidia, quando a duplicação acontece por meio da hibridização de espécies diferentes (SOLTIS et al., 2015). Para os softwares montadores, essa duplicação pode aumentar o número de montagens errôneas, reduzindo assim a precisão do genoma montado. Para superar esses problemas, os pesquisadores usam a separação dos cromossomos por citometria de fluxo e sequenciam cada cromossomo separadamente, a fim de facilitar análises posteriores, como mapeamento e montagem do genoma (SCHNABLE et al., 2009b). Outra técnica que vem sendo bastante utilizada para minimizar os erros de montagens de genomas poliploides é a abordagem Hi-C, que busca obter informações das interações de cromatina feitas no núcleo, permitindo que os *contigs* e *scaffolds* obtidos em uma montagem sejam agrupados com as informações de proximidade obtidas pela técnica, levando a montagem ao nível cromossômico (HOSHINO et al., 2017).

Alguns montadores foram desenvolvidos para lidar com as complexidades inerentes de um genoma, seja (i) usando *reads* curtas, como o SOAPdenovo2 (LUO et al., 2012), SSAKE (WARREN et al., 2007) e All-Paths (BUTLER et al., 2008); mais recentemente, (ii) combinando dados de mapeamento óptico e de Hi-C, como o Hifiasm (CHENG et al., 2021); (iii) usando leituras longas, como o MaSuRCA (ZIMIN et al., 2013), Canu (KOREN et al., 2017), HINGE (KAMATH et al., 2017) e Flye (KOLMOGOROV et al., 2019); e (iv) específicos a genomas de eucariotos, como o Platanus (KAJITANI et al., 2014) e Falcon (CHIN et al., 2016).

Muitos esforços já foram gastos na tentativa de obter a melhor montagem de um grande genoma eucariótico. As Tabelas 1 (HAMILTON; ROBIN BUELL, 2012) e 2 (KYRIAKIDOU et al., 2018) resumem as primeiras montagens de genoma de plantas mais importantes que alcançaram o status de genoma completo, juntamente com a plataforma de sequenciamento, tamanho do genoma e da montagem e a ploidia do genoma.

Tabela 1 - Grandes marcos históricos na montagem de genomas de plantas.

Ano	Espécie	Plataforma	Tamanho da montagem	Referências
1999	Cromossomos 2 e 4 da *Arabidposis thaliana*	Sanger	19,6 e 17,4 Mb	(CORRECTION

				MAYER et al., 1999; LIN et al., 1999)
2000	Genoma completo da *A. thaliana*	Sanger	115,4 Mb	(INITIATIVE, 2000)
2002	Genoma da *Oryza sativa* L.	Sanger	390 Mb	(GOFF et al., 2002; YU et al., 2002)
2003	*Zea mays* L.; Draft	Sanger	132 Mb	(WHITELAW et al., 2003)
2005	Genoma da *Oryza sativa*	Sanger	370,7 Mb	(MATSUMOTO et al., 2005)
2007	20 genomas de *A. thaliana*	Perlegen	20×119 Mb (2.4 Gb)	(CLARK et al., 2007)
2008	Três genomas de *A. thaliana*	Illumina	3×119Mb (357Mb)	(OSSOWSKI et al., 2008)
2009	Genoma de *Z. mays*	Sanger	2.3 Gb	(SCHNABLE et al., 2009b)
2009	Genoma de *Cucumis sativus* L.	Sanger e Illumina	243.5 Mb	(HUANG et al., 2009)
2011	Genoma de *Fragaria vesca* L.	454; Illumina; SoLiD	220 Mb	(SHULAEV et al., 2011)
2011	80 genomas de *A. thaliana*	Illumina	80 ×119 Mb (9.5 Gb)	(CAO et al., 2011)

Fonte: Adaptado da Tabela 1 de Hamilton & Buell (2012)

Como mostra a Tabela 1, de 1999 a 2009, a plataforma de sequenciamento mais utilizada para genomas de plantas foi o Sanger. A partir de 2009, a plataforma Illumina começou a ser mais utilizada o que refletiu enormemente na quantidade de genomas sequenciados e montados, devido ao seu alto rendimento. Na Tabela 2 são mostrados os primeiros genomas completos de plantas com ploidia diferente de diploide, juntamente com o tamanho do genoma.

Tabela 2 - Primeiros genomas completos obtidos de espécies de plantas com ploidia maior que 2

Espécie	Tamanho do genoma	Ploidia	Referência
Glycine max (L.) Merr.	979 MB	Tetraploide	(SCHMUTZ et al., 2010)

Triticum aestivum L.	15,34 GB	Hexaploide	(CHOULET et al., 2010)
Camelina sativa (L.) Crantz	641 MB	Hexaploide	(KAGALE et al., 2014)
Brassica napus L.	976 MB	Tetraploide	(CHALHOUB et al., 2014)
Gossypium hirsutum L.	2,18 GB	Tetraploide	(LI et al., 2015)
Utricularia gibba L.	101 MB	16-ploide	(LAN et al., 2017)

Fonte: Adaptado da Tabela 1 de Kyriakidou (2018)

Nas Tabelas 1 e 2 pode ser verificado que foram realizados vários projetos de montagem de genomas de plantas, cada um com ploidias, coberturas de sequenciamento e tamanhos de montagem diferentes. Além disso, as plataformas de sequenciamento usada por cada grupo de pesquisa também diferem. Portanto, torna-se difícil comparar as estratégias e como elas evoluíram com o tempo. No entanto, com todas as informações apresentadas nas Tabelas 1 e 2, pode parecer possível decidir qual estratégia usar nas novas plantas sequenciadas, considerando todos os parâmetros e estratégias dos projetos bem-sucedidos de sequenciamento e montagem. Até a presente data, apenas o Assemblathon 2 (BRADNAM et al., 2013) e o GAGE (SALZBERG et al., 2012) compararam o desempenho de diferentes montadores em genomas grandes de aves, peixes e mamíferos, o que também evidencia a necessidade de uma comparação e avaliação recentes apenas de montadores de genomas de plantas.

As ciências genômicas envolvendo plantas endêmicas e raras apresentam um caráter multidisciplinar e integrativo, sendo foco das atividades de conservação, manejo (LANES et al., 2018) e Recuperação de Áreas Degradadas (RAD). Nesse contexto, áreas de conservação na Amazônia, como a Floresta Nacional de Carajás e o Parque Nacional dos Campos Ferruginosos, despertam interesse para o uso e desenvolvimento sustentável de recursos naturais, sendo foco de projetos de diversos institutos de pesquisas e iniciativas, como o Instituto Tecnológico Vale (ITV) e o Museu Paraense Emílio Goeldi (MPEG). Análises dos genomas de algumas plantas dos gêneros *Isoetes* L. e *Ipomoea* L., por exemplo, permitiram sua caracterização genética quando apenas a morfologia não foi suficiente para fins de distinção de espécies e sistemática (BABIYCHUK et al., 2017; NUNES et al., 2018).

Atividades de RAD vêm sendo feitas também por meio dos estudos genéticos de plantas nativas e de como elas podem facilmente se adaptar à uma região degradada após a conclusão das atividades de mineração (MIJANGOS et al., 2015). Estudos já foram capazes de identificar marcadores genéticos associados à capacidade de algumas plantas se adaptarem a regiões altamente degradadas (CARVALHO et al., 2020), porém sem saber quais genes seriam os responsáveis por tal adaptação. Assim, a montagem e anotação dos genomas dessas plantas poderiam ajudar na identificação de genes de interesse.

Quanto mais informações genômicas forem obtidas com relação às espécies de plantas raras, endêmicas ou até mesmo plantas nativas, mais estudos de adaptação dessas plantas a novas áreas poderão ser realizados, auxiliando em projetos que envolvem a recuperação de áreas degradadas e a preservação de espécies. Além disso, se fazem necessários os esforços feitos na melhoria e escalabilidade das montagens de genomas de plantas para redução dos erros de montagem e de comparações genômicas errôneas (EXPOSITO-ALONSO et al., 2020).

1.3 Justificativa

Até agosto de 2022, das mais de 410.000 espécies de plantas conhecidas no mundo, apenas 479 possuem seu genoma nuclear completo montado e depositado no banco de dados do NCBI/GenBank. A razão pela qual apenas 0,00085% das espécies de plantas conhecidas possuírem seu genoma nuclear montado à nível de cromossomo é devido à complexidade e tamanho desses genomas, demandando recursos e tempo para sua total análise. Portanto, se torna importante conhecer como esses genomas foram sequenciados e montados para que se possa entender e sugerir novas ideias de como melhorar o processo de montagem.

Realizar um estudo comparativo entre os *softwares* montadores disponíveis e recentes para genomas grandes e de plantas, permite escolher qual montador é adequado para um determinado genoma, considerando todas as dificuldades e complexidades inerentes. Além disso, a falta na literatura de comparações recentes entre montadores de genomas de planta faz necessária a realização desse estudo comparativo.

Obter genomas bem montados permite que genes, nucleares ou organelares, sejam identificados. Esses genes anotados podem garantir mais robustez nos estudos evolutivos ou de identificação taxonômica, sendo também importantes para as análises de *DNA barcoding* e *metabarcoding*.

Para os estudos evolutivos, é importante se ter ferramentas capazes de otimizar as análises feitas com múltiplos genes, facilitando as etapas do seu alinhamento, concatenação e sendo capazes de lidar com informações e genes faltantes (*missing data*) para a geração das supermatrizes. Para a identificação taxonômica, é importante que ferramentas de *barcoding* e *metabarcoding* sejam desenvolvidas a fim de se permitir que os grupos de pesquisa possam usar bancos de referência com sequências geradas pelos próprios grupos, sem se limitar à existência de bancos de referências públicos.

Os estudos comparativos de pipelines desenvolvidos para montagem de genomas nucleares de plantas e as ferramentas de bioinformática desenvolvidas para otimizar os estudos evolutivos, filogenômicos e de identificação taxonômica de amostras ambientais podem ajudar na escolha da melhor abordagem a ser utilizada para se analisar genomas de novos sequenciamentos, assim como também auxiliar na tomada de decisão das atividades de recuperação de áreas degradadas e na conservação de espécies.

2 OBJETIVOS

2.1 Objetivo Geral

Desenvolver ferramentas para cobrir lacunas em pipelines e abordagens ômicas utilizadas nas atividades de conservação da diversidade vegetal.

2.2 Objetivos específicos

- Desenvolver uma ferramenta que otimize a geração de supermatrizes para análises filogenômicas, lidando com *missing data*;
- Desenvolver uma ferramenta que permita a utilização de bancos de dados de referência genética próprios nas análises de identificação taxonômica;
- Reunir informações sobre montagem de genomas completos de plantas já realizadas e quais ferramentas foram usadas;
- Testar ferramentas de montagem de genomas grandes em dados de sequenciamento simulados.

3 ARTIGOS PUBLICADOS E SUBMETIDOS

3.1 Artigo *“SPLACE: A tool to automatically SPLit, Align and ConcatenatE genes for phylogenomic inference of several organisms.”*

Este artigo (**APÊNDICE A**) foi publicado em 8 de dezembro de 2022, na revista *Frontiers in Bioinformatics* (OLIVEIRA; VASCONCELOS; OLIVEIRA, 2022) e também gerou um registro de software pelo INPI sob o número de registro BR512019002834. Nele é descrito um novo software que provê uma forma automatizada de se gerar supermatrizes para análises filogenômicas, por meio do alinhamento separado e posterior concatenação dos genes a serem incluídos na análise, além de também lidar com possíveis dados faltantes.

O SPLACE se mostrou a única ferramenta capaz de realizar essas tarefas mencionadas anteriormente. Em apenas 36 minutos o SPLACE conseguiu gerar a supermatriz contendo 83 genes codificantes de 270 organismos.

3.2 Artigo *“PIMBA: A PIpeline for MetaBarcoding Analysis.”*

Este artigo (**APÊNDICE B**) foi publicado em 23 de novembro de 2021 como capítulo no livro *“Advances in Bioinformatics and Computational Biology”* (OLIVEIRA et al., 2021), publicado pela Springer. Nele é descrito um novo pipeline para análise de dados de *metabarcoding* que realiza o tratamento de qualidade dos dados brutos (pimba_prepare), podendo analisar dados Illumina ou que tenham sido sequenciados por tecnologias de *single-reads* via *single* ou *dual-index* no *pool* de amostras, realizando assim sua demultiplexação de forma simples e organizada.

Após o tratamento de qualidade, o usuário pode usar abordagens de OTU ou ASV na etapa de clusterização (pimba_run) para encontras as unidades taxonômicas, podendo escolher uma gama de bancos de dados ou até mesmo especificar um banco de dados próprio. Os resultados obtidos com o PIMBA mostraram que a ferramenta obtém ótimos valores de sensibilidade ao analisar dados de *metabarcoding*. Este artigo já possui 8 citações, incluindo das revistas *Plants*, *Processes* e *Microorganisms*.

3.3 Manuscrito submetido *"A review and benchmarking of assembling nuclear genome of plants."*

Neste manuscrito (**APÊNDICE C**) foi feita uma revisão sobre o processo de montagem de genomas nucleares de plantas, desde as primeiras espécies de plantas sequenciadas até agosto de 2022. Essa revisão reuniu informações sobre tamanhos dos genomas, conteúdo GC, número cromossômico e nível de ploidia das 479 espécies de plantas que apresentem o genoma nuclear completo depositado no NCBI, totalizando 856 registros. Além dessas informações, também foram obtidas as plataformas de sequenciamento e os softwares montadores utilizados em cada um dos 856 genomas depositados, permitindo assim se ter um conhecimento sobre a evolução do uso do sequenciamento e dos softwares utilizados ao longo do tempo.

Também foi realizada no artigo uma comparação entre pipelines de montagem de genomas nucleares de plantas, usados em estudos recentes. Essa comparação entre os pipelines permitiu verificar como os grupos de pesquisa vem combinando softwares nas etapas de montagem de genomas nucleares, e por meio da simulação do sequenciamento, foi possível identificar softwares que apresentavam erros no processo de montagem. O manuscrito foi submetido à revista *Briefings in Bioinformatics* no dia 01/05/2023.

4 DISCUSSÃO INTEGRADORA

Os avanços recentes nas tecnologias de sequenciamento de DNA causaram um crescimento exponencial no número de genomas depositados em bancos de dados públicos. Contudo, montar genomas de plantas, que podem apresentar genomas grandes e complexos, ainda apresenta dificuldades devido às características inerentes a estes genomas, como poliploidia, tamanho dos genomas e regiões repetitivas.

Diversos softwares e estratégias foram desenvolvidas para resolver estes problemas. O artigo apresentado na seção 3.1 reuniu informações sobre as tecnologias de sequenciamento e os softwares utilizados na análise de 856 genomas de plantas (referentes a 479 espécies) até agosto de 2022, além de informações como nível de ploidia, tamanho dos genomas e conteúdo GC. Essas informações permitiram entender quais softwares e tecnologias de sequenciamento estão sendo mais utilizadas. Além disso, um estudo comparativo entre pipelines recentes de montagem de genomas nucleares de plantas foi realizado, onde por meio da simulação de sequenciamento, foi identificado um software que talvez não esteja desempenhando bem o seu papel de montagem. O software WTDBG2 cometeu muitos erros de montagem nos dados do sequenciamento simulado das espécies *Setaria italica* e *Oryza sativa*. Também foi identificado que os softwares CANU e SOAPdenovo executam bem sua função de montar leituras longas e curtas, respectivamente, apesar de que o uso do software SOAPdenovo gere uma montagem mais fragmentada.

Também foi identificado que o software Quickmerge realmente é capaz de realizar a junção de montagens diferentes, sem causar erros de montagem, e que as ferramentas SSPACE e GapCloser são capazes de fomar *scaffolds* e fechar uma grande quantidade de gaps, respectivamente.

Esse estudo comparativo de pipelines de montagem de genomas nucleares de plantas permitiu a identificação de ferramentas que talvez não devam ser utilizadas em certas análises, assim como na identificação de ferramentas e pipelines que podem ser adotadas por grupos de pesquisa. Mais especificamente, o Pipeline 3 mostrado no artigo da seção 3.1 foi utilizado nos processos de montagem de genomas nucleares de plantas do Instituto Tecnológico Vale, uma vez que faz uso tanto das leituras longas quanto das leituras curtas para a realização das montagens. A aplicação desse pipeline permitiu uma melhoria em alguns genomas de plantas analisados, cujos artigos ainda estão em preparação.

Ter genomas bem montados permite que as informações contidas nestes genomas estejam corretas, em especial genes que possam ser identificados e usados como *barcodes*. Nesse contexto, o uso de ferramentas de *barcoding* e *metabarcoding* cresceu bastante como métodos de monitoramento da biodiversidade a partir de amostras ambientais, apesar de algumas limitações quanto ao uso de bancos de dados próprios e personalizados. Visando resolver essas limitações, o pipeline PIMBA foi desenvolvido e publicado, permitindo análises de *metabarcoding* que utilizam tanto a abordagem OTU quanto a ASV e também tornando possível a utilização de bancos de dados de referência próprios, além dos já estabelecidos bancos SILVA, RDP, UNITE e NCBI.

O PIMBA tem sido bastante utilizado em vários estudos do Instituto Tecnológico Vale, tendo sido citado em oito publicações (incluindo artigos publicados nas revistas científicas *Plants*, *Microorganisms* e *Processes*). No trabalho de Nascimento e colaboradores (DO NASCIMENTO et al., 2022), o PIMBA foi utilizado para a identificação das comunidades de fungos e bactérias associadas à rizosfera da planta *Dioclea apurensis* Kunth, comparando o crescimento da planta em áreas de canga e áreas recuperadas pela atividade da mineração. Essa mesma comparação foi feita no trabalho de Costa e colaboradores (COSTA et al., 2021a), porém avaliando as comunidades associadas à rizosfera da planta *Mimosa acutistipula* var. *férrea* Barneby. No trabalho de Cardoso e colaboradores (CARDOSO et al., 2023), o PIMBA foi utilizado para a obtenção do perfil taxonômico das bactérias presentes em diferentes amostras de água e solo coletadas na Floresta Nacional de Carajás, às quais foram utilizados meios específicos para obter uma cultura enriquecida de microrganismos capazes de realizar a redução de ferro.

Além dos artigos científicos mencionados anteriormente, o PIMBA também foi utilizado em relatórios técnico-científicos elaborados por pesquisadores do Instituto Tecnológico Vale. No relatório elaborado por Catarina e colaboradores (CATARINA et al., 2021), o PIMBA foi utilizado como ferramenta para a identificação simultânea de espécies de plantas e monitoramento da flora da canga da Serra dos Carajás, a partir de amostras do solo da região e amplificação da região ITS2 presente no DNA extraído das amostras de solo. Esse estudo mostrou que a técnica de *metabarcoding* e o uso do PIMBA foram capazes de identificar um maior número de espécies e gêneros que métodos tradicionais por morfologia e taxonomistas. O trabalho desenvolvido por Costa e colaboradores (COSTA et al., 2021b) mostra a importância de se ter ferramentas que permitam a utilização de bancos de dados próprios e personalizados, como o PIMBA, pois foram capazes de identificar a ictiofauna de lagoas presentes na região da Serra de Carajás por meio da coleta de amostras de água e amplificação do gene 12S nos rastros de DNA presentes nas amostras.

Os genes identificados nos genomas bem montados também podem ser utilizados em estudos evolutivos, por meio de análises filogenômicas. A ferramenta SPLACE contribuiu para que supermatrizes contendo diversos genes de diversos organismos pudessem ser rapidamente geradas, por meio da automatização de todo o processo de separação dos genes, alinhamento e concatenação dos genes alinhados em uma supermatriz que posteriormente pode ser usada em ferramentas filogenéticas para a inferência das árvores de espécies. Em apenas 36 minutos, o SPLACE foi capaz de gerar uma árvore de espécies contendo 270 espécies de plantas, representando 91 famílias e contendo a informação genética de 83 genes codificantes.

5 CONCLUSÃO

Este trabalho possibilitou o desenvolvimento, publicação e disponibilização de ferramentas ômicas para os estudos de conservação da biodiversidade, além de também ter sido realizada a coleta de diversas informações sobre genomas completos de planta depositados no NCBI até agosto de 2022, juntamente com um estudo comparativo de pipelines utilizados em publicações recentes.

O SPLACE foi mostrado na Seção 3.1, sendo uma ferramenta que automatiza todas as etapas envolvidas na obtenção de supermatrizes para estudos evolutivos e filogenômicos. A utilização do SPLACE garantiu celeridade nos estudos evolutivos desenvolvidos, se mostrando uma ferramenta eficiente no processo de obtenção de árvores de espécies.

Na Seção 3.2 foi mostrado a ferramenta PIMBA, voltada para as análises de *metabarcoding*, permitindo a identificação taxonômica de organismos em amostras ambientais, usando abordagens de OTU ou ASV e a utilização de bancos de dados próprios. Os resultados mostrados no artigo e a intensa utilização da ferramenta em artigos científicos e relatórios técnico-científicos mostram a importância de se ter uma ferramenta como o PIMBA auxiliando nas análises de *metabarcoding* voltadas para a conservação da biodiversidade.

No artigo da Sessão 3.3 buscou-se então reunir informações de ploidia, tamanho do genoma, número cromossômico, conteúdo GC, total de genes, tecnologias de sequenciamento e *softwares* montadores utilizados em 856 registros de genomas nucleares de plantas depositados em bancos de dados públicos, como o NCBI. Tais informações podem ajudar a entender melhor todo o processo de montagem de genomas nucleares de plantas, dado o alto número de ferramentas disponíveis. Além da reunião de informações, buscou-se também comparar alguns pipelines de montagem de genomas de plantas utilizados nos dois últimos anos. Três pipelines foram escolhidos para serem executados em duas espécies diploides (*Setaria itálica* e *Oryza sativa*), que tiveram seus genomas sequenciados de forma simulada e passaram pelos três pipelines estabelecidos.

Com os resultados, observou-se que as ferramentas WTDBG2 utilizada no Pipeline 1 ocasiona muitos erros de montagem, que acabam se perpetuando ao longo do pipeline. Os Pipelines 2 e 3 foram o que menos geraram erros de montagem, apesar de o Pipeline 3 gerar uma montagem mais fragmentada pelo fato de utilizar leituras curtas para realizar a montagem, além das leituras longas. Além da comparação, este estudo possibilitou entender a linha do tempo dos sequenciadores e programas de montagem de genomas utilizados nos 856 registros de genomas nucleares completos de plantas depositados no banco público NCBI.

Todos os trabalhos mostrados nesta tese ressaltam que apesar da grande diversidade de ferramentas ômicas desenvolvidas pela comunidade científica, sempre há espaço para novas ferramentas que consigam suprir necessidades ainda não contempladas em certas atividades. Desenvolver novas ferramentas e realizar benchmarking de ferramentas existentes garante que a comunidade vai estar sempre amparada nas atividades a serem desenvolvidas por ferramentas que estejam corretamente desempenhando suas funções.

6 PRODUÇÃO CIENTÍFICA E OUTRAS ATIVIDADES

Ao longo desses quatro anos de doutorado (mais precisamente desde agosto de 2018), alguns artigos foram publicados e submetidos pelo autor como primeiro autor e coautoria, *softwares* foram registrados, congressos e eventos foram atendidos e atividades em alguns projetos foram realizadas. Tais feitos serão mostrados nas seções seguintes.

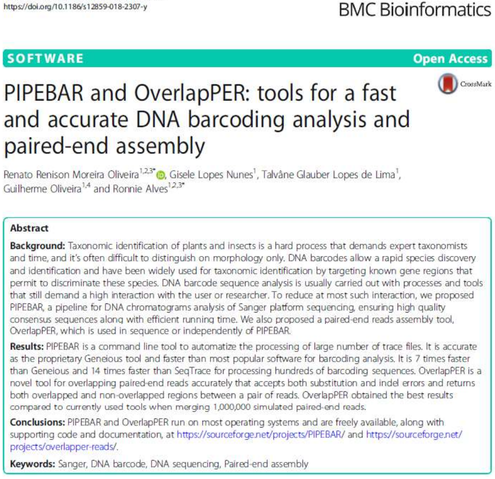
6.1 Atividades realizadas em Projetos

Os projetos "Genômica para o monitoramento da biodiversidade e serviços de ecossistema" e "Cavidades", executados pelo Instituto Tecnológico Vale, fizeram uso de um pipeline desenvolvido e mantido pelo autor para análise de dados de *metabarcoding*: PIMBA (*PIpeline for MetaBarcoding Analysis*).

No projeto "Rede de pesquisa para o sequenciamento genômico do SARS-Cov-2, causador da Covid-19", o autor foi responsável pelo desenvolvimento do pipeline PipeCov (OLIVEIRA et al., 2022), que faz o tratamento de qualidade e montagem dos dados Illumina, resultando ao final do processo o genoma de SARS-CoV-2 montado e anotado.

6.2 Artigos publicados e Registros de *Software*

Oliveira, R. R. M., Nunes, G. L., de Lima, T. G. L., Oliveira, G., & Alves, R. (2018). PIPEBAR and OverlapPER: tools for a fast and accurate DNA barcoding analysis and paired-end assembly. *BMC bioinformatics*, *19*(1), 297. **Citações: 11.**



Oliveira *et al. BMC Bioinformatics* (2018) 19:297
https://doi.org/10.1186/s12859-018-2307-y

BMC Bioinformatics

SOFTWARE **Open Access**

PIPEBAR and OverlapPER: tools for a fast and accurate DNA barcoding analysis and paired-end assembly

Renato Renison Moreira Oliveira[1,2,3*], Gisele Lopes Nunes[1], Talvâne Glauber Lopes de Lima[1], Guilherme Oliveira[1,4] and Ronnie Alves[1,2,3*]

Abstract

Background: Taxonomic identification of plants and insects is a hard process that demands expert taxonomists and time, and it's often difficult to distinguish on morphology only. DNA barcodes allow a rapid species discovery and identification and have been widely used for taxonomic identification by targeting known gene regions that permit to discriminate these species. DNA barcode sequence analysis is usually carried out with processes and tools that still demand a high interaction with the user or researcher. To reduce at most such interaction, we proposed PIPEBAR, a pipeline for DNA chromatograms analysis of Sanger platform sequencing, ensuring high quality consensus sequences along with efficient running time. We also proposed a paired-end reads assembly tool, OverlapPER, which is used in sequence or independently of PIPEBAR.

Results: PIPEBAR is a command line tool to automatize the processing of large number of trace files. It is accurate as the proprietary Geneious tool and faster than most popular software for barcoding analysis. It is 7 times faster than Geneious and 14 times faster than SeqTrace for processing hundreds of barcoding sequences. OverlapPER is a novel tool for overlapping paired-end reads accurately that accepts both substitution and indel errors and returns both overlapped and non-overlapped regions between a pair of reads. OverlapPER obtained the best results compared to currently used tools when merging 1,000,000 simulated paired-end reads.

Conclusions: PIPEBAR and OverlapPER run on most operating systems and are freely available, along with supporting code and documentation, at https://sourceforge.net/projects/PIPEBAR/ and https://sourceforge.net/projects/overlapper-reads/.

Keywords: Sanger, DNA barcode, DNA sequencing, Paired-end assembly

Nunes, G. L., **Oliveira, R. R. M.**, Guimarães, J. T. F., Giulietti, A. M., Caldeira, C., Vasconcelos, S., Pires, E., Dias, M., Watanabe, M., Pereira, J., Jaffé, R., Bandeira, C. H. M. M., Carvalho-Filho, N., da Silva, E. F., Rodrigues, T. M., dos Santos, F. M. G., Fernandes, T., Castilho, A., Souza-Filho, P. W. M., Fonseca, V., Siqueira, J. O., Alves, R. & Oliveira, G. (2018). Quillworts from the Amazon: A multidisciplinary populational study on Isoetes serracarajensis and Isoetes cangae. *PloS one*, *13*(8), e0201417. **Citações: 27.**

PLOS | ONE

RESEARCH ARTICLE

Quillworts from the Amazon: A multidisciplinary populational study on *Isoetes serracarajensis* and *Isoetes cangae*

Gisele Lopes Nunes[1], Renato Renison Moreira Oliveira[1], José Tasso Felix Guimarães[2], Ana Maria Giulietti[3], Cecílio Caldeira[4], Santelmo Vasconcelos[1], Eder Pires[1], Mariana Dias[1], Maurício Watanabe[3], Jovani Pereira[3], Rodolfo Jaffé[3], Cinthia Helena M. M. Bandeira[4], Nelson Carvalho-Filho[1], Edilson Freitas da Silva[5], Tarcísio Magevski Rodrigues[6], Fernando Marino Gomes dos Santos[7], Taís Fernandes[8], Alexandre Castilho[9], Pedro Walfir M. Souza-Filho[4], Vera Imperatriz-Fonseca[3], José Oswaldo Siqueira[10], Ronnie Alves[1], Guilherme Oliveira[1]*

1 Environmental Genomics Group, Instituto Tecnológico Vale, Belém, PA, Brazil, **2** Environmental Geology and Water Resources Group, Instituto Tecnológico Vale, Belém, PA, Brazil, **3** Biodiversity and Ecosystems Services Group, Instituto Tecnológico Vale, Belém, PA, Brazil, **4** Environmental Technology Group, Instituto Tecnológico Vale, Belém, PA, Brazil, **5** Botany Coordination, Museu Paraense Emilio Goeldi, Belém, PA, Brazil, **6** Zoobotanical Park, Vale, Parauapebas, PA, Brazil, **7** Environmental Studies, Amplo Engenharia, MG, Brazil, **8** Environmental Studies Office, Vale, Belo Horizonte, MG, Brazil, **9** North Ferrous Environmental Office, Vale, Parauapebas, PA, Brazil, **10** Director, Instituto Tecnológico Vale, Belém, PA, Brazil

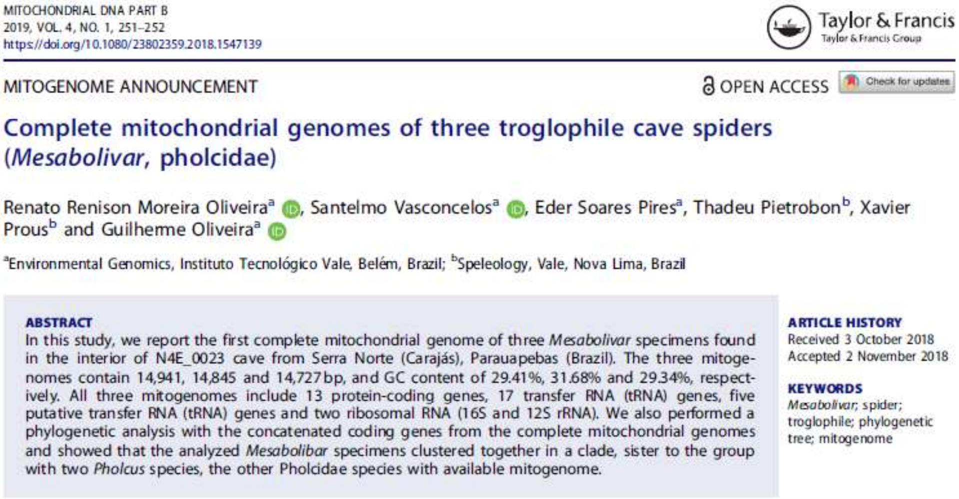
* guilherme.oliveira@itv.org

Check for updates

OPEN ACCESS

Citation: Nunes GL, Oliveira RRM, Guimarães JTF,

Oliveira, R. R. M., Vasconcelos, S., Pires, E. S., Pietrobon, T., Prous, X., & Oliveira, G. (2019). Complete mitochondrial genomes of three troglophile cave spiders (Mesabolivar, pholcidae). *Mitochondrial DNA Part B*, *4*(1), 251-252. **Citações: 10.**



MITOCHONDRIAL DNA PART B
2019, VOL. 4, NO. 1, 251–252
https://doi.org/10.1080/23802359.2018.1547139

Taylor & Francis
Taylor & Francis Group

MITOGENOME ANNOUNCEMENT

OPEN ACCESS Check for updates

Complete mitochondrial genomes of three troglophile cave spiders (*Mesabolivar*, pholcidae)

Renato Renison Moreira Oliveira[a], Santelmo Vasconcelos[a], Eder Soares Pires[a], Thadeu Pietrobon[b], Xavier Prous[b] and Guilherme Oliveira[a]

[a]Environmental Genomics, Instituto Tecnológico Vale, Belém, Brazil; [b]Speleology, Vale, Nova Lima, Brazil

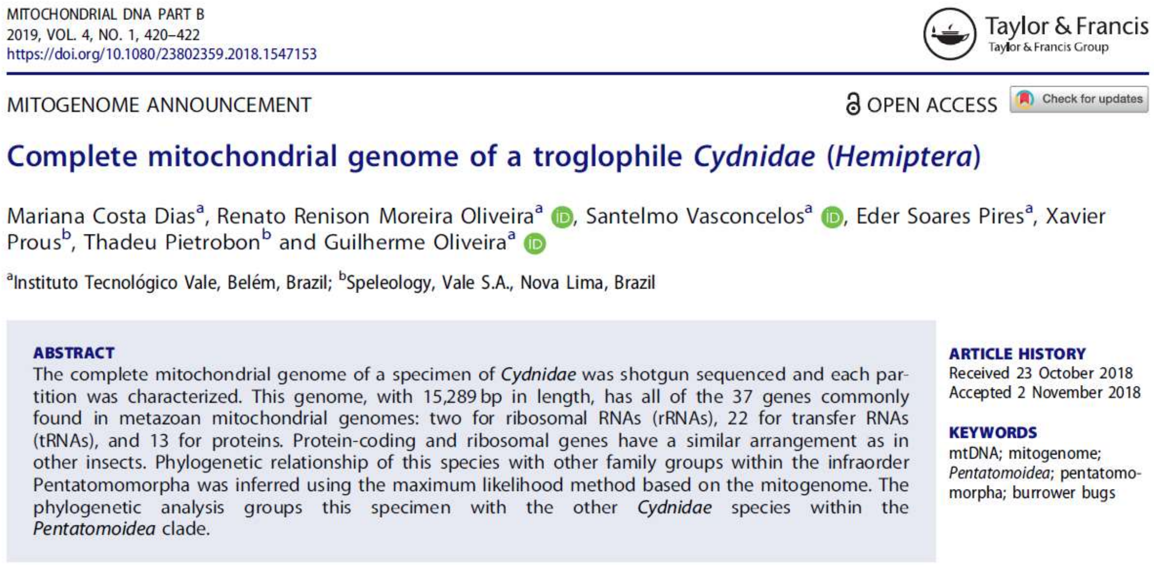
ABSTRACT

In this study, we report the first complete mitochondrial genome of three *Mesabolivar* specimens found in the interior of N4E_0023 cave from Serra Norte (Carajás), Parauapebas (Brazil). The three mitogenomes contain 14,941, 14,845 and 14,727 bp, and GC content of 29.41%, 31.68% and 29.34%, respectively. All three mitogenomes include 13 protein-coding genes, 17 transfer RNA (tRNA) genes, five putative transfer RNA (tRNA) genes and two ribosomal RNA (16S and 12S rRNA). We also performed a phylogenetic analysis with the concatenated coding genes from the complete mitochondrial genomes and showed that the analyzed *Mesabolibar* specimens clustered together in a clade, sister to the group with two *Pholcus* species, the other Pholcidae species with available mitogenome.

ARTICLE HISTORY
Received 3 October 2018
Accepted 2 November 2018

KEYWORDS
Mesabolivar; spider; troglophile; phylogenetic tree; mitogenome

Costa Dias, M., **Oliveira, R. R. M.**, Vasconcelos, S., Pires, E. S., Prous, X., Pietrobon, T., & Oliveira, G. (2019). Complete mitochondrial genome of a troglophile Cydnidae (Hemiptera). *Mitochondrial DNA Part B*, *4*(1), 420-422. **Citações: 1.**



MITOCHONDRIAL DNA PART B
2019, VOL. 4, NO. 1, 420–422
https://doi.org/10.1080/23802359.2018.1547153

Taylor & Francis
Taylor & Francis Group

MITOGENOME ANNOUNCEMENT

OPEN ACCESS | Check for updates

Complete mitochondrial genome of a troglophile *Cydnidae* (*Hemiptera*)

Mariana Costa Dias[a], Renato Renison Moreira Oliveira[a], Santelmo Vasconcelos[a], Eder Soares Pires[a], Xavier Prous[b], Thadeu Pietrobon[b] and Guilherme Oliveira[a]

[a]Instituto Tecnológico Vale, Belém, Brazil; [b]Speleology, Vale S.A., Nova Lima, Brazil

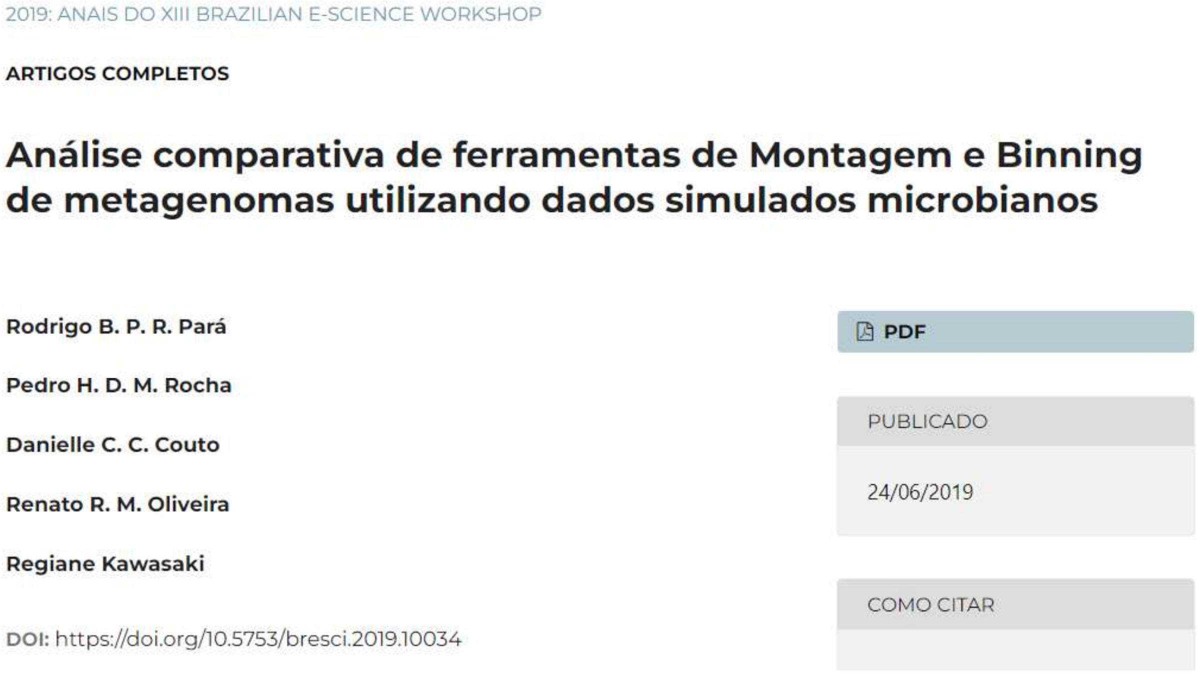
ABSTRACT

The complete mitochondrial genome of a specimen of *Cydnidae* was shotgun sequenced and each partition was characterized. This genome, with 15,289 bp in length, has all of the 37 genes commonly found in metazoan mitochondrial genomes: two for ribosomal RNAs (rRNAs), 22 for transfer RNAs (tRNAs), and 13 for proteins. Protein-coding and ribosomal genes have a similar arrangement as in other insects. Phylogenetic relationship of this species with other family groups within the infraorder Pentatomomorpha was inferred using the maximum likelihood method based on the mitogenome. The phylogenetic analysis groups this specimen with the other *Cydnidae* species within the *Pentatomoidea* clade.

ARTICLE HISTORY
Received 23 October 2018
Accepted 2 November 2018

KEYWORDS
mtDNA; mitogenome; *Pentatomoidea*; pentatomomorpha; burrower bugs

Pará, R. B., Rocha, P. H., Couto, D. C., **Oliveira, R. R.**, & Kawasaki, R. (2019, June). Análise comparativa de ferramentas de Montagem e Binning de metagenomas utilizando dados simulados microbianos. In *Anais do XIII Brazilian e-Science Workshop*. SBC.



2019: ANAIS DO XIII BRAZILIAN E-SCIENCE WORKSHOP

ARTIGOS COMPLETOS

Análise comparativa de ferramentas de Montagem e Binning de metagenomas utilizando dados simulados microbianos

Rodrigo B. P. R. Pará

Pedro H. D. M. Rocha

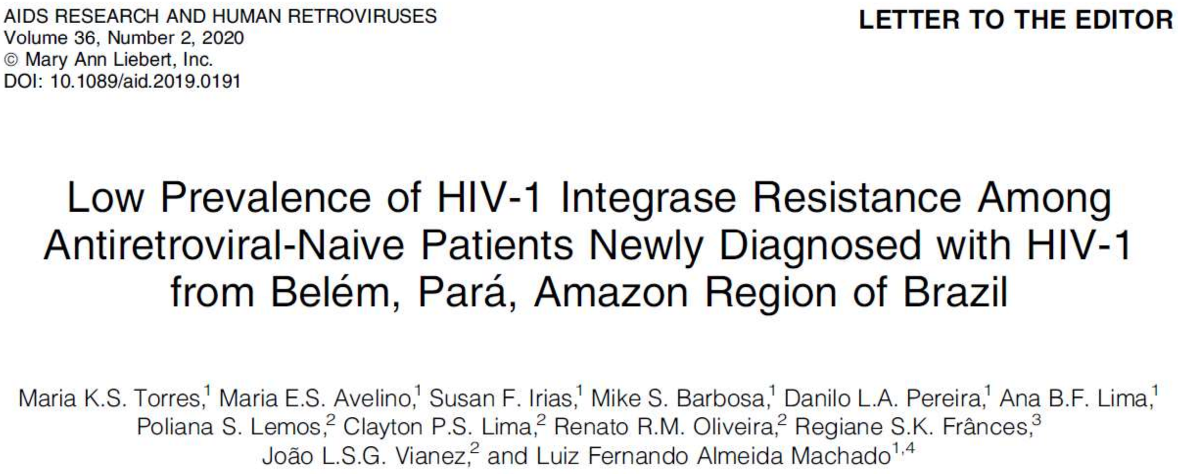
Danielle C. C. Couto

Renato R. M. Oliveira

Regiane Kawasaki

DOI: https://doi.org/10.5753/bresci.2019.10034

Torres, M. K., Avelino, M. E., Irias, S. F., Barbosa, M. S., Pereira, D. L., Lima, A. B., Lemos, P. S., Lima, C. P. S., **Oliveira, R. R. M.**, Frânces, R. S. K., Vianez, J. L. S. G. & Machado, L. F. A. (2020) (2020). Low prevalence of HIV-1 integrase resistance among antiretroviral-naive patients newly diagnosed with HIV-1 from Belém, Pará, Amazon Region of Brazil. *AIDS Research and Human Retroviruses*, *36*(2), 97-98. **Citações: 2.**



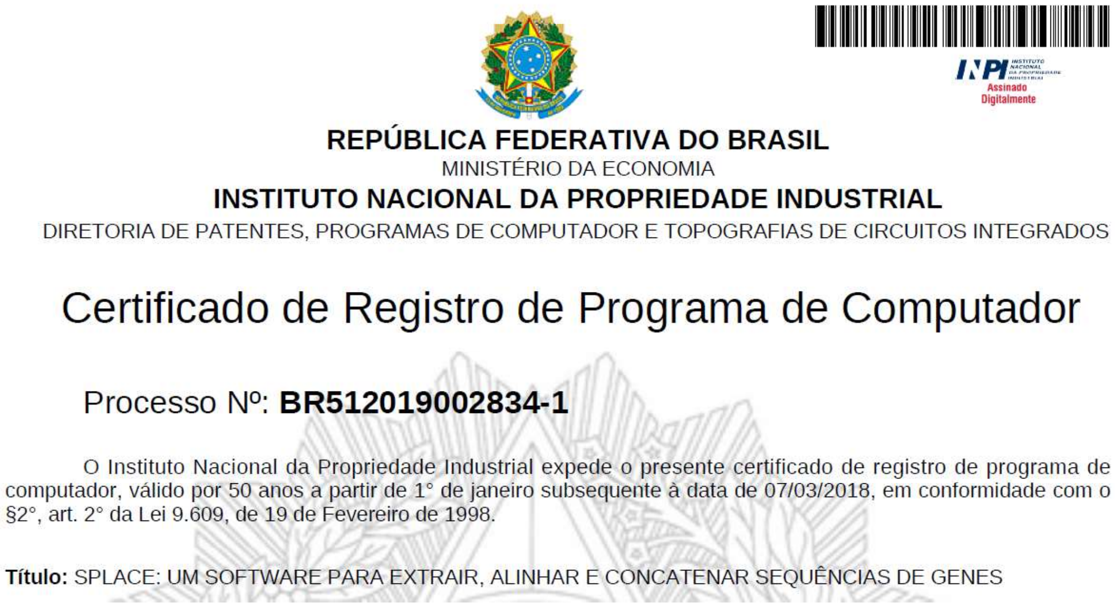
AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 36, Number 2, 2020
© Mary Ann Liebert, Inc.
DOI: 10.1089/aid.2019.0191

LETTER TO THE EDITOR

Low Prevalence of HIV-1 Integrase Resistance Among Antiretroviral-Naive Patients Newly Diagnosed with HIV-1 from Belém, Pará, Amazon Region of Brazil

Maria K.S. Torres,[1] Maria E.S. Avelino,[1] Susan F. Irias,[1] Mike S. Barbosa,[1] Danilo L.A. Pereira,[1] Ana B.F. Lima,[1] Poliana S. Lemos,[2] Clayton P.S. Lima,[2] Renato R.M. Oliveira,[2] Regiane S.K. Frânces,[3] João L.S.G. Vianez,[2] and Luiz Fernando Almeida Machado[1,4]

SPLACE: UM SOFTWARE PARA EXTRAIR, ALINHAR E CONCATENAR SEQUÊNCIAS DE GENES. Número do Processo no INPI: BR512019002834



INPI INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
Assinado Digitalmente

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA ECONOMIA
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

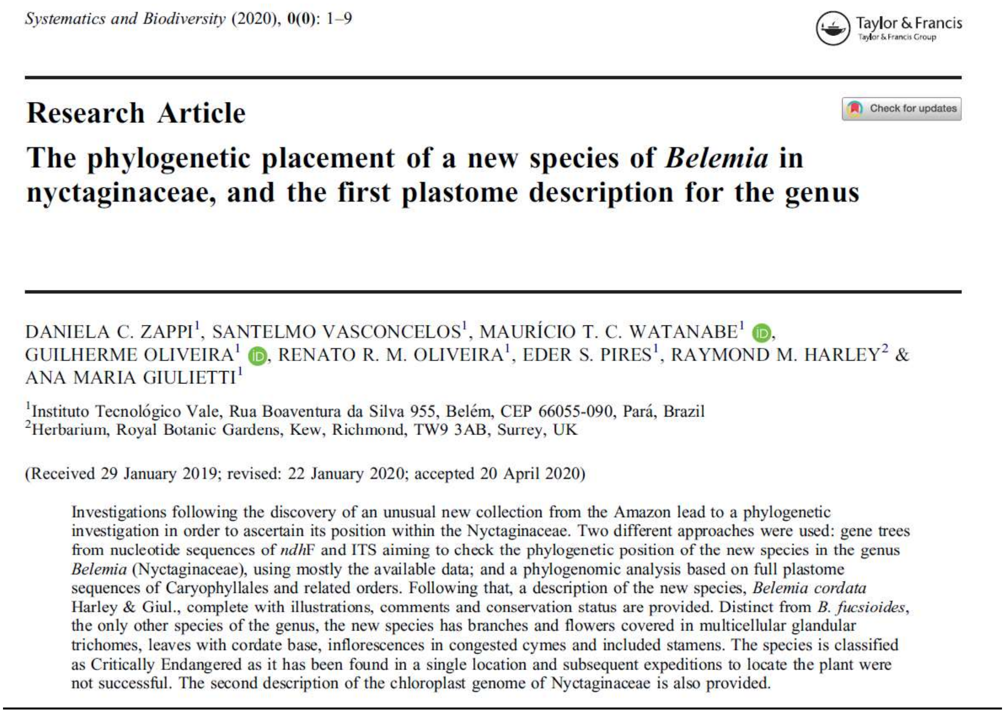
Certificado de Registro de Programa de Computador

Processo Nº: **BR512019002834-1**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 07/03/2018, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: SPLACE: UM SOFTWARE PARA EXTRAIR, ALINHAR E CONCATENAR SEQUÊNCIAS DE GENES

Zappi, D. C., Vasconcelos, S., Watanabe, M. T., Oliveira, G., **Oliveira, R. R**., Pires, E. S., Harley, R. M. & Giulietti, A. M. (2020). The phylogenetic placement of a new species of Belemia in nyctaginaceae, and the first plastome description for the genus. *Systematics and Biodiversity*, 1-9.



Systematics and Biodiversity (2020), **0**(**0**): 1–9

Taylor & Francis
Taylor & Francis Group

Research Article

Check for updates

The phylogenetic placement of a new species of *Belemia* in nyctaginaceae, and the first plastome description for the genus

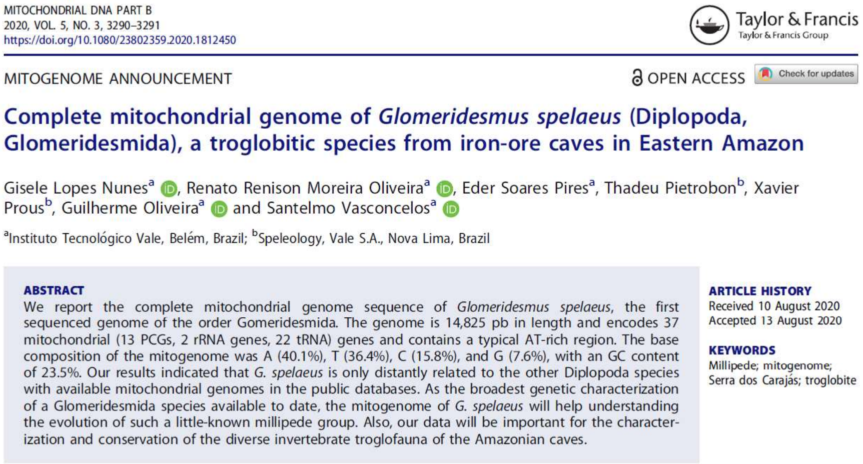
DANIELA C. ZAPPI[1], SANTELMO VASCONCELOS[1], MAURÍCIO T. C. WATANABE[1], GUILHERME OLIVEIRA[1], RENATO R. M. OLIVEIRA[1], EDER S. PIRES[1], RAYMOND M. HARLEY[2] & ANA MARIA GIULIETTI[1]

[1]Instituto Tecnológico Vale, Rua Boaventura da Silva 955, Belém, CEP 66055-090, Pará, Brazil
[2]Herbarium, Royal Botanic Gardens, Kew, Richmond, TW9 3AB, Surrey, UK

(Received 29 January 2019; revised: 22 January 2020; accepted 20 April 2020)

Investigations following the discovery of an unusual new collection from the Amazon lead to a phylogenetic investigation in order to ascertain its position within the Nyctaginaceae. Two different approaches were used: gene trees from nucleotide sequences of *ndh*F and ITS aiming to check the phylogenetic position of the new species in the genus *Belemia* (Nyctaginaceae), using mostly the available data; and a phylogenomic analysis based on full plastome sequences of Caryophyllales and related orders. Following that, a description of the new species, *Belemia cordata* Harley & Giul., complete with illustrations, comments and conservation status are provided. Distinct from *B. fucsioides*, the only other species of the genus, the new species has branches and flowers covered in multicellular glandular trichomes, leaves with cordate base, inflorescences in congested cymes and included stamens. The species is classified as Critically Endangered as it has been found in a single location and subsequent expeditions to locate the plant were not successful. The second description of the chloroplast genome of Nyctaginaceae is also provided.

Nunes, G. L., **Oliveira, R. R. M.**, Pires, E. S., Pietrobon, T., Prous, X., Oliveira, G., & Vasconcelos, S. (2020). Complete mitochondrial genome of Glomeridesmus spelaeus (Diplopoda, Glomeridesmida), a troglobitic species from iron-ore caves in Eastern Amazon. *Mitochondrial DNA Part B*, *5*(3), 3272-3273. **Citações: 3.**



MITOCHONDRIAL DNA PART B
2020, VOL. 5, NO. 3, 3290–3291
https://doi.org/10.1080/23802359.2020.1812450

Taylor & Francis
Taylor & Francis Group

MITOGENOME ANNOUNCEMENT

OPEN ACCESS Check for updates

Complete mitochondrial genome of *Glomeridesmus spelaeus* (Diplopoda, Glomeridesmida), a troglobitic species from iron-ore caves in Eastern Amazon

Gisele Lopes Nunes[a], Renato Renison Moreira Oliveira[a], Eder Soares Pires[a], Thadeu Pietrobon[b], Xavier Prous[b], Guilherme Oliveira[a] and Santelmo Vasconcelos[a]

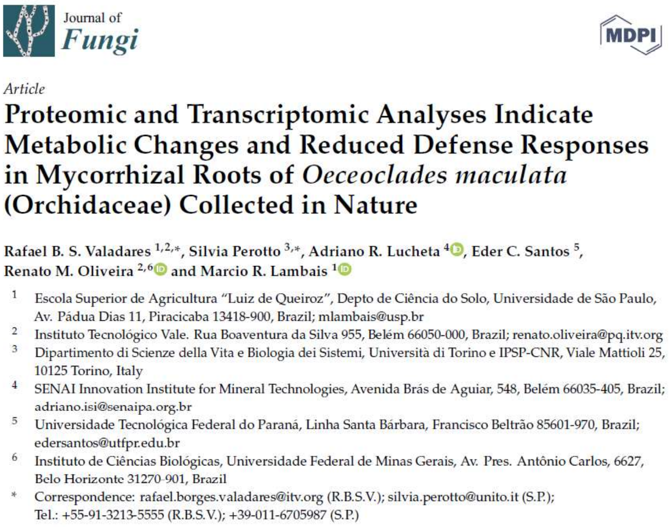
[a]Instituto Tecnológico Vale, Belém, Brazil; [b]Speleology, Vale S.A., Nova Lima, Brazil

ABSTRACT
We report the complete mitochondrial genome sequence of *Glomeridesmus spelaeus*, the first sequenced genome of the order Gomeridesmida. The genome is 14,825 pb in length and encodes 37 mitochondrial (13 PCGs, 2 rRNA genes, 22 tRNA) genes and contains a typical AT-rich region. The base composition of the mitogenome was A (40.1%), T (36.4%), C (15.8%), and G (7.6%), with an GC content of 23.5%. Our results indicated that *G. spelaeus* is only distantly related to the other Diplopoda species with available mitochondrial genomes in the public databases. As the broadest genetic characterization of a Glomeridesmida species available to date, the mitogenome of *G. spelaeus* will help understanding the evolution of such a little-known millipede group. Also, our data will be important for the characterization and conservation of the diverse invertebrate troglofauna of the Amazonian caves.

ARTICLE HISTORY
Received 10 August 2020
Accepted 13 August 2020

KEYWORDS
Millipede; mitogenome; Serra dos Carajás; troglobite

Valadares, R. B., Perotto, S., Lucheta, A. R., Santos, E. C., **Oliveira, R. M.**, & Lambais, M. R. (2020). Proteomic and transcriptomic analyses indicate metabolic changes and reduced defense responses in mycorrhizal roots of Oececlades maculata (Orchidaceae) collected in nature. *Journal of Fungi*, *6*(3), 148. **Citações: 11.**



Journal of *Fungi*

MDPI

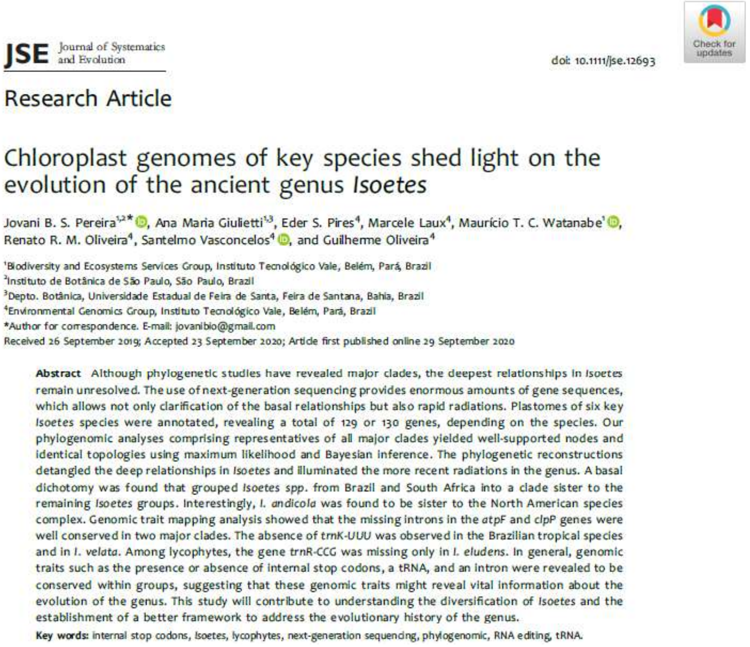
Article

Proteomic and Transcriptomic Analyses Indicate Metabolic Changes and Reduced Defense Responses in Mycorrhizal Roots of *Oeceoclades maculata* (Orchidaceae) Collected in Nature

Rafael B. S. Valadares [1,2,*], Silvia Perotto [3,*], Adriano R. Lucheta [4], Eder C. Santos [5], Renato M. Oliveira [2,6] and Marcio R. Lambais [1]

[1] Escola Superior de Agricultura "Luiz de Queiroz", Depto de Ciência do Solo, Universidade de São Paulo, Av. Pádua Dias 11, Piracicaba 13418-900, Brazil; mlambais@usp.br
[2] Instituto Tecnológico Vale. Rua Boaventura da Silva 955, Belém 66050-000, Brazil; renato.oliveira@pq.itv.org
[3] Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino e IPSP-CNR, Viale Mattioli 25, 10125 Torino, Italy
[4] SENAI Innovation Institute for Mineral Technologies, Avenida Brás de Aguiar, 548, Belém 66035-405, Brazil; adriano.isi@senaipa.org.br
[5] Universidade Tecnológica Federal do Paraná, Linha Santa Bárbara, Francisco Beltrão 85601-970, Brazil; edersantos@utfpr.edu.br
[6] Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627, Belo Horizonte 31270-901, Brazil
* Correspondence: rafael.borges.valadares@itv.org (R.B.S.V.); silvia.perotto@unito.it (S.P.); Tel.: +55-91-3213-5555 (R.B.S.V.); +39-011-6705987 (S.P.)

Pereira, J. B., Giulietti, A. M., Pires, E. S., Laux, M., Watanabe, M. T., **Oliveira, R. R.**, Vasconcelos, S. & Oliveira, G. (2021). Chloroplast genomes of key species shed light on the evolution of the ancient genus Isoetes. *Journal of Systematics and Evolution*, *59*(3), 429-441. **Citações: 9**



JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12693

Check for updates

Research Article

Chloroplast genomes of key species shed light on the evolution of the ancient genus *Isoetes*

Jovani B. S. Pereira[1,2*], Ana Maria Giulietti[1,3], Eder S. Pires[4], Marcele Laux[4], Maurício T. C. Watanabe[1], Renato R. M. Oliveira[4], Santelmo Vasconcelos[4], and Guilherme Oliveira[4]

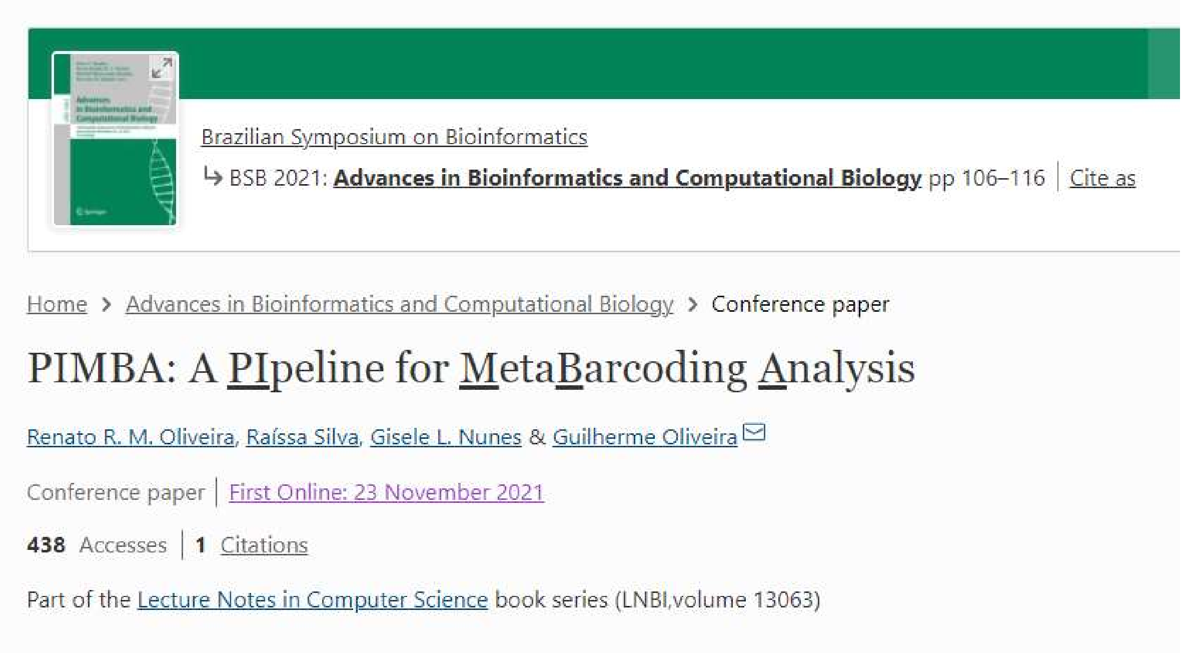
[1]Biodiversity and Ecosystems Services Group, Instituto Tecnológico Vale, Belém, Pará, Brazil
[2]Instituto de Botânica de São Paulo, São Paulo, Brazil
[3]Depto. Botânica, Universidade Estadual de Feira de Santa, Feira de Santana, Bahia, Brazil
[4]Environmental Genomics Group, Instituto Tecnológico Vale, Belém, Pará, Brazil
*Author for correspondence. E-mail: jovanibio@gmail.com

Received 26 September 2019; Accepted 23 September 2020; Article first published online 29 September 2020

Abstract Although phylogenetic studies have revealed major clades, the deepest relationships in *Isoetes* remain unresolved. The use of next-generation sequencing provides enormous amounts of gene sequences, which allows not only clarification of the basal relationships but also rapid radiations. Plastomes of six key *Isoetes* species were annotated, revealing a total of 129 or 130 genes, depending on the species. Our phylogenomic analyses comprising representatives of all major clades yielded well-supported nodes and identical topologies using maximum likelihood and Bayesian inference. The phylogenetic reconstructions detangled the deep relationships in *Isoetes* and illuminated the more recent radiations in the genus. A basal dichotomy was found that grouped *Isoetes spp*. from Brazil and South Africa into a clade sister to the remaining *Isoetes* groups. Interestingly, *I. andicola* was found to be sister to the North American species complex. Genomic trait mapping analysis showed that the missing introns in the *atp*F and *clp*P genes were well conserved in two major clades. The absence of *trnK-UUU* was observed in the Brazilian tropical species and in *I. velata*. Among lycophytes, the gene *trnR-CCG* was missing only in *I. eludens*. In general, genomic traits such as the presence or absence of internal stop codons, a tRNA, and an intron were revealed to be conserved within groups, suggesting that these genomic traits might reveal vital information about the evolution of the genus. This study will contribute to understanding the diversification of *Isoetes* and the establishment of a better framework to address the evolutionary history of the genus.

Key words: internal stop codons, *Isoetes*, lycophytes, next-generation sequencing, phylogenomic, RNA editing, tRNA.

Oliveira, R. R., Silva, R., Nunes, G. L., & Oliveira, G. (2021). PIMBA: A PI peline for M eta B arcoding A nalysis. In *Advances in Bioinformatics and Computational Biology: 14th Brazilian Symposium on Bioinformatics, BSB 2021, Virtual Event, November 22–26, 2021, Proceedings 14* (pp. 106-116). Springer International Publishing. **Citações: 12.**



Brazilian Symposium on Bioinformatics

↳ BSB 2021: **Advances in Bioinformatics and Computational Biology** pp 106–116 | Cite as

Home > Advances in Bioinformatics and Computational Biology > Conference paper

PIMBA: A PIpeline for MetaBarcoding Analysis

Renato R. M. Oliveira, Raíssa Silva, Gisele L. Nunes & Guilherme Oliveira

Conference paper | First Online: 23 November 2021

438 Accesses | **1** Citations

Part of the Lecture Notes in Computer Science book series (LNBI,volume 13063)

Valadares, R. B., Marroni, F., Sillo, F., **Oliveira, R. R**., Balestrini, R., & Perotto, S. (2021). A transcriptomic approach provides insights on the mycorrhizal symbiosis of the mediterranean orchid Limodorum abortivum in nature. *Plants*, *10*(2), 251. **Citações: 9.**

plants

MDPI

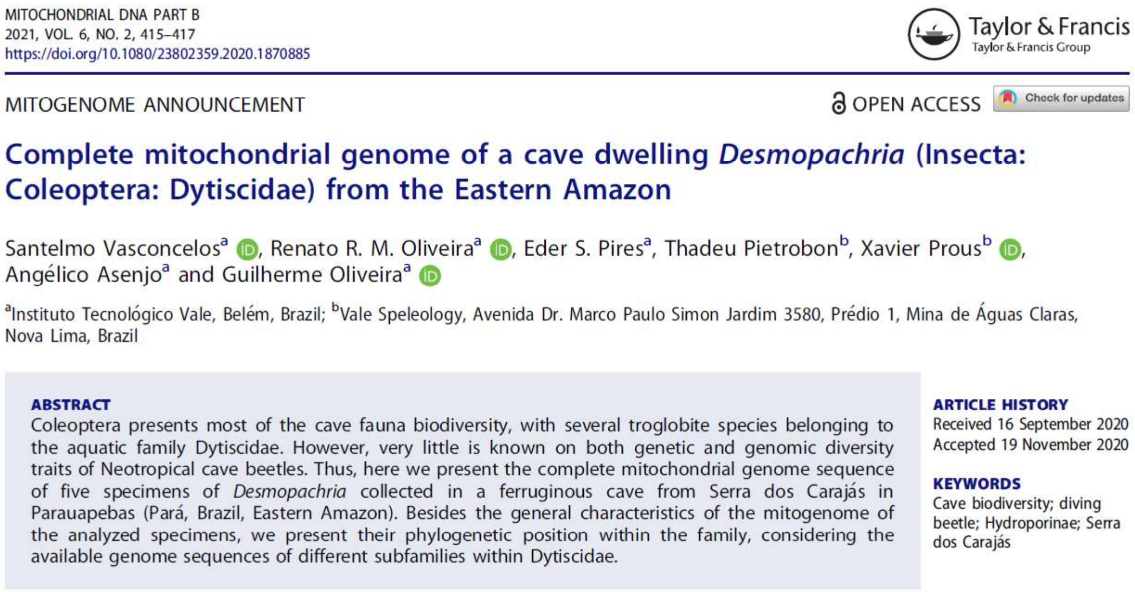
Article

A Transcriptomic Approach Provides Insights on the Mycorrhizal Symbiosis of the Mediterranean Orchid *Limodorum abortivum* in Nature

Rafael B. S. Valadares [1,†], Fabio Marroni [2,3,†], Fabiano Sillo [4,†], Renato R. M. Oliveira [1,5], Raffaella Balestrini [4,*] and Silvia Perotto [6,*]

[1] Instituto Tecnológico Vale, Rua Boaventura da Silva 955, 66050-000 Belém, Pará, Brazil; rafaelbsvaladares@gmail.com (R.B.S.V.); renato.renison@gmail.com (R.R.M.O.)
[2] Dipartimento di Scienze Agroalimentari, Ambientali e Animali, Università di Udine, Via delle Scienze, I-33100 Udine, Italy; marroni@appliedgenomics.org
[3] Istituto di Genomica Applicata, Via Linussio 51, I-33100 Udine, Italy
[4] Consiglio Nazionale Delle Ricerche-Istituto per la Protezione Sostenibile Delle Piante, Viale P.A. Mattioli 25, I-10125 Torino, Italy; fabiano.sillo@ipsp.cnr.it
[5] Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627, 31270-901 Belo Horizonte, Minas Gerais, Brazil
[6] Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, Viale Mattioli 25, I-10125 Torino, Italy
[*] Correspondence: raffaella.balestrini@ipsp.cnr.it (R.B.); silvia.perotto@unito.it (S.P.); Tel.: +39-011-6502927 (R.B.); +39-011-6705987 (S.P.)
[†] These authors equally contributed to the paper as first authors.

Vasconcelos, S., **Oliveira, R. R.**, Pires, E. S., Pietrobon, T., Prous, X., Asenjo, A., & Oliveira, G. (2021). Complete mitochondrial genome of a cave dwelling Desmopachria (Insecta: Coleoptera: Dytiscidae) from the Eastern Amazon. *Mitochondrial DNA Part B*, *6*(2), 415-417.



MITOCHONDRIAL DNA PART B
2021, VOL. 6, NO. 2, 415–417
https://doi.org/10.1080/23802359.2020.1870885

Taylor & Francis
Taylor & Francis Group

MITOGENOME ANNOUNCEMENT

OPEN ACCESS Check for updates

Complete mitochondrial genome of a cave dwelling *Desmopachria* (Insecta: Coleoptera: Dytiscidae) from the Eastern Amazon

Santelmo Vasconcelos[a], Renato R. M. Oliveira[a], Eder S. Pires[a], Thadeu Pietrobon[b], Xavier Prous[b], Angélico Asenjo[a] and Guilherme Oliveira[a]

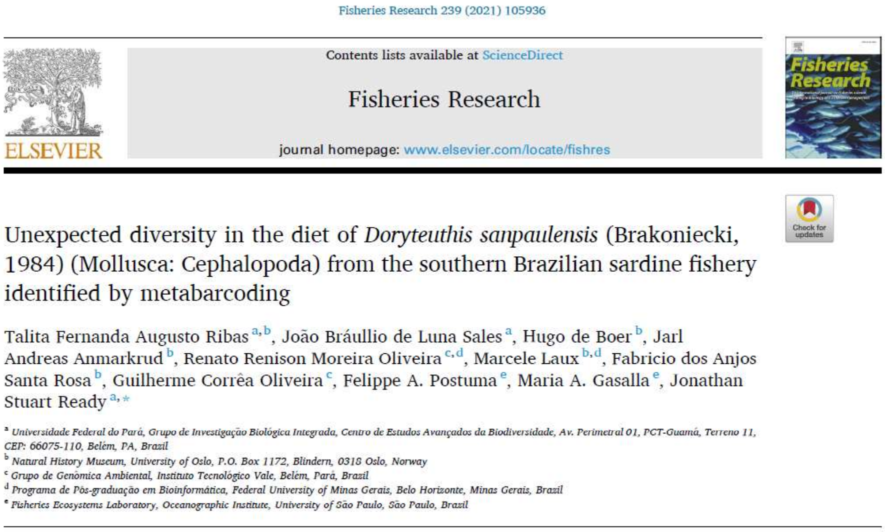
[a]Instituto Tecnológico Vale, Belém, Brazil; [b]Vale Speleology, Avenida Dr. Marco Paulo Simon Jardim 3580, Prédio 1, Mina de Águas Claras, Nova Lima, Brazil

ABSTRACT
Coleoptera presents most of the cave fauna biodiversity, with several troglobite species belonging to the aquatic family Dytiscidae. However, very little is known on both genetic and genomic diversity traits of Neotropical cave beetles. Thus, here we present the complete mitochondrial genome sequence of five specimens of *Desmopachria* collected in a ferruginous cave from Serra dos Carajás in Parauapebas (Pará, Brazil, Eastern Amazon). Besides the general characteristics of the mitogenome of the analyzed specimens, we present their phylogenetic position within the family, considering the available genome sequences of different subfamilies within Dytiscidae.

ARTICLE HISTORY
Received 16 September 2020
Accepted 19 November 2020

KEYWORDS
Cave biodiversity; diving beetle; Hydroporinae; Serra dos Carajás

Ribas, T. F. A., de Luna Sales, J. B., de Boer, H., Anmarkrud, J. A., **Oliveira, R. R. M.**, Laux, M., Rosa, F. dos A. S., Oliveira, G. C., Postuma, F. A., Gasalla, M. A. & Ready, J. S. (2021). Unexpected diversity in the diet of Doryteuthis sanpaulensis (Brakoniecki, 1984)(Mollusca: Cephalopoda) from the southern Brazilian sardine fishery identified by metabarcoding. *Fisheries Research*, *239*, 105936. **Citações: 1.**



Fisheries Research 239 (2021) 105936

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

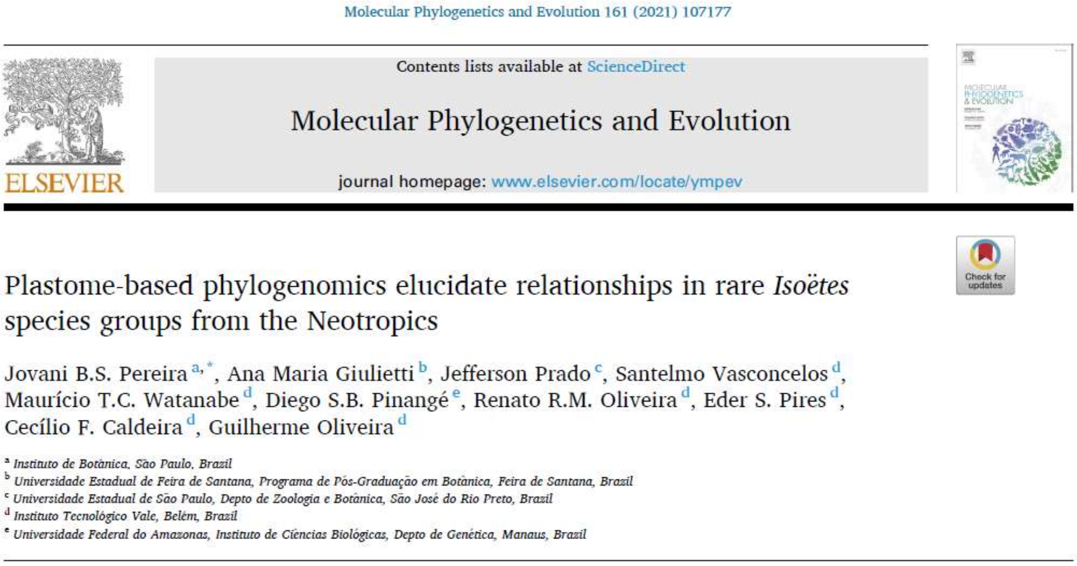
ELSEVIER

Unexpected diversity in the diet of *Doryteuthis sanpaulensis* (Brakoniecki, 1984) (Mollusca: Cephalopoda) from the southern Brazilian sardine fishery identified by metabarcoding

Talita Fernanda Augusto Ribas[a,b], João Bráullio de Luna Sales[a], Hugo de Boer[b], Jarl Andreas Anmarkrud[b], Renato Renison Moreira Oliveira[c,d], Marcele Laux[b,d], Fabricio dos Anjos Santa Rosa[b], Guilherme Corrêa Oliveira[c], Felippe A. Postuma[e], Maria A. Gasalla[e], Jonathan Stuart Ready[a,*]

[a] *Universidade Federal do Pará, Grupo de Investigação Biológica Integrada, Centro de Estudos Avançados da Biodiversidade, Av. Perimetral 01, PCT-Guamá, Terreno 11, CEP: 66075-110, Belém, PA, Brazil*
[b] *Natural History Museum, University of Oslo, P.O. Box 1172, Blindern, 0318 Oslo, Norway*
[c] *Grupo de Genômica Ambiental, Instituto Tecnológico Vale, Belém, Pará, Brazil*
[d] *Programa de Pós-graduação em Bioinformática, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil*
[e] *Fisheries Ecosystems Laboratory, Oceanographic Institute, University of São Paulo, São Paulo, Brazil*

Pereira, J. B., Giulietti, A. M., Prado, J., Vasconcelos, S., Watanabe, M. T., Pinangé, D. S., **Oliveira, R. R. M.**, Pires, E., Caldeira, C. F., & Oliveira, G. (2021). Plastome-based phylogenomics elucidate relationships in rare Isoëtes species groups from the Neotropics. *Molecular Phylogenetics and Evolution*, *161*, 107177. **Citações: 14.**



Molecular Phylogenetics and Evolution 161 (2021) 107177

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

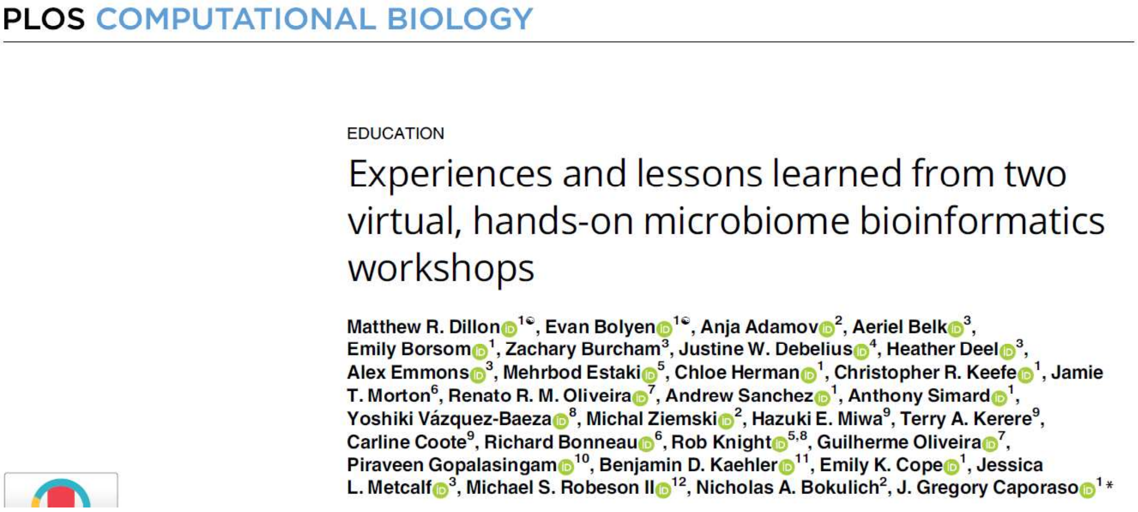
ELSEVIER

Plastome-based phylogenomics elucidate relationships in rare *Isoëtes* species groups from the Neotropics

Jovani B.S. Pereira[a,*], Ana Maria Giulietti[b], Jefferson Prado[c], Santelmo Vasconcelos[d], Maurício T.C. Watanabe[d], Diego S.B. Pinangé[e], Renato R.M. Oliveira[d], Eder S. Pires[d], Cecílio F. Caldeira[d], Guilherme Oliveira[d]

[a] *Instituto de Botânica, São Paulo, Brazil*
[b] *Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Botânica, Feira de Santana, Brazil*
[c] *Universidade Estadual de São Paulo, Depto de Zoologia e Botânica, São José do Rio Preto, Brazil*
[d] *Instituto Tecnológico Vale, Belém, Brazil*
[e] *Universidade Federal do Amazonas, Instituto de Ciências Biológicas, Depto de Genética, Manaus, Brazil*

Dillon, M. R., Bolyen, E., Adamov, A., Belk, A., Borsom, E., Burcham, Z., Debelius, J. W., Deel, H., Emmons, A., Estaki, M., Herman, C., Keefe, C. R., Morton, J. T., **Oliveira, R. R. M.**, Sanchez, A., Simard, A., Vázquez-Baeza, Y., Ziemski, M., Miwa, H. E., ... & Caporaso, J. G. (2021). Experiences and lessons learned from two virtual, hands-on microbiome bioinformatics workshops. *PLoS computational biology*, *17*(6), e1009056. **Citações: 2.**



PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Experiences and lessons learned from two virtual, hands-on microbiome bioinformatics workshops

Matthew R. Dillon[1☯], Evan Bolyen[1☯], Anja Adamov[2], Aeriel Belk[3], Emily Borsom[1], Zachary Burcham[3], Justine W. Debelius[4], Heather Deel[3], Alex Emmons[3], Mehrbod Estaki[5], Chloe Herman[1], Christopher R. Keefe[1], Jamie T. Morton[6], Renato R. M. Oliveira[7], Andrew Sanchez[1], Anthony Simard[1], Yoshiki Vázquez-Baeza[8], Michal Ziemski[2], Hazuki E. Miwa[9], Terry A. Kerere[9], Carline Coote[9], Richard Bonneau[6], Rob Knight[5,8], Guilherme Oliveira[7], Piraveen Gopalasingam[10], Benjamin D. Kaehler[11], Emily K. Cope[1], Jessica L. Metcalf[3], Michael S. Robeson II[12], Nicholas A. Bokulich[2], J. Gregory Caporaso[1]*

Vasconcelos, S., Nunes, G. L., Dias, M. C., Lorena, J., **Oliveira, R. R.**, Lima, T. G., Pires, E. S., Valadares, R. B. S., Alves, R., Watanabe, M. T. C., Zappi, D. C., Hiura, A. L., Pastore, M., Vasconcelos, L. v., Mota, N. F. O., Viana, P. L., Gil, A. S. B., Simões, A. O., Imperatriz-Fonseca, V. L., … & Oliveira, G. (2021). Unraveling the plant diversity of the Amazonian canga through DNA barcoding. *Ecology and Evolution*, *11*(19), 13348-13362. **Citações: 4.**

Received: 17 July 2021 | Revised: 3 August 2021 | Accepted: 11 August 2021

DOI: 10.1002/ece3.8057

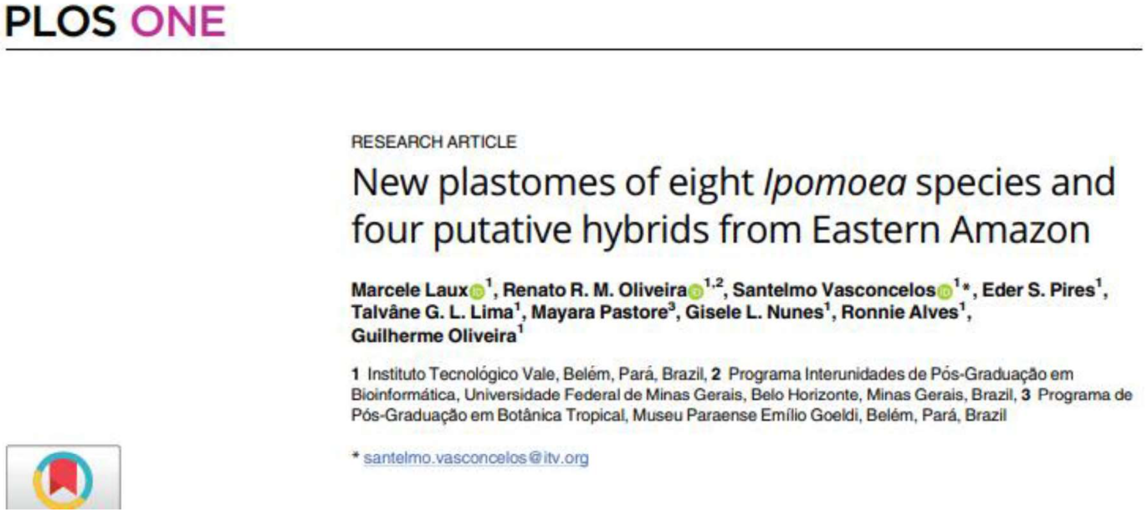
ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

Unraveling the plant diversity of the Amazonian *canga* through DNA barcoding

Santelmo Vasconcelos[1] | Gisele L. Nunes[1] | Mariana C. Dias[1,2] | Jamily Lorena[1] | Renato R. M. Oliveira[1,2] | Talvâne G. L. Lima[1] | Eder S. Pires[1] | Rafael B. S. Valadares[1] | Ronnie Alves[1] | Maurício T. C. Watanabe[1] | Daniela C. Zappi[1,3] | Alice L. Hiura[1] | Mayara Pastore[1,4] | Liziane V. Vasconcelos[1,5] | Nara F. O. Mota[1,4] | Pedro L. Viana[4] | André S. B. Gil[4] | André O. Simões[6] | Vera L. Imperatriz-Fonseca[1,7] | Raymond M. Harley[8] | Ana M. Giulietti[1,9] | Guilherme Oliveira[1]

Laux, M., **Oliveira, R. R.**, Vasconcelos, S., Pires, E. S., Lima, T. G., Pastore, M., Nunes, G. L., Alves, R., & Oliveira, G. (2022). New plastomes of eight Ipomoea species and four putative hybrids from Eastern Amazon. *Plos one*, *17*(3), e0265449. **Citações: 2**



PLOS ONE

RESEARCH ARTICLE

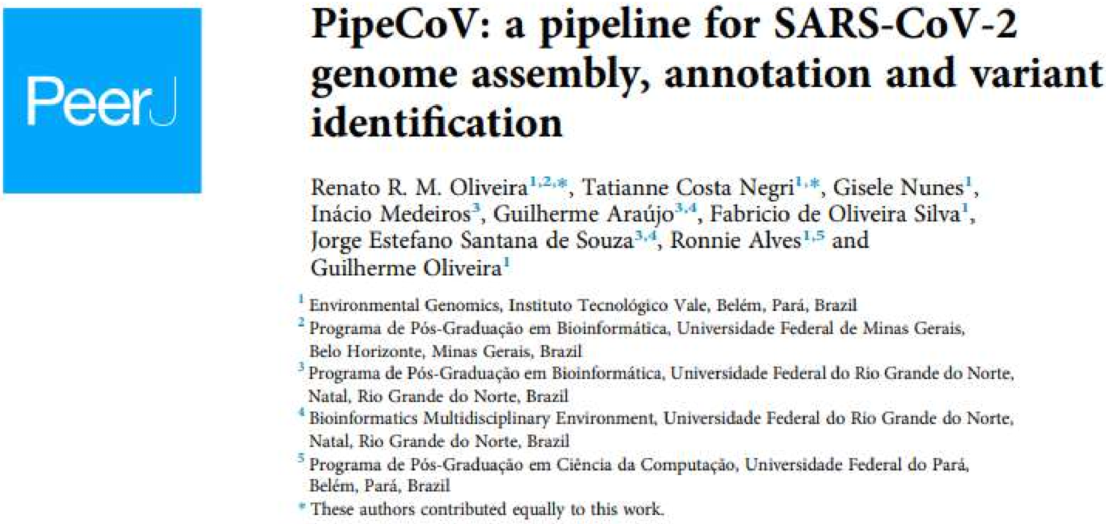
New plastomes of eight *Ipomoea* species and four putative hybrids from Eastern Amazon

Marcele Laux[1], Renato R. M. Oliveira[1,2], Santelmo Vasconcelos[1]*, Eder S. Pires[1], Talvâne G. L. Lima[1], Mayara Pastore[3], Gisele L. Nunes[1], Ronnie Alves[1], Guilherme Oliveira[1]

1 Instituto Tecnológico Vale, Belém, Pará, Brazil, **2** Programa Interunidades de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** Programa de Pós-Graduação em Botânica Tropical, Museu Paraense Emílio Goeldi, Belém, Pará, Brazil

* santelmo.vasconcelos@itv.org

Oliveira, R. R., Negri, T. C., Nunes, G., Medeiros, I., Araújo, G., de Oliveira Silva, F., de Souza, J. E. S., Alves, R., & Oliveira, G. (2022). PipeCoV: a pipeline for SARS-CoV-2 genome assembly, annotation and variant identification. *PeerJ*, *10*, e13300. **Citações: 4. (*Top-5 most viewed paper in PeerJ*)**



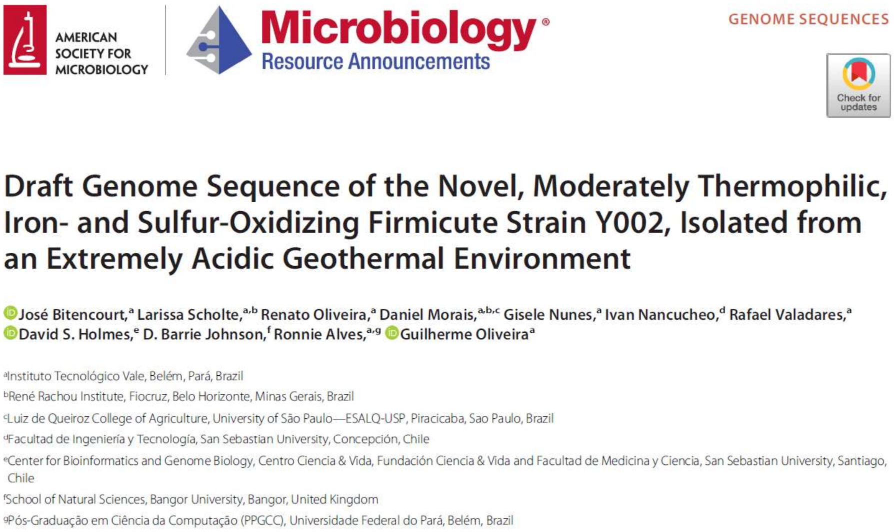
PeerJ

PipeCoV: a pipeline for SARS-CoV-2 genome assembly, annotation and variant identification

Renato R. M. Oliveira[1,2,*], Tatianne Costa Negri[1,*], Gisele Nunes[1], Inácio Medeiros[3], Guilherme Araújo[3,4], Fabricio de Oliveira Silva[1], Jorge Estefano Santana de Souza[3,4], Ronnie Alves[1,5] and Guilherme Oliveira[1]

[1] Environmental Genomics, Instituto Tecnológico Vale, Belém, Pará, Brazil
[2] Programa de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[3] Programa de Pós-Graduação em Bioinformática, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil
[4] Bioinformatics Multidisciplinary Environment, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil
[5] Programa de Pós-Graduação em Ciência da Computação, Universidade Federal do Pará, Belém, Pará, Brazil
[*] These authors contributed equally to this work.

Bitencourt, J., Scholte, L., **Oliveira, R.**, Morais, D., Nunes, G., Nancucheo, I., Valadares, R., Holmes, D. S., Johnson, D. B., Alves, R., & Oliveira, G. (2022). Draft Genome Sequence of the Novel, Moderately Thermophilic, Iron-and Sulfur-Oxidizing Firmicute Strain Y002, Isolated from an Extremely Acidic Geothermal Environment. *Microbiology Resource Announcements*, *11*(6), e00149-22. **Citações: 1.**



AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology® Resource Announcements

GENOME SEQUENCES

Check for updates

Draft Genome Sequence of the Novel, Moderately Thermophilic, Iron- and Sulfur-Oxidizing Firmicute Strain Y002, Isolated from an Extremely Acidic Geothermal Environment

José Bitencourt,[a] Larissa Scholte,[a,b] Renato Oliveira,[a] Daniel Morais,[a,b,c] Gisele Nunes,[a] Ivan Nancucheo,[d] Rafael Valadares,[a] David S. Holmes,[e] D. Barrie Johnson,[f] Ronnie Alves,[a,g] Guilherme Oliveira[a]

[a]Instituto Tecnológico Vale, Belém, Pará, Brazil
[b]René Rachou Institute, Fiocruz, Belo Horizonte, Minas Gerais, Brazil
[c]Luiz de Queiroz College of Agriculture, University of São Paulo—ESALQ-USP, Piracicaba, Sao Paulo, Brazil
[d]Facultad de Ingeniería y Tecnología, San Sebastian University, Concepción, Chile
[e]Center for Bioinformatics and Genome Biology, Centro Ciencia & Vida, Fundación Ciencia & Vida and Facultad de Medicina y Ciencia, San Sebastian University, Santiago, Chile
[f]School of Natural Sciences, Bangor University, Bangor, United Kingdom
[g]Pós-Graduação em Ciência da Computação (PPGCC), Universidade Federal do Pará, Belém, Brazil

Frederico, T. D., Nancucheo, I., Santos, W. C. B., **Oliveira, R. R. M.**, Buzzi, D. C., Pires, E. S., Silva, P. M. P., Lucheta, A. R., Alves, J. O., Oliveira, G., & Bitencourt, J. A. P. (2022). Comparison of two acidophilic sulfidogenic consortia for the treatment of acidic mine water. *Frontiers in Bioengineering and Biotechnology*, *10*.

frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Original Research
PUBLISHED 29 November 2022
DOI 10.3389/fbioe.2022.1048412

Check for updates

OPEN ACCESS

EDITED BY
Rajesh K. Sani,
South Dakota School of Mines and Technology, United States

REVIEWED BY
Pedro Antonio Galleguillos,
Centro de Investigación Científico y Tecnológico para la Minería (CICITEM), Chile
Javier Sánchez-España,
Spanish National Research Council (CSIC), Spain

*CORRESPONDENCE
Ivan Nancucheo,
ivan.nancucheo@uss.cl
José Augusto Pires Bitencourt,
jose.augusto.bitencourt@itv.org

SPECIALTY SECTION

Comparison of two acidophilic sulfidogenic consortia for the treatment of acidic mine water

Tayná Diniz Frederico[1], Ivan Nancucheo[2]*, Werica Colaço Barros Santos[1], Renato Renison Moreira Oliveira[1], Daniella Cardoso Buzzi[3], Eder Soares Pires[1], Patricia Magalhães Pereira Silva[4], Adriano Reis Lucheta[4], Joner Oliveira Alves[4], Guilherme Corrêa de Oliveira[1] and José Augusto Pires Bitencourt[1]*

[1]Instituto Tecnológico Vale, Belém, Brazil, [2]Facultad de Ingeniería, Arquitectura y Diseño. Universidad San Sebastián, Concepción, Chile, [3]REDEMAT/Universidade Federal de Ouro Preto (UFOP), Ouro Preto, Brazil, [4]Instituto SENAI de Inovação em Tecnologias Minerais, Belém, Brazil

Oliveira, R. R., Vasconcelos, S., & Oliveira, G. (2022). SPLACE: A tool to automatically SPLit, Align, and ConcatenatE genes for phylogenomic inference of several organisms. *Frontiers in Bioinformatics*, *2*.

frontiers | Frontiers in Bioinformatics

TYPE Brief Research Report
PUBLISHED 08 December 2022
DOI 10.3389/fbinf.2022.1074802

Check for updates

OPEN ACCESS

EDITED BY
Arne Seitz,
Swiss Federal Institute of Technology Lausanne, Switzerland

REVIEWED BY
Yafei Mao,
Shanghai Jiao Tong University, China
Guanglin Li,
Shaanxi Normal University, China

*CORRESPONDENCE
Renato R. M. Oliveira,
renato.renison@gmail.com

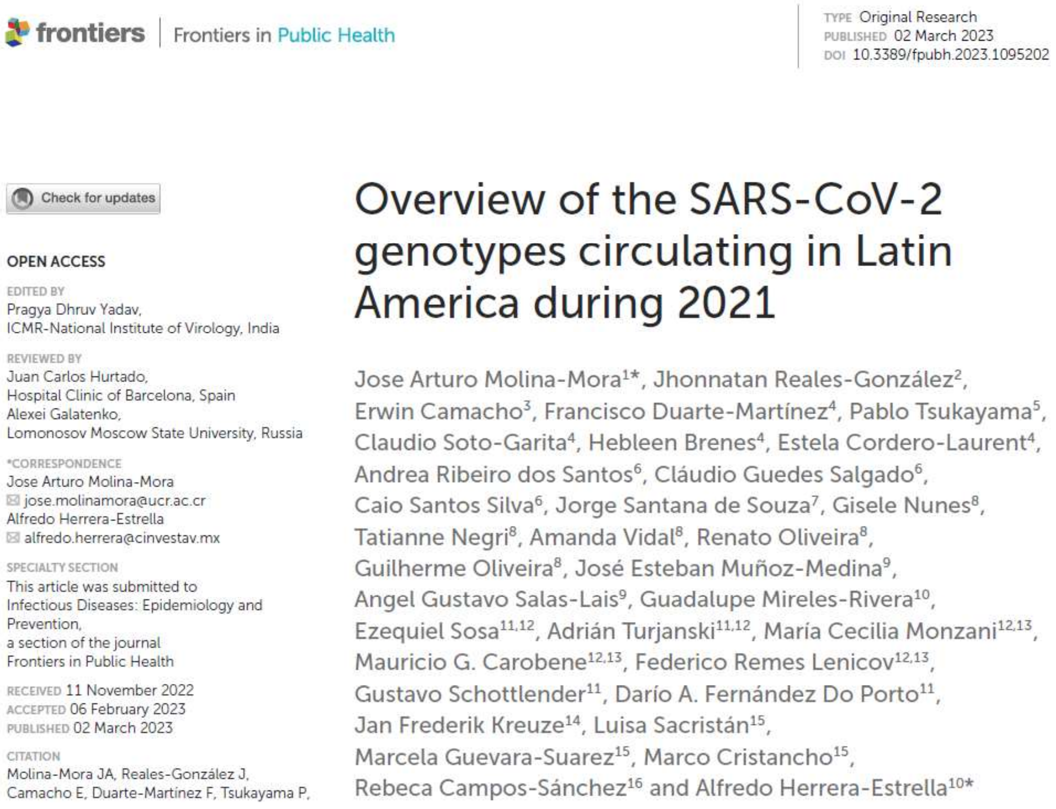
SPECIALTY SECTION
This article was submitted to

SPLACE: A tool to automatically SPLit, Align, and ConcatenatE genes for phylogenomic inference of several organisms

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1] and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Molina-Mora, J. A., Reales-Gonzalez, J., Camacho, E., Duarte-Martinez, F., Tsukayama, P., Soto-Garita, C., **Oliveira, R. R. M.**, ... & Herrera-Estrella, A. (2022). Overview of the SARS-CoV-2 genotypes circulating in Latin America during 2021. *Frontiers in Public Health, 11.*
Citation: 1.



frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 02 March 2023
DOI 10.3389/fpubh.2023.1095202

Check for updates

Overview of the SARS-CoV-2 genotypes circulating in Latin America during 2021

Jose Arturo Molina-Mora[1]*, Jhonnatan Reales-González[2], Erwin Camacho[3], Francisco Duarte-Martínez[4], Pablo Tsukayama[5], Claudio Soto-Garita[4], Hebleen Brenes[4], Estela Cordero-Laurent[4], Andrea Ribeiro dos Santos[6], Cláudio Guedes Salgado[6], Caio Santos Silva[6], Jorge Santana de Souza[7], Gisele Nunes[8], Tatianne Negri[8], Amanda Vidal[8], Renato Oliveira[8], Guilherme Oliveira[8], José Esteban Muñoz-Medina[9], Angel Gustavo Salas-Lais[9], Guadalupe Mireles-Rivera[10], Ezequiel Sosa[11,12], Adrián Turjanski[11,12], María Cecilia Monzani[12,13], Mauricio G. Carobene[12,13], Federico Remes Lenicov[12,13], Gustavo Schottlender[11], Darío A. Fernández Do Porto[11], Jan Frederik Kreuze[14], Luisa Sacristán[15], Marcela Guevara-Suarez[15], Marco Cristancho[15], Rebeca Campos-Sánchez[16] and Alfredo Herrera-Estrella[10]*

OPEN ACCESS

EDITED BY
Pragya Dhruv Yadav,
ICMR-National Institute of Virology, India

REVIEWED BY
Juan Carlos Hurtado,
Hospital Clinic of Barcelona, Spain
Alexei Galatenko,
Lomonosov Moscow State University, Russia

*CORRESPONDENCE
Jose Arturo Molina-Mora
✉ jose.molinamora@ucr.ac.cr
Alfredo Herrera-Estrella
✉ alfredo.herrera@cinvestav.mx

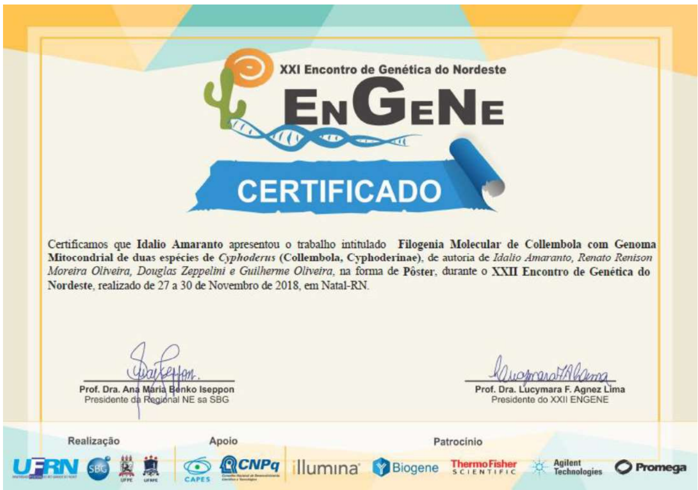
SPECIALTY SECTION
This article was submitted to Infectious Diseases: Epidemiology and Prevention,
a section of the journal
Frontiers in Public Health

RECEIVED 11 November 2022
ACCEPTED 06 February 2023
PUBLISHED 02 March 2023

CITATION
Molina-Mora JA, Reales-González J,
Camacho E, Duarte-Martinez F, Tsukayama P,

6.3 Participação em Congressos e Eventos

6.3.1 XXI Encontro de Genética do Nordeste



XXI Encontro de Genética do Nordeste

ENGENE

CERTIFICADO

Certificamos que **Idalio Amaranto** apresentou o trabalho intitulado **Filogenia Molecular de Collembola com Genoma Mitocondrial de duas espécies de *Cyphoderus* (Collembola, Cyphoderinae)**, de autoria de *Idalio Amaranto, Renato Renison Moreira Oliveira, Douglas Zeppelini e Guilherme Oliveira*, na forma de **Pôster**, durante o **XXII Encontro de Genética do Nordeste**, realizado de 27 a 30 de Novembro de 2018, em Natal-RN.

Prof. Dra. Ana Maria Benko Iseppon
Presidente da Regional NE sa SBG

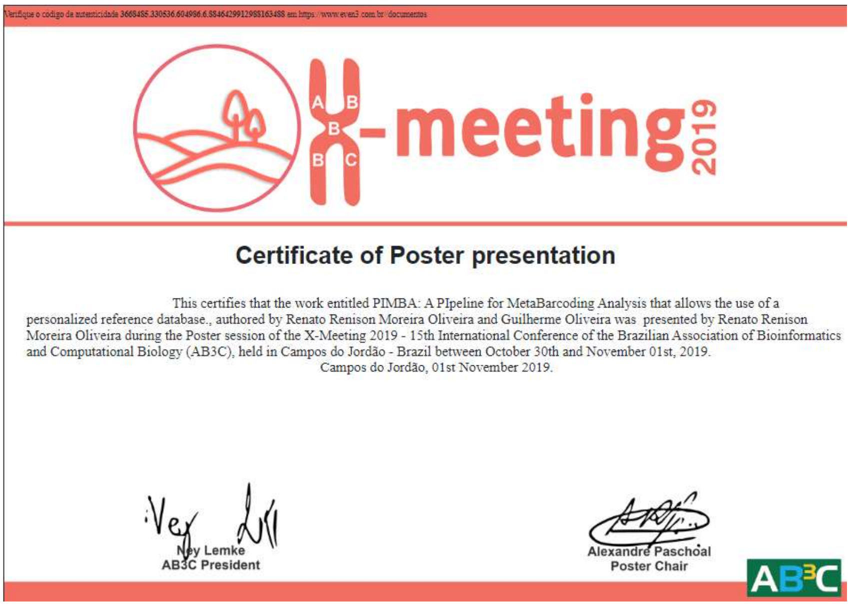
Prof. Dra. Lucymara F. Agnez Lima
Presidente do XXII ENGENE

Realização

Apoio

Patrocínio

6.3.2 X-meeting 2019 - 15th International Conference of the AB3C



Verifique o codigo de autenticidade 3668485.330536.604986.6.8846429912988163488 em https://www.even3.com.br/documentos

ABBBC
X-meeting 2019

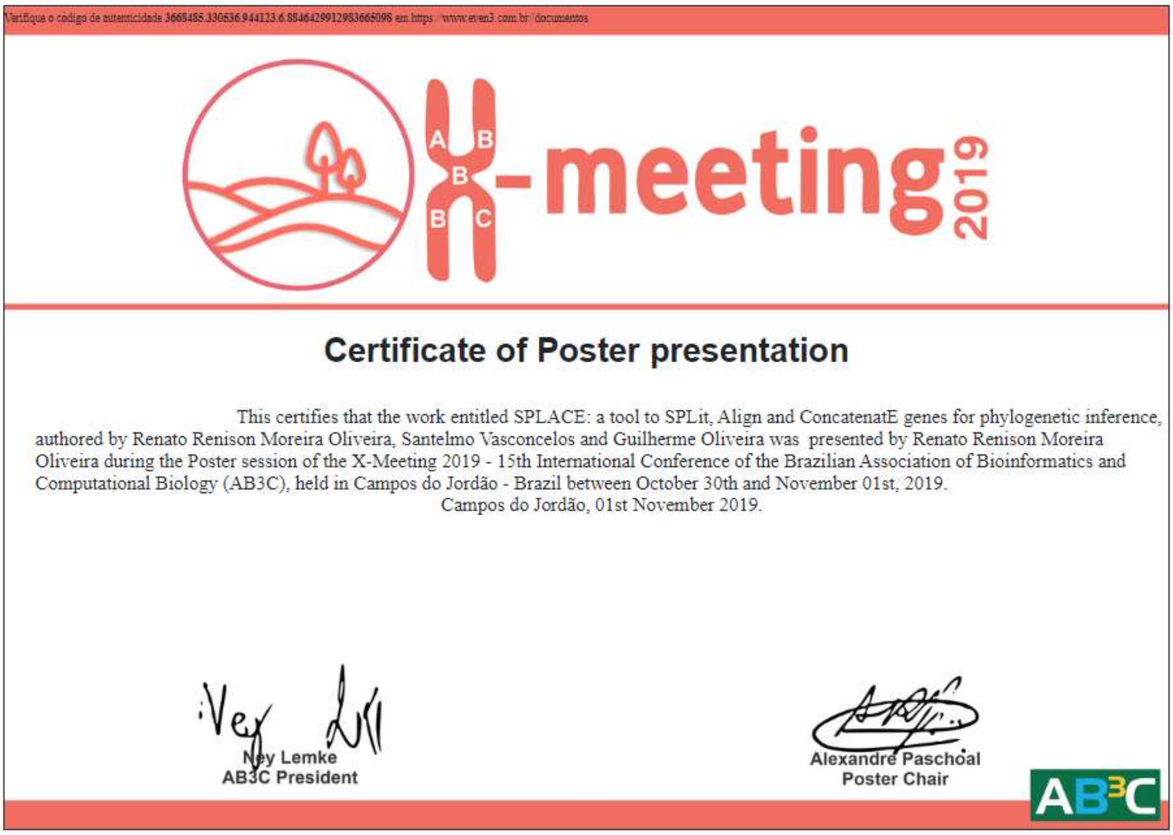
Certificate of Poster presentation

This certifies that the work entitled PIMBA: A PIpeline for MetaBarcoding Analysis that allows the use of a personalized reference database., authored by Renato Renison Moreira Oliveira and Guilherme Oliveira was presented by Renato Renison Moreira Oliveira during the Poster session of the X-Meeting 2019 - 15th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Campos do Jordão - Brazil between October 30th and November 01st, 2019.

Campos do Jordão, 01st November 2019.

Ney Lemke
AB3C President

Alexandre Paschoal
Poster Chair



Verifique o código de autenticidade 3668488.330536.944123.6.8846429912983665098 em https://www.even3.com.br/documentos

X-meeting 2019

Certificate of Poster presentation

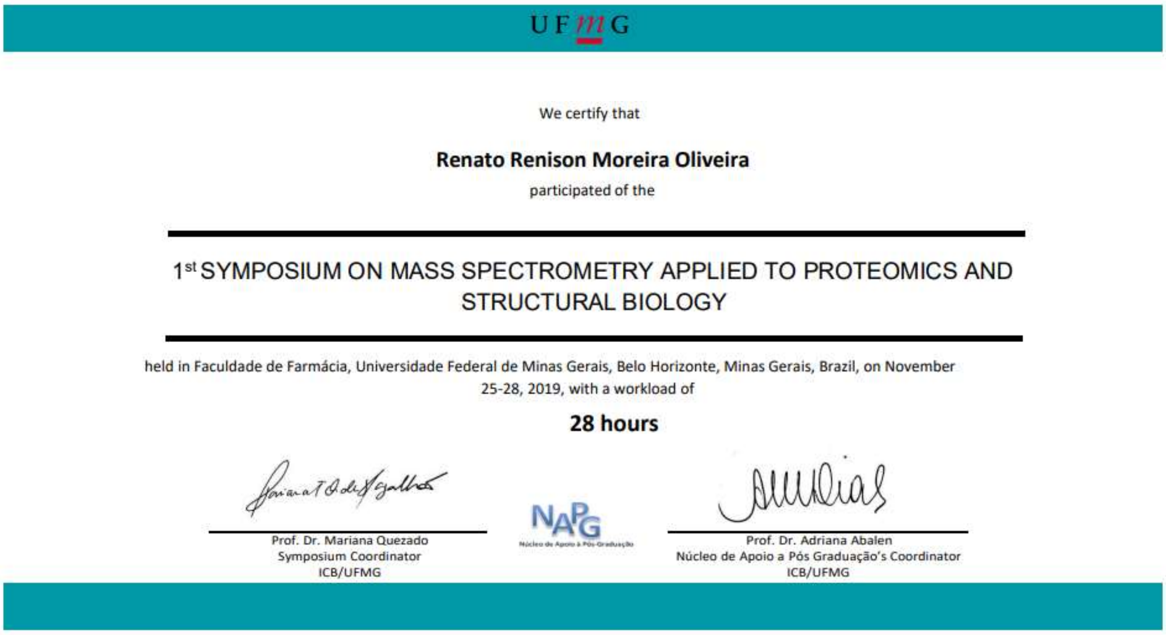
This certifies that the work entitled SPLACE: a tool to SPLit, Align and ConcatenatE genes for phylogenetic inference, authored by Renato Renison Moreira Oliveira, Santelmo Vasconcelos and Guilherme Oliveira was presented by Renato Renison Moreira Oliveira during the Poster session of the X-Meeting 2019 - 15th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Campos do Jordão - Brazil between October 30th and November 01st, 2019.

Campos do Jordão, 01st November 2019.

Ney Lemke
AB3C President

Alexandre Paschoal
Poster Chair

6.3.3 1st SYMPOSIUM ON MASS SPECTROMETRY APPLIED TO PROTEOMICS AND STRUCTURAL BIOLOGY



We certify that

Renato Renison Moreira Oliveira

participated of the

1st SYMPOSIUM ON MASS SPECTROMETRY APPLIED TO PROTEOMICS AND STRUCTURAL BIOLOGY

held in Faculdade de Farmácia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, on November 25-28, 2019, with a workload of

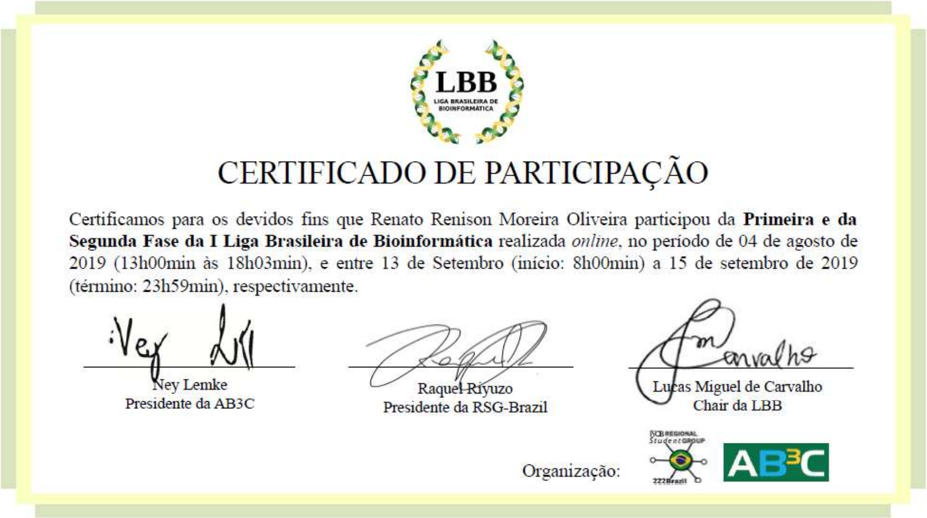
28 hours

Prof. Dr. Mariana Quezado
Symposium Coordinator
ICB/UFMG

Prof. Dr. Adriana Abalen
Núcleo de Apoio a Pós Graduação's Coordinator
ICB/UFMG

6.3.4 I Liga Brasileira de Bioinformática

Participação na Primeira e Segunda fase da Liga Brasileira de Bioinformática



LBB
LIGA BRASILEIRA DE BIOINFORMÁTICA

CERTIFICADO DE PARTICIPAÇÃO

Certificamos para os devidos fins que Renato Renison Moreira Oliveira participou da **Primeira e da Segunda Fase da I Liga Brasileira de Bioinformática** realizada *online*, no período de 04 de agosto de 2019 (13h00min às 18h03min), e entre 13 de Setembro (início: 8h00min) a 15 de setembro de 2019 (término: 23h59min), respectivamente.

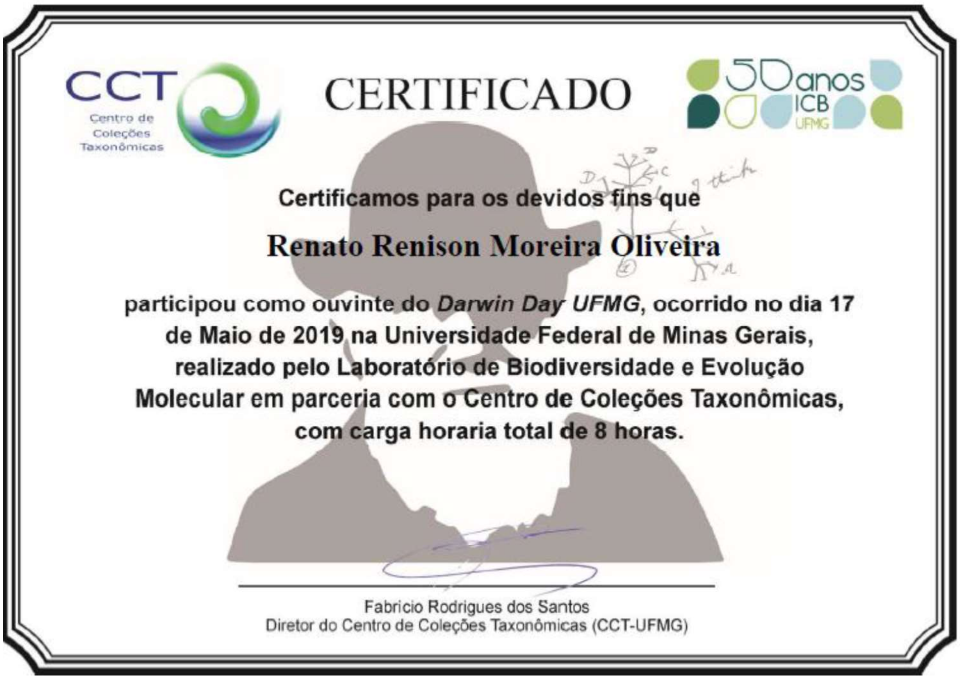
Ney Lemke
Presidente da AB3C

Raquel Riyuzo
Presidente da RSG-Brazil

Lucas Miguel de Carvalho
Chair da LBB

Organização: ISCB REGIONAL Student GROUP RSG Brazil AB3C

6.3.5 Darwin Day 2019



CCT Centro de Coleções Taxonômicas

CERTIFICADO

50 anos ICB UFMG

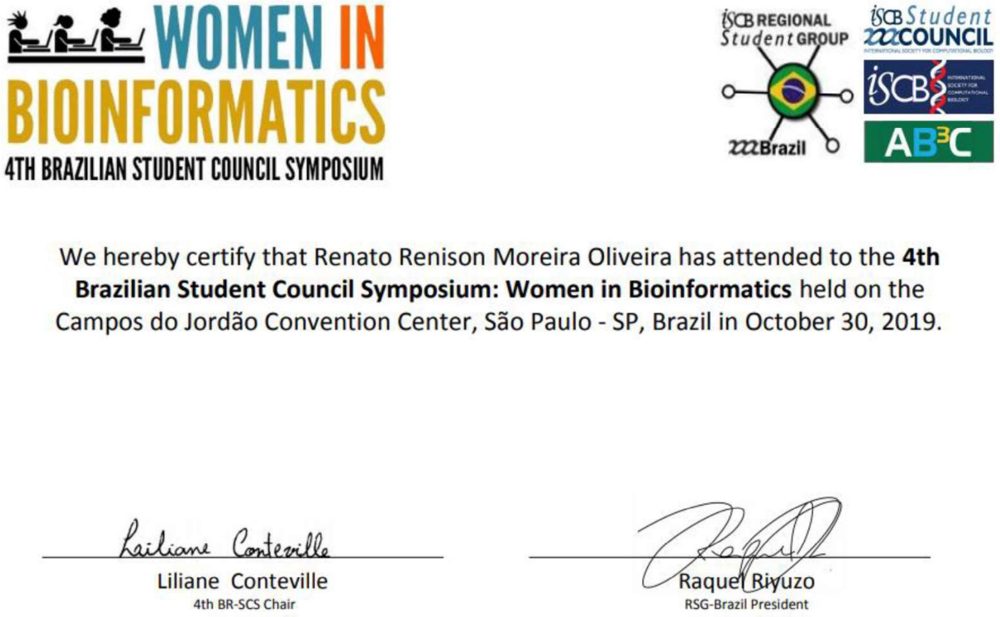
Certificamos para os devidos fins que

Renato Renison Moreira Oliveira

participou como ouvinte do *Darwin Day UFMG*, ocorrido no dia 17 de Maio de 2019 na Universidade Federal de Minas Gerais, realizado pelo Laboratório de Biodiversidade e Evolução Molecular em parceria com o Centro de Coleções Taxonômicas, com carga horaria total de 8 horas.

Fabricio Rodrigues dos Santos
Diretor do Centro de Coleções Taxonômicas (CCT-UFMG)

6.3.6 4th Brazilian Student Council Symposium: Women in Bioinformatics



WOMEN IN BIOINFORMATICS
4TH BRAZILIAN STUDENT COUNCIL SYMPOSIUM

ISCB REGIONAL Student GROUP Brazil

ISCB Student COUNCIL

ISCB

AB³C

We hereby certify that Renato Renison Moreira Oliveira has attended to the **4th Brazilian Student Council Symposium: Women in Bioinformatics** held on the Campos do Jordão Convention Center, São Paulo - SP, Brazil in October 30, 2019.

Liliane Conteville

Liliane Conteville
4th BR-SCS Chair

Raquel Riyuzo
RSG-Brazil President

7 REFERÊNCIAS BIBLIOGRÁFICAS

ABARENKOV, K. et al. **The UNITE database for molecular identification of fungi – recent updates and future perspectives**. **The New Phytologist**WileyNew Phytologist Trust, , 2010. Disponível em: <https://www.jstor.org/stable/27797548>. Acesso em: 22 ago. 2020

ALLENDORF, F. W.; HOHENLOHE, P. A.; LUIKART, G. Genomics and the future of conservation genetics. **Nature Reviews Genetics 2010 11:10**, v. 11, n. 10, p. 697–709, 17 set. 2010a.

ALLENDORF, F. W.; HOHENLOHE, P. A.; LUIKART, G. Genomics and the future of conservation genetics. **Nature Reviews Genetics 2010 11:10**, v. 11, n. 10, p. 697–709, 17 set. 2010b.

BABIYCHUK, E. et al. Natural history of the narrow endemics Ipomoea cavalcantei and I. marabaensis from Amazon Canga savannahs. **Scientific Reports**, v. 7, n. 1, p. 1–15, 1 dez. 2017.

BALVANERA, P. et al. Quantifying the evidence for biodiversity effects on ecosystem functioning and services. **Ecology Letters**, v. 9, n. 10, p. 1146–1156, 1 out. 2006.

BANKEVICH, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, 7 maio 2012.

BAUM, D. A.; SMITH, S. D. **Tree thinking : an introduction to phylogenetic biology**. [s.l.] Roberts, 2012.

BELSER, C. et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. **Nature plants**, v. 4, n. 11, p. 879–887, 1 nov. 2018.

BENTLEY, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. **Nature**, v. 456, n. 7218, p. 53–59, 6 nov. 2008.

BENTO, M. et al. Size matters in Triticeae polyploids: larger genomes have higher remodeling. **Genome**, v. 54, n. 3, p. 175–183, mar. 2011.

BESSE, P.; DA SILVA, D.; GRISONI, M. Plant DNA Barcoding Principles and Limits: A Case Study in the Genus Vanilla. **Methods in Molecular Biology**, v. 2222, p. 131–148, 2021.

BONAN, G. B. Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. **Science**, v. 320, n. 5882, p. 1444–1449, 13 jun. 2008.

BOYER, F. et al. obitools : a unix -inspired software package for DNA metabarcoding. **Molecular Ecology Resources**, v. 16, n. 1, p. 176–182, 1 jan. 2016.

BRADNAM, K. R. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. **GigaScience**, v. 2, n. 1, p. 10, 22 dez. 2013.

BRÄUTIGAM, A.; GOWIK, U. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. **Plant Biology**, v. 12, n. 6, p. 831–841, 1 nov. 2010.

BROOKS, T. M. et al. Habitat Loss and Extinction in the Hotspots of Biodiversity. **Conservation Biology**, v. 16, n. 4, p. 909–923, 1 ago. 2002.

BRUMMITT, N.; ARAÚJO, A. C.; HARRIS, T. Areas of plant diversity—What do we know? **Plants, People, Planet**, v. 3, n. 1, p. 33–44, 1 jan. 2021.

BUTLER, J. et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. **Genome research**, v. 18, n. 5, p. 810–20, 1 maio 2008.

CAO, J. et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. **Nature Genetics**, v. 43, n. 10, p. 956–965, 28 out. 2011.

CAPORASO, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. **Nature Methods**, v. 7, n. 5, p. 335–336, 11 maio 2010.

CARDINALE, B. J. et al. Biodiversity loss and its impact on humanity. **Nature 2012 486:7401**, v. 486, n. 7401, p. 59–67, 6 jun. 2012.

CARDOSO, A. F. et al. Acquiring Iron-Reducing Enrichment Cultures: Environments, Methods and Quality Assessments. **Microorganisms**, v. 11, n. 2, p. 448, 1 fev. 2023.

CARVALHO, C. S. et al. Combining genotype, phenotype, and environmental data to delineate site-adjusted provenance strategies for ecological restoration. **Molecular Ecology Resources**, p. 1755– 0998.13191, 23 jun. 2020.

CATARINA, V. et al. **MONITORAMENTO DA BIODIVERSIDADE DA FLORA DE CANGA, SERRA DOS CARAJÁS, PARÁ, ATRAVÉS DE DNA METABARCODING RELATÓRIO PARCIAL OU FINAL DO PROJETO**. [s.l: s.n.]. . Acesso em: 7 mar. 2023.

CBOL PLANT WORKING GROUP, C. P. W. et al. A DNA barcode for land plants. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 31, p. 12794–7, 4 ago. 2009.

CHALHOUB, B. et al. Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. **Science**, v. 345, n. 6199, p. 950–953, 22 ago. 2014.

CHENG, H. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. **Nature Methods 2021 18:2**, v. 18, n. 2, p. 170–175, 1 fev. 2021.

CHIN, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. **Nature Methods**, v. 13, n. 12, p. 1050–1054, 1 dez. 2016.

CHOULET, F. et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. **Plant Cell**, v. 22, n. 6, p. 1686–1701, 1 jun. 2010.

CLARK, R. M. et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. **Science**, v. 317, n. 5836, p. 338–342, 20 jul. 2007.

CLAROS, M. G. et al. Why Assembling Plant Genome Sequences Is So Challenging. **Biology**, v. 1, n. 2, p. 439, 2012.

COISSAC, E.; RIAZ, T.; PUILLANDRE, N. Bioinformatic challenges for DNA metabarcoding of plants and animals. **Molecular Ecology**, v. 21, n. 8, p. 1834–1847, 1 abr. 2012.

COLE, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. **Nucleic Acids Research**, v. 42, n. D1, p. D633–D642, 1 jan. 2014.

CORRECTION MAYER, K. et al. Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. **Nature**, v. 402, n. 6763, p. 769–777, 16 dez. 1999.

COSTA, P. H. DE O. et al. Non-Specific Interactions of Rhizospheric Microbial Communities Support the Establishment of Mimosa acutistipula var. ferrea in an Amazon Rehabilitating Mineland. **Processes**, v. 9, n. 11, p. 2079, 1 nov. 2021a.

COSTA, P. H. O. et al. **MANUAL PARA O MONITORAMENTO DA ICTIOFAUNA POR MEIO DE DNA AMBIENTAL (eDNA)**. [s.l: s.n.]. . Acesso em: 7 mar. 2023b.

COSTANZA, R. et al. The value of the world's ecosystem services and natural capital. **Nature 1997 387:6630**, v. 387, n. 6630, p. 253–260, 15 maio 1997.

CREER, S. et al. The ecologist's field guide to sequence-based identification of biodiversity. **Methods in Ecology and Evolution**, v. 7, n. 9, p. 1008–1018, 14 set. 2016.

DASZKOWSKA-GOLEC, A.; MASCHER, M.; ZHANG, R. Editorial: Applications of long-read sequencing in plant genomics and transcriptomics. **Frontiers in Plant Science**, v. 14, p. 1141429, 31 jan. 2023.

DE MEDEIROS AZEVEDO, T. et al. The endophytome (plant-associated microbiome): methodological approaches, biological aspects, and biotech applications. **World Journal of Microbiology and Biotechnology 2021 37:12**, v. 37, n. 12, p. 1–25, 28 out. 2021.

DEINER, K. et al. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. **Molecular Ecology**, v. 26, n. 21, p. 5872–5895, 1 nov. 2017.

DESANTIS, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Applied and Environmental Microbiology**, v. 72, n. 7, p. 5069–5072, 1 jul. 2006.

DIRZO, R.; CEBALLOS, G.; EHRLICH, P. R. Circling the drain: the extinction crisis and the future of humanity. **Philosophical Transactions of the Royal Society B**, v. 377, n. 1857, 15 ago. 2022.

DO NASCIMENTO, S. V. et al. Proteomic Profiling and Rhizosphere-Associated Microbial Communities Reveal Adaptive Mechanisms of Dioclea apurensis Kunth in Eastern Amazon's Rehabilitating Minelands. **Plants**, v. 11, n. 5, p. 712, 1 mar. 2022.

EREN, A. M. et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. **Methods in Ecology and Evolution**, v. 4, n. 12, p. 1111–1119, 1 dez. 2013.

EXPOSITO-ALONSO, M. et al. The Earth BioGenome project: opportunities and challenges for plant genomics and conservation. **The Plant Journal**, v. 102, n. 2, p. 222–229, 29 abr. 2020.

FINDLEY, K. et al. Topographic diversity of fungal and bacterial communities in human skin. **Nature**, v. 498, n. 7454, p. 367–370, 2013.

FRANKHAM, R.; BALLOU, J. D.; BRISCOE, D. A. **Introduction to Conservation Genetics**. [s.l.] Cambridge University Press, 2010.

FUNK, W. C. et al. Harnessing genomics for delineating conservation units. **Trends in Ecology & Evolution**, v. 27, n. 9, p. 489–496, 1 set. 2012.

GADAGKAR, S. R.; ROSENBERG, M. S.; KUMAR, S. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. **Journal of Experimental Zoology Part B: Molecular and Developmental Evolution**, v. 304B, n. 1, p. 64–74, 15 jan. 2005.

GOFF, S. A. et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). **Science**, v. 296, n. 5565, p. 92–100, 5 abr. 2002.

GUILLOU, L. et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. **Nucleic Acids Research**, v. 41, n. D1, p. D597–D604, 1 jan. 2013.

HAMILTON, J. P.; ROBIN BUELL, C. Advances in plant genome sequencing. **The Plant Journal**, v. 70, n. 1, p. 177–190, 1 abr. 2012.

HEBERT, P. D. N. et al. Biological identifications through DNA barcodes. **Proceedings of the Royal Society of London B: Biological Sciences**, v. 270, n. 1512, 2003.

HOSHINO, A. et al. Hi-C Revolution: From a Snapshot of DNA–DNA Interaction in a Single Cell to Chromosome-Scale &lt;i&gt;De Novo&lt;/i&gt; Genome Assembly. **CYTOLOGIA**, v. 82, n. 3, p. 223–226, 25 jun. 2017.

HUANG, S. et al. The genome of the cucumber, Cucumis sativus L. **Nature Genetics**, v. 41, n. 12, p. 1275–1281, 1 dez. 2009.

HULSE-KEMP, A. M. et al. Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. **Horticulture Research**, v. 5, n. 1, p. 4, 1 dez. 2018.

INITIATIVE, T. A. G. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. **Nature**, v. 408, n. 6814, p. 796–815, 14 dez. 2000.

JAIN, M. et al. Improved data analysis for the MinION nanopore sequencer. **Nature Methods**, v. 12, n. 4, p. 351–356, 31 mar. 2015.

JUMP, A. S.; MARCHANT, R.; PEÑUELAS, J. Environmental change and the option value of genetic diversity. **Trends in Plant Science**, v. 14, n. 1, p. 51–58, 1 jan. 2009.

KAGALE, S. et al. The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. **Nature Communications**, v. 5, n. 1, p. 1–11, 23 abr. 2014.

KAJITANI, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. **Genome research**, v. 24, n. 8, p. 1384–95, 1 ago. 2014.

KAMATH, G. M. et al. HINGE: long-read assembly achieves optimal repeat resolution. **Genome research**, v. 27, n. 5, p. 747–756, 1 maio 2017.

KELLY, L. J. et al. Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. **New Phytologist**, v. 208, n. 2, p. 596–607, 1 out. 2015.

KOLMOGOROV, M. et al. Assembly of long, error-prone reads using repeat graphs. **Nature Biotechnology**, v. 37, n. 5, p. 540–546, 1 maio 2019.

KOREN, S. et al. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. **Genome Research**, v. 27, n. 5, p. 722–736, 1 maio 2017.

KRESS, W. J. et al. Use of DNA barcodes to identify flowering plants. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 23, p. 8369–8374, 7 jun. 2005.

KRESS, W. J. et al. DNA barcodes for ecology, evolution, and conservation. **Trends in Ecology & Evolution**, v. 30, n. 1, p. 25–35, 1 jan. 2015.

KRESS, W. J. Plant DNA barcodes: Applications today and in the future. **Journal of Systematics and Evolution**, v. 55, n. 4, p. 291–307, 1 jul. 2017.

KYRIAKIDOU, M. et al. **Current strategies of polyploid plant genome sequence assembly**. **Frontiers in Plant Science**Frontiers Media S.A., , 21 nov. 2018. . Acesso em: 10 fev. 2020

LAN, T. et al. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 22, p. E4435–E4441, 30 maio 2017.

LANES, É. C. et al. Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from amazonian savannas. **Frontiers in Plant Science**, v. 9, p. 532, 7 maio 2018.

LEITCH, I. J.; LEITCH, A. R. Genome size diversity and evolution in land plants. **Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes**, p. 307–322, 1 jan. 2013.

LI, F. et al. Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. **Nature Biotechnology**, v. 33, n. 5, p. 524–530, 12 maio 2015.

LI, Z. et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. **Briefings in Functional Genomics**, v. 11, n. 1, p. 25–37, 1 jan. 2012.

LIN, X. et al. Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. **Nature**, v. 402, n. 6763, p. 761–765, 16 dez. 1999.

LOWE, A. et al. Using DNA Metabarcoding to Identify Floral Visitation by Pollinators. **Diversity**, v. 14, n. 4, p. 236, 1 abr. 2022.

LUO, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. **GigaScience**, v. 1, n. 1, p. 18, 27 dez. 2012.

MACHIDA, R. J. et al. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. **Scientific Data 2017 4:1**, v. 4, n. 1, p. 1–7, 14 mar. 2017.

MACK, R. N. et al. **Issues in Ecology BIOTIC INVASIONS: CAUSES, EPIDEMIOLOGY, GLOBAL CONSEQUENCES, AND CONTROLEcological Applications**. [s.l: s.n.].

MARKS, R. A. et al. Representation and participation across 20 years of plant genome sequencing. **Nature Plants 2021 7:12**, v. 7, n. 12, p. 1571–1578, 29 nov. 2021.

MATSUMOTO, T. et al. The map-based sequence of the rice genome. **Nature**, v. 436, n. 7052, p. 793–800, 11 ago. 2005.

MCCOUCH, S. **Agriculture: Feeding the future**. **Nature**Nature Publishing Group, , 3 jul. 2013. Disponível em: <https://www.nature.com/articles/499023a>. Acesso em: 23 jul. 2020

MEHROTRA, S.; GOYAL, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. **Genomics, Proteomics & Bioinformatics**, v. 12, n. 4, p. 164–171, 1 ago. 2014.

MEIER, R. et al. DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success. **Systematic Biology**, v. 55, n. 5, p. 715–728, 1 out. 2006.

MEYERS, L. A.; LEVIN, D. A. ON THE ABUNDANCE OF POLYPLOIDS IN FLOWERING PLANTS. **Evolution**, v. 60, n. 6, p. 1198, 2 set. 2006.

MIJANGOS, J. L. et al. Contribution of genetics to ecological restoration. **Molecular Ecology**, v. 24, n. 1, p. 22–37, 1 jan. 2015.

MYERS, E. W. Toward Simplifying and Accurately Formulating Fragment Assembly. **Journal of Computational Biology**, v. 2, n. 2, p. 275–290, 1 jan. 1995.

NAGARAJAN, N.; POP, M. Sequence assembly demystified. **Nature Reviews Genetics**, v. 14, n. 3, p. 157–167, 29 jan. 2013.

NIC LUGHADHA, E. et al. Extinction risk and threats to plants and fungi. **Plants, People, Planet**, v. 2, n. 5, p. 389–408, 1 set. 2020.

NUNES, G. L. et al. Quillworts from the Amazon: A multidisciplinary populational study on Isoetes serracarajensis and Isoetes cangae. **PLoS ONE**, v. 13, n. 8, 2018.

OLIVEIRA, R. R. M. et al. PIMBA: A PIpeline for MetaBarcoding Analysis. p. 106–116, 2021.

OLIVEIRA, R. R. M. et al. PipeCoV: a pipeline for SARS-CoV-2 genome assembly, annotation and variant identification. **PeerJ**, v. 10, p. e13300, 13 abr. 2022.

OLIVEIRA, R. R. M.; VASCONCELOS, S.; OLIVEIRA, G. SPLACE: A tool to automatically SPLit, Align, and ConcatenatE genes for phylogenomic inference of several organisms. **Frontiers in Bioinformatics**, v. 2, p. 109, 8 dez. 2022.

OSSOWSKI, S. et al. Sequencing of natural strains of Arabidopsis thaliana with short reads. **Genome Research**, v. 18, n. 12, p. 2024–2033, 1 dez. 2008.

OUBORG, N. J. et al. Conservation genetics in transition to conservation genomics. **Trends in Genetics**, v. 26, n. 4, p. 177–187, 1 abr. 2010.

PATERSON, A. H. et al. Insights from the Comparison of Plant Genome Sequences. **Annual Review of Plant Biology**, v. 61, n. 1, p. 349–372, 2 jun. 2010.

PEVZNER, P. A.; TANG, H.; WATERMAN, M. S. An Eulerian path approach to DNA fragment assembly. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 17, p. 9748–53, 14 ago. 2001.

PIMM, S. L. et al. The Future of Biodiversity. **Science**, v. 269, n. 5222, p. 347–350, 21 jul. 1995.

POWO. **State of the World's Plants**. [s.l: s.n.]. Disponível em: <www.plantsoftheworldonline.org>.

QUAIL, M. et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. **BMC Genomics**, v. 13, n. 1, p. 341, 24 jul. 2012.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. **Nucleic Acids Research**, v. 41, n. D1, p. D590–D596, 1 jan. 2013.

RATNASINGHAM, S. mBRAVE: The Multiplex Barcode Research And Visualization Environment. **Biodiversity Information Science and Standards**, v. 3, p. e37986, 10 jul. 2019.

RATNASINGHAM, S.; HEBERT, P. D. N. BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). **Molecular Ecology Notes**, v. 7, n. 3, p. 355–364, 24 jan. 2007.

RHOADS, A.; AU, K. F. PacBio Sequencing and Its Applications. **Genomics, Proteomics & Bioinformatics**, v. 13, n. 5, p. 278–289, 1 out. 2015.

SALZBERG, S. L. et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. **Genome research**, v. 22, n. 3, p. 557–67, 1 mar. 2012.

SATAM, H. et al. Next-Generation Sequencing Technology: Current Trends and

Advancements. **Biology 2023, Vol. 12, Page 997**, v. 12, n. 7, p. 997, 13 jul. 2023.

SCHLOSS, P. D. et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. **Applied and Environmental Microbiology**, v. 75, n. 23, p. 7537–7541, 1 dez. 2009.

SCHMUTZ, J. et al. Genome sequence of the palaeopolyploid soybean. **Nature**, v. 463, n. 7278, p. 178–183, 14 jan. 2010.

SCHNABLE, P. S. et al. The B73 maize genome: Complexity, diversity, and dynamics. **Science**, v. 326, n. 5956, p. 1112–1115, 20 nov. 2009a.

SCHNABLE, P. S. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. **Science**, v. 326, n. 5956, p. 1112–1115, 20 nov. 2009b.

SHAFER, A. B. A. et al. Genomics and the challenging translation into conservation practice. **Trends in Ecology & Evolution**, v. 30, n. 2, p. 78–87, 1 fev. 2015.

SHULAEV, V. et al. The genome of woodland strawberry (Fragaria vesca). **Nature Genetics**, v. 43, n. 2, p. 109–116, 26 fev. 2011.

SIMPSON, J. T. et al. ABySS: a parallel assembler for short read sequence data. **Genome research**, v. 19, n. 6, p. 1117–23, 1 jun. 2009.

SOLTIS, P. S. et al. Polyploidy and genome evolution in plants. **Current Opinion in Genetics & Development**, v. 35, p. 119–125, 1 dez. 2015.

STUART CHAPIN, F.; MATSON, P. A.; VITOUSEK, P. M. Principles of terrestrial ecosystem ecology. **Principles of Terrestrial Ecosystem Ecology**, p. 1–529, 1 jan. 2012.

SUTTON, G. G. et al. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. **Genome Science and Technology**, v. 1, n. 1, p. 9–19, 9 jan. 1995.

TILMAN, D. et al. Global food demand and the sustainable intensification of agriculture. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 50, p. 20260–20264, 13 dez. 2011.

VALLIYODAN, B.; LEE, S.-H.; NGUYEN, H. T. Sequencing, Assembly, and Annotation of the Soybean Genome. Em: [s.l.] Springer, Cham, 2017. p. 73–82.

VASCONCELOS, S. et al. Unraveling the plant diversity of the Amazonian canga through DNA barcoding. **Ecology and Evolution**, v. 11, n. 19, p. 13348–13362, 1 out. 2021.

VELEBA, A. et al. Genome size and genomic GC content evolution in the miniature genome-sized family Lentibulariaceae. **New Phytologist**, v. 203, n. 1, p. 22–28, 1 jul. 2014.

WARREN, R. L. et al. Assembling millions of short DNA sequences using SSAKE. **Bioinformatics**, v. 23, n. 4, p. 500–501, 15 fev. 2007.

WENGER, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. **Nature Biotechnology 2019 37:10**, v. 37, n. 10, p. 1155–1162, 12 ago. 2019.

WHEELER, D. L. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 36, n. Database, p. D13–D21, 23 dez. 2007.

WHITELAW, C. A. et al. Enrichment of Gene-Coding Sequences in Maize by Genome Filtration. **Science**, v. 302, n. 5653, p. 2118–2120, 19 dez. 2003.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics 2007 8:12**, v. 8, n. 12, p. 973–982, 2007.

YU, J. et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). **Science**, v. 296, n. 5565, p. 79–92, 5 abr. 2002.

ZAFEIROPOULOS, H. et al. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. **GigaScience**, v. 9, n. 3, p. 1–12, 1 mar. 2020.

ZERBINO, D.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome research**, 2008.

ZIMIN, A. V. et al. The MaSuRCA genome assembler. **Bioinformatics**, v. 29, n. 21, p. 2669–2677, 1 nov. 2013.

APÊNDICE A - *A tool to automatically SPLit, Align and ConcatenatE genes for phylogenomic inference of several organisms.*

frontiers | Frontiers in Bioinformatics

TYPE Brief Research Report
PUBLISHED 08 December 2022
DOI 10.3389/fbinf.2022.1074802

OPEN ACCESS

EDITED BY
Arne Seitz,
Swiss Federal Institute of Technology Lausanne, Switzerland

REVIEWED BY
Yafei Mao,
Shanghai Jiao Tong University, China
Guanglin Li,
Shaanxi Normal University, China

*CORRESPONDENCE
Renato R. M. Oliveira,
renato.renison@gmail.com

SPECIALTY SECTION
This article was submitted to Integrative Bioinformatics, a section of the journal Frontiers in Bioinformatics

RECEIVED 20 October 2022
ACCEPTED 21 November 2022
PUBLISHED 08 December 2022

CITATION
Oliveira RRM, Vasconcelos S and Oliveira G (2022), SPLACE: A tool to automatically SPLit, Align, and ConcatenatE genes for phylogenomic inference of several organisms.
Front. Bioinform. 2:1074802.
doi: 10.3389/fbinf.2022.1074802

SPLACE: A tool to automatically SPLit, Align, and ConcatenatE genes for phylogenomic inference of several organisms

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1] and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

The reconstruction of phylogenomic trees containing multiple genes is best achieved by using a supermatrix. The advent of NGS technology made it easier and cheaper to obtain multiple gene data in one sequencing run. When numerous genes and organisms are used in the phylogenomic analysis, it is difficult to organize all information and manually align the gene sequences to further concatenate them. This study describes SPLACE, a tool to automatically SPLit, Align, and ConcatenatE the genes of all species of interest to generate a supermatrix file, and consequently, a phylogenetic tree, while handling possible missing data. In our findings, SPLACE was the only tool that could automatically align gene sequences and also handle missing data; and, it required only a few minutes to produce a supermatrix FASTA file containing 83 aligned and concatenated genes from the chloroplast genomes of 270 plant species. It is an open-source tool and is publicly available at https://github.com/reinator/splace.

KEYWORDS

phylogenomics, pipeline, supermatrix, concatenation, tree

Introduction

Inferring a species tree using multiple genes is often achieved through two approaches: 1) concatenating aligned genes into a supermatrix or 2) generating a consensus tree (supertree) from separate gene trees. The latter approach looks for congruence among all individual gene trees. The reconstruction of phylogenomic trees containing multiple genes is best achieved by concatenating all aligned genes into a supermatrix. It has greater phylogenetic accuracy because it uses a greater number of sites than examples in which

Abbreviations: NGS, next-generation sequencing; GUI, graphical user Interface; INPI, *Instituto Nacional da Propriedade Industrial*; NCBI, National Center for Biotechnology Information; RAM, random access memory.

only a single gene is used (Gadagkar et al., 2005). The steady development of next-generation sequencing (NGS) technology makes it easier and cheaper to obtain information from multiple genes from many organisms of interest, resulting in a more robust supermatrix. This supermatrix can then be used in phylogenetic reconstructions to create a species tree.

Many studies have used the supermatrix strategy to infer the phylogeny among species, such as when analyzing genomes from prokaryotic organisms and eukaryotic organelle genomes (mitogenomes and plastomes) (Sims and Kim, 2011). Building a supermatrix can be very time-consuming, especially if there is a large number of genes from many organisms to be used in the analysis. Some published tools, such as SequenceMatrix (Vaidya et al., 2011), TaxMan (Jones and Blaxter, 2006), ScaFoS (Roure et al., 2007), and Phyutility (Smith and Dunn, 2008) aim to concatenate gene files, while others are focused on using the supertree approach, such as TNT (Goloboff and Catalano, 2016), or both supermatrix and supertree methods, such as TREEasy (Mao et al., 2020). Tools that are focused on other strategies, such as NetRax (Lutteropp et al., 2022), which uses the phylogenetic network inference approach, and GeneRax (Morel et al., 2020), which uses the species-tree-aware (STA) approach, were also developed. In this work, we focused on tools that use the supermatrix and supertree strategies.

SequenceMatrix utilizes a graphical user interface (GUI), which may facilitate use by allowing the drag and drop of files with gene alignments in FASTA, TNT, or NEXUS formats. SequenceMatrix can concatenate gene sequences, but it is not able to automatically detect missing data. Its last update was in May 2021, which demonstrates that the software has been continuously maintained since its first release in 2011. As SequenceMatrix was developed in Java, memory consumption may be a problem, particularly if many genes from several organisms are to be analyzed. However, the main limitation that SequenceMatrix presents is that it requires already aligned input files, making the user responsible for detecting missing data in input files. For example, to generate a supermatrix containing 80 genes from 200 organisms, in addition to downloading and organizing the sequences locally, it would also be necessary to group each of the 80 gene sequences across the 200 organisms into 80 different files, align those files separately, and then drag and drop those files into SequenceMatrix, after checking to see if the taxa names were consistent across the 200 sequence entries in the 80 files. This task would be very time consuming and susceptible to errors, leading to delays in the analysis because of the need to review and correct the input information.

TaxMan was developed to facilitate phylogenetic studies by automating sequence acquisition, consensus building, alignment, and taxon selection. It was developed in Perl 5.8.6 and requires a set of prerequisites to be installed in the environment, such as BLAST (Tatusova and Madden, 1999), PostgreSQL (Momjian, 2001), Emboss (Rice et al., 2000), PHRAP (de la Bastide and McCombie, 2007), and POA (Lee et al., 2002). TaxMan accepts GenBank files of the taxa to be analyzed and a file with gene synonyms to be considered, which will be used to extract the gene information automatically from the GenBank files. It cannot automatically detect missing data, so the user is responsible for handling the possible lack of genes. The last TaxMan release was dated September 2006 and it is deprecated to be installed, although the paper is still online.

SCaFoS (Selection, Concatenation, and Fusion of Sequences) is a GUI tool developed in Perl that selects sequences, species, and genes, while dealing with paralogous and xenologous genes, and allows the use of partial genes in the absence of full sequences. It handles missing data by creating chimeric sequences according to the proportion of missing data that the user allows. SCaFoS accepts FASTA, PHYLIP, or Nexus file formats as inputs. The last update of SCaFoS was in October 2007, which means that it no longer has support from the development team. Since the software requires some tools, such as Perl-tk and tree-puzzle, and those libraries evolved through time, it is impossible to execute SCaFoS without a recent code update.

TNT provides a GUI tool for Windows users and command-line tools for Linux and Mac users, allowing the user to run an enormous variety of phylogenetic analyses, simulations, and methods for diagnosing trees and exploring character evolution without automatically handling missing data. The last update of TNT was in October 2022, which means that the tool is still being supported by the developers. The TNT limitation is the same as for SequenceMatrix, i.e., it requires that all the gene files be already aligned, which can be a time-consuming and error-prone activity when performed manually.

Phyutility is a command-line program developed in Java that automates phylogenetic tree, molecular sequence, and alignment manipulation. It accepts FASTA, Newick, and Nexus formats as input to perform tree and sequence manipulation, and it handles missing data by removing or trimming regions of the alignment according to the percentage of missing data allowed by the user. The last update of Phyutility was in September 2012, and it also requires that gene sequences be aligned for use.

TREEasy is the most recent tool to infer phylogeny by concatenating gene sequences. Its last update was in June 2020, allowing options for both GUI and command-line usage. For input, TREEasy needs FASTA files containing the nucleotide sequences of the genes to be included in the analysis and the corresponding amino acid sequence to generate output results for individual gene trees and species trees with supertree and supermatrix approaches. Although TREEasy can automate the alignment of gene sequences using MAFFT, it cannot handle missing data, so the user is responsible for selecting genes shared by all taxa included in the analysis. For the supertree approach, the authors of TREEasy also mention that the tool is only appropriate for working with a few taxa and not hundreds of taxa. Another limitation is that the use of TREEasy requires the installation of eight additional software modules as dependencies, which can be a bottleneck in the analysis if one of the dependencies has an update issue.

TABLE 1 Summary of tools developed to aid in phylogenetic/phylogenomic analyses. The approach used by each tool is given in parentheses.

Tool (approach)	Last update	Programming language	S.O.	Dependency	Whether aligns sequences	Whether detects missing data
TaxMan (SM)	September 2006	Perl 5.8.6	Linux	Blast, PostgreeSQL, Emboss, and PHRAP	Yes	No
SCaFoS (SM)	October. 2007	Perl 5.8.0	Linux, Windows XP, and Mac OS X	Perl-tk and tree-puzzle	No	Yes
Phyutility (SM)	September. 2012	Java	Linux, Windows, and Mac OS	Java VM	No	Yes
SequenceMatrix (SM)	May 2021	Java	Linux, Windows, and Mac OS	Java VM	No	No
TNT (ST)	October. 2022	Own language	Linux, Windows, and Mac OS	None	No	No
TREEasy (SM; ST)	July. 2020	Python	Linux, Windows, and Mac OS	MAFFT, IQ-TREE, RAxML-NG, ASTRAL, MP-EST, STELLS2, PhyloNet, and SNaQ	Yes	No
SPLACE (SM)	August. 2022	Python and Bash	Linux, Windows, and Mac OS	Docker	Yes	Yes

SM, supermatrix; ST, supertree.

Although there is a variety of tools that were developed to aid in phylogenetic/phylogenomic studies, some of them require aligned gene files, which can be a time-consuming and error-prone task if many genes are used, and only a few utilities can handle missing data, although being deprecated. Here we present SPLACE, a tool to automatically SPLit, Align, and ConcatenatE the genes from the species of interest, and generate a supermatrix file and a phylogenetic tree. It can automatically identify and handle missing data, reducing preparation time and the probability of errors in the data to be analyzed. SPLACE can be run with one single command line and is compiled into Docker containers to avoid errors in the installation of dependencies. It is open-source and publicly available at https://github.com/reinator/splace, and its patent is deposited at INPI under the accession #BR512019002834-1. Table 1 summarizes all the tools mentioned previously so that their features can be compared with SPLACE.

Methods

For the creation of a supermatrix of n organisms, SPLACE will need a text file listing n FASTA files, each containing all the g genes from a particular organism (Figure 1A). SPLACE can operate in two modes: 1) handling missing data, by specifying a list of genes to consider in the analysis, or 2) considering only the shared genes among the n FASTA files of the organisms.

First, SPLACE splits the genes from an organism, gathering genes that have the same name from the n organisms into a single FASTA file, generating g new FASTA files, each containing the same gene from different organisms (Figure 1B). Then, SPLACE aligns each one of the g FASTA files using the MAFFT (Katoh et al., 2002) aligner (with the default parameter –auto), generating $g´$ new aligned FASTA files (Figure 1C). Finally, the different genes in the aligned $g´$ FASTA files that came from the same organism are concatenated into a single sequence, generating a unique FASTA file with a supermatrix containing n sequences, each representing one of the n organisms (Figure 1D). The same gene order is followed in all n sequences of the supermatrix.

If a list of genes is given as input, SPLACE will be able to handle missing data and if a particular gene is not present in an organism, the space of this missing gene in the concatenated alignment is automatically filled with a "?", indicating the missing data. If no list of genes is given, the analysis is carried out only with shared genes. Phylogeny can then be reconstructed using the FASTA file with the supermatrix and the method of choice by the user (Figure 1E). SPLACE also generates some reports at the end of the analysis, containing the genes shared among the organisms and a table with the genes found in each organism.

The main limitation of SPLACE is that it requires that the names of the genes be the same in the FASTA files of the different organisms. To facilitate this checking step, the table generated by SPLACE, containing the genes found in each organism may help the researcher determine if there are genes with different names and representations.

Results and discussion

We intended to benchmark SPLACE with other tools developed to aid in phylogenomic/phylogenetic analysis, but TaxMan, SCaFoS, and Phyutility are obsolete and cannot be installed. SequenceMatrix, TNT, and TREEasy are available for

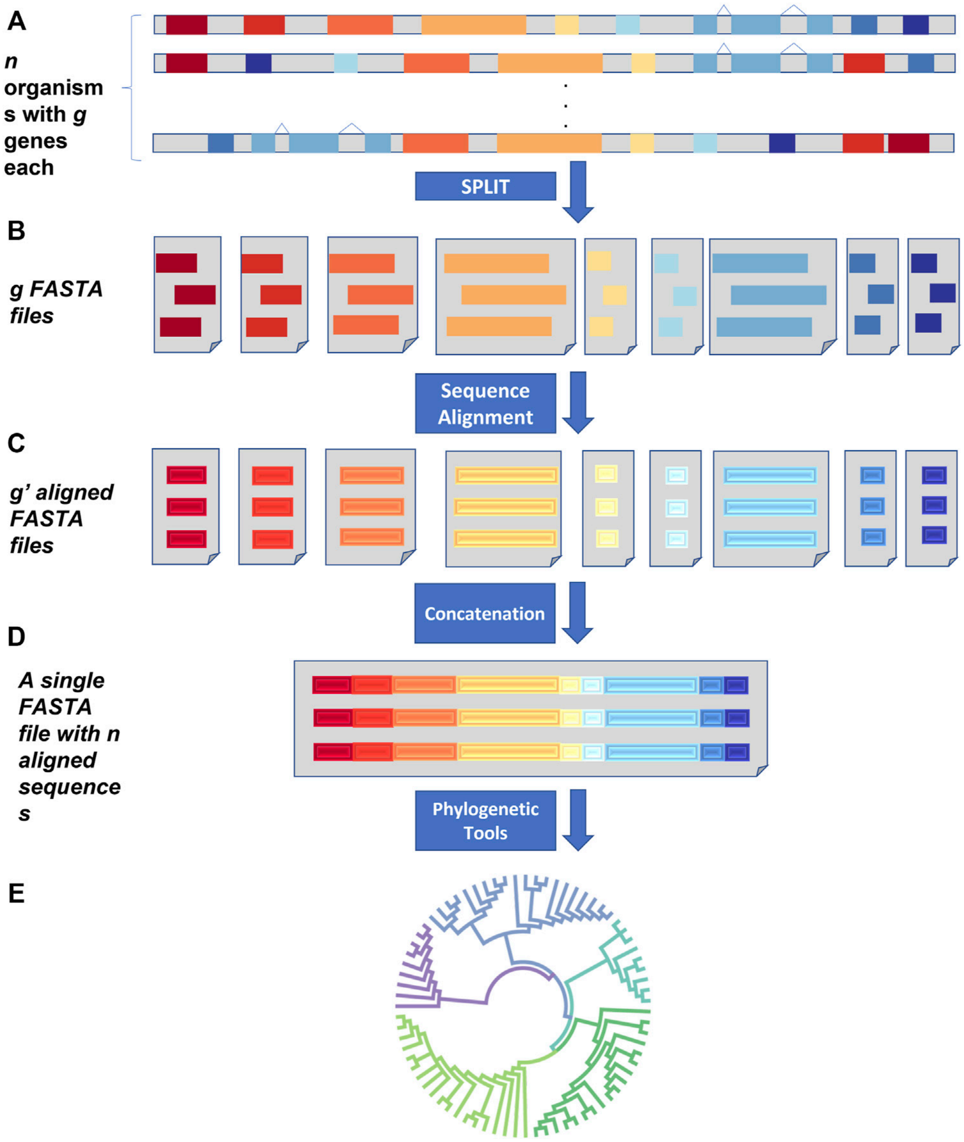


FIGURE 1
SPLACE workflow. **(A)** SPLACE accepts *n* FASTA files for each organism containing a maximum of *g* genes as input. **(B)** All genes were split into *g* FASTA files containing the same genes from the different organisms and then aligned **(C)**, generating *g*´ -aligned FASTA files. **(D)** SPLACE then concatenates the *g*´ -aligned genes from the same organism, generating a supermatrix with the resulting *n* sequences, which can then be used to infer phylogeny **(E)**.

installation, but they are not able to automatically align the sequences or detect missing data. Therefore, we decided not to compare these tools with SPLACE, which is the only utility that can automatically align sequences and detect missing data. Thus, SPLACE was used to build a phylogenomic tree for all plant species with a complete nuclear genome deposited on NCBI (270 species, until April 2022), using their respective chloroplast genes. We downloaded all 270 GenBank files with chloroplast annotation and extracted the gene sequences to compose FASTA files to be used as input. We created a text file containing a list of all FASTA files of the 270 plant species, and we also provided another text file as input containing the 83 genes that we wished to be present in the final results. The supermatrix generated by SPLACE was then submitted to the CIPRES portal (Miller et al., 2010) to generate a maximum likelihood phylogenetic tree using RAxML (Stamatakis, 2014), with the GTRGAMMA model, a bootstrap of 1,000 replicates, and *Physcomitrella patens* (NC_005087) as outgroups. The resulting phylogenomic tree (Figure 2) mainly showed the expected relationships within and among families (Chase et al., 2016), considering the incomplete

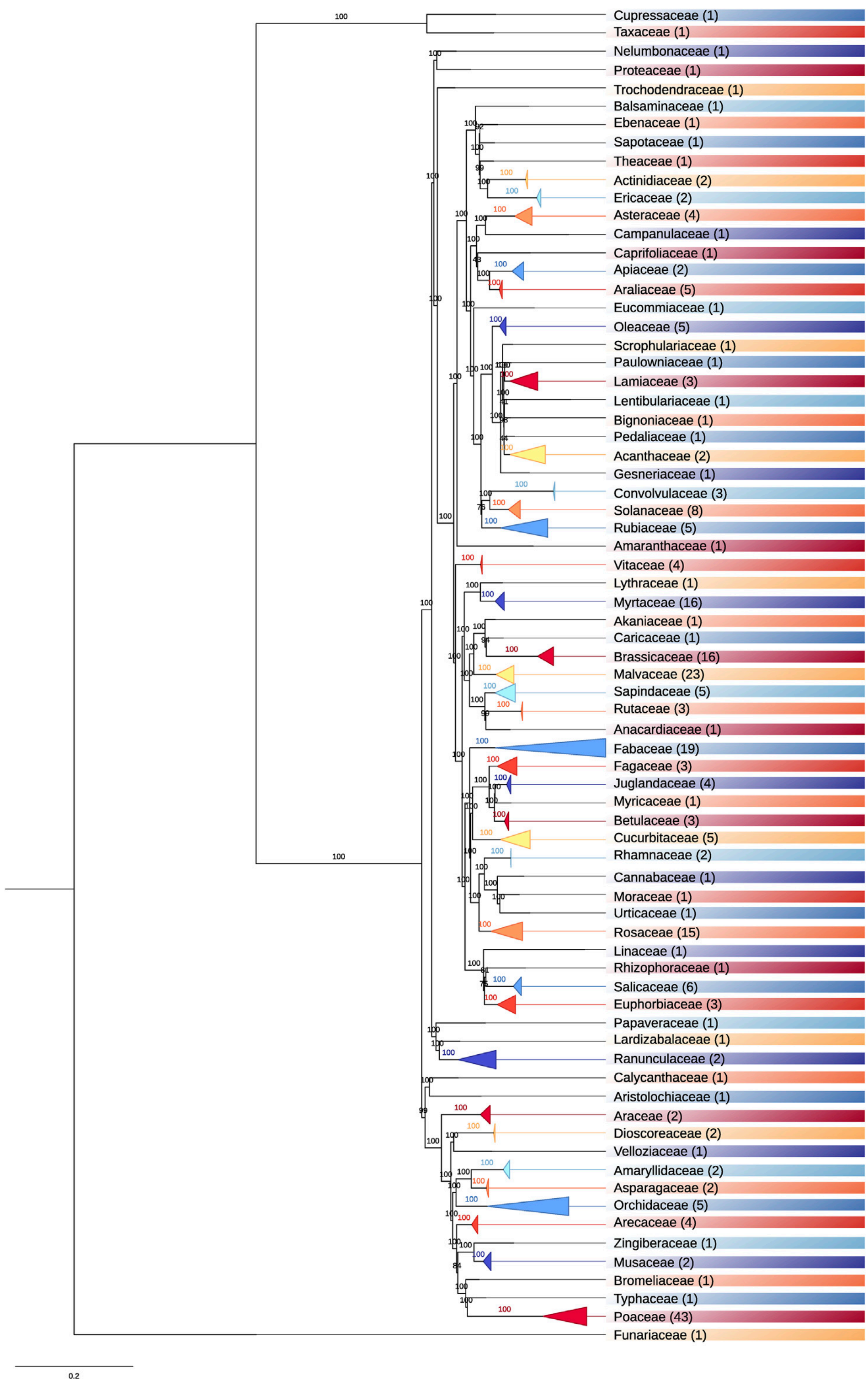


FIGURE 2
Plastome-based phylogenomic tree for 270 plant species with complete nuclear genomes deposited on NCBI. The number of species a family comprises is given in parentheses. The colors indicate different families.

sampling of taxa due to the selection criteria (species with a complete nuclear genome available). The phylogenomic tree without collapsed branches can be found in Supplementary Figure S1, and all the accessions for the species used and the listed genes are in Supplementary Tables S1, S2, respectively.

The phylogenomic tree represented in Figure 2 comprises 73 families, of which 35 contain more than one species. Poaceae is the family with the most species (43), followed by Malvaceae (23), Fabaceae (19), Brassicaceae (16), Myrtaceae (16), and Rosaceae (15). The remaining families comprise less than eight species. These results show where most efforts are being expended when assembling complete plant nuclear genomes.

SPLACE took only 36 minutes to generate the supermatrix file in a Core i5 2.20 Ghz computer with 12 Gb RAM and using four threads, with a dataset size of 20 MB for the 270 FASTA files containing up to 83 coding genes. If it was necessary to manually group the same genes in FASTA files to further align and manually concatenate them, the time to obtain a supermatrix would be longer. The results show that SPLACE took only a few minutes to extract, align, and generate the phylogeny tree when analyzing many genes from several organisms.

Nowadays, computational clusters provide hundreds of threads and terabytes of RAM memory to run bioinformatics analyses. In this environment, the time taken by SPLACE to generate the supermatrix of 270 plant species could be less than the 36 min required in the previously described computational environment.

Conclusion

SPLACE is the most recent tool to automatically SPLit, Align, and ConcatenatE gene sequences from several organisms, and also detect missing data. The FASTA files used as input for SPLACE might include either nucleotide or amino acid sequences, since the alignment step with MAFFT automatically recognizes the type of sequence. In addition, the researcher can choose whether the supermatrix generated at the end of the analysis will contain missing data or only shared genes. At the end of the analysis, SPLACE provides a FASTA file containing the supermatrix, which can be used with other utilities, such as tools to select the best evolution model and consequently generate a phylogenomic tree. A table with the genes found in each organism and a table with the shared genes are also generated in the output.

We believe that SPLACE will facilitate phylogenomic analysis by reducing the time needed to separate many genes from several organisms and also reduce the risk of errors.

Data availability statement

The original contributions presented in the study are included in the Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

RO developed SPLACE, structured, and wrote the manuscript. SV had the idea for SPLACE, made important suggestions, tested the scripts, and reviewed the manuscript. GO made important suggestions and reviewed the manuscript.

Funding

This work received financial support from Vale S.A. (Biodiversity Genomics project, R100603.85.0X.CB01). GO is a CNPq (Conselho Nacional de Desenvolvimento Científico) fellow (444227/2018-0 and 307479/2016-1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2022.1074802/full#supplementary-material

References

Chase, M. W., Christenhusz, M. J., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., et al. (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181, 1–20. doi:10.1111/BOJ.12385

de la Bastide, M., and McCombie, W. R. (2007). Assembling genomic DNA sequences with PHRAP. *Curr. Protoc. Bioinforma.* 17, Unit11.4. doi:10.1002/0471250953.BI1104S17

Gadagkar, S. R., Rosenberg, M. S., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J. Exp. Zool.* 304B, 64–74. doi:10.1002/jez.b.21026

Goloboff, P. A., and Catalano, S. A. (2016). Tnt version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* 32, 221–238. doi:10.1111/cla.12160

Jones, M., and Blaxter, M. (2006). TaxMan: a taxonomic database manager. *BMC Bioinforma.* 7, 536. doi:10.1186/1471-2105-7-536

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436

Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics* 18, 452–464. doi:10.1093/BIOINFORMATICS/18.3.452

Lutteropp, S., Scornavacca, C., Kozlov, A. M., Morel, B., and Stamatakis, A. (2022). Netrax: Accurate and fast maximum likelihood phylogenetic network inference. *Bioinformatics* 38, 3725–3733. doi:10.1093/bioinformatics/btac396

Mao, Y., Hou, S., Shi, J., and Economo, E. P. (2020). Treeasy: An automated workflow to infer gene trees, species trees, and phylogenetic networks from multilocus data. *Mol. Ecol. Resour.* 20. [Dataset]. doi:10.1111/1755-0998.13149

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). “Creating the CIPRES Science Gateway for inference of large phylogenetic trees,” in 2010 Gateway Computing Environments Workshop (GCE), New Orleans, LA, November 14, 2010. (IEEE), 1–8. doi:10.1109/GCE.2010.5676129

Momjian, B. (2001). *PostgreSQL: Introduction and concepts* 192, New York: Addison-Wesley.

Morel, B., Kozlov, A. M., Stamatakis, A., and Szöllősi, G. J. (2020). Generax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* 37, 2763–2774. doi:10.1093/molbev/msaa141

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi:10.1016/S0168-9525(00)02024-2

Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* 7, S2. doi:10.1186/1471-2148-7-S1-S2

Sims, G. E., and Kim, S.-H. (2011). Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U. S. A.* 108, 8329–8334. doi:10.1073/pnas.1105168108

Smith, S. A., and Dunn, C. W. (2008). Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24, 715–716. doi:10.1093/bioinformatics/btm619

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Tatusova, T. A., and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–250. doi:10.1111/j.1574-6968.1999.tb13575.x

Vaidya, G., Lohman, D. J., and Meier, R. (2011). SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* 27, 171–180. doi:10.1111/J.1096-0031.2010.00329.X

APÊNDICE B - *PIMBA: A PIpeline for MetaBarcoding Analysis.*

PIMBA: A PIpeline for MetaBarcoding Analysis

Renato R. M. Oliveira[1,2](✉), Raíssa Silva[1], Gisele L. Nunes[1],
and Guilherme Oliveira[1](✉)

[1] Instituto Tecnológico Vale, Belém, Pará, Brazil
guilherme.oliveira@itv.org
[2] Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas,
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract. DNA metabarcoding is an emerging monitoring method capable of assessing biodiversity from environmental samples (eDNA). Advances in computational tools have been required due to the increase of Next-Generation Sequencing data. Tools for DNA metabarcoding analysis, such as MOTHUR, QIIME, Obitools, PEMA, and mBRAVE have been widely used in ecological studies, however, some difficulties are encountered when there is a need to use custom databases. Here we present PIMBA, a PIpeline for MetaBarcoding Analysis, which allows the use of customized databases, as well as other reference databases used by the software mentioned here. PIMBA is an open-source and user-friendly pipeline that consolidates all analyses in just three command lines. PIMBA's implementation is available at https://github.com/reinator/pimba.

Keywords: DNA metabarcoding · Flexible pipeline · OTU · ASV

1 Introduction

DNA metabarcoding is a powerful tool that has been widely used for biodiversity monitoring and ecosystem assessment from environmental DNA samples (eDNA). The technique based on high-throughput sequencing (HTS) allows the multispecies detection from specific molecular markers in a group (plants, animals, fungi, and bacteria) [1]. Next-Generation Sequencing (NGS) results in millions of DNA sequences (reads) that allow deciphering the genetic code of species, answering taxonomic and functional questions. Nowadays, different sequencing platforms are available and can generate either paired-end or single-end reads. The technologies associated with paired-end reads allow to sequence a pool of different samples and automatically demultiplex them by using different indexes. Such a pool of samples can also be achieved with single-end technologies, but they still did not automatize the demultiplexing steps. If single-end reads are being sequenced for a metabarcoding analysis, the pool of samples might be done by using either single or dual-indexes multiplexing, this latter allowing to pool a larger number of samples in the same sequencing run, reducing costs and time.

The recent increase of NGS data has caused the development of new tools for DNA metabarcoding analysis, making the metabarcoding method more accessible and user-friendly [2]. Mothur [3], Qiime [4], Obitools [5], mBRAVE [6], and PEMA [7] are

P. F. Stadler et al. (Eds.): BSB 2021, LNBI 13063, pp. 106–116, 2021.
https://doi.org/10.1007/978-3-030-91814-9_10

currently the most used tools for metabarcoding analysis. Most pipelines use operational taxonomic units (OTUs) as the clustering method, except Pema which works for both OTU clustering (1) and Amplicon Sequence Variants (ASVs) inference (2). In the first approach (1) the reads are grouped into OTUs to differentiate species or taxa based on the similarity of sequences [8, 9]. A similarity of 97% is commonly used as a cutoff, but this value depends on the group to be evaluated [10]. The second approach (2) infers ASVs, which are reads that differ in 1 nucleotide (or more than) [11].

The biggest restriction of these pipelines is to make it difficult of using a customized database for taxonomy assignments. Among the available tools, Mothur is useful when analyzing 16S/18S rRNA, Influenza viral, and fungal ITS regions, using Greengenes [12], Influenza Virus, and SILVA [13] databases, respectively. Qiime (and even its updated version, Qiime2) is optimized to analyze metabarcoding data from 16S rRNA, 18S rRNA, and fungal ITS marker genes, using Greengenes, SILVA, and UNITE [14] databases, respectively. Qiime2 allows the user to train a classifier with a NaiveBayes model, but they report tests by using only a 16S example and with some constraints to the use of this classifier with other marker genes. Obitools is optimized to analyze data from 16S (SILVA and PR2) and it also allows the use of the NCBI database for taxonomic assignment. mBRAVE is optimized to use only the BOLD [15] database as a reference, allowing the researcher to use a personalized database only after BOLD submission. PEMA allows the analysis of metabarcoding data from 16S/18S rRNA, fungal ITS, and metazoan COI, using SILVA, UNITE, and MIDORI [16] databases, respectively.

To allow the researcher to use a customized database in the metabarcoding analyses as well as reference database such as NCBI/Genbank, we developed PIMBA, a PIpeline for MetaBarcoding Analysis, which adapts the Qiime/BMP [17] pipeline for OTUs clustering with additional and optional OTU corrections based on the algorithm LULU [18]. PIMBA accepts both single and paired-end reads, with both single and dual-index. PIMBA also allows inferring ASVs using Swarm [19]. A preliminary abundance and diversity analysis are also automatically delivered. The main innovation of this pipeline is, in just three command lines, the ease of using both standard and customized databases, minimizing errors in taxonomic assignments.

2 Implementation

PIMBA is fully containerized in docker images, being more platform-independent and easy to maintain and update. Besides implementing all the features provided by the other metabarcoding tools, PIMBA also allows the user to apply different databases and not only those commonly used by most of the available software. PIMBA can be used with single or paired-end reads and is divided into three steps: (1) preprocessing, which promotes the demultiplexing and quality treatment of reads (Fig. 1A); (2) taxonomy assignment, in which reads are clustered into OTUs or ASVs are inferred, along with errors correction (Fig. 1B), and (3) plotting, in which alfa and beta diversity plots are built by Phyloseq [20], including rarefaction curves and Principal Coordinates Analysis (PCoA), respectively. A metadata file is required for this last step (all PIMBA commands are available at https://github.com/reinator/pimba).

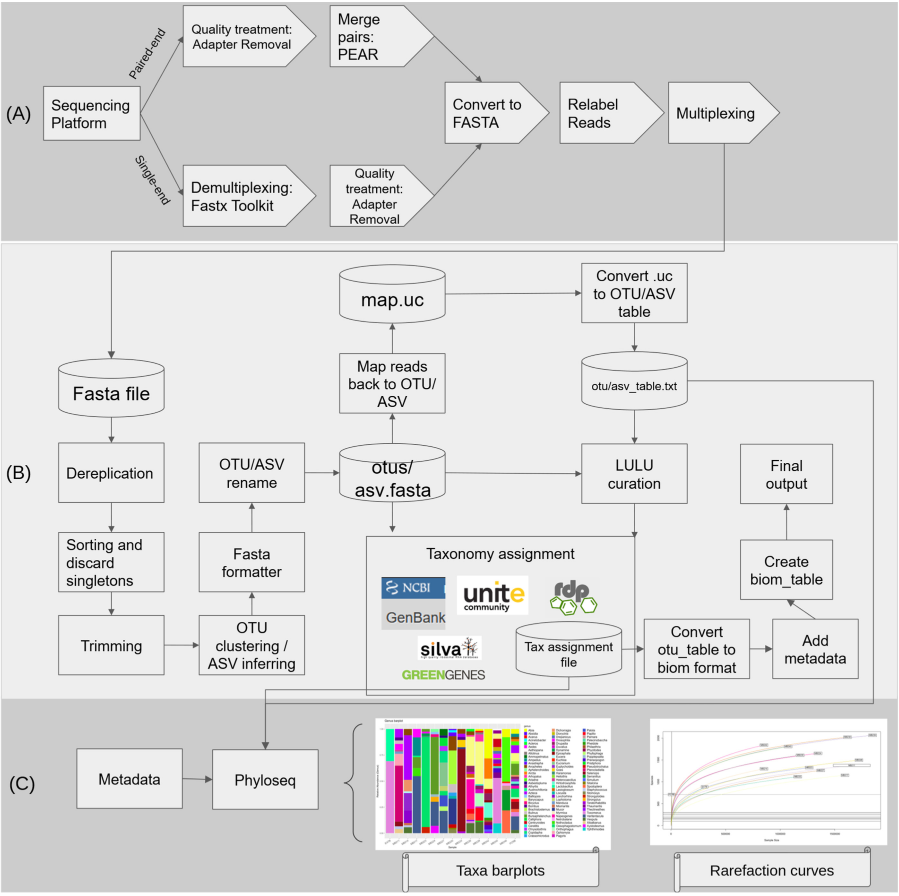


Fig. 1. PIMBA workflow. (A) preprocessing, (B) OTU clustering or ASV inference, and (C) plotting.

2.1 Preprocessing

PIMBA (pimba_prepare) can process either single-end or paired-end reads. Depending on the sequencing strategy, a few steps for demultiplexing or merging reads are needed. For paired-end reads, first AdapterRemoval v2.2.3 [21] will trim the adapters used in the sequencing (-mm 5, allowing 10% difference in the adapters sequence) with additional quality treatment, by using a 10bp window (--trimwindows), removing all reads with mean quality below PHRED 20 (--minquality) and with length less than 100bp (--minlength). Then, all read pairs will be merged with PEAR [22], using default parameters.

For single-end reads, a demultiplex step is performed. PIMBA allows the demultiplex of both single and dual-index libraries. In both cases, PIMBA will use Fastx-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to detect the 5' or 3' index (in the case of single-index) or the 5' and 3' indexes (in the case of dual-index), generating at the end all the demultiplexed reads. Then, AdapterRemoval is used to perform the quality treatment with the same parameters as mentioned for paired-end reads.

Both high-quality paired-end and single-end demultiplexed reads are converted to FASTA with Prinseq v0.20.4 [23], relabeled and concatenated to a single FASTA file, that will be used in the next step, for OTU clustering or ASV inference. All the steps above for preprocessing (Fig. 1A) illumina paired-end datasets with 10 threads, for example, can be run in the following command-line:./pimba_prepare.sh illumina < rawdata_dir > < output_reads > < num_threads > < adapters.txt > < min_length > < min_phred > (e.g.: pimba_prepare.sh illumina rawdata_dir/ AllSamples_16S_hqdata 10 adapters.txt 100 20).

2.2 Taxonomy Assignment

With the multiplexed FASTA file resulting from the preprocessing step, PIMBA (pimba_run) will use VSEARCH v2.15.2 [24] to dereplicate, discard singletons and trim the reads to a given length (–l 0, if no trim is desired). Then, depending on the approach, PIMBA will cluster the reads into OTUs (-w out) at a given similarity threshold (-s) using VSEARCH or infer the ASVs with Swarm (-w asv), accepting difference in only one nucleotide. For both OTUs/ASVs, PIMBA will use VSEARCH to remove chimeras (--uchime_denovo), Fastx_toolkit to format the FASTA file, and a Perl script from BMP to rename the OTUs/ASVs. VSEARCH will also be used to map back the reads to the OTUs/ASVs and then a script from QIIME will be used to generate an OTU/ASV table. PIMBA will optionally use LULU to curate all the found OTUs/ASVs (-x).

Depending on the marker gene the user is analyzing (-g), PIMBA will use different databases to taxonomically assign the OTUs/ASVs. To work properly, the user will need to pass a database file as a parameter (-d), where the location from the desired database will be set. Currently, PIMBA allows the analysis of 16S rRNA (SILVA, Greengenes, RDP [25] or Genbank [26]), fungal ITS (UNITE or Genbank), and for any other desired marker gene (e.g., metazoan COI, plant ITS) with BLAST [27] assignment to the Genbank database. Also, PIMBA allows the user to generate a customized database for assignments. (see https://github.com/reinator/pimba). When the user desires to use the SILVA, Greengenes, RDP, or UNITE databases, PIMBA will use scripts adapted from QIIME/BMP pipeline, and the user will also need to define the similarity for the assignment (-a). In the case of analyzing fungal ITS, PIMBA will also use ITSx [28] to discard the ribosomal regions flanking the ITS regions. When the user desires to use the NCBI Genbank database, a set of PIMBA scripts will be used, and besides the assignment similarity, the user will need to define the minimum alignment coverage (-c), the maximum e-value allowed (-e), and the number of hits per sequence that BLAST needs to return (-h). When -h is 1, only the best hit is returned. If -h is greater than 1, PIMBA will perform a voting system to properly assign the taxonomy. In case of a tie, the taxon with greater similarity will be chosen.

Finally, PIMBA will use Biom v2.1.10 [29] to convert the OTU/ASV table to biom format and add the taxonomy assignments, generating a summarized biom table file, needed in the next step.

All the steps above for running taxonomy assignment (Fig. 1B) can be run in the following command-line:./pimba_run.sh -i < input_reads > -o < output_dir > -w < approach > -s < otu_similarity > -a < assign_similarity > -c < coverage > -l < otu_length > -h < hits_per_subject > -g < marker_gene > -t < num_threads > -e < E-value > -d < databases.txt > -x < lulu > (e.g.: pimba_run.sh -i 16S_hqdata.fasta -o run_OTU_NCBI -w otu -s 0.97 -a 0.9 -c 0.9 -l 200 -h 5 -g 16S-NCBI -t 10 -d databases.txt).

2.3 Plotting

In the end, PIMBA (pimba_plot) will use Phyloseq to plot alpha and beta diversity results, such as rarefaction curves and PCoA plots. The user only needs to give as parameters the OTU/ASV table (-t), the taxonomy assignment file (-a), and a metadata file (-m). Depending on the metadata file, the user will also be able to group the results according to a given attribute from the samples (-g). To perform the plotting (Fig. 1C), the following command-line can be used: pimba_plot.sh -t < otu_table > -a < tax_assignment > -m < metadata > -g < group_by > (e.g.: pimba_plot.sh -t 16S_otu_table.txt -a 16S_otus_tax_assignments.txt -m mapping_file.csv -g Description).

3 Results and Discussion

To demonstrate that PIMBA is effective at analyzing a metabarcoding dataset, we used the same benchmark used by PEMA [7]: three mock communities sequencing. PIMBA's results are also being compared to the results presented in the PEMA publication.

From the mock community sequencing, the first dataset is from the 16S rRNA gene, comprising 20 bacterial species [30]. The second is a dataset from fungal ITS, comprising 19 fungal species [31]. The third is a dataset from metazoan COI amplicons, comprising 14 species [32]. Information regarding the datasets mentioned above is summarized in Table 1.

To evaluate the mock communities' results, we ran an extensive comparative benchmark, varying parameters such as truncation length, minimum assignment similarity, taxon database, and strategy (with OTU clustering or ASV inference). For each test, we calculated the True-Positives (TP), False-Positives (FP), and False-Negatives (FN) obtained by PIMBA, at both Genus and Species levels. Then, we were able to calculate precision (to check how many correct results PIMBA returned), recall (to check how much of the known taxa in the mock communities PIMBA can recover), and F1 score (which combines the precision and recall values) [33]. All TP, FP, FN, and F1 values obtained in the tests we performed are available at https://github.com/reinator/pimba.

For all tests, we fixed the cluster similarity for OTUs (-s 0.97) and maximum difference for ASV (-d 1). Besides, only taxa existing in all replicates were considered as a hit, except for the COI dataset, where we accepted as a hit, taxon occurring in at least two replicates, given its low depth. We also decided not to run LULU curation, as we saw that in all tests, the F1 scores were lower than when LULU was not used. In the next sections, the results from the mock datasets will be described and discussed.

Table 1. Mock community datasets and accessions. Total reads, bases, and sequencing read length are also shown.

Marker gene	SRA	Total reads	Total bases (Mb)	Read length (bp)
16S	SRR3163904 SRR3163905 SRR3163906	895,113	471.3	2 × 300
Fungal ITS	SRR5838515 SRR5838516 SRR5838522	162,841	81.5	2 × 250
Metazoan COI	ERR2181459 ERR2181468 ERR2181466	228,019	113.8	2 × 300

3.1 16S rRNA Mock Community

The 16S rRNA mock community was sequenced with Illumina MiSeq using the v3 reagent kit (2x300 cycles), targeting the V4 region (~252 bp) [30]. After quality treatment and pair merging, a total of 810,981 amplicon reads were used as input to pimba_run. We varied the strategy (OTU or ASV), the minimum assignment similarity (0.90, 0.97, and 0.99), the truncation length (200 bp, 250 bp), and the taxon database (SILVA or Genbank/NCBI). The F1 scores obtained at the Genus level are shown in Table 2. The best F1 scores are highlighted in all the tables that follow.

Table 2. F1 scores for each one of PIMBA's 16S rRNA results at Genus level, when varying assignment similarity, truncation length, strategy, and taxon database.

Min assign similarity	0.90		0.97		0.99	
Truncation length	200 bp	250 bp	200 bp	250 bp	200 bp	250 bp
OTU - SILVA	0.89	0.89	0.89	0.91	0.88	0.90
ASV - SILVA	0.95	0.95	0.95	0.93	**0.98**	0.95
OTU – Genbank	0.90	0.93	0.90	0.93	0.90	0.93
ASV - Genbank	0.93	0.95	0.88	0.95	0.90	0.95

PIMBA performed better (F1 score = 0.98) when running ASV inference, truncating the sequences at 200bp and assigning similarity only above 99% against the SILVA database. This configuration returned only one false positive (*Prevotella*) and recovered all 20 bacterial taxa, being 1.18-fold better than PEMA's results (F1 = 0.83, see [7]) when analyzing the same dataset. At the Species level, PIMBA performed better when running OTU clustering at 250bp and assigning the taxa at any of the selected similarities (Table 3).

Table 3. F1 scores for each one of PIMBA's 16S rRNA results at Species-level, when varying assignment similarity, truncation length, strategy, and taxon database.

Min assign similarity	0.90		0.97		0.99	
Truncation length	200 bp	250 bp	200 bp	250 bp	200 bp	250 bp
OTU - SILVA	0.28	0.39	0.21	0.33	0.22	0.37
ASV - SILVA	0.58	0.58	0.41	0.61	0.60	0.59
OTU – Genbank	0.78	**0.80**	0.78	**0.80**	0.78	**0.80**
ASV - Genbank	0.62	0.68	0.61	0.68	0.60	0.68

PIMBA recovered 17 species of the 20 bacterial taxa in the mock community when used with OTU strategy, being 5.6-fold better than PEMA, which recovered only 3 species.

3.2 Fungal ITS Mock Community

The fungal mock community targeted the ITS2 region (~327bp, ± 40.1) [34] and was sequenced with Illumina MiSeq, using v2 reagent kit (2 × 250 cycles) [31]. The pimba_prepare script outputted a total of 155,691 amplicon reads, which were used by pimba_run. We compared the results by varying the strategy (OTU or ASV), the minimum assign similarity (0.90, 0.95, and 0.97), the truncation length (100 bp, 130 bp, and 160 bp), and the taxon database (UNITE or Genbank/NCBI). The F1 scores obtained at the Genus levels are shown in Table 4.

Table 4. F1 scores for each one of PIMBA's ITS results at Genus level, when varying assignment similarity, truncation length, strategy, and taxon database.

Min assign similarity	0.90			0.95			0.97		
Truncation length	100 bp	130 bp	160 bp	100 bp	130 bp	160 bp	100 bp	130 bp	160 bp
OTU - UNITE	0.85	0.88	0.64	0.88	0.85	0.64	0.85	0.88	0.64
ASV -UNITE	0.85	0.85	0.59	0.85	0.85	0.64	0.81	0.81	0.69
OTU - Genbank	**0.94**	**0.94**	0.85	**0.94**	**0.94**	0.85	**0.94**	**0.94**	**0.94**
ASV - Genbank	**0.94**	**0.94**	0.85	**0.94**	**0.94**	0.85	**0.94**	0.94	0.85

For ITS, PIMBA performed better (F1 = 0.94) when using the Genbank database for taxonomy assignment, being 1.09-fold better than PEMA (F1 = 0.86, see [7]).

Both OTU and ASV strategies used by PIMBA had the same F1 scores in almost all configurations, except for truncation at 160bp, with 0.97 of assignment similarity, where the OTU strategy outperformed ASV's. At the Species level, PIMBA performed better when running OTU clustering at 100bp and assigning the taxa at any of the selected similarities (Table 5) using the Genbank database.

Table 5. F1 scores for each one of PIMBA's ITS results at Species-level, when varying assignment similarity, truncation length, strategy, and taxon database.

Min assign similarity	0.90			0.95			0.97		
Truncation length	100 bp	130 bp	160 bp	100 bp	130 bp	160 bp	100 bp	130 bp	160 bp
OTU - UNITE	038	0.37	0.33	0.44	0.38	0.32	0.43	0.36	0.31
ASV -UNITE	0.38	0.37	0.26	0.43	0.38	0.33	0.37	0.36	0.31
OTU - Genbank	**0.74**	0.72	0.63	**0.74**	0.72	0.61	**0.74**	0.72	0.72
ASV - Genbank	0.67	0.67	0.61	065	0.67	061	0.65	0.67	0.63

PIMBA recovered 14 species of the 19 bacterial taxa in the mock community when using either OTU or ASV strategy and truncating at 100 bp, being 2.8-fold better than PEMA, which recovered only 5 species. However, the number of false positives increased when the ASV strategy was used (10 False Positives), in comparison to OTU (5 False Positives).

3.3 Metazoan COI Mock Community

This dataset comprises a 3' region from the Cytochrome oxidase I gene (~450bp), sequenced with Illumina MiSeq, using v2 reagent kit (2x250 cycles)[32]. After performing preprocessing in the paired-end raw data, pimba_prepare outputted a total of 141,283 amplicon reads. We compared the results by varying the strategy (OTU or ASV), the minimum assign similarity (0.97, 0.98, and 0.99), the truncation length (250 bp, 350 bp, and 450 bp). PIMBA does not use a specific database for metazoan COI, so the taxon database used was Genbank/NCBI. The F1 scores obtained at the Genus levels and species levels were the same and are shown in Table 6.

PIMBA's performance was quite homogenous when varying OTU and ASV, getting an incredible F1 score of 1 in almost all configurations, being 1.35-fold better than PEMA (F1 = 0.74, see [7]). PIMBA recovered all 13 invertebrate species from the mock community.

Table 6. F1 scores for each one of PIMBA's COI results at Genus and Species-level, when varying assignment similarity, truncation length, strategy, and taxon database.

Min assign similarity	0.97			0.98			0.99		
Truncation length	250 bp	350 bp	450 bp	250 bp	350 bp	450 bp	250 bp	350 bp	450 bp
OTU	0.96	**1**	**1**	0.96	**1**	**1**	0.96	**1**	**1**
ASV	**1**	**1**	0.96	**1**	**1**	**1**	**1**	0.96	**1**

4 Conclusion

In contrast to the pipelines mentioned above, PIMBA allows the use of some specific or commonly used databases, such as Genbank, for taxonomy assignment. This feature is of paramount importance when there is a need to work with private and non-public databases. Another advantage of PIMBA is the freedom to use different forms of grouping sequences (ASVs or OTUs) within the same pipeline (most available pipelines apply a unique grouping approach). Regarding the results, it was possible to see how accurate PIMBA is in obtaining taxon for both Genus and Species levels and how flexible it is in the use of different strategies, parameters, and databases. Using as a comparison the PEMA pipeline, which applies similar strategies to PIMBA, we show that our results (both OTUs and ASVs) were superior concerning the expected taxonomy since we used a mock community as a dataset. Regarding the choice of the best grouping strategy (OTU or AVS) for our dataset, OTU presented a better resolution at the Species level, especially when using the Genbank database, while the ASV approach showed better results for analysis at the level of Genus.

References

1. Creer, S., et al.: The ecologist's field guide to sequence-based identification of biodiversity. Meth. Ecol. Evol. **7**, 1008–1018 (2016). https://doi.org/10.1111/2041-210X.12574
2. Alberdi, A., Aizpurua, O., Gilbert, M.T.P., Bohmann, K.: Scrutinizing key steps for reliable metabarcoding of environmental samples. Meth. Ecol. Evol. **9**, 134–147 (2018). https://doi.org/10.1111/2041-210X.12849
3. Schloss, P.D., et al.: Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75**, 7537–7541 (2009). https://doi.org/10.1128/AEM.01541-09
4. Caporaso, J.G., et al.: QIIME allows analysis of high-throughput community sequencing data. Nat. Meth. **7**, 335–336 (2010). https://doi.org/10.1038/nmeth.f.303
5. Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., Coissac, E.: Obitools : a unix -inspired software package for DNA metabarcoding. Mol. Ecol. Resour. **16**, 176–182 (2016). https://doi.org/10.1111/1755-0998.12428
6. Ratnasingham, S.: mBRAVE: the multiplex barcode research and visualization environment. Biodivers. Inf. Sci. Stand. **3**, e37986 (2019). https://doi.org/10.3897/biss.3.37986

7. Zafeiropoulos, H., et al.: PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. Gigascience **9**, 1–12 (2020). https://doi.org/10.1093/GIGASCIENCE/GIAA022
8. Cristescu, M.E.: From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol. Evol. **29**(10), 566-571 (2014). https://doi.org/10.1016/j.tree.2014.08.001
9. Hering, D., et al.: Implementation options for DNA-based identification into ecological status assessment under the European water framework directive. Water Res. **138**, 192–205 (2018). https://doi.org/10.1016/j.watres.2018.03.003
10. Deiner, K., et al.: Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol. Ecol. **26**, 5872–5895 (2017). https://doi.org/10.1111/mec.14350
11. Callahan, B.J., McMurdie, P.J., Holmes, S.P.: Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. **11**(12), 2639–2643 (2017). https://doi.org/10.1038/ismej.2017.119
12. DeSantis, T.Z., et al.: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72**, 5069–5072 (2006). https://doi.org/10.1128/AEM.03006-05
13. Quast, C., et al.: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. **41**, D590–D596 (2013). https://doi.org/10.1093/nar/gks1219
14. Abarenkov, K., et al.: The UNITE database for molecular identification of fungi – recent updates and future perspectives. https://www.jstor.org/stable/27797548. (2010). https://doi.org/10.2307/27797548
15. Ratnasingham, S., Hebert, P.D.N.: BARCODING: bold: the barcode of life data system (http://www.barcodinglife.org). Mol. Ecol. Notes. **7**, 355–364 (2007). https://doi.org/10.1111/j.1471-8286.2007.01678.x
16. Machida, R.J., Leray, M., Ho, S.-L., Knowlton, N.: Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci. Data **41**(4), 1–7 (2017). https://doi.org/10.1038/sdata.2017.27
17. Pylro, V.S., et al.: Brazilian microbiome project: revealing the unexplored microbial diversity—challenges and prospects. Microb. Ecol. **67**(2), 237–241 (2013). https://doi.org/10.1007/s00248-013-0302-4
18. Frøslev, T.G., et al.: Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nat. Commun. **8**, 1–11 (2017). https://doi.org/10.1038/s41467-017-01312-x
19. Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M.: Swarm v2: highly-scalable and high-resolution amplicon clustering. Peer J. **3**, e1420 (2015). https://doi.org/10.7717/PEERJ.1420
20. McMurdie, P.J., Holmes, S.: phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE **8**, e61217 (2013). https://doi.org/10.1371/journal.pone.0061217
21. Schubert, M., Lindgreen, S., Orlando, L.: AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes **91**(9), 1–7 (2016). https://doi.org/10.1186/S13104-016-1900-2
22. Zhang, J., Kobert, K., Flouri, T., Stamatakis, A.: PEAR: a fast and accurate Illumina paired-End reAd mergeR. Bioinformatics **30**, 614–620 (2014). https://doi.org/10.1093/bioinformatics/btt593
23. Schmieder, R., Edwards, R.: Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**, 863–864 (2011). https://doi.org/10.1093/bioinformatics/btr026
24. Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F.: VSEARCH: a versatile open source tool for metagenomics. Peer J. **4**, e2584 (2016). https://doi.org/10.7717/PEERJ.2584

25. Cole, J.R., et al.: Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. **42**, D633–D642 (2014). https://doi.org/10.1093/NAR/GKT1244
26. Benson, D.A., et al.: GenBank. Nucleic Acids Res. **41**, D36–D42 (2013). https://doi.org/10.1093/NAR/GKS1195
27. Tatusova, T.A., Madden, T.L.: BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174**, 247–250 (1999). https://doi.org/10.1111/j.1574-6968.1999.tb13575.x
28. Bengtsson-Palme, J., et al.: Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Meth. Ecol. Evol. **4**, 914–919 (2013). https://doi.org/10.1111/2041-210X.12073
29. McDonald, D., et al.: The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience **1**(1), 2047-217X (2012). https://doi.org/10.1186/2047-217X-1-7
30. Gohl, D.M., et al.: Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat. Biotechnol. **349**(34), 942–949 (2016). https://doi.org/10.1038/nbt.3601
31. Bakker, M.G.: A fungal mock community control for amplicon sequencing experiments. Mol. Ecol. Resour. **18**, 541–556 (2018). https://doi.org/10.1111/1755-0998.12760
32. Bista, I., et al.: Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Mol. Ecol. Resour. **18**, 1020–1034 (2018). https://doi.org/10.1111/1755-0998.12888
33. Encyclopedia of Machine Learning: Encycl. Mach. Learn. (2010). https://doi.org/10.1007/978-0-387-30164-8
34. Toju, H., Tanabe, A.S., Yamamoto, S., Sato, H.: High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. PLoS ONE **7**, e40863 (2012). https://doi.org/10.1371/JOURNAL.PONE.0040863

APÊNDICE C - *A review and benchmarking of assembling nuclear genome of plants.*

OXFORD UNIVERSITY PRESS | Briefings In Bioinformatics

A review and benchmark of assembling nuclear genomes of plants.

Journal:	*Briefings in Bioinformatics*
Manuscript ID	Draft
Manuscript Type:	Review
Date Submitted by the Author:	n/a
Complete List of Authors:	Oliveira, Renato; Instituto Tecnológico Vale Desenvolvimento Sustentável, Environmental Genomics; Universidade Federal de Minas Gerais, Programa Interunidades de Pós-graduação em Bioinformática Vasconcelos, Santelmo; Instituto Tecnológico Vale Desenvolvimento Sustentável, Environmental Genomics Nunes, Gisele; Instituto Tecnológico Vale Desenvolvimento Sustentável, Environmental Genomics Petersen, Bent; University of Copenhagen Globe Institute, Center for Evolutionary Hologenomics; Asian Institute of Medicine Science and Technology University, Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio) Sicheritz-Pontén , Thomas; University of Copenhagen Globe Institute, Center for Evolutionary Hologenomics; Asian Institute of Medicine Science and Technology University, Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio) Oliveira, Guilherme; Instituto Tecnológico Vale Desenvolvimento Sustentável, Environmental Genomics
Keywords:	Plant genome, Benchmark, pipeline comparison, genome assembly

Journal Title Here, 2022, pp. 1–9

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Paper

PAPER

A Review and Benchmark of assembling nuclear genomes of plants.

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1], Gisele Nunes[1], Bent Petersen[3,4], Thomas Sicheritz-Pontén[3,4], and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, [3] Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Denmark, [4] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia.

*Corresponding author. renato.renison@gmail.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Advancements in sequencing technologies have allowed exponential growth of genomes deposited in public databases, with data generation outpacing the decrease in sequencing costs. Assembling large eukaryotes genomes, particularly plants, remains complex and expensive due to inherent characteristics like polyploidy, genome size, and repetitive regions. Many assembly software have been developed to address these issues, but a fair comparison of their effectiveness in assembling plant nuclear genomes is lacking in the literature. To address this gap, we collected information on 856 complete plant nuclear genomes deposited in the NCBI database by August 2022, along with associated data such as ploidy, genome size, chromosome number, GC content, sequencing technologies, and assembly software. We sequenced by simulation two diploid plant species (*Setaria italica* and *Oryza sativa*), generating short and long reads. Three pipelines used in recent publications of complete plant nuclear genomes were compared to identify optimal strategies and areas for improvement. WTDBG2 and SMARTdenovo generated assemblies with larger contig sizes and higher N50 values but with many assembly errors compared to the CANU and SOAPdenovo. SOAPdenovo generated fragmented assemblies, even when combined with long-read assemblers. CANU had fewer assembly errors, with good contig length and N50 values. Quickmerge joined different assemblies, increasing N50 values without introducing many errors. PurgeHaplotigs identified syntenic contigs from highly heterozygous regions, increasing the final assembly's N50 values. SSPACE and GapCloser formed new scaffolds and filled over 90% gaps. Our results highlight areas for improvement in existing pipelines and suggest opportunities for developing new assembly strategies.

Keywords: Plant genome, benchmark, pipeline comparison, genome assembly.

Key points: The study addresses the lack of fair comparison among available assembly software for plant genome assembly by collecting and analyzing data from 856 complete nuclear genome records of plants deposited in the NCBI database;

Among the compared software, WTDBG2 and SMARTdenovo produced assemblies with larger contig sizes and higher N50 values but had many assembly errors, while CANU yielded fewer errors with balanced contig length and N50 values;

The use of supplementary software like Quickmerge, PurgeHaplotigs, SSPACE, and GapCloser can significantly improve assembly quality, increasing N50 values, separating syntenic contigs, and filling gaps in the assemblies, providing insights for the optimization of existing pipelines and suggestions for the creation of new ones.

Introduction

The importance of bioinformatics has been heightened with the advent of novel NSG sequencing technologies and genome assembly algorithms [1]. Public databases such as NCBI/GenBank [2] house thousands of assembled genomes from microorganisms, animals, and plants. Except for DNA/RNA viruses of up to 30kb, there is still no method capable of obtaining the complete genetic information of a DNA molecule of an organism. To address this challenge, the shotgun sequencing method was proposed. This method randomly fragments the DNA molecule into smaller pieces that are subsequently sequenced, generating individual sequencing reads. Generally, if these reads overlap (share similar nucleotide sequences), they are grouped into larger structures called contigs, just as contigs can be grouped into even larger structures called scaffolds. The different ways of assembling these reads constitute the well-known Fragment Assembly Problem [1]. Many programs and algorithms can assemble a set of reads using different assembly methods. Most of these are based on well-known methods: the Overlap-Layout-Consensus (OLC) approach and the de Bruijn graph, which use Hamiltonian and Eulerian graphs properties, respectively [3]. These methods were intensively used to assemble prokaryotic and small genomes.

Assembling eukaryotic genomes is still a challenge for bioinformatics. Properties inherent to these genomes, such as size, duplication, and repetition of genome regions, make the assembly into large contigs or scaffolds difficult [4]. Since plant genomes are commonly reported as larger than animal genomes or other eukaryotes, many efforts have been made to overcome the difficulties associated with this fact. The problem of genome size is being solved by the evolution of sequencing machines, which can now generate a high volume of sequence data and can fully cover a large genome, e.g., Illumina (NextSeq, HiSeq, NovaSeq), producing reads in the range of 75-300 bp and up to 6 Tb of throughput. Repetitive sequences in plant genomes are still an issue for assembly algorithms. Two types of sequence repetitions are easily recognized: (i) short-period interspersion (where the single copy sequences range from 300 to 1,200 bp and are separated by repeated sequences ranging from 50 to 2,000 bp) and (ii) long-period interspersion (single copy sequences ranging from 2,000 to 6,000 bp separated by longer repeat sequences). The low complexity and small size of these regions make it difficult to produce reliable assembly results [4]. Furthermore, whole genome duplication events (polyploidization) are commonly observed in angiosperms, being characterized by the presence of at least an additional copy of the whole chromosome set within the nucleus of a gametic cell [5], and may result in poor genome assembly quality. The emergence of Third Generation Sequencing (TGS) machines producing long reads, such as those from PacBio (up to 25 Kb long) and Oxford Nanopore (up to 100 Kb), also contributed to solving these problems of repeat regions, allowing for more contiguity in the assembly. Many current genome assemblers have been developed to handle large genomes, repetitive DNA, and polyploidy. Nevertheless, the availability of many assemblers results in challenging choices due to the need for more direct efficiency and accuracy comparisons.

Of the almost 410,000 known plant species worldwide (https://www.catalogueoflife.org/data/metadata), only 349 had their complete nuclear genome assembled and deposited in the NCBI/GenBank database until August 2022. Only 0.00085% of the known plant species have their nuclear genome assembled at the chromosome level due to the complexity and size of these genomes, which require resources and time for their complete analysis. With current efforts to sequence biodiversity genomes in scale, it is essential to know how these deposited genomes were sequenced and assembled to understand and suggest new ideas to improve the assembly process. Thus, a comparative study among currently employed genome assemblers used for large plant genomes allows the choice of which assembler is appropriate for a given genome, considering all inherent difficulties and complexities. Furthermore, the lack in the literature of recent benchmarks among assemblers for plant genomes makes such a comparative study necessary. In this article, we present a survey with sequencers and assemblers that were and are currently being used to analyze plant nuclear genomes. Moreover, we have compared the various plant genome assemblers available and used by the scientific community.

The history of sequencing and assembling plant nuclear genomes

According to the Royal Botanic Gardens, Kew, and the World Checklist of Vascular Plants (WCVP version 10), there are 382,332 known and accepted vascular plant species worldwide [6]. The Catalogue of Life Checklist counts 407,329 species of vascular plants and bryophytes. Of the known plant species, 217,322 have at least one DNA sequence deposited in GenBank/NCBI (as of April 13th, 2023). The number of represented species has grown in the last five years, mainly due to the advance and decreasing cost of new sequencing technologies [4]. Hence, analyzing and assembling plant genomes is extremely important because it allows for answering relevant biological questions [7], knowing their gene content, discovering new genes, breeding, and conducting evolutionary and adaptive studies [8].

The first nuclear genome of a plant species (*Arabidopsis thaliana* (L.) Heynh.) completely assembled was achieved using the Sanger sequencing approach more than two decades ago [9]. Despite the high quality of the bases sequenced by this technology, it had some disadvantages, such as high cost and low throughput, thus being a considerably slower sequencing approach [10]. These disadvantages were partially overcome with Next Generation Sequencing (NGS) and Third-Generation Sequencing (TGS) platforms (Illumina [11], Ion Torrent [12], PacBio [13], and MinION Oxford Nanopore [14]), which allowed for much higher data throughput, fast sequencing (providing up to 1.5 TB of data in just a few days), low costs, and longer reads. Initially, PacBio and Oxford Nanopore sequencers, despite being able to generate longer reads, presented some disadvantages compared to the Illumina platform, such as low sequencing coverage and quality of reads. However, new methodologies have been developed and implemented for long-read sequencer data, which have increased the quality of their sequenced bases [15]. In line with these advancements, PacBio has recently launched its latest generation of sequencing machines - Sequel II, Sequel IIe, and Revio - designed to deliver highly accurate long reads (HiFi reads) with enhanced throughput and cost-effectiveness.

The reads sequenced by the Sanger method were assembled by the first developed genome assemblers: TIGR [16] and Celera [17]. As the NGS and TGS platforms evolved and sequencing throughput increased, new assemblers were designed to handle the large amount of data generated. The de Bruijn Graph method [18] aimed at simplifying the representation of millions of reads in smaller structures called *k-mers*, reducing the computational complexity of the analyses for the assemblers, such as Velvet [19], ABySS [20] and SPAdes [21]. Although dealing with large datasets became easier, these early assemblers needed to change their assembly process to analyze datasets originating from more complex genomes, such as animal and plant genomes, as they were first designed to assemble prokaryotic genomes.

As previously mentioned, the main challenges for plant nuclear genome assembly are their large genome size, polyploidy, highly repetitive regions, and duplications [22]. As the NGS and TGS platforms were being developed, the size of the plant nuclear genome was no longer considered an issue but rather the complexity of the genome itself. This complexity is often associated with highly repetitive regions of the genomes, which are challenging to assemble since identical or nearly identical reads could come from different locations in the genome, generating gaps, ambiguities, and collapses in the assembly graphs. However, depending on the extension of the region, the problem of repetitive DNA in a genome can be addressed with the intensive use of long-read sequencers, such as PacBio and Oxford Nanopore, in which a single sequence can span both sides of a repeat region [4].

Polyploidy (or whole genome duplication) is a condition where there is at least one additional copy of the whole chromosome set within the nucleus of a gametic cell, being a major driving force in plant evolution, speciation, and diversity [23,24]. For the assembly software, this duplication can increase the number of erroneous assemblies, thus reducing the accuracy of the assembled genome. Researchers used chromosome sorting with flow cytometry to overcome these problems and sequenced each separately to facilitate further analysis, such as mapping and genome assembly [25]. More recently, another technique widely used to minimize errors in polyploid genome assemblies is the Hi-C approach, which seeks to obtain information from chromatin interactions in the nucleus, allowing the contigs and scaffolds obtained in an assembly to be grouped with the proximity information obtained by the method, bringing the assembly to the chromosome level [26].

Some assemblers were developed to deal with the inherent complexities of a genome, (i) using short reads, such as SOAPdenovo2 [27], SSAKE [28], and AllPaths [29] and more recently, (ii) combining optical and Hi-C mapping data, such as Supernova [30], (iii) using long reads, such as MaSurCA [31], Canu [32], HINGE [33], Flye [34], and Racon [35], and (iv) specific to eukaryotic genomes, such as Platanus [36] and Falcon [37].

Many efforts have already been made to achieve the best assembly of a large eukaryotic genome. Table 1 [38] and Table 2 [39] summarize the first significant plant genome assemblies that have achieved whole genome status, along with the sequencing platform, genome size, assembly size, and genome ploidy. Additionally, Table 1 shows that from 1999 to 2009, the most used sequencing platform for plant genomes was Sanger, with the Illumina platform becoming more used afterward, enormously reflecting the number of genomes sequenced and assembled due to its high throughput.

Table 2 shows the first complete genomes of plants with ploidy other than diploid, along with their genome sizes. Tables 1 and 2 show that there have been several plant genome assembly projects, each with different ploidies, sequencing coverage, and assembly sizes. In addition, the sequencing platforms used by each research project also differ. Therefore, comparing the strategies and how they have evolved becomes difficult. However, updating all the information in Tables 1 and 2 with recent data allows research groups to decide which approach to use in new plant sequencing projects, considering all the parameters and methods of successful sequencing and assembly projects.

Only Assemblathon 2 [57] and GAGE [58] have compared the performance of different assemblers on large bird, fish, and mammalian genomes, highlighting the need for a recent benchmark of plant genome assemblers. In addition, efforts to improve and scale plant genome assemblies are needed to reduce assembly errors and erroneous genomic comparisons [40].

Global initiatives like the Earth Biogenome Project [41] are increasingly sequencing a large number of plant species. Due to their genome sizes and ploidy levels, analyzing these newly sequenced species might be challenging. The sequences for those plant species must be correctly generated, producing high-quality genomes which can be used for evolutive innovations and to benefit humanity while the conservation of those species also happens.

Methodology

We selected plant species deposited in NCBI that had complete genome status (chromosome level), extracted related metadata (genome size, number of chromosomes, ploidy, GC content, sequencers, and assembly software), and inferred their phylogenetic distribution. In addition, we also captured which genome assembly pipelines and software were involved in the most recent publications of complete plant genomes and which of these pipelines is considered the most appropriate for each of the chosen species.

Retrieving information from GenBank/NCBI

The following options were selected in the NCBI genome browser to obtain information regarding plants with complete genomes: Eukaryotes > Filters > Plants > Land Plants and Other Plants> Chromosome. The table was downloaded in the .tsv format, containing the following columns: Organism Name, Organism Groups, Strain, BioSample, BioProject, Assembly, Level, Size(Mb), GC%, Replicons, WGS_ID, Number of Scaffolds, Number of CDS, Release Date, GenBank FTP, RefSeq FTP, Number of Genes, Modify Date, and Number of tRNAs. The chromosome number information was obtained from the replicons column, which has an identifier for each chromosome present in the genome.

As new genomes can be deposited at any time on NCBI, the Python script merge_database.py was developed to take an old and new table and automatically check which records are new. Also, since the submitting group can update already deposited records, the script can check which records were updated. In addition to the information in the table downloaded from NCBI, it was also necessary to obtain more specific information about the methodology used in each deposited genome, such as the sequencers and software used in the assembly pipeline. The Python script retrieve_information.py was developed to obtain this information, using the E-utilities v. 13.7 tool from NCBI in the background [42].

To capture information about the ploidy of the plant species listed in the downloaded table, we automated the search by developing a Python script that searches for the ploidy information using the Plant DNA C-value database from the Royal Botanical Gardens, Kew (https://cvalues.science.kew.org/). Unfortunately, not all species had their ploidy successfully recorded in the database. Therefore, searching for the remaining ploidy information was done manually using the Google Scholar website (https://scholar.google.com.br/). The Venn Diagrams created to visualize the collected information were generated with Interactivenn [43].

The three scripts previously mentioned are available at https://github.com/reinator/plant_review.

Reconstructing the Phylogenetic Tree

Aiming to understand the phylogenetic arrangement of the plant species with complete genomes deposited in NCBI, a supermatrix of nucleotides was obtained from the coding gene sequences. Therefore, the nucleotide sequences of the chloroplast genomes were downloaded using Geneious Prime, which also extracted the nucleotide sequences of the coding genes from all downloaded chloroplasts.

We used SPLACE [44] to automate the entire process of supermatrix obtention and phylogenetic reconstruction, using as input the FASTA files containing the coding genes from the chloroplasts of the desired plant species. The supermatrix obtained by SPLACE was submitted to the RAxML v8 tool [45] to generate phylogenies with a support value of 1,000 replicates using the GTRGAMMA model. The spreading earthmoss *Physcomitrella patens* (accession NC_005087), model species for plant evolutionary and physiological studies [46], was used as an outgroup. The resulting phylogenomic tree then allowed a better understanding of which plant groups concentrate efforts on sequencing and assembly, showing the families and orders with representative species. The following sections will describe the process of sequencing simulation and assembly pipelines to be compared.

Sequencing simulation

The DNA sequencing platforms most used in recent publications of complete plant nuclear genomes are the Illumina platform, which generates many short reads and high coverage, and the PacBio platform, which produces long reads but usually with less coverage.

We used reads from simulated sequencing to compare the deployed pipelines for genome assembly. Thus, the comparison is fairer since the reads are generated from the same reference genome, allowing us to compare the results of the pipelines with the factual information we wish them to return (the reference genome itself). The software NEAT v.2 [47] was chosen to perform the sequencing simulation, mainly because it can simulate sequencing of short (Illumina) and long reads (PacBio), allowing to set the ploidy of the organism to be sequenced, which is closer to the reality of plant genome sequencing.

For the Illumina sequencing simulation, the following parameters were used: -R 151 (read size), --pe 300 30 (specifies that pairs of reads should be generated, with average insert size at 300bp and deviation of 30bp), -c 100 (sequencing coverage), -E 0. 0001 (probability of a base being wrong), --bam (generates the BAM file with the source region of each read), --vcf (generates a file with the variants arising from the sequencing error), -p # (where # is the ploidy of the organism you wish to sequence. 2 for diploid, 3 for triploid, 4 for tetraploid, and so on). The following parameters were used for the PacBio sequencing simulation: -R 15000, -c 30, -E 0.10, --bam, --vcf, -p #. All simulations were run using 40 threads and 180GB of RAM.

Pipelines for plant nuclear genome assembly

For the selection of pipelines for plant nuclear genome assembly, we analyzed complete plant nuclear genomes published from 2018 to date. We chose the pipelines that differed the most regarding the assembly software and approach used (Figure 1).

Pipeline 1 was used to publish the complete nuclear genome of *Cannabis sativa* L. [48]. Initially, the long reads generated by PacBio are subjected to an auto-correction process in Canu v.2.0. Only the reads with coverage greater than or equal to 20 are corrected (corOutCoverage=20). The corrected reads are termed unitigs and are generated through the following procedure: only overlaps between reads with up to 30% error (corErrorRate=0.3) and a 3% deviation in error (-dg 3) are considered, and bubbles found in the overlaps graph with up to 3% deviation in error (-db 3) are merged. Additionally, unitigs are broken if alternative overlaps of at least 500 bp (-ca 500) or up to 50% (-cp 50) of the size of the best overlap with up to 1% deviation in error (-dr 1) are found.

The unitigs generated by Canu in the self-correction step are used in two different assemblers: WTDBG2 v2.5 [49] and SMARTdenovo [50]. In the assembly with WTDBG2, the parameters are preset as recommended for assembling PacBio data (-x rs) according to the author, where the kmer size used is 21 (-p 21), only ¼ of the generated kmers are used (-S 4), 5% minimum similarity between kmers (-s 0.05) and only reads longer than 5,000 bp are used (-L 5000). In addition, the genome size to be assembled is also indicated (-g), allowing coverage calculations. In the assembly with SMARTdenovo, the overlapping algorithm used is "dmo" (Dot matrix overlapper, -e dmo), the kmer size used is 17 (-k 17), only reads larger than 5,000 bp will be used (-J 5000), and consensus generation is enabled (-c 1).

Next, the results of the assemblies made with WTDBG2 and SMARTdenovo are joined with the Quickmerge v0.3 genome reconciliation tool [51], where the WTDBG2 result is used as a hybrid assembly, and the SMARTdenovo result is used as a self-assembly, where only contigs greater than 100 bp are used (-l 100. Contigs with overlaps whose rate between the aligned region and the sum of unaligned regions is greater than 5 (-hco 5) were overlapped, and only contigs that have the same rate greater than 1.5 are extended (-c 1.5), considering only overlaps that are greater than 5,000 bp (-ml 5000).

The last step of Pipeline 1 is to correct the output generated by Quickmerge using the tools Minimap2 v2.17 [52] and Pilon v1.23 [53]. Minimap2 mapped the long reads from PacBio to the Quickmerge contigs with the "-x map-pb" parameter, which automatically enables homopolymer compression (-H) and kmer size to 19 (-k), generating a SAM file as output (-a). In this step, the short reads sequenced in Illumina are also mapped to the resulting Quickmerge assembly, ensuring a more significant correction of possible assembly errors, as Illumina sequencing has a higher coverage (~100X) and a lower error rate (~0.1%) than sequences produced by PacBio (30X and 10%, respectively). With Illumina data, the parameter "-x sr" is used, which automatically sets the kmer size to 21 (-k 21), the compression window of consecutive kmers to 11 (-w 11), enables fragmentation mode (--frag=yes), increments two points when identifying alignment hit (-A 2), applies eight penalty points in cases of alignment error (-B 8), 12 and 32 penalty points for opening gaps (-O 12,32), uses two and one penalty points for extending gaps (-E 2,1), sets the maximum gap size to 50 bp (-r 50) and generates an output file in SAM format (-a). Other parameters are configured as the tool's author recommends for short reads.

The SAM format outputs generated for the short and long reads are converted to the BAM format using Samtools [54] and then used as input into Pilon, along with the resulting Quickmerge assembly, using the default parameters and enabling the "--diploid" parameter. Pilon used its heuristics to correct base errors, small indels, misassemblies, and fill gaps, then generated the final assembly.

Pipeline 2 was used to assemble the complete nuclear genome of *Mangifera indica* L. [55]. Besides performing the self-correction of the long reads in this pipeline, Canu also assembles the corrected reads. The parameters used to correct the reads are the same as in Pipeline 1. In the assembly step, besides having informed the genome size with the parameter "genomeSize=", only reads larger than 1,000 bp (minReadLength=), a minimum overlap of 500 bp (minOverlapLength=), and overlaps with less than 4.5% error (correctedErrorRate=) are assembled, using the "falcon" algorithm to generate the consensus sequence (corConsensus=).

The Illumina short reads and the long reads corrected by Canu are used as input to Minimap2 and Pilon to correct the contigs generated by Canu, with the same parameters used in Pipeline 1. Finally, the Pilon-corrected assembly goes through a final correction step. PurgeHaplotigs [56] identifies the highly syntenic contigs, choosing one to compose the haploid genome assembly and associating it with its syntenic contig, thus reducing the assembly redundancy in the highly heterozygous regions. To perform haplotype identification, the uncorrected long reads are mapped against the resulting Pilon correction with the parameter "-x map-pb", generating a BAM (-a) format file. PurgeHaplotigs first generates a histogram with the coverage of the genome reads using the BAM file and the FASTA file of the Pilon correction. Figure 2, taken from the PurgeHaplotigs user manual, shows a typical histogram generated for a genome with different coverage in heterozygotic regions. Visualization of the histogram allows us to choose three essential parameters: the lowest (-l), middle (-m), and highest (-h) coverage limit, shown in Figure 2 as low-cutoff, mid-point, and high-cutoff, respectively. According to the previous parameters, contigs with 80% or more of their coverage as low or high will be discarded. Those with 80% or less of their coverage between the low and medium threshold are marked as suspect contigs. Then, haplotype contigs (haplotigs) are identified and extracted from the final assembly, thus decreasing redundancy.

Pipeline 3 was adapted from the one used in the publication of the complete nuclear genome of *Morella rubra* Lour. [57]. Originally the pipeline used Falcon to perform the self-correction and assembly steps of the PacBio reads and HABOT2 (https://github.com/asarum/HABOT2) to merge the assemblies, but for reasons of lack of user support, these tools were replaced by others that fulfill the same purpose. The long reads are self-corrected and assembled by Canu using the same parameters as in

Pipelines 1 and 2. The novelty in Pipeline 3 is that Illumina reads are assembled using SOAPdenovo2 v2.04 software with minimum and maximum kmer sizes of 23 and 63, respectively. The resulting contigs from the assemblies with Canu and SOAPdenovo2 are joined by Quickmerge using the same parameters as in Pipeline 1. Next, the Illumina reads are mapped by SSPACE [58] to the result generated by Quickmerge to try to form new scaffolds using the default parameters. The gaps in the result generated by SSPACE are closed with GapCloser [27] using Illumina reads and default parameters. Finally, the assembly resulting from the gap closure is analyzed by PurgeHaplotigs, using the raw PacBio reads and the same parameters explained in Pipeline 2, generating the final assembly.

All assembly pipelines were run with 40 threads and 1TB of RAM on Computerome2 ("EOSC-Nordic | Computerome 2.0 - HPC", [n.d.]), the current Danish National Life Science Supercomputer.

Assembly metrics evaluation

The quality of the assemblies generated by the pipelines described in the previous section was assessed using quantitative metrics that measure sequence size and contiguity, such as N50 value, largest and smallest contig, number of contigs, total bases, and number of gaps. These metrics were obtained by the PRINSEQ tool [59] using the "-stats_all" parameter.

QUAST-LG [60] was used to generate qualitative metrics, including the number of misassemblies and the fraction of the assembled genome. The tool used the Illumina and PacBio reads aligned to the reference genome to produce the UpperBound assembly, representing the highest assembly quality possible. The circular visualization of the assembly results made by the pipelines aligned to the reference genome was generated by CIRCOS [61], indicating in red lines where misassemblies occurred.

Genome completeness was analyzed by BUSCO v. 5.4.4 [62], which evaluated the gene content of near-universal single-copy orthologs using the Poales_odb10 database with 4896 ortholog genes.

Results and Discussion

Ploidy, Sequencers, and Assemblers software

As of August 2022, NCBI contained 856 records of complete nuclear genomes of plants, representing 481 unique species. Of those species, 378 have a strict ploidy: 330 are diploid, one is triploid, 36 are tetraploid, seven are hexaploid, three are octaploid, and one has a ploidy level of 16. We could not find the ploidy for 81 species, and 20 species might be found in multiple ploidies (see Supplementary Table S1).

From the 481 plant species, we found the ploidy information for 286 of them using the Plant DNA C-value database from the Royal Botanical Gardens, Kew. The ploidy information for 114 plant species was found manually, searching on Google Scholar for publications. We also sent those manually found information to the Kew curating team, contributing to increasing their database. Public genomic databases must keep the information well curated while correctly gathering the associated metadata. The difficulty when searching the ploidy level of plant species raises concerns about metadata organization. Recently, the GOAT database [63] has gathered the most information for the genomes deposited in public databases, including ploidy levels. An official database dedicated to recording information on plant species is the CCDB (Chromosome Counts Database) [64], but no ploidy information can be found on it, just chromosome counts.

Of the 330 diploid species, 49 species are from the Poaceae family, which includes important crops, such as rice (*Oryza sativa* L., and other 14 species of the genus), maize (*Zea mays* L.), sorghum (*Sorghum bicolor* (L.) Moench), and some other grass species, such as goat-grass (eight species of the genus *Aegilops*), panicgrass (*Panicum hallii* Vasey), and foxtails (*Setaria italica* and *S. viridis* (L.) P.Beauv). In Myrtaceae, out of the 33 species with complete genomes, 31 are from *Eucalyptus*. Also, several food crops from Fabaceae have been fully sequenced, totalizing 30 legume species, including soybeans (*Glycine max* (L.) Merr., and other two species of the genus), peanuts (four *Arachis* species), chickpea (*Cicer arietinum* L., and one more species of the genus), common bean (*Phaseolus vulgaris* L., and another species of the genus), cowpea (*Vigna unguiculata* (L.) Walp and four other species of the genus), pea (*Pisum sativum* L.), and some medicinal plants (e.g., *Spatholobus suberectus* Dunn, and *Trifolium pratense* L.). Both Malvaceae and Brassicaceae have 21 sequenced species each, including 19 cotton species (*Gossypium* spp.) and cocoa (*Theobroma cacao* L.), from the former, and the model plant *Arabidopsis thaliana* (L.) Heynh., black mustard (*Brassica nigra* (L.) W.D.J. Koch), cabbage (*Brassica oleracea* L.), and radish (*Raphanus sativus* L.), from the latter family.

The only triploid species with a complete nuclear genome deposited on NCBI is *Ananas comosus* (L.) Merr. (pineapple), the only representant of the Bromeliaceae. Of the 35 tetraploid species, some Fabaceae species are also represented, including three peanut species (*Arachis* spp.). Eight Poaceae species are also represented among the tetraploids, including the wild emmer wheat (*Triticum dicoccoides* (Körn. ex Asch. & Graebn.) Schweinf), pasta wheat (*Triticum durum* L.), the wild rice (*Oryza minuta* J. Presl), teff (*Eragrostis tef* (Zucc.) Trotter), and the wild oat (*Avena insularis* Ladiz.). Moreover, five cotton species (*Gossypium* spp.) and coffee (*Coffea arabica* L.) are tetraploid representatives of Malvaceae and Rubiaceae, respectively, with their complete nuclear genome deposited on NCBI. Among the seven hexaploid species, some important food crops are represented, such as the sweet potato (*Ipomoea batatas* (L.) Lam., Convolvulaceae), oat (*Avena sativa* L., Poaceae), and kiwifruit (*Actinidia deliciosa* A. Chev., Actinidiaceae). Finally, considering the octaploid species, the strawberry (*Fragaria x ananassa* Duch., Rosaceae) and sugar cane (*Saccharum officinarum* L., Poaceae) are represented, and the only 16-ploid is a carnivorous plant (*Utricularia gibba* L., Lentibulariaceae), a species with an estimated genome of less than 100 Mbp (see Supplementary Table S1).

Notably, the plant species that have concentrated most efforts on sequencing and assembly are important for society as food crops, most of which are diploid species.

Figure 3 shows two Venn Diagrams containing the six most used sequencing technologies (Figure 3A) and the six most used assembly software (Figure 3B) in the 856 records for plants with complete nuclear genomes. PacBio was the leading platform with which the data were generated, with 416 plant genomes being sequenced, followed by Illumina short reads with 355, Oxford Nanopore with 151, 38 with Illumina Hi-C, 28 with 454 pyrosequencing, and 23 with Sanger sequencing. It is important to mention that these values may complement each other, i.e., one plant genome could be sequenced with two or more technologies (Figure 3A). For the assembly software, CANU was the most used in 213 plant genomes, followed by Falcon, used in 143, 67 used MECAT [65], 59 used SOAPdenovo, 35 used wtdbg, and 33 used AllPaths-LG, Smartdenovo, and hifiasm (Figure 3B). The remaining genomes were assembled with tools that no longer have support (Newbler, Celera, Arachne, PhredPhrapConsed) or used in fewer records (NextDenovo, HiRise, MaSurCA, Flye, DenovoMAGIC, Abyss, Supernova, Racon, Platanaus, miniasm, SPAdes, necat, Hera, shasta, Discovar, Tritex, HABOT, DBG2OLC, and SparseAssembler).

All software used to assemble the 856 complete plant nuclear genomes are listed in Table 3.

Figure 4 shows how the top six sequencing technologies and the top eight assembly software were used throughout the analyzed period. The Sanger sequencing technology was intensively used in the first years of genome sequencing, but its use has reduced with time (Figure 4A). The use of Sanger technology as the workhorse for data production was last observed in 2021 in a project to sequence and assemble the genome of the *Oryza sativa* Japonica Group, together with PacBio and Illumina sequencing platforms [66]. Another sequencing technology not used since 2018 is the 454 Roche sequencer, used for the first time in 2009. PacBio and Illumina sequencers started to be used in 2009 and are now intensively used on complete plant nuclear genomes projects. The Oxford Nanopore and Hi-C technologies began being used in 2015 and are now intensively used for plant sequencing. Although Sanger and 454 sequencers are no longer used for plant genome assembly, they were essential for sequencing the first plant genomes (e.g., [67,68]). Developing new sequencing technologies demands new analytical tools, which might explain the high diversity of assembly software used across time and the need for developing new ones.

Figure 4B shows that two short-read assemblers (SOAPdenovo and AllPaths-LG) and two long-read assemblers (Falcon and Canu) started being used in 2009 until now, except for AllPaths-LG, that has been last used in 2020 (BioProject PRJNA237957 on NCBI). Wtdbg, Smartdenovo, and Hifiasm are long-read assemblers that started being used in 2018 [69], 2019 [70], and 2021 [71], respectively. Mecat also started being used in 2011 [72] and has been used until now. Interestingly, 2020 was the year with more genomes sequenced by Illumina and PacBio, and these sequenced genomes were assembled mostly with Falcon and Canu assemblers (e.g., [73,74]).

Of the 481 species with complete nuclear genomes, only 349 had a chloroplast genome deposited on NCBI. Figure 5 shows the phylogenomic tree containing those 349 species, with alternating colors indicating the families and the orders alongside braces. Ninety-one families are represented in the phylogenomic tree, where 41 have more than one species, with Poaceae being the family with the most species (52), followed by Malvaceae (24), Rosaceae (23), Fabaceae (21), and Brassicaceae (19). Lamiales is the order with the highest number of represented families (nine), followed by Ericales (six), Caryophyllalles (five), and Rosales (five), from a total of 40 represented orders.

Of the 1,085 accepted plant families, according to The Catalogue of Life Checklist, only 91 (~8%) have at least one species with a complete nuclear genome assembled and deposited on NCBI. Out of the 235 accepted plant orders (https://www.catalogueoflife.org/data/metadata), only 40 (17%) have at least one species with a complete nuclear genome.

Sequencing simulation and pipeline comparison

Since most efforts are concentrated on diploid species, we searched pipelines that were applied to them. We performed the Illumina and PacBio CLR sequencing simulation of the diploid species *Setaria italica* (L.) P. Beauv and *Oryza sativa* L. Table 4 shows the results of the sequencing simulation for those species.

For the benchmarking of the pipelines, we show the results obtained for each species along with the assembly metrics (Figure 6). At each step of the pipelines, the assembly metrics are shown in Tables S1 and S2, allowing us to check if there were improvements during the workflow execution (see Supplementary material). We also show the main assembly metrics with QUAST-LG for each pipeline result.

Setaria italica

The assembly metrics obtained by the three pipelines when analyzing the simulated sequencing dataset for *Setaria italica* are shown in Figure 6A. Pipeline 1 initially obtained the highest N50 values, largest contig size, and the fewest number of contigs with the WTDBG2 and SMARTdenovo assemblers, despite having a smaller fraction of the reference genome, with the fewest total bases assembled. Quickmerge joined the two assemblies, increasing the values of N50, the largest contig size, and the total assembled bases (see Supplementary Table S2). In Pipeline 2, the Purge Haplotigs software increased the N50 of the Pilon-corrected assembly, possibly removing redundancies but reducing the fraction of the assembled genome (Table S2). In Pipeline 3, the assembly with SOAPdenovo came closest to the reference genome size. Still, it caused a very fragmented assembly that was perpetuated during the pipeline until the execution of Purge Haplotigs, which was able to increase the N50 and decrease the number of contigs and consequently reduce the total number of bases of the final assembly. Nevertheless, Pipeline 3 produced the highest number of contigs. As for the running time, Pipeline 1 was the fastest to run (7.84 h), followed by Pipeline 2 (30.06 h), with Pipeline 3 being the most extended (56.12 h).

The results in Figure 6A show metrics referring to the size of the contigs in the assemblies, not necessarily meaning that the assembly performed well since there may be misassemblies and chimeric contigs. Figure 7A shows the result generated by QUAST when aligning the contigs against the reference genome. It is apparent that the assembly resulting from Pipeline 1 produced the most misassemblies, while Pipeline 2 produced the fewest assembly errors, and Pipeline 3 assembled a greater fraction of the genome. Figure 7B shows the circular alignment generated by CIRCOS in conjunction with QUAST-LG, where the assembly errors that occurred in the contigs aligned to the reference genome chromosomes are shown as red lines. It is evident in Figure 6B that the assembly generated by Pipeline 1 has caused many assembly errors, even though it was the pipeline with the largest contigs and the least number of them.

Oryza sativa

Figure 6B shows the results obtained by the three pipelines when assembling the *O. sativa* genome. Again, Pipeline 1 initially obtained the highest N50 values, largest contig size, and the fewest number of contigs with the WTDBG2 and SMARTdenovo assemblers, and this time getting closer to the actual size of the reference genome with the SMARTdenovo assembly. Quickmerge joined the two assemblies, increasing the values of N50, the largest contig size, and the total assembled bases (see Supplementary Table S3). In Pipeline 2, the Purge Haplotigs software increased the N50 of the Pilon-corrected assembly. The assembly generated by Canu exceeded the total bases of the reference genome, but when haplotigs were removed, the assembly was closer to the actual reference size. In Pipeline 3, the assembly with SOAPdenovo was again very fragmented. The fragmentation was reduced by Purge Haplotigs, which increased the N50 and significantly decreased the number of contigs. Notably, the software GapCloser and Purge Haplotigs reduced by 98% the number of gaps in a single contig and by 99.99% the number of contigs with gaps at the end of the assembly (see Supplementary Table S3).

Figure 8A shows the result generated by QUAST when aligning the contigs generated by the pipelines with the *O. sativa* reference genome. The assembly resulting from Pipeline 1 caused the most errors. The pipeline which caused the fewest assembly errors was Pipeline 2. Pipeline 3 did not generate as many assembly errors as Pipeline 1 but generated many mismatches. All pipelines have assembled more than 95% of the reference genome. This occurred due to the lower fragmentation presented by the assembly of the *O. sativa* genome, allowing a larger fraction of its genome to be sequenced in the step of sequencing simulation. Figure 8B shows the alignment of the contigs from each pipeline with the *O. sativa* reference genome. Figure 8B also shows that the assembly made by Pipeline 1 has many assembly errors, despite having the best N50 values, the largest contig, and the fewest number of contigs. It can also be seen that Pipeline 2 has the fewest assembly errors.

Table 5 shows BUSCO results, where for *S. italica*, Pipeline 3 obtained the best completeness value, reaching 95.2% and around 8% more than the completeness obtained by the other pipelines. Although the single-copy, fragmented and missing BUSCOs obtained by Pipeline 3 were better, the duplicated BUSCOs from Pipeline 2 were the best. For *O. sativa*, Pipeline 2 generated the best completeness result (97.6%), but the other pipelines also reached more than 95.5% of completeness.

To investigate why Pipeline 1 obtained the worst results in the benchmarking, we used the Smartdenovo and the Wtdbg2 assemblies generated for *Setaria italica* (Figure 9A) and *Oryza sativa* (Figure 9B) and used them as input for QUAST-LG. In *S. italica* results, Smartdenovo slightly causes more assembly errors, possibly compromising all genome projects that used it as an assembler, more precisely 31 genome records since 2019 and until August 2022. However, in *O. sativa* results, Wtdbg2 causes more erros, what might compromise 36 genome projects that used it as assembler, from 2018 until August 2022. The details of comparing Smartdenovo and Wtdbg2 with QUAST-LG when assembling the genomes of *Setaria italica* and *Oryza sativa* can be found in Supplementary Tables S4 and S5, respectively.

Of the eight assembly software most used to analyze plant genomes, only two (SOAPdenovo2 and AllPaths-LG) use short reads as input. This less frequent use might be due to the high genome fragmentation the short reads can cause when assembled. Corroborating this observation, results for Pipeline 3 were the most fragmented. Short reads were also used to correct long reads errors, but since HiFi long reads started being used [75], the generation of short reads for nuclear genome assembly may decrease. Although HiFi reads are currently the technology of choice, the combination with Hi-C data can provide final assemblies with more contiguous and accurate information [76–78]. Reads generated by Oxford Nanopore Technologies (ONT) are also starting to be increasingly used to achieve Telomere-to-Telomere (T2T) genome assemblies.

Conclusions and perspectives

This study gathered information on complete plant genomes deposited on NCBI that cannot be easily found in public databases and compared three recent pipelines used to assemble nuclear plant genomes. The collected data on ploidy, genome size, chromosome number, GC content, total genes, sequencing technologies, and used assembly software can guide other research groups to understand better the whole process of genome assembly for plant species, given the high number of analytical tools aiming to obtain the best results.

The pipeline comparison revealed that Pipelines 2 and 3 had produced fewer errors, although Pipeline 3 has produced a more fragmented assembly, possibly because it also uses an assembler for short reads. The tool WTDBG2 used in Pipeline 1 generated many assembly errors, causing them to spread throughout the pipeline.

The comparative analysis presented here allowed us to know which pipelines used in the literature analyze the data correctly, raising the importance of performing more comparative and benchmark studies. The gathered information may lead to improving current tools and developing new ones. In future work, we intend to find pipelines used in plant species with different ploidy levels. We also will include HiFi, ONT, and Hi-C reads in future benchmark work.

Competing interests

No competing interest is declared.

Author contributions statement

R.R.M.O. idealized the research, structured it, and wrote the manuscript. S.V. made important suggestions and reviewed the manuscript. B.P. structured the access to Computerome2, made important suggestions, and reviewed the manuscript. T.S.P. granted access to Computerome2, made important suggestions, and reviewed the manuscript. G.O. made important suggestions and reviewed the manuscript.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work received financial support from Vale S.A. (Biodiversity Genomics project, R100603.85.0X.CB01). We also thank Dr. Ilia Leitch, from the Royal Botanic Gardens, Kew, for providing the updated C-Value Plant database.

References

1. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet 2013; 14:157–167
2. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2007; 36:D13–D21
3. Li Z, Chen Y, Mu D, et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics 2012; 11:25–37
4. Claros MG, Bautista R, Guerrero-Fernández D, et al. Why Assembling Plant Genome Sequences Is So Challenging. Biology (Basel) 2012; 1:439
5. Meyers LA, Levin DA. ON THE ABUNDANCE OF POLYPLOIDS IN FLOWERING PLANTS. Evolution (N Y) 2006; 60:1198
6. Govaerts R. World Checklist of Vascular Plants (WCVP) – Version 10. 2022;
7. McCouch S. Agriculture: Feeding the future. Nature 2013; 499:23–24
8. Hulse-Kemp AM, Maheshwari S, Stoffel K, et al. Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. Hortic Res 2018; 5:4
9. Initiative TAG. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000; 408:796–815
10. Bräutigam A, Gowik U. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol 2010; 12:831–841
11. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456:53–59
12. Quail M, Smith ME, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 2012; 13:341
13. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 2015; 13:278–289
14. Jain M, Fiddes IT, Miga KH, et al. Improved data analysis for the MinION nanopore sequencer. Nat Methods 2015; 12:351–356
15. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 2019; 37:1155–1162
16. SUTTON GG, WHITE O, ADAMS MD, et al. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Sci Technol 1995; 1:9–19
17. MYERS EW. Toward Simplifying and Accurately Formulating Fragment Assembly. Journal of Computational Biology 1995; 2:275–290
18. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences 2001; 98:9748–9753
19. Zerbino D, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;
20. Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data. Genome Res 2009; 19:1117–23
21. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 2012; 19:455–477

22. Valliyodan B, Lee S-H, Nguyen HT. Sequencing, Assembly, and Annotation of the Soybean Genome. 2017; 73–82
23. Bento M, Gustafson JP, Viegas W, et al. Size matters in Triticeae polyploids: larger genomes have higher remodeling. Genome 2011; 54:175–183
24. Paterson AH, Freeling M, Tang H, et al. Insights from the Comparison of Plant Genome Sequences. Annu Rev Plant Biol 2010; 61:349–372
25. Schnable PS, Ware D, Fulton RS, et al. The B73 maize genome: Complexity, diversity, and dynamics. Science (1979) 2009; 326:1112–1115
26. Hoshino A, Matsunaga TM, Sakamoto T, et al. Hi-C Revolution: From a Snapshot of DNA–DNA Interaction in a Single Cell to Chromosome-Scale &lt;i&gt;De Novo&lt;/i&gt; Genome Assembly. Cytologia (Tokyo) 2017; 82:223–226
27. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 2012; 1:18
28. Warren RL, Sutton GG, Jones SJM, et al. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 2007; 23:500–501
29. Butler J, MacCallum I, Kleber M, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 2008; 18:810–20
30. Weisenfeld NI, Kumar V, Shah P, et al. Direct determination of diploid genome sequences. Genome Res 2017; 27:757–767
31. Zimin A v., Marçais G, Puiu D, et al. The MaSuRCA genome assembler. Bioinformatics 2013; 29:2669–2677
32. Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017; 27:722–736
33. Kamath GM, Shomorony I, Xia F, et al. HINGE: long-read assembly achieves optimal repeat resolution. Genome Res 2017; 27:747–756
34. Kolmogorov M, Yuan J, Lin Y, et al. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019; 37:540–546
35. Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 2017; 27:737–746
36. Kajitani R, Toshimoto K, Noguchi H, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 2014; 24:1384–95
37. Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 2016; 13:1050–1054
38. Hamilton JP, Robin Buell C. Advances in plant genome sequencing. The Plant Journal 2012; 70:177–190
39. Kyriakidou M, Tai HH, Anglin NL, et al. Current strategies of polyploid plant genome sequence assembly. Front Plant Sci 2018; 871:1660
40. Exposito-Alonso M, Drost H, Burbano HA, et al. The Earth BioGenome project: opportunities and challenges for plant genomics and conservation. The Plant Journal 2020; 102:222–229
41. Lewin HA, Richards S, Aiden EL, et al. The Earth BioGenome Project 2020: Starting the clock. Proc Natl Acad Sci U S A 2022; 119:e2115635118
42. . Entrez Direct: E-utilities on the UNIX Command Line - Entrez Programming Utilities Help - NCBI Bookshelf.
43. Heberle H, Meirelles VG, da Silva FR, et al. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 2015; 16:169
44. Oliveira RRM, Vasconcelos S, Oliveira G. SPLACE: A tool to automatically SPLit, Align, and ConcatenatE genes for phylogenomic inference of several organisms. Frontiers in Bioinformatics 2022; 2:109
45. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30:1312–1313
46. Rensing SA, Lang D, Zimmer AD, et al. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science (1979) 2008; 319:64–69
47. Stephens ZD, Hudson ME, Mainzer LS, et al. Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. PLoS One 2016; 11:e0167047
48. Gao S, Wang B, Xie S, et al. A high-quality reference genome of wild Cannabis sativa. Hortic Res 2020; 7:1–11
49. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020; 17:155–158
50. Liu H, Wu S, Li A, et al. SMARTdenovo: a de novo assembler using long noisy reads. GigaByte 2021; 2021:1–9
51. Chakraborty M, Baldwin-Brown JG, Long AD, et al. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res 2016; 44:e147
52. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018; 34:3094–3100
53. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014; 9:e112963
54. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–2079
55. Wang P, Luo Y, Huang J, et al. The genome evolution and domestication of tropical fruit mango. Genome Biol 2020; 21:1–17
56. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 2018; 19:460
57. Jia H-M, Jia H-J, Cai Q-L, et al. The red bayberry genome and genetic basis of sex determination. Plant Biotechnol J 2019; 17:397–409
58. Boetzer M, Henkel C v, Jansen HJ, et al. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2010; 27:578–579
59. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011; 27:863–864
60. Mikheenko A, Prjibelski A, Saveliev V, et al. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 2018; 34:i142–i150
61. Krzywinski M, Schein J, Birol I, et al. Circos: An information aesthetic for comparative genomics. Genome Res 2009; 19:1639–1645
62. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015; 31:3210–3212
63. Challis R, Kumar S, Sotero-Caio C, et al. Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life [version 1; peer review: 2 approved]. 2023;
64. Rice A, Glick L, Abadi S, et al. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. New Phytologist 2015; 206:19–26
65. Xiao C le, Chen Y, Xie SQ, et al. MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods 2017; 14:1072–1074
66. Jain R, Jenkins J, Shu S, et al. Genome sequence of the model rice variety KitaakeX. BMC Genomics 2019; 20:1–9
67. Tuskan GA, DiFazio S, Jansson S, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science (1979) 2006; 313:1596–1604
68. Velasco R, Zharkikh A, Affourtit J, et al. The genome of the domesticated apple (Malus × domestica Borkh.). Nature Genetics 2010 42:10 2010; 42:833–839
69. Yin D, Ji C, Ma X, et al. Genome of an allotetraploid wild peanut Arachis monticola: a de novo assembly. Gigascience 2018; 7:
70. Yang J, Wariss HM, Tao L, et al. De novo genome assembly of the endangered Acer yangbiense, a plant species with extremely small populations endemic to Yunnan Province, China. Gigascience 2019; 8:
71. Lovell JT, Bentley NB, Bhattarai G, et al. Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. Nature Communications 2021 12:1 2021; 12:1–12
72. Chen J, Huang Q, Gao D, et al. Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. Nature Communications 2013 4:1 2013; 4:1–9
73. Zhou Y, Chebotarov D, Kudrna D, et al. A platinum standard pan-genome resource that represents the population structure of Asian rice. Scientific Data 2020 7:1 2020; 7:1–11
74. Jones A, Torkel C, Stanley D, et al. High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. PLoS One 2021; 16:e0253830
75. Hon T, Mars K, Young G, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific Data 2020 7:1 2020; 7:1–11
76. Yekefenhazi D, He Q, Wang X, et al. Chromosome-level genome assembly of Nibea coibor using PacBio HiFi reads and Hi-C technologies. Scientific Data 2022 9:1 2022; 9:1–8
77. Field MA, Rosen BD, Dudchenko O, et al. Canfam_GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (Canis lupus familiaris) using a combination of long reads, optical mapping, and Hi-C. Gigascience 2020; 9:
78. Han B, Jing Y, Dai J, et al. A Chromosome-Level Genome Assembly of Dendrobium Huoshanense Using Long Reads and Hi-C Data. Genome Biol Evol 2020; 12:2486–2490
79. Correction Mayer K, Schüller C, Wambutt R, et al. Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature 1999; 402:769–777
80. Lin X, Kaul S, Rounsley S, et al. Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. Nature 1999; 402:761–765
81. Goff SA, Ricke D, Lan TH, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science (1979) 2002; 296:92–100
82. Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science (1979) 2002; 296:79–92
83. Whitelaw CA, Barbazuk WB, Pertea G, et al. Enrichment of Gene-Coding Sequences in Maize by Genome Filtration. Science (1979) 2003; 302:2118–2120
84. Matsumoto T, Wu J, Kanamori H, et al. The map-based sequence of the rice genome. Nature 2005; 436:793–800
85. Clark RM, Schweikert G, Toomajian C, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science (1979) 2007; 317:338–342
86. Ossowski S, Schneeberger K, Clark RM, et al. Sequencing of natural strains of

Arabidopsis thaliana with short reads. Genome Res 2008; 18:2024–2033

87. Huang S, Li R, Zhang Z, et al. The genome of the cucumber, Cucumis sativus L. Nat Genet 2009; 41:1275–1281

88. Shulaev V, Sargent DJ, Crowhurst RN, et al. The genome of woodland strawberry (Fragaria vesca). Nat Genet 2011; 43:109–116

89. Cao J, Schneeberger K, Ossowski S, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 2011; 43:956–965

90. Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature 2010; 463:178–183

91. Choulet F, Wicker T, Rustenholz C, et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 2010; 22:1686–1701

92. Kagale S, Koh C, Nixon J, et al. The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. Nat Commun 2014; 5:1–11

93. Chalhoub B, Denoeud F, Liu S, et al. Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. Science (1979) 2014; 345:950–953

94. Li F, Fan G, Lu C, et al. Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol 2015; 33:524–530

95. Lan T, Renner T, Ibarra-Laclette E, et al. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. Proc Natl Acad Sci U S A 2017; 114:E4435–E4441

96. Gnerre S, Maccallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011; 108:1513–8

97. Cheng H, Concepcion GT, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 2021 18:2 2021; 18:170–175

98. . Nextomics/NextDenovo: Fast and accurate de novo assembler for long reads.

99. Myers EW, Sutton GG, Delcher AL, et al. A whole-genome assembly of Drosophila. Science (1979) 2000; 287:2196–2204

100. . De Novo Genome Assembly - Dovetail Genomics.

101. Kolmogorov M, Yuan J, Lin Y, et al. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology 2019 37:5 2019; 37:540–546

102. . NRGene – Growing the future. Together. DeNovoMAGIC - NRGene - Growing the future. Together.

103. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005 437:7057 2005; 437:376–380

104. . Assembly Process -Software -De Novo Assembly -Official 10x Genomics Support.

105. Batzoglou S, Jaffe DB, Stanley K, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res 2002; 12:177–89

106. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 2016; 32:2103–2110

107. Chen Y, Nie F, Xie SQ, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nature Communications 2021 12:1 2021; 12:1–10

108. Ye C, Hill CM, Wu S, et al. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Scientific Reports 2016 6:1 2016; 6:1–9

109. Du H, Liang C. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nature Communications 2019 10:1 2019; 10:1–10

110. Shafin K, Pesout T, Lorig-Roach R, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nature Biotechnology 2020 38:9 2020; 38:1044–1053

111. . asarum/HABOT2: Hybrid genome assembly using 2nd and 3rd generation sequencing data.

112. Weisenfeld NI, Yin S, Sharpe T, et al. Comprehensive variation discovery in single human genomes. Nature Genetics 2014 46:12 2014; 46:1350–1355

113. Monat C, Padmarasu S, Lux T, et al. TRITEX: Chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 2019; 20:1–18

114. Ye C, Ma ZS, Cannon CH, et al. Exploiting sparseness in de novo genome assembly. BMC Bioinformatics 2012; 13 Suppl 6:1–8

Renato R.M. Oliveira received Bachelor's and Master's Degree in Computer Science (2017) from the Universidade Federal do Pará (UFPA). Currently, he is a Ph.D. Candidate at the Universidade Federal de Minas Gerais (UFMG), comparing and developing omics pipelines. He is also a Bioinformatician at the Instituto Tecnológico Vale, Belém, Pará, Brazil. As of April 2023, his 25 scientific contributions indexed in Scopus hold 126 citations from 113 different documents (h-index = 7).

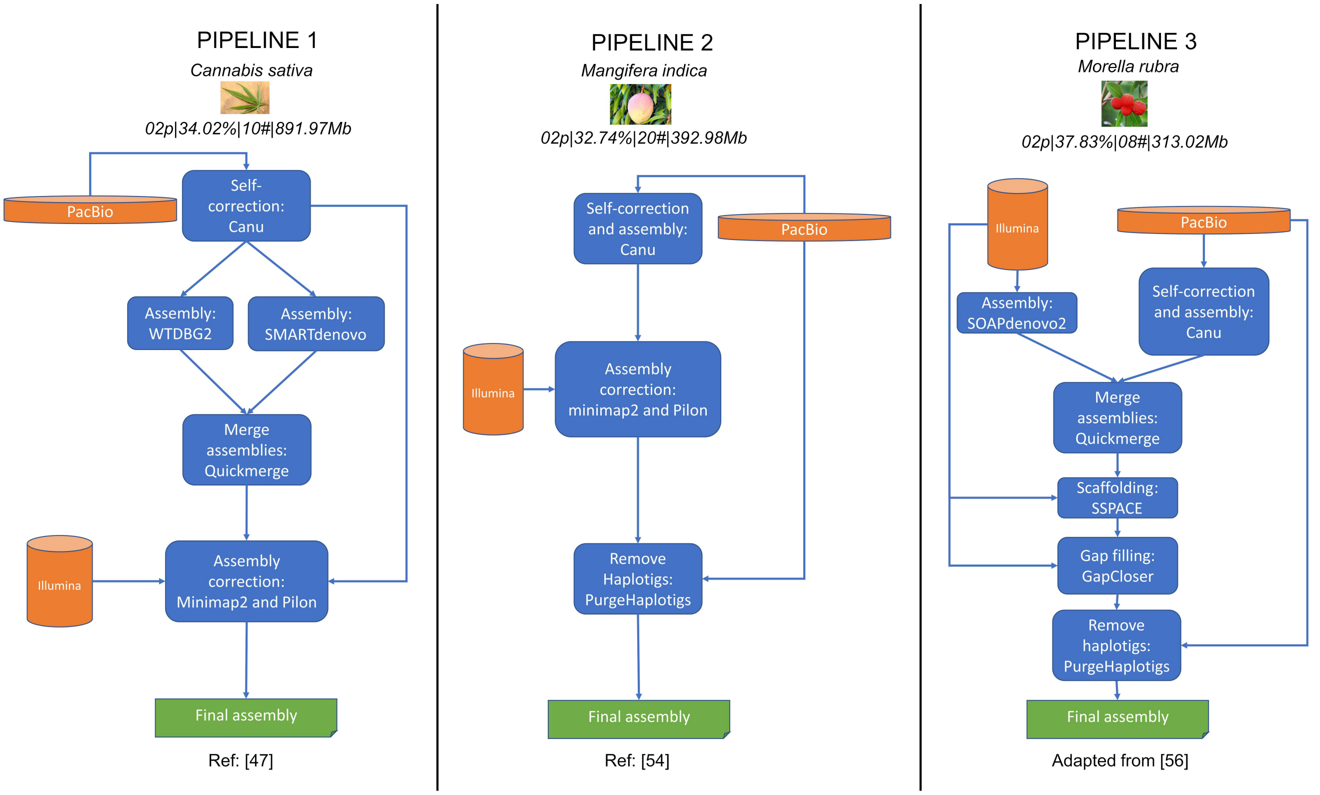
Santelmo Vasconcelos is a researcher in the Environmental Genomics Group at the Instituto Tecnológico Vale, with a Ph.D. in Plant Biology from the Universidade Federal de Pernambuco, currently conducting biodiversity and evolutionary studies focusing on the Amazon basin, and supervising graduate students in genomics and molecular systematics of plant and animal taxa. As of April 2023, his 39 scientific contributions indexed in Scopus hold 250 citations from 212 different documents (h-index = 10).

Gisele Nunes is a Bachelor of Science and a Ph.D. in Agricultural Microbiology. She is a bioinformatics researcher at the Vale Institute of Technology and works with integrative genomics focused on mining and sustainable development. Her research line includes environmental microbiology and plant and animal genomics. As of April 2023, her 18 scientific contributions indexed in Scopus hold 108 citations from 96 different documents (h-index = 7).

Bent Pettersen has a Ph.D. in Bioinformatics and has been working in the Bioinformatics field since 2007. He is currently a senior researcher in the Computational Biodiscovery Group at the University of Copenhagen and joint deputy director for the Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio) at AIMST University in Malaysia. His research includes precision farming, bacteriophage discovery, and large-scale genomics sequencing and assembly. As of April 2023, his 61 scientific contributions indexed in Scopus hold 4,771 citations from 4,277 different documents (h-index = 26).

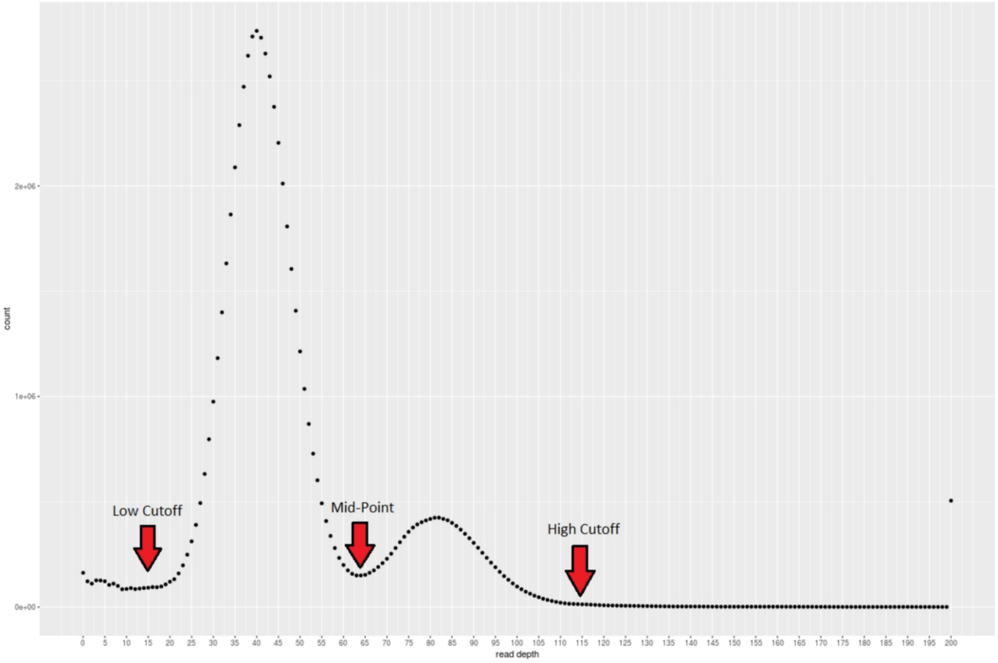
Thomas Sicheritz-Pontén holds a Ph.D. in Bioinformatics and has been actively involved in the field since 1996. He is currently a Professor of Computational Phage Biodiscovery at the University of Copenhagen and Joint Director for the Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio) at AIMST University in Malaysia. His research interests span metagenomics, comparative genomics, artificial intelligence, supercomputing, and phage discovery. As of April 2023, his 128 scientific contributions indexed in Scopus hold 31,863 citations from 27,070 different documents (h-index = 53).

Guilherme Oliveira obtained his Ph.D. from Texas A&M University and, for 13 years, was a researcher at FIOCRUZ. He is currently the head of the Vale Institute of Technology and conducts research in the environmental genomics field. As of April 2023, his 207 scientific contributions indexed in Scopus hold 6,015 citations from 4,515 different documents (h-index = 35).



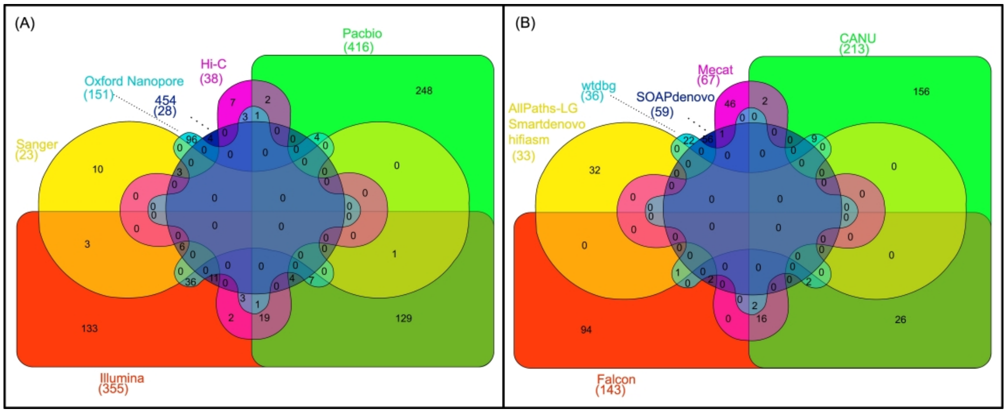
Published pipelines from the assembly of complete nuclear genomes of Cannabis sativa, Mangifera indica, and Morella rubra. Below each plant picture are shown the ploidy, GC content, number of chromosomes, and genome size.

530x320mm (300 x 300 DPI)



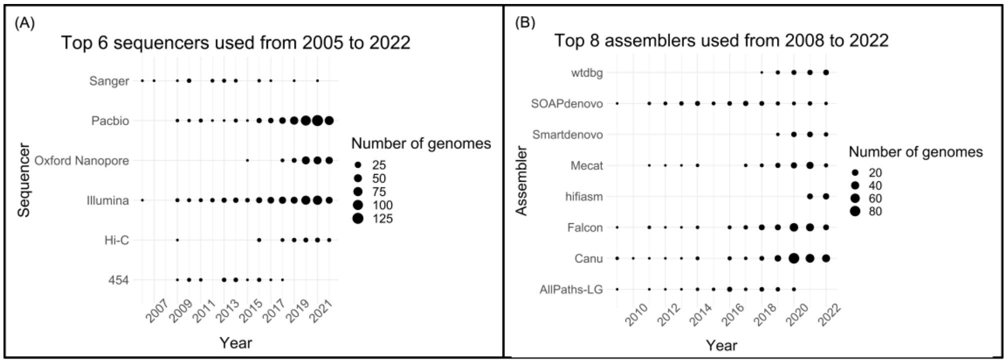
Example of a histogram generated by Purge Haplotigs, showing choice points for the low, medium, and high coverage limits. Source: https://bitbucket.org/mroachawri/purge_haplotigs

317x211mm (300 x 300 DPI)



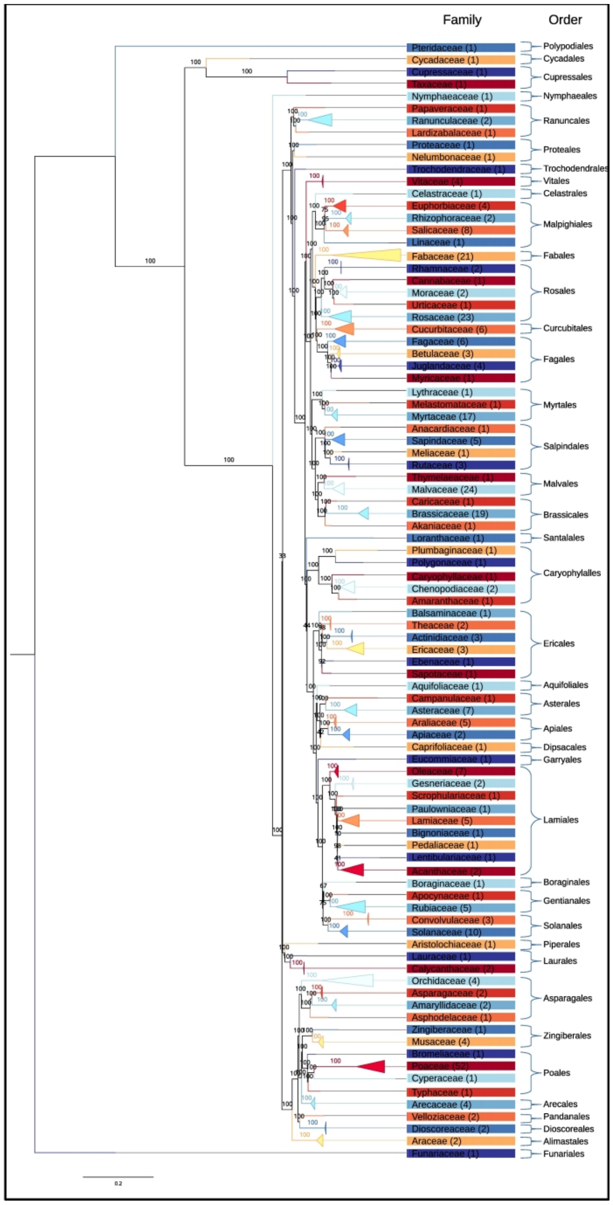
Venn Diagrams showing the top six sequencers (A) and the top eight assembly software (B) used in 856 records of plant nuclear genomes deposited on NCBI.

837x339mm (300 x 300 DPI)



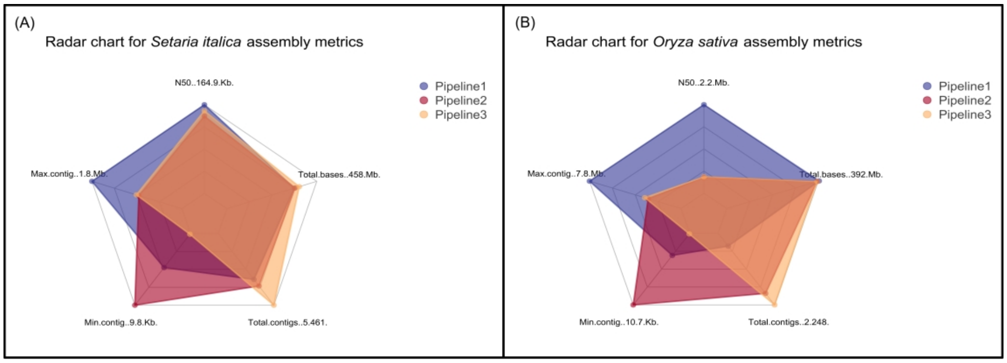
Timeline visualizations for (A) the top 6 sequencers and for (B) the top 6 assemblers software used on complete nuclear plant genomes deposited on NCBI across the years.

837x297mm (300 x 300 DPI)



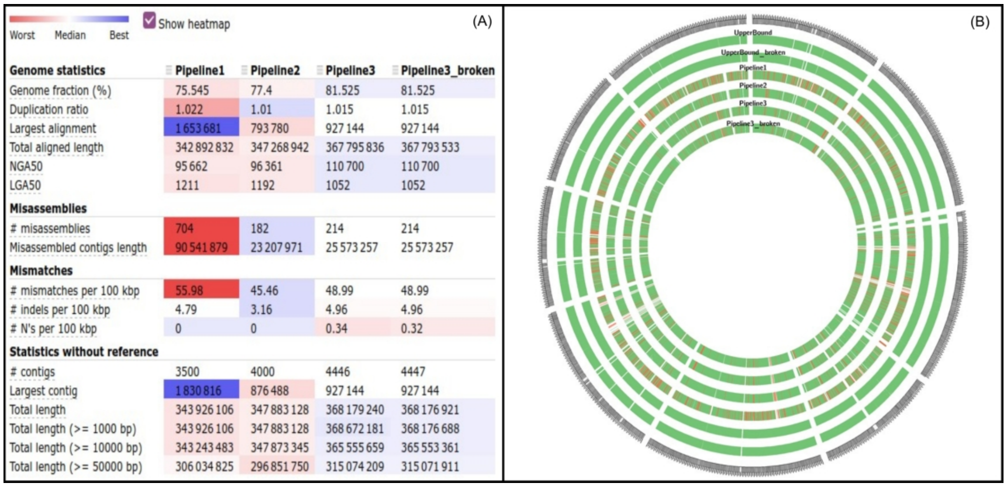
Phylogenomic tree with 349 plant species with a complete genome deposited on NCBI. In different colors are represented 91 families, with the number of species in parentheses. The orders are alongside the braces.

292x578mm (300 x 300 DPI)



Radar charts for the assembly metrics of (A) Setaria italica and (B) Oryza sativa.

837x297mm (300 x 300 DPI)

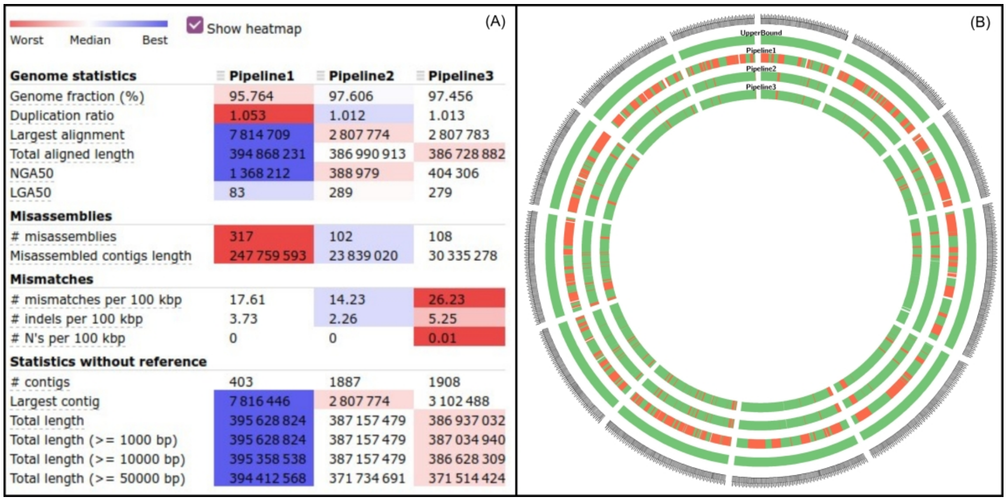


Worst Median Best ☑ Show heatmap

Genome statistics	Pipeline1	Pipeline2	Pipeline3	Pipeline3_broken
Genome fraction (%)	75.545	77.4	81.525	81.525
Duplication ratio	1.022	1.01	1.015	1.015
Largest alignment	1 653 681	793 780	927 144	927 144
Total aligned length	342 892 832	347 268 942	367 795 836	367 793 533
NGA50	95 662	96 361	110 700	110 700
LGA50	1211	1192	1052	1052
Misassemblies				
# misassemblies	704	182	214	214
Misassembled contigs length	90 541 879	23 207 971	25 573 257	25 573 257
Mismatches				
# mismatches per 100 kbp	55.98	45.46	48.99	48.99
# indels per 100 kbp	4.79	3.16	4.96	4.96
# N's per 100 kbp	0	0	0.34	0.32
Statistics without reference				
# contigs	3500	4000	4446	4447
Largest contig	1 830 816	876 488	927 144	927 144
Total length	343 926 106	347 883 128	368 179 240	368 176 921
Total length (>= 1000 bp)	343 926 106	347 883 128	368 672 181	368 176 688
Total length (>= 10000 bp)	343 243 483	347 873 345	365 555 659	365 553 361
Total length (>= 50000 bp)	306 034 825	296 851 750	315 074 209	315 071 911

(A) Assembly metrics found by QUAST for the Setaria italica results. In red are the worst values and in blue are the best metric values. (B) The alignment of the assemblies generated by the pipelines along with the mapping of the reads against the Setaria italica reference genome (UpperBound). Description with "broken" indicates that the alignment was made with contigs; without "broken" indicates that it was made with scaffolds.

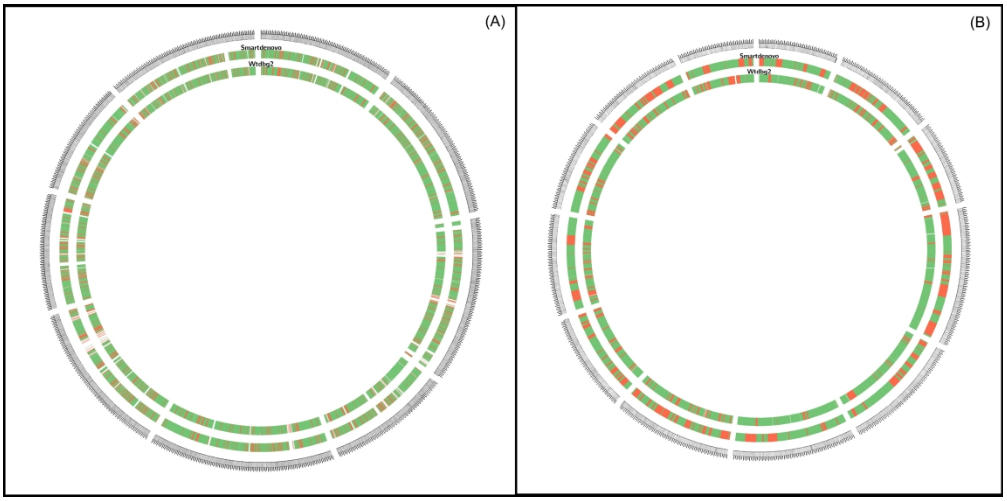
279x134mm (300 x 300 DPI)



Genome statistics	Pipeline1	Pipeline2	Pipeline3
Genome fraction (%)	95.764	97.606	97.456
Duplication ratio	1.053	1.012	1.013
Largest alignment	7 814 709	2 807 774	2 807 783
Total aligned length	394 868 231	386 990 913	386 728 882
NGA50	1 368 212	388 979	404 306
LGA50	83	289	279
Misassemblies			
# misassemblies	317	102	108
Misassembled contigs length	247 759 593	23 839 020	30 335 278
Mismatches			
# mismatches per 100 kbp	17.61	14.23	26.23
# indels per 100 kbp	3.73	2.26	5.25
# N's per 100 kbp	0	0	0.01
Statistics without reference			
# contigs	403	1887	1908
Largest contig	7 816 446	2 807 774	3 102 488
Total length	395 628 824	387 157 479	386 937 032
Total length (>= 1000 bp)	395 628 824	387 157 479	387 034 940
Total length (>= 10000 bp)	395 358 538	387 157 479	386 628 309
Total length (>= 50000 bp)	394 412 568	371 734 691	371 514 424

Assembly metrics found by QUAST for the Oryza sativa results. In red are the worst values and in blue are the best metric values. (B) Alignment of the assemblies generated by the pipelines along with the mapping of the reads against the Oryza sativa reference genome (UpperBound).

267x131mm (300 x 300 DPI)



Circos plot for the alignment of the assemblies generated by Smartdenovo and Wtdbg2 in Pipeline 1 for the Setaria italica (A) and Oryza sativa (B) assemblies. Red dashes indicate the assembly errors found.

267x132mm (300 x 300 DPI)

Table 1. Significant milestones in plant genome assembly.

Year	Species	Platform	Assembly size	Refs.
1999	Chromosomes 2 and 4 from *Arabidposis thaliana*	Sanger	19.6 and 17.4 Mb	[79,80]
2000	*A. thaliana*	Sanger	115.4 Mb	[9]
2002	*Oryza sativa* L.	Sanger	390.Mb	[81,82]
2003	*Zea mays* L.; Draft	Sanger	132 Mb	[83]
2005	*O. sativa*	Sanger	370.7 Mb	[84]
2007	20 genomes of *A. thaliana*	Perlegen	20×119 Mb (2.4 Gb)	[85]
2008	Three genomes of *A. thaliana*	Illumina	3×119Mb (357Mb)	[86]
2009	*Z. mays*	Sanger	2.3 Gb	[25]
2009	*Cucumis sativus* L.	Sanger and Illumina	243.5 Mb	[87]
2011	*Fragaria vesca* L.	454; Illumina; SoLiD	220 Mb	[88]
2011	80 genomes of A. thaliana	Illumina	80 ×119 Mb (9.5 Gb)	[89]

Source: Adapted from [38]

Table 2. Major milestones in plant genome assembly.

Species	Genome size	Ploidy	References
Glycine max (L.) Merr.	979 Mb	Tetraploid	[90]
Triticum aestivum L.	15.34 Gb	Hexaploid	[91]
Camelina sativa (L.) Crantz	641 Mb	Hexaploid	[92]
Brassica napus L.	976 MB	Tetraploid	[93]
Gossypium hirsutum L.	2.18 GB	Tetraploid	[94]
Utricularia gibba L.	101 MB	16-ploid	[95]

Source: Adapted from [39]

Table 3. List of assembly software used in the 856 complete plant nuclear genomes records.

Assembler	Total count of usage	Read type	Assembly method	Refs
CANU	213	Long	Hybrid	[32]
Falcon	143	Long	String-graph	[37]
MECAT	67	Long	Hybrid	[65]
SOAPdenovo	59	Short	De Bruijn	[27]
wtdbg	36	Long	Hybrid	[49]
AllPaths-LG	33	Short	De Bruijn	[96]
Smartdenovo	33	Long	OLC	[50]
Hifiasm	33	Long	OLC	[97]
NextDenovo	23	Long	String-graph	[98]
Celera	22	Short	OLC	[99]
HiRise	19	Long	String-graph	[100]
Masurca	19	Hybrid	Hybrid	[31]
Flye	18	Long	De Bruijn	[101]
DenovoMAGIC	17	Hybrid	Hybrid	[102]
Newbler	17	Short	OLC	[103]
Abyss	17	Short	De Bruijn	[20]
Supernova	13	Hybrid	Hybrid	[104]
Arachne	12	Short	Hybrid	[105]
Racon	11	Long	OLC	[35]
Platanus	10	Short	De Bruijn	[36]
miniasm	9	Long	OLC	[106]
SPAdes	7	Short	De Bruijn	[21]
necat	6	Long	String-graph	[107]
DBG2OLC	6	Long	Hybrid	[108]
Hera	3	Long	OLC	[109]
Shasta	2	Long	Overlap-based	[110]
HABOT	2	Hybrid	Hybrid	[111]
Discovar	2	Short	Overlap-based	[112]
Tritex	1	Hybrid	De Bruijn	[113]
SparseAssembler	1	Short	Hybrid	[114]

Table 4. Results of the sequencing simulation for Illumina and Pacbio platforms.

Species	Platform	Read pairs	Total bases	Cov	Running time (h)
Setaria italica (443 Mb)	Illumina	145,923,953	44,069,033,806	99.48	6.34
	Pacbio	706,772	10,601,580,000	23.93	63.43
Oryza sativa (379 MB)	Illumina	128,349,444	38,761,532,088	102.27	5.56
	Pacbio	772,219	11,583,285,000	30.56	55.35

Table 5. BUSCO results for Setaria italica and Oryza sativa in each pipeline. The percentage represents the total ortholog genes found (n=4896), considering the Poales_odb10 database. In bold are described the best values.

Species	***Setaria italica***					***Oryza sativa***				
Busco metrics	***C %***	***S %***	***D %***	***F %***	***M %***	***C %***	***S %***	***D %***	***F %***	***M %***
Pipeline 1	87.8	84.7	3.1	1.0	11.2	95.6	88.3	7.3	0.5	3.9
Pipeline 2	88.7	86.1	**2.6**	1.0	10.3	**97.6**	**95.0**	2.6	0.6	**1.8**
Pipeline 3	**95.2**	**92.3**	2.9	**0.7**	**4.1**	97.2	94.7	**2.5**	**0.5**	2.3

C = Complete; S = Complete and single-copy; D = Complete and duplicated; F = Fragmented; M = Missing.

Organism/Name	Ploidy	Assembly Accession	chromosom	Size (Mb)
Acer yangbiense	2	GCA_008009225.1	13	665.888
Actinidia chinensis	2	GCA_009663005.1	29	653.926
Actinidia chinensis var. chinensis	2	GCA_003024255.1	29	553.842
Actinidia eriantha	2	GCA_004150315.1	29	690.611
Actinidia eriantha	2	GCA_019202715.1	29	657,097
Aegilops bicornis	2	GCA_021605145.1	7	5902.72
Aegilops longissima	2	GCA_904067125.1	7	6689.48
Aegilops longissima	2	GCA_021605205.1	7	5796.09
Aegilops searsii	2	GCA_021605185.1	7	5336.42
Aegilops sharonensis	2	GCA_021641835.1	7	5892.84
Aegilops speltoides	2	GCA_021437245.1	7	4110.19
Aegilops speltoides subsp. speltoid	2	GCA_944222845.1	7	5116.92
Aegilops tauschii	2	GCA_002105435.1	7	247,197
Aegilops tauschii	2	GCA_000347335.2	7	4310.350
Aegilops tauschii subsp. strangulat	2	GCA_002575655.1	7	4224.920
Aegilops tauschii subsp. strangulat	2	GCA_002575655.2	7	4218.18
Aeschynomene evenia	2	GCA_013621005.1	10	375.94
Akebia trifoliata	2	GCA_017979445.1	16	652,797
Allium cepa	2	GCA_905187595.1	8	14937.4
Allium sativum	2	GCA_014155895.2	9	16243.2
Alloteropsis semialata	2	GCA_004135705.1	9	747.772
Amaranthus cruentus	2	GCA_019425755.1	17	370,914
Amaranthus hypochondriacus	2	GCA_000753965.2	16	417.46
Amorphophallus konjac	2	GCA_022559845.1	13	5598.56
Amphicarpaea edgeworthii	2	GCA_014843725.1	11	299,059
Andrographis paniculata	2	GCA_004354405.1	24	269.408
Apium graveolens	2	GCA_009905375.1	11	3332.58
Arabidopsis thaliana	2	GCA_000211275.1	5	93.655
Arabidopsis thaliana	2	GCA_000001735.2	5	119.669
Arabidopsis thaliana	2	GCA_001651475.1	5	118.891
Arabidopsis thaliana	2	GCA_900660825.1	5	119.627
Arabidopsis thaliana	2	GCA_902460285.1	5	120.338
Arabidopsis thaliana	2	GCA_902460305.1	5	122.202
Arabidopsis thaliana	2	GCA_902460275.1	5	119.75
Arabidopsis thaliana	2	GCA_902460295.1	5	120.29
Arabidopsis thaliana	2	GCA_902460315.1	5	120.795
Arabidopsis thaliana	2	GCA_902460265.3	5	120.13
Arabidopsis thaliana	2	GCA_902825305.1	5	120
Arabidopsis thaliana	2	GCA_903064325.1	5	120,693
Arabidopsis thaliana	2	GCA_903064285.1	5	118,889
Arabidopsis thaliana	2	GCA_903064295.1	5	119,797
Arabidopsis thaliana	2	GCA_903064275.1	5	117,836
Arabidopsis thaliana	2	GCA_904420315.1	5	130,173
Arabidopsis thaliana	2	GCA_020911765.1	5	133.266
Arabidopsis thaliana	2	GCA_933208065.1	5	138.836
Arabidopsis thaliana	2	GCA_023115395.1	5	132.081
Arabidopsis thaliana	2	GCA_024498475.1	5	123.948
Arabidopsis thaliana	2	GCA_024498435.1	5	127.151
Arabidopsis thaliana	2	GCA_024498455.1	5	126.09
Arabidopsis thaliana	2	GCA_024498555.1	5	124.618
Arabidopsis thaliana	2	GCA_024498495.1	5	122.66
Arabidopsis thaliana	2	GCA_024498515.1	5	120.087
Arabis alpina	2	GCA_000733195.1	8	308.033
Arabis alpina	2	GCA_900128785.1	8	311.642
Arabis montbretiana	2	GCA_001484125.2	8	257,692

Arachis cardenasii	2	GCA_018493915.1	10	1238.08
Arachis duranensis	2	GCA_000817695.2	10	1084.260
Arachis duranensis	2	GCA_014805325.1	10	1106.33
Arachis duranensis	2	GCA_018207795.1	10	1099.87
Arachis ipaensis	2	GCA_000816755.2	10	1353.500
Arachis ipaensis	2	GCA_013265535.1	10	1438.65
Arachis stenosperma	2	GCA_014773155.1	10	1328.99
Arctium lappa	2	GCA_023525745.1	18	1727.36
Asparagus officinalis	2	GCA_001876935.1	10	1187.540
Asparagus setaceus	2	GCA_012295165.1	10	735.53
Avena longiglumis	2	GCA_023614385.1	7	3736.64
Avicennia marina	2	GCA_019155195.1	31	457,335
Avicennia marina subsp. marina	2	GCA_013168755.1	32	457
Bauhinia variegata	2	GCA_022379115.2	14	326.365
Bauhinia variegata	2	GCA_022379115.1		326,359
Benincasa hispida	2	GCA_009727055.1	12	912.951
Beta vulgaris subsp. vulgaris	2	GCA_000510975.1	9	568.609
Beta vulgaris subsp. vulgaris	2	GCA_000511025.2	9	566.55
Beta vulgaris subsp. vulgaris	2	GCA_002917755.1	9	540.534
Boechera stricta	2	GCA_018361395.1	7	190,536
Boechera stricta	2	GCA_018361405.1	7	189.46
Brachypodium distachyon	2	GCA_000005505.4	5	271.299
Brassica nigra	2	GCA_001682895.1	8	402.145
Brassica nigra	2	GCA_016432835.1	8	534,239
Brassica oleracea	2	GCA_900416815.2	9	554.977
Brassica oleracea var. oleracea	2	GCA_000695525.1	9	488.954
Brassica rapa	2	GCA_000309985.3	10	353
Brassica rapa	2	GCA_900412535.2	10	401.927
Brassica rapa	2	GCA_900412535.3	10	443,954
Brassica rapa	2	GCA_003434825.1	10	314.865
Brassica rapa	2	GCA_016163755.1	10	370,897
Brassica rapa subsp. pekinensis	2	GCA_008629595.1	10	234.688
Bruguiera parviflora	2	GCA_019804595.1	18	210,391
Buddleja alternifolia	2	GCA_019426215.1	19	853,755
Cajanus cajan	2	GCA_000340665.1	11	592.971
Cajanus cajan	2	GCA_000340665.2	11	590.524
Camelina hispida	2	GCA_023657505.1	7	283.324
Camelina hispida	2	GCA_023864115.1	7	339.775
Camelina laxa	2	GCA_024034495.1	6	213.074
Camelina neglecta	2	GCA_023864065.1	6	210.446
Camellia oleifera	2	GCA_022316695.1	15	2889.51
Camellia sinensis	2	GCA_013676235.1	15	3113.46
Camellia sinensis	2	GCA_020536595.1	15	3062.75
Camellia sinensis	2	GCA_020536515.1	15	3062.86
Camellia sinensis var. assamica	2	GCA_020536865.1	15	3062.77
Camellia sinensis var. assamica	2	GCA_020536795.1	15	3062.62
Camellia sinensis var. assamica	2	GCA_020536855.1	15	3062.8
Camellia sinensis var. assamica	2	GCA_020536565.1	15	3062.8
Camellia sinensis var. lasiocalyx	2	GCA_020536555.1	15	3062.77
Camellia sinensis var. sinensis	2	GCA_017311205.1	15	3062.88
Camellia sinensis var. sinensis	2	GCA_020536495.1	15	3062.74
Cannabis sativa	2	GCA_000230575.5	10	891.965
Cannabis sativa	2	GCA_003417725.2	10	1009.670
Cannabis sativa	2	GCA_900626175.2	10	876
Cannabis sativa	2	GCA_013030365.1	10	813
Cannabis sativa	2	GCA_016165845.1	10	914,397

Capsicum annuum	2	GCA_000512255.2	12	3063.860
Capsicum annuum	2	GCA_000710875.1	12	2935.880
Capsicum annuum	2	GCA_002878395.2	12	3212.120
Capsicum annuum	2	GCA_002878395.3	12	3211.82
Capsicum annuum	2	GCA_021292125.1	12	3077.74
Capsicum baccatum	2	GCA_002271885.2	12	3215.610
Capsicum chinense	2	GCA_002271895.2	12	3070.910
Carex littledalei	2	GCA_011114355.1	29	373.853
Carica papaya	2	GCA_001310045.1	9	807,745
Carica papaya	2	GCA_021527605.1	10	350,833
Carica papaya	2	GCA_022788785.1	10	349.878
Carya illinoinensis	2	GCA_016808215.1	16	649,961
Carya illinoinensis	2	GCA_018687715.1	16	674,433
Carya illinoinensis	2	GCA_018688675.1	16	668,996
Carya illinoinensis	2	GCA_018689175.1	16	656,687
Catharanthus roseus	2	GCA_024505715.1	8	572.922
Cenchrus americanus	2	GCA_002174835.2	7	1816.950
Cenchrus americanus	2	GCA_020739585.1	7	1950.23
Cenchrus americanus	2	GCA_020739525.1	7	1891.08
Cenchrus americanus	2	GCA_020739575.1	7	1937.98
Cenchrus americanus	2	GCA_020739535.1	7	1973.74
Cenchrus americanus	2	GCA_020739565.1	7	1911.91
Cenchrus americanus	2	GCA_021560375.1	7	1908.26
Centella asiatica	2	GCA_014636745.1	9	430,217
Ceratopteris richardii	2	GCA_020310875.1	39	7462.46
Chloranthus sessilifolius	2	GCA_021018995.1	15	2168.75
Chrysanthemum lavandulifolium	2	GCA_022545495.1	9	2670.47
Cicer arietinum	2	GCA_000331145.1	8	530.894
Cicer arietinum	2	GCA_000347275.4	8	511.684
Cicer arietinum	2	GCA_006151565.1	8	347.247
Cicer arietinum	2	GCA_006345785.1	8	347.247
Cicer reticulatum	2	GCA_003689015.2	8	416.904
Citrullus lanatus	2	GCA_000238415.2	11	365.45
Citrullus lanatus	2	GCA_004801215.2	11	397.83
Citrullus lanatus subsp. cordophan	2	GCA_018142915.1	11	367,122
Citrus maxima	2	GCA_002006925.1	9	345.757
Citrus sinensis	2	GCA_000317415.1	9	327.83
Citrus sinensis	2	GCA_018105775.1	9	334,323
Citrus sinensis	2	GCA_019144155.1	9	346,492
Citrus sinensis	2	GCA_019144195.1	9	315,067
Citrus sinensis	2	GCA_019143665.1	9	310,567
Citrus sinensis	2	GCA_019144185.1	9	322.59
Citrus sinensis	2	GCA_019144245.1	9	328,733
Citrus sinensis	2	GCA_019144225.1	9	330,225
Citrus sinensis	2	GCA_022201065.1	9	299,603
Citrus sinensis	2	GCA_022201045.1	9	298,978
Citrus trifoliata	2	GCA_018350135.1	9	303,067
Cocos nucifera	2	GCA_008124465.1	16	2202.46
Coffea canephora	2	GCA_900059795.1	11	568.612
Coffea eugenioides	2	GCA_003713205.1	11	1094.450
Coffea humblotiana	2	GCA_023065735.1	11	420.722
Coix aquatica	2	GCA_009725075.1	10	1615.47
Coix lacryma-jobi var. lacryma-jobi	2	GCA_009763385.1	10	1731.46
Coptis chinensis	2	GCA_015680905.1	9	935.66
Corylus avellana	2	GCA_901000735.2	11	369,779
Corylus heterophylla	2	GCA_016403345.1	11	370,751

Cucumis melo	2	GCA_011762645.1	12	378
Cucumis melo	2	GCA_020920065.1	12	365,223
Cucumis melo	2	GCA_020920055.1	12	367,304
Cucumis melo subsp. agrestis	2	GCA_014525375.1	12	366,171
Cucumis melo var. inodorus	2	GCA_009760825.1	12	386.497
Cucumis sativus	2	GCA_000004075.3	7	226.641
Cucumis sativus	2	GCA_016161885.1	7	238,422
Cucumis sativus	2	GCA_016161765.1	7	243,737
Cucumis sativus	2	GCA_016163705.1	7	239,412
Cucumis sativus	2	GCA_016163745.1	7	240,112
Cucumis sativus	2	GCA_016161775.1	7	234,565
Cucumis sativus	2	GCA_016161785.1	7	251,094
Cucumis sativus	2	GCA_016161805.1	7	232,305
Cucumis sativus	2	GCA_016161715.1	7	241,867
Cucumis sativus	2	GCA_016163735.1	7	247,122
Cucumis sativus	2	GCA_016161855.1	7	237,397
Cucumis sativus	2	GCA_016161875.1	7	242,878
Cucurbita argyrosperma subsp. arg	2	GCA_004115005.2	20	228,921
Cucurbita pepo subsp. pepo	2	GCA_002806865.2	20	261.355
Daucus carota subsp. sativus	2	GCA_001625215.1	9	421.539
Dendrobium chrysotoxum	2	GCA_019925795.1	19	1368.17
Dendrobium huoshanense	2	GCA_016618105.1	19	1284.29
Dendrobium nobile	2	GCA_022539455.1	19	1199.12
Dendrobium officinale	2	GCA_019514585.1	19	1228.67
Digitaria exilis	2	GCA_902806635.1	19	716
Digitaria exilis	2	GCA_902859565.1	19	716,471
Dimocarpus longan	2	GCA_020457875.1	15	483,436
Dimocarpus longan	2	GCA_022984855.1	15	454.299
Diospyros lotus	2	GCA_014633355.1	15	617,726
Diospyros lotus	2	GCA_014633365.1	15	630,099
Elaeis guineensis	2	GCA_000442705.1	16	1535.180
Elaeis guineensis	2	GCA_015461965.1	16	1209.42
Ensete glaucum	2	GCA_021527575.1	9	494,938
Erigeron canadensis	2	GCA_010389155.1	9	426.383
Erysimum cheiranthoides	2	GCA_011420285.1	8	177.181
Eucalyptus albens	2	GCA_014182695.1	11	607,093
Eucalyptus brandiana	2	GCA_014182725.1	11	507,241
Eucalyptus caleyi	2	GCA_014182885.2	11	589,484
Eucalyptus camaldulensis	2	GCA_014182705.1	11	558,609
Eucalyptus cladocalyx	2	GCA_017140615.1	11	544,245
Eucalyptus cloeziana	2	GCA_014182715.1	11	480,235
Eucalyptus coolabah	2	GCA_014182585.1	11	606,466
Eucalyptus curtisii	2	GCA_017140595.1	11	435,421
Eucalyptus dawsonii	2	GCA_016097615.1	11	707,061
Eucalyptus decipiens	2	GCA_014182575.1	11	591,117
Eucalyptus erythrocorys	2	GCA_014182555.1	11	539,357
Eucalyptus fibrosa	2	GCA_017140475.1	11	590,072
Eucalyptus globulus	2	GCA_014182545.1	11	545,185
Eucalyptus grandis	2	GCA_016545825.1	11	616.53
Eucalyptus guilfoylei	2	GCA_016097605.1	11	472,517
Eucalyptus lansdowneana	2	GCA_017140395.1	11	633,712
Eucalyptus leucophloia subsp. euro	2	GCA_017140325.1	11	568,636
Eucalyptus marginata	2	GCA_014182565.1	11	513,053
Eucalyptus microcorys	2	GCA_014182515.1	11	441,067
Eucalyptus paniculata subsp. matu	2	GCA_017140255.1	11	589.01
Eucalyptus polyanthemos subsp. p	2	GCA_017140185.1	11	603.44

Eucalyptus pumila	2	GCA_016097595.1	11	529,916
Eucalyptus regnans	2	GCA_014182855.1	11	495,129
Eucalyptus salubris	2	GCA_014182395.1	11	508,094
Eucalyptus shirleyi	2	GCA_017140165.1	11	597.34
Eucalyptus sideroxylon	2	GCA_014182405.1	11	592,318
Eucalyptus sideroxylon x Eucalyptu	2	GCA_016097485.1	11	603,738
Eucalyptus tenuipes	2	GCA_014182365.1	11	397,939
Eucalyptus victrix	2	GCA_016097545.1	11	557,319
Eucalyptus viminalis	2	GCA_014182385.1	11	558,867
Eucalyptus virginea	2	GCA_014182375.1	11	532,948
Eucommia ulmoides	2	GCA_016647705.1	17	947.85
Eutrema salsugineum	2	GCA_000325905.2	7	231.893
Eutrema salsugineum	2	GCA_016617915.1	7	295,494
Fagopyrum tataricum	2	GCA_002319775.1	8	505.883
Fagus sylvatica	2	GCA_907173295.1	12	540,344
Fragaria iinumae	2	GCA_009720345.1	7	240.582
Fragaria nilgerrensis	2	GCA_010134655.1	7	772.253
Fragaria vesca subsp. vesca	2	GCA_000184155.1	7	214.373
Fraxinus excelsior	2	GCA_019097785.1	23	807,608
Fraxinus pennsylvanica	2	GCA_912172775.1	23	756,791
Gardenia jasminoides	2	GCA_013103745.1	11	536
Gillenia trifoliata	2	GCA_018257905.1	9	296,281
Glycine latifolia	2	GCA_013407115.1	20	939,492
Glycine max	2	GCA_000004515.4	20	979.046
Glycine max	2	GCA_002905335.2	20	985.26
Glycine max	2	GCA_003349995.2	20	1011.38
Glycine max	2	GCA_012273815.1	20	1116.18
Glycine max	2	GCA_012273815.2	20	1000.03
Glycine max	2	GCA_014282085.1	20	988.84
Glycine max	2	GCA_014282095.1	20	995,708
Glycine max	2	GCA_014282065.1	20	987.26
Glycine max	2	GCA_014282185.1	20	993,002
Glycine max	2	GCA_014282035.1	20	996.72
Glycine max	2	GCA_014282075.1	20	1001.33
Glycine max	2	GCA_014282145.1	20	985,988
Glycine max	2	GCA_015227745.1	20	1020.98
Glycine max	2	GCA_019321705.1	20	992,121
Glycine max	2	GCA_020497155.1	20	933,123
Glycine max	2	GCA_021733175.1	20	995,269
Glycine max	2	GCA_022114995.1	20	1011.4
Glycine max	2	GCA_000004515.5		978,942
Glycine soja	2	GCA_002907465.1	20	1016.28
Glycine soja	2	GCA_004193775.2	20	1013.770
Glycine soja	2	GCA_014282345.1	20	975,919
Gossypioides kirkii	2	GCA_002818315.1	12	528.715
Gossypioides kirkii	2	GCA_005610355.1	12	538.061
Gossypium anomalum	2	GCA_019455425.1	13	1193.34
Gossypium arboreum	2	GCA_013265605.1	13	94,637
Gossypium arboreum	2	GCA_000612285.2	13	1694.600
Gossypium aridum	2	GCA_013487665.1	13	739,119
Gossypium armourianum	2	GCA_013677265.1	13	780,951
Gossypium australe	2	GCA_005393395.2	13	1743.39
Gossypium davidsonii	2	GCA_013677245.1	13	704,224
Gossypium gossypioides	2	GCA_013467495.1	13	664,724
Gossypium harknessii	2	GCA_013677255.1	13	732,155
Gossypium klotzschianum	2	GCA_013677235.1	13	670,958

Gossypium laxum	2	GCA_013511315.1	13	833,895
Gossypium lobatum	2	GCA_013467485.1	13	744,535
Gossypium longicalyx	2	GCA_010883175.1	13	1190.21
Gossypium raimondii	2	GCA_000327365.1	13	761.565
Gossypium raimondii	2	GCA_000327365.2	13	761.252
Gossypium raimondii	2	GCA_005931075.1	13	734.884
Gossypium raimondii	2	GCA_013467475.1	13	615,033
Gossypium schwendimanii	2	GCA_013677275.1	13	729,429
Gossypium stocksii	2	GCA_020496765.1	13	1448.11
Gossypium thurberi	2	GCA_004027125.1	13	582.007
Gossypium trilobum	2	GCA_013467465.1	13	655,377
Gossypium turneri	2	GCA_008044935.1	13	755.203
Gynostemma pentaphyllum	2	GCA_020536105.1	11	582,948
Helianthus annuus	2	GCA_002127325.1	17	3027.840
Helianthus annuus	2	GCA_002127325.2	17	3010.05
Hemerocallis citrina	2	GCA_017893485.1	11	3775.58
Hevea brasiliensis	2	GCA_010458925.1	18	1473.45
Hibiscus mutabilis	2	GCA_019671005.1	46	2675.93
Hordeum vulgare	2	GCA_017309745.1	7	4123.26
Hordeum vulgare	2	GCA_916098225.1	7	4051.28
Hordeum vulgare	2	GCA_024137805.1	7	5111
Hordeum vulgare subsp. spontaneu	2	GCA_907165085.1	7	4498.6
Hordeum vulgare subsp. vulgare	2	GCA_902375235.1	7	4830.08
Hordeum vulgare subsp. vulgare	2	GCA_902498975.1	7	4340.66
Hordeum vulgare subsp. vulgare	2	GCA_903813605.1	7	4257.71
Hordeum vulgare subsp. vulgare	2	GCA_903970725.1	7	4837.64
Hordeum vulgare subsp. vulgare	2	GCA_903970685.1	7	4278.66
Hordeum vulgare subsp. vulgare	2	GCA_903994145.1	7	4139.83
Hordeum vulgare subsp. vulgare	2	GCA_904849725.1	7	4225.71
Hordeum vulgare subsp. vulgare	2	GCA_905310525.1	7	4064.89
Hordeum vulgare subsp. vulgare	2	GCA_907165075.1	7	4439.13
Hordeum vulgare subsp. vulgare	2	GCA_902499585.1	8	4342.74
Impatiens glandulifera	2	GCA_907164915.1	9	653,879
Ipomoea trifida	2	GCA_003576665.1	15	492.376
Ipomoea trifida	2	GCA_004706985.1	15	460.934
Ipomoea triloba	2	GCA_003576645.1	16	461.827
Jacaranda mimosifolia	2	GCA_018894105.1	18	707,412
Juglans mandshurica	2	GCA_002916435.2	16	538,616
Juglans mandshurica	2	GCA_022457165.1	16	528,151
Juglans regia	2	GCA_001411555.2	16	572,947
Juglans regia	2	GCA_002916465.2	16	525,075
Lagenaria siceraria	2	GCA_002890555.2	11	297.879
Leersia perrieri	2	GCA_000325765.3	12	266.688
Lemna minuta	2	GCA_024174645.1	21	360.444
Limonium bicolor	2	GCA_023374045.1	8	2925.44
Linum usitatissimum	2	GCA_000224295.2	15	316.167
Linum usitatissimum	2	GCA_010665275.2	15	306,379
Litchi chinensis	2	GCA_019925255.1	15	470,369
Litchi chinensis	2	GCA_020101655.1	15	455,362
Litchi chinensis	2	GCA_020101635.1	15	450,294
Litsea cubeba	2	GCA_012931725.1	12	1325.68
Lolium rigidum	2	GCA_022539505.1	7	2438.47
Luffa acutangula	2	GCA_012295215.1	13	710
Luffa aegyptiaca	2	GCA_017139565.1	13	656,028
Lupinus albus	2	GCA_009771035.1	25	450.972
Lupinus albus	2	GCA_010261695.1	25	558.896

Lupinus angustifolius	2	GCA_001865875.1	20	609.203
Lupinus angustifolius	2	GCA_002285895.2	20	557.909
Lycium barbarum	2	GCA_019175385.1	12	1669.72
Macadamia integrifolia	2	GCA_013358625.1	14	744,937
Macadamia tetraphylla	2	GCA_022985045.1	14	750.867
Malus sieversii	2	GCA_020795835.1	17	683,351
Malus sylvestris	2	GCA_916048215.2	17	640.97
Mangifera indica	2	GCA_020138855.1	20	411,308
Mangifera indica	2	GCA_011075055.1	20	392.983
Mangifera indica	2	GCA_016746415.1	20	374,839
Mangifera indica	2	GCA_021014495.1	20	370.64
Mangifera indica	2	GCA_021014955.1	20	368,776
Manihot esculenta	2	GCA_013618965.1	17	710,029
Manihot esculenta	2	GCA_001659605.1	18	582.279
Manihot esculenta	2	GCA_001659605.2	18	640,431
Manihot esculenta	2	GCA_020916425.1	18	706,331
Manihot esculenta	2	GCA_020916445.1	18	762,395
Medicago ruthenica	2	GCA_018208015.1	8	904.13
Medicago truncatula	2	GCA_000219495.2	8	412.924
Medicago truncatula	2	GCA_003473485.2	8	429.612
Mentha longifolia	2	GCA_001642375.2	12	468,947
Miscanthus floridulus	2	GCA_019320115.1	19	2684.45
Morella rubra	2	GCA_003952965.1	8	313.02
Morella rubra	2	GCA_003952965.2	8	313,009
Morus alba	2	GCA_012066045.3	14	336,456
Morus alba	2	GCA_012066045.1	16	357
Musa balbisiana	2	GCA_004837865.1	11	492.775
Musa beccarii	2	GCA_024322285.1	9	569.618
Nelumbo nucifera	2	GCA_003033685.1	8	817.268
Nelumbo nucifera	2	GCA_003033695.1	8	799.479
Nelumbo nucifera	2	GCA_014319735.1	8	821,286
Nephelium lappaceum	2	GCA_021234005.1	16	409,258
Nicotiana attenuata	2	GCA_001879085.1	12	2365.680
Nymphaea colorata	2	GCA_008831285.1	14	409.014
Olea europaea subsp. cuspidata	2	GCA_023089605.1	23	1197.68
Olea europaea subsp. europaea	2	GCA_902713445.1	23	1316.68
Olea europaea var. sylvestris	2	GCA_002742605.1	23	1141.150
Oryza barthii	2	GCA_000182155.3	12	308.272
Oryza barthii	2	GCA_000182155.4	12	347,716
Oryza brachyantha	2	GCA_000710545.1	12	144,404
Oryza brachyantha	2	GCA_000231095.2	12	259.908
Oryza brachyantha	2	GCA_000231095.3	12	263.33
Oryza glaberrima	2	GCA_000147395.3	12	347,321
Oryza glumipatula	2	GCA_000576495.1	12	372.86
Oryza glumipatula	2	GCA_000576495.2	12	388,593
Oryza longistaminata	2	GCA_001514335.2	12	362.064
Oryza longistaminata	2	GCA_009805545.1	12	371.348
Oryza meridionalis	2	GCA_000338895.2	12	335.668
Oryza meridionalis	2	GCA_000338895.3	12	393,639
Oryza nivara	2	GCA_000710535.2	12	194,244
Oryza nivara	2	GCA_000576065.2	12	395,534
Oryza officinalis	2	GCA_000717455.1	12	261,885
Oryza rufipogon	2	GCA_000700045.1	12	127,409
Oryza rufipogon	2	GCA_023541355.1	12	462.581
Oryza sativa	2	GCA_014636015.1	12	404,811
Oryza sativa	2	GCA_014636035.1	11	399,284

Oryza sativa	2	GCA_003865215.1	12	395.354
Oryza sativa	2	GCA_004007595.1	12	377.604
Oryza sativa	2	GCA_004348155.2	12	415.393
Oryza sativa	2	GCA_009829725.1	12	490.155
Oryza sativa	2	GCA_009829375.1	12	382.007
Oryza sativa	2	GCA_009914875.1	12	387.484
Oryza sativa	2	GCA_018853525.1	12	409,523
Oryza sativa	2	GCA_019137765.1	12	407,498
Oryza sativa aromatic subgroup	2	GCA_009831255.1	12	391.87
Oryza sativa aus subgroup	2	GCA_001952365.2	12	372.203
Oryza sativa aus subgroup	2	GCA_001952365.3	12	388,175
Oryza sativa aus subgroup	2	GCA_009831335.1	12	383243
Oryza sativa Indica Group	2	GCA_000004655.2	12	426.337
Oryza sativa Indica Group	2	GCA_000725085.2	12	389.753
Oryza sativa Indica Group	2	GCA_001305255.1	12	352.121
Oryza sativa Indica Group	2	GCA_001618785.1	12	398.762
Oryza sativa Indica Group	2	GCA_001618795.1	12	386.486
Oryza sativa Indica Group	2	GCA_001623345.2	12	387.326
Oryza sativa Indica Group	2	GCA_001623365.2	12	387.424
Oryza sativa Indica Group	2	GCA_001889745.1	12	389.088
Oryza sativa Indica Group	2	GCA_002151415.1	12	390.984
Oryza sativa Indica Group	2	GCA_009831025.1	12	392.847
Oryza sativa Indica Group	2	GCA_009829395.1	12	405.399
Oryza sativa Indica Group	2	GCA_009831295.1	12	399.249
Oryza sativa Indica Group	2	GCA_009831355.1	12	394.602
Oryza sativa Indica Group	2	GCA_009831045.1	12	381.57
Oryza sativa Indica Group	2	GCA_001623345.3	12	391.696
Oryza sativa Indica Group	2	GCA_019338905.1	12	378,288
Oryza sativa Japonica Group	2	GCA_004295705.1	12	433,424
Oryza sativa Japonica Group	2	GCA_000149285.1	12	391.148
Oryza sativa Japonica Group	2	GCA_000005425.2	12	382.778
Oryza sativa Japonica Group	2	GCA_000164945.1	12	382.151
Oryza sativa Japonica Group	2	GCA_000321445.1	12	382.627
Oryza sativa Japonica Group	2	GCA_000817615.1	12	342.028
Oryza sativa Japonica Group	2	GCA_000817635.1	12	337.74
Oryza sativa Japonica Group	2	GCA_001433935.1	12	374.423
Oryza sativa Japonica Group	2	GCA_003865235.1	12	379.626
Oryza sativa Japonica Group	2	GCA_009797565.1	12	381.571
Oryza sativa Japonica Group	2	GCA_009830595.1	12	395.947
Oryza sativa Japonica Group	2	GCA_014526345.1	12	378,861
Oryza sativa tropical japonica subgroup	2	GCA_009831275.1	12	380.354
Oryza sativa tropical japonica subgroup	2	GCA_009831315.1	12	384.204
Osmanthus fragrans	2	GCA_019395295.1	23	733,264
Oxytropis ochrocephala	2	GCA_020916435.1	8	958.909
Panax notoginseng	2	GCA_014296215.1	12	2263.67
Panax notoginseng	2	GCA_016801055.1	12	2660.68
Panax stipuleanatus	2	GCA_020205555.1	12	1965.48
Panicum hallii	2	GCA_002211085.2	9	535.889
Panicum hallii var. hallii	2	GCA_003061485.1	9	487.474
Papaver somniferum	2	GCA_010119955.1	11	190.07
Papaver somniferum	2	GCA_003573695.1	11	2715.530
Papaver somniferum	2	GCA_010119995.1	11	270.303
Paspalum notatum	2	GCA_022530915.1	10	540.95
Paulownia fortunei	2	GCA_019321725.1	20	511,772
Pharus latifolius	2	GCA_019359835.1	12	1002.92
Phaseolus lunatus	2	GCA_013389735.1	11	546.42

Phaseolus vulgaris	2	GCA_000499845.1	11	521.077
Phaseolus vulgaris	2	GCA_001517995.1	11	549.748
Phaseolus vulgaris	2	GCA_015708805.1	11	423,736
Phoenix dactylifera	2	GCA_000181215.3	18	854.664
Phoenix dactylifera	2	GCA_009389715.1	18	772.474
Pisum sativum	2	GCA_900700895.2	7	3920.13
Pisum sativum	2	GCA_024323335.1	7	3796.64
Platycodon grandiflorus	2	GCA_016624345.1	9	574,706
Pogostemon cablin	2	GCA_023678885.1	63	1940.59
Populus simonii	2	GCA_007827005.2	19	441.407
Populus tomentosa	2	GCA_018804465.1	38	739,803
Populus trichocarpa	2	GCA_000002775.3	19	434.29
Populus trichocarpa	2	GCA_024362645.1	19	392.256
Primulina eburnea	2	GCA_022965805.1	18	812.379
Primulina huaijiensis	2	GCA_012295235.1	18	470
Prunus armeniaca	2	GCA_018524995.1	8	243,369
Prunus armeniaca	2	GCA_020226305.1	8	249,727
Prunus armeniaca	2	GCA_020424065.1	8	251.33
Prunus avium	2	GCA_013416215.1	8	271,649
Prunus avium	2	GCA_014155035.1	8	344,342
Prunus davidiana	2	GCA_020226225.1	8	243,914
Prunus dulcis	2	GCA_902201215.1	8	227.599
Prunus dulcis	2	GCA_021292205.2	8	257.659
Prunus dulcis	2	GCA_021292205.1		257,221
Prunus mira	2	GCA_020226265.1	8	239,885
Prunus mume	2	GCA_000346735.1	8	234.03
Prunus persica	2	GCA_024337555.1	8	243.543
Prunus persica	2	GCA_000346465.2	8	227.569
Prunus persica	2	GCA_015730445.1	8	239,053
Prunus persica	2	GCA_018340835.1	8	257,171
Prunus persica	2	GCA_022343065.2	8	206,428
Prunus salicina	2	GCA_014863905.1	8	284,209
Prunus salicina	2	GCA_019277915.1	8	282,437
Prunus salicina	2	GCA_020226455.1	8	267,139
Psidium guajava	2	GCA_016432845.1	11	443,756
Punica granatum	2	GCA_007655135.2	8	320.494
Pyrus betulifolia	2	GCA_007844245.1	17	532.747
Quercus aquifolioides	2	GCA_019022515.1	12	926,488
Quercus gilva	2	GCA_023621385.1	12	889.843
Quercus glauca	2	GCA_023736055.1	12	903.126
Quercus lobata	2	GCA_001633185.2	12	1277.010
Quercus lobata	2	GCA_001633185.5	12	845,947
Quercus mongolica	2	GCA_011696235.1	12	809.993
Quercus robur	2	GCA_932294415.1	12	789.739
Raphanus sativus	2	GCA_002197605.1	9	382.79
Raphanus sativus	2	GCA_902824885.1	9	440.318
Raphanus sativus	2	GCA_019705875.1	9	461,741
Raphanus sativus	2	GCA_019705855.1	9	487,865
Raphanus sativus	2	GCA_019705955.1	9	466,463
Raphanus sativus	2	GCA_019703475.1	9	459,818
Raphanus sativus var. caudatus	2	GCA_019705895.1	9	473,763
Raphanus sativus var. niger	2	GCA_019705885.1	9	465,105
Raphanus sativus var. oleiformis	2	GCA_019705865.1	9	514,662
Raphanus sativus var. raphanistroi	2	GCA_019705965.1	9	486,119
Rhododendron griersonianum	2	GCA_018127125.1	13	674,982
Rhododendron henanense subsp. l	2	GCA_020567845.1	13	634,288

Rhododendron ovatum	2	GCA_019656835.1	13	549.69
Rhododendron simsii	2	GCA_014282245.1	13	528,637
Rhododendron williamsianum	2	GCA_009746105.1	13	532.293
Ricinus communis	2	GCA_019578655.1	10	316,113
Rosa chinensis	2	GCA_002994745.1	7	513.854
Rosa chinensis	2	GCA_002994745.2	7	515,119
Salix brachista	2	GCA_009078335.1	19	339.588
Salix dunnii	2	GCA_015731905.1	19	328,089
Salix suchowensis	2	GCA_017552425.1	19	355,718
Salvia hispanica	2	GCA_023119035.1	6	321.469
Schrenkiella parvula	2	GCA_000218505.1	7	137.073
Scutellaria baicalensis	2	GCA_005771605.1	9	386.674
Secale cereale	2	GCA_016097815.1	7	7735.06
Secale cereale	2	GCA_902687465.1	8	6735.23
Senna tora	2	GCA_014851425.1	13	526,357
Sequoiadendron giganteum	2	GCA_007115665.2	11	8125.6
Sesamum indicum	2	GCA_000512975.1	16	275.059
Setaria italica	2	GCA_000263155.2	9	405.868
Setaria italica	2	GCA_001652605.1	9	477.542
Setaria viridis	2	GCA_005286985.1	9	395.732
Setaria viridis	2	GCA_012934335.1	9	397
Solanum commersonii	2	GCA_018258275.1	12	731,618
Solanum lycopersicum	2	GCA_000188115.3	12	828.349
Solanum lycopersicum	2	GCA_000188115.4	12	827.963
Solanum lycopersicum	2	GCA_900008105.1	12	926.426
Solanum lycopersicum	2	GCA_002954035.1	12	824.01
Solanum lycopersicum	2	GCA_012431665.1	12	813
Solanum lycopersicum	2	GCA_915070445.1	12	833,004
Solanum lycopersicum	2	GCA_022405115.1	12	797,153
Solanum pennellii	2	GCA_001406875.2	12	2768.130
Solanum pinnatisectum	2	GCA_009887355.1	12	724.558
Solanum stenotomum	2	GCA_019186545.1	12	846,249
Solanum tuberosum	2	GCA_009827155.1	12	724.894
Solanum tuberosum	2	GCA_009827175.1	12	724.882
Solanum tuberosum	2	GCA_010127505.1	12	2637.75
Solanum tuberosum	2	GCA_014182995.2	12	740,244
Solanum tuberosum	2	GCA_014182985.2	12	748,219
Solanum tuberosum	2	GCA_014189305.1	12	863,432
Solanum tuberosum	2	GCA_014189475.1	12	810,123
Solanum tuberosum	2	GCA_015076265.1	12	716,171
Solanum tuberosum	2	GCA_020169575.1	12	774,674
Solanum tuberosum	2	GCA_020169585.1	12	764,145
Solanum tuberosum	2	GCA_020169535.1	12	754,204
Solanum tuberosum	2	GCA_020169555.1	12	774,056
Sorghum bicolor	2	GCA_000003195.3	10	709.345
Sorghum bicolor	2	GCA_015952705.1	10	729.38
Spatholobus suberectus	2	GCA_004329165.1	9	798.47
Spinacia oleracea	2	GCA_020520425.1	6	894,256
Stellera chamaejasme	2	GCA_024586325.1	9	439.664
Syzygium aromaticum	2	GCA_024500025.1	11	370.258
Telopea speciosissima	2	GCA_018873765.1	11	823,061
Theobroma cacao	2	GCA_000403535.1	10	345.994
Theobroma cacao	2	GCA_000208745.2	10	324.88
Thinopyrum elongatum	2	GCA_011799875.1	7	4634.14
Thlaspi arvense	2	GCA_018983045.1	7	527,299
Thlaspi arvense	2	GCA_911865555.2	7	525,555

Trifolium pratense	2	GCA_020283565.1	7	414,027
Trifolium pratense	2	GCA_900079335.1	7	345.991
Trifolium pratense	2	GCA_900292005.1	7	351.622
Triticum urartu	2	GCA_003073215.1	7	4851.900
Triticum urartu	2	GCA_003073215.2	7	4849.19
Typha latifolia	2	GCA_019914945.1	15	214,326
Urochloa ruziziensis	2	GCA_015476505.1	9	604,562
Vaccinium darrowii	2	GCA_020921065.1	12	582,668
Vaccinium darrowii	2	GCA_020921045.1	12	480,504
Vaccinium macrocarpon	2	GCA_022606695.1	12	484,919
Vaccinium myrtillus	2	GCA_016920895.1	12	524,293
Vicia sativa	2	GCA_021764765.1	6	1653.55
Vigna angularis	2	GCA_001190045.1	11	467.301
Vigna angularis	2	GCA_016808095.1	11	447,806
Vigna angularis var. angularis	2	GCA_004320505.1	11	522.761
Vigna angularis var. angularis	2	GCA_001723775.1	11	444.439
Vigna mungo	2	GCA_013427195.1	11	498,912
Vigna mungo	2	GCA_023940565.1	11	454.426
Vigna radiata var. radiata	2	GCA_000741045.2	11	463.638
Vigna unguiculata	2	GCA_004118075.1	11	519.067
Vigna unguiculata	2	GCA_004118075.2	11	518.748
Vigna unguiculata	2	GCA_003958685.2	11	597.52
Vitellaria paradoxa	2	GCA_019916065.1	12	667,213
Vitis amurensis	2	GCA_016071775.1	19	603,559
Vitis riparia	2	GCA_004353265.1	19	500.106
Vitis vinifera	2	GCA_000003745.2	19	486.197
Xanthoceras sorbifolium	2	GCA_003430845.1	15	504.383
Xanthoceras sorbifolium	2	GCA_020796215.1	15	469,996
Zea mays	2	GCA_017315365.1	10	121,199
Zea mays	2	GCA_000005005.6	10	2135.080
Zea mays	2	GCA_003185045.1	10	2182.610
Zea mays	2	GCA_003704525.1	10	2198.500
Zea mays	2	GCA_003709335.1	10	2288.190
Zea mays	2	GCA_009176585.1	10	2160.69
Zea mays	2	GCA_902166955.1	10	2164.76
Zea mays	2	GCA_902167015.1	10	2176.5
Zea mays	2	GCA_902167105.1	10	2215.86
Zea mays	2	GCA_902167095.1	10	2219.25
Zea mays	2	GCA_902167135.1	10	2224.9
Zea mays	2	GCA_902167065.1	10	2124.54
Zea mays	2	GCA_902167085.1	10	2140.95
Zea mays	2	GCA_902166985.1	10	2190.8
Zea mays	2	GCA_902373975.1	10	2307.69
Zea mays	2	GCA_902167165.1	10	2223.23
Zea mays	2	GCA_902167025.1	10	2231.26
Zea mays	2	GCA_902167205.1	10	2215.81
Zea mays	2	GCA_902167035.1	10	2227.42
Zea mays	2	GCA_902167055.1	10	2162.44
Zea mays	2	GCA_902167145.1	10	2182.08
Zea mays	2	GCA_902167175.1	10	2192.4
Zea mays	2	GCA_902166975.1	10	2214.75
Zea mays	2	GCA_902167155.1	10	2300.77
Zea mays	2	GCA_902166995.1	10	2184.33
Zea mays	2	GCA_902167185.1	10	2271.03
Zea mays	2	GCA_902167045.1	10	2171.65
Zea mays	2	GCA_902167005.1	10	2290.5

Zea mays	2	GCA_902167075.1	10	2193.12
Zea mays	2	GCA_902166965.1	10	2138.71
Zea mays	2	GCA_902167115.1	10	2273.84
Zea mays	2	GCA_902167375.1	10	2214.05
Zea mays	2	GCA_902714155.1	10	2243.62
Zea mays	2	GCA_014529475.1	10	2246.85
Zea mays	2	GCA_905067065.1	10	2147.75
Zea mays	2	GCA_016432965.1	10	2285.79
Zea mays	2	GCA_019095955.1	10	2131.38
Zea mays	2	GCA_019096025.1	10	2125.15
Zea mays	2	GCA_019095975.1	10	2181.86
Zea mays	2	GCA_019096015.1	10	2155.6
Zea mays	2	GCA_019095995.1	10	2153.79
Zea mays	2	GCA_910593975.1	10	2160.3
Zea mays	2	GCA_021307875.1	10	2207.74
Zea mays	2	GCA_022117705.1	10	2178.6
Zea mays	2	GCA_024505845.1	10	2345.81
Zea mays subsp. mays	2	GCA_001644905.2	10	2133.880
Zea mays subsp. mays	2	GCA_001984235.2	10	2455.260
Zea mays subsp. mays	2	GCA_001990705.1	10	2392.800
Zea mays subsp. mays	2	GCA_002237485.1	10	456.675
Zea mays subsp. mays	2	GCA_002682915.2	10	2197.970
Zingiber officinale	2	GCA_018446385.1	22	3090.43
Ziziphus jujuba	2	GCA_000826755.1	12	437.754
Ziziphus jujuba	2	GCA_001835785.2	12	362.583
Ananas comosus	3	GCA_001540865.1	25	382.056
Ananas comosus	3	GCA_902162155.2	25	381,905
Arabidopsis suecica	4	GCA_019202805.1	13	272,254
Arachis hypogaea	4	GCA_003086295.2	20	2557.07
Arachis hypogaea	4	GCA_003713155.1	20	2506.710
Arachis hypogaea	4	GCA_004170445.1	20	2551.680
Arachis hypogaea	4	GCA_016103905.1	20	2490.84
Arachis hypogaea subsp. fastigiata	4	GCA_022829005.1	20	2535.29
Arachis monticola	4	GCA_003063285.2	20	2618.650
Avena insularis	4	GCA_023614405.1	14	7519.2
Brassica carinata	4	GCA_016771965.1	17	1086.99
Brassica juncea	4	GCA_015484525.1	18	884,898
Brassica juncea	4	GCA_018703725.1	18	933,495
Brassica juncea	4	GCA_020002505.1	18	894.19
Brassica juncea	4	GCA_020002515.1	18	904,412
Brassica juncea var. tumida	4	GCA_001687265.1	18	954.861
Brassica napus	4	GCA_000686985.2	19	976.191
Brassica napus	4	GCA_020379485.1	19	1001.5
Cenchrus purpureus	4	GCA_022644695.1	14	2018
Coffea arabica	4	GCA_003713225.1	22	699.904
Eragrostis tef	4	GCA_024500355.1	20	575.076
Gossypium barbadense	4	GCA_008761655.1	26	2195.8
Gossypium barbadense	4	GCA_018997955.1	26	2210.13
Gossypium darwinii	4	GCA_007990325.1	26	2182.96
Gossypium hirsutum	4	GCA_002504345.1	26	169,285
Gossypium hirsutum	4	GCA_000987745.1	26	2189.140
Gossypium hirsutum	4	GCA_006980745.1	26	2287.87
Gossypium hirsutum	4	GCA_006980775.1	26	2308.22
Gossypium hirsutum	4	GCA_007990345.1	26	2306.07
Gossypium hirsutum	4	GCA_018997965.1	26	2291.75
Gossypium hirsutum	4	GCA_021461695.1	26	2354.68

Gossypium hirsutum	4	GCA_021461685.1	26	2422.29
Gossypium hirsutum	4	GCA_024600755.1	26	2331.22
Gossypium mustelinum	4	GCA_007990455.1	26	2315.09
Gossypium mustelinum	4	GCA_017165895.1	26	2297.22
Gossypium tomentosum	4	GCA_007990485.1	26	2193.56
Gossypium tomentosum	4	GCA_018144435.1	26	2226.97
Isatis tinctoria	4	GCA_010577795.1	7	293.865
Mikania micrantha	4	GCA_009363875.1	19	1790.64
Miscanthus sacchariflorus	4	GCA_002993905.1	19	2074.920
Oryza minuta	4	GCA_000632695.1	24	451,659
Oryza punctata	4	GCA_000710525.1	24	224,654
Panax ginseng	4	GCA_020205605.1	24	3355.15
Panax japonicus	4	GCA_020205505.1	24	2028.86
Panax quinquefolius	4	GCA_020205615.1	24	3565.32
Panicum miliaceum	4	GCA_002895445.2	18	848.352
Panicum miliaceum	4	GCA_003046395.2	18	854.793
Polygonum aviculare	4	GCA_934048045.1	10	352.071
Potentilla anserina	4	GCA_933775445.1	7	237.424
Prunus fruticosa	4	GCA_018703695.1	8	375,294
Salvia splendens	4	GCA_004379255.2	22	806,486
Spirodela polyrhiza	4	GCA_001981405.1	20	136.67
Triadica sebifera	4	GCA_023653625.1	22	739.398
Trifolium occidentale	4	GCA_012979555.1	8	501
Trifolium repens	4	GCA_005869975.1	16	919.831
Triticum dicoccoides	4	GCA_002162155.3	14	10677.1
Triticum dicoccoides	4	GCA_002162155.2	14	10677.900
Triticum dicoccoides	4	GCA_900184675.1	14	10495.000
Triticum turgidum subsp. durum	4	GCA_900231445.1	14	9964.34
Actinidia deliciosa	6	GCA_024454175.1	29	621.991
Avena sativa	6	GCA_022788535.1	21	10839.2
Avena sativa	6	GCA_023646675.1	21	10757.5
Avena sativa	6	GCA_916181665.1	22	10840.7
Camelina sativa	6	GCA_000633955.1	20	641.356
Dendrocalamus latiflorus	6	GCA_017311315.1	70	2748.73
Echinochloa crus-galli	6	GCA_020466025.1	27	1340.74
Ipomoea batatas	6	GCA_002525835.2	15	837.013
Triticum aestivum	6	GCA_000210335.1	21	126,608
Triticum aestivum	6	GCA_900411305.1	21	563,502
Triticum aestivum	6	GCA_002220415.3	21	15418.8
Triticum aestivum	6	GCA_900519105.1	21	14547.300
Triticum aestivum	6	GCA_903993975.1	21	14281.3
Triticum aestivum	6	GCA_903993985.1	21	14645.5
Triticum aestivum	6	GCA_903993795.1	21	14538.3
Triticum aestivum	6	GCA_903994185.1	21	14884.6
Triticum aestivum	6	GCA_903994175.1	21	14350.8
Triticum aestivum	6	GCA_903994195.1	21	14385
Triticum aestivum	6	GCA_903994155.1	21	14463.5
Triticum aestivum	6	GCA_903995565.1	21	14433.2
Triticum aestivum	6	GCA_904066035.1	21	14910.4
Triticum aestivum	6	GCA_018294505.1	21	14567
Triticum aestivum	6	GCA_907166925.1	21	14702.9
Triticum aestivum	6	GCA_920937835.1	21	14256.7
Triticum aestivum	6	GCA_937894285.1	21	14195.6
Triticum aestivum	6	GCA_918797515.1	21	14679.4
Triticum aestivum	6	GCA_910594105.1	22	14677.2
Fragaria x ananassa	8	GCA_019022445.1	28	805.68

Saccharum officinarum	8	GCA_020631735.1	80	6804.89
Saccharum spontaneum	8	GCA_003544955.1	32	3133.290
Saccharum spontaneum	8	GCA_022457205.1	40	2761.16
Utricularia gibba	16	GCA_002189035.1	14	100.689
Abrus pulchellus subsp. cantoniens	-	GCA_023634445.1	11	381.268
Abrus pulchellus subsp. cantoniens	-	GCA_024086825.1	11	381.268
Acanthochlamys bracteata	-	GCA_019914995.1	20	197,966
Angophora floribunda	-	GCA_014182895.1	11	388,367
Aquilegia kansuensis	-	GCA_020826895.1	7	293,206
Arabidopsis thaliana x Arabidopsis	-	GCA_019202795.1	13	268,996
Areca catechu	-	GCA_021397845.1	16	2823.08
Aristolochia contorta	-	GCA_022405105.1	7	210,541
Aristolochia fimbriata	-	GCA_019845555.1	7	257,862
Artemisia tridentata subsp. tridenta	-	GCA_023558565.1	9	4198.57
Azadirachta indica	-	GCA_022749755.1	14	281.703
Begonia darthvaderiana	-	GCA_022432945.1	15	785,386
Begonia loranthoides	-	GCA_022433065.1	19	671,672
Begonia masoniana	-	GCA_022432975.1	15	799,687
Begonia peltatifolia	-	GCA_022433055.1	15	309,997
Boehmeria nivea	-	GCA_021020685.1	14	266,599
Boehmeria nivea var. tenacissima	-	GCA_018132145.1	14	270,213
Brassica oleracea var. capitata	-	GCA_018177695.1	9	565,471
Brassica rapa subsp. trilocularis	-	GCA_017639395.1	10	346,507
Bretschneidera sinensis	-	GCA_018105755.1	9	1170.88
Bretschneidera sinensis	-	GCA_023935145.1	9	1213.74
Capsicum annuum var. glabriusculı	-	GCA_000950795.1	12	760.067
Carpinus fangiana	-	GCA_006937295.1	8	381.949
Ceratodon purpureus	-	GCA_014871385.1	13	362,511
Ceratodon purpureus	-	GCA_014871845.1	13	349,464
Ceriops tagal	-	GCA_021533255.1	18	231,919
Chimonanthus praecox	-	GCA_022113865.1	11	737,025
Chimonanthus salicifolius	-	GCA_013350335.1	11	853,434
Cichorium endivia	-	GCA_023376185.1	9	886.978
Cichorium intybus	-	GCA_023525715.1	9	1278.75
Codonopsis lanceolata	-	GCA_013146195.2	8	1273.26
Corymbia calophylla	-	GCA_014182845.1	11	394,897
Corymbia citriodora subsp. variega	-	GCA_014858505.1	11	544,192
Corymbia maculata	-	GCA_014182735.1	11	403,979
Cucurbita argyrosperma subsp. sor	-	GCA_018691285.1	20	255,123
Cycas panzhihuaensis	-	GCA_023213395.1	11	10482.7
Cymbidium sinense	-	GCA_021442155.1	20	3525.77
Cynara cardunculus var. scolymus	-	GCA_001531365.1	17	725.198
Cynara cardunculus var. scolymus	-	GCA_001531365.2	17	724.962
Echium plantagineum	-	GCA_003412495.2	8	349.028
Eragrostis curvula	-	GCA_007726485.1	7	603.072
Eucalyptus melliodora	-	GCA_004368105.3	11	639.598
Eucalyptus melliodora	-	GCA_004368105.2		624,609
Flaveria linearis	-	GCA_024085815.1	18	1654.55
Forsythia suspensa	-	GCA_020510225.1	14	737,551
Forsythia suspensa	-	GCA_023638005.1	14	737.526
Gastrodia elata	-	GCA_016760335.1	18	1046.14
Gentiana dahurica var. dahurica	-	GCA_024500145.1	13	1416.54
Gynochthodes officinalis	-	GCA_020080225.1	11	484,869
Ilex asprella	-	GCA_023539305.1	19	804.072
Juglans microcarpa x Juglans regia	-	GCA_004785585.1	16	534.672
Juglans microcarpa x Juglans regia	-	GCA_004785595.1	16	527.896

Juglans nigra	-	GCA_002916485.2	16	531,995
Kandelia obovata	-	GCA_021464305.1	18	190.32
Lactuca sativa	-	GCA_002870075.3	9	2388.97
Lactuca sativa	-	GCA_002870075.2		2391.58
Leptodermis oblonga	-	GCA_016801395.1	11	497,295
Melastoma candidum	-	GCA_023653495.1	12	256.218
Microstegium vimineum	-	GCA_022036555.1	23	1118.67
Musa troglodytarum	-	GCA_023547065.1	10	603.588
Nyssa sinensis	-	GCA_008638375.1	22	1001.45
Papilionanthe hookeriana x Papilio	-	GCA_022702705.1	19	2570.13
Pohlia nutans	-	GCA_022496805.1	23	698,196
Populus deltoides	-	GCA_014885025.1	19	431,876
Populus deltoides	-	GCA_015852605.2	19	424.59
Pugionium cornutum	-	GCA_018901935.1	11	550.39
Pyrus ussuriensis x Pyrus commun	-	GCA_008932095.1	17	510.637
Raphanus raphanistrum subsp. lan	-	GCA_019706005.1	9	418,302
Raphanus raphanistrum subsp. rap	-	GCA_019706035.1	9	421,544
Raphanus raphanistrum x Raphanu	-	GCA_019705995.1	9	488,876
Rhamnella rubrinervis	-	GCA_007844105.2	12	245.348
Rhodamnia argentea	-	GCA_020921035.1	11	346,714
Saccharum hybrid cultivar	-	GCA_020102875.1	10	903.62
Salix arbutifolia	-	GCA_021905355.1	19	324,001
Sindora glabra	-	GCA_020226215.1	12	1113.92
Smallanthus sonchifolius	-	GCA_023525975.1	29	2716.52
Solanum lycopersicoides	-	GCA_022817965.1	12	1287.24
Sphagnum fallax	-	GCA_021442195.1	20	395,135
Sphagnum magellanicum	-	GCA_021904315.1	20	439,011
Spirodela intermedia	-	GCA_902729315.2	18	136.68
Syntrichia caninervis	-	GCA_016097705.1	13	329,824
Taxillus chinensis	-	GCA_023512835.1	9	521.908
Taxus chinensis	-	GCA_019776745.2	12	10238
Taxus wallichiana var. yunnanensis	-	GCA_018340775.1	12	10738.3
Tetracentron sinense	-	GCA_015143295.1	24	1161.36
Thymus quinquecostatus	-	GCA_024222315.1	13	528.675
Tripterygium wilfordii	-	GCA_013401445.1	23	348,533
Tripterygium wilfordii	-	GCA_016880815.1	23	342.58
Triticum aestivum subsp. tibeticum	-	GCA_014338645.1	21	14708.2
Vitis rotundifolia	-	GCA_022557335.1	20	393,821
Ziziphus jujuba var. spinosa	-	GCA_020796205.1	12	406,164
Physcomitrella patens	1;2	GCA_000002425.2	27	472.081
Colocasia esculenta	2;3	GCA_014218235.1	14	2405.85
Dioscorea rotundata	2;3	GCA_002240015.2	21	2155.820
Ficus carica	2;3	GCA_009761775.1	13	333.44
Malus domestica	2;3	GCA_000148765.2	17	1874.770
Malus domestica	2;3	GCA_002114115.1	17	702.961
Malus domestica	2;3	GCA_004115385.1	17	660.463
Malus domestica	2;3	GCA_916050505.1	17	648,233
Malus domestica	2;3	GCA_916612005.1	17	652,806
Malus domestica	2;3	GCA_916615385.1	17	646.79
Malus domestica	2;3	GCA_916615275.2	17	642,631
Malus domestica	2;3	GCA_022606005.1	17	754,684
Musa acuminata subsp. malaccens	2;3	GCA_000313855.2	11	472.231
Vanilla planifolia	2;3;4	GCA_016413895.1	14	736,753
Vanilla planifolia	2;3;4	GCA_016413885.1	14	744,192
Vanilla planifolia	2;3;4	GCA_023846275.1	14	1416.36
Vanilla planifolia	2;3;4	GCA_023853775.1	14	1967.63

Adiantum capillus-veneris	2;4	GCA_014529385.2	30	4822.57
Arabidopsis arenosa	2;4	GCA_905216605.1	8	149,659
Bothriochloa decipiens	2;4	GCA_023333625.1	20	1213.27
Dianthus caryophyllus	2;4	GCA_023091065.1	15	636.302
Hordeum marinum	2;4	GCA_022496015.1	7	3815.97
Lolium perenne	2;4	GCA_019359855.1	7	2277.55
Lonicera japonica	2;4	GCA_021464415.1	9	886,132
Oryza punctata	2;4	GCA_000573905.1	12	393.817
Oryza punctata	2;4	GCA_000573905.2	12	422,391
Solanum pimpinellifolium	2;4	GCA_014964335.1	12	808.1
Mercurialis annua	2;4;6	GCA_937616625.1	8	453.169
Chenopodium formosanum	2;4;6;10	GCA_024500155.1	27	1629.91
Dioscorea cayenensis subsp. rotun	4;6;8	GCA_009730915.2	28	584,309
Panicum virgatum	4;6;8	GCA_016808335.1	18	1130
Dioscorea alata	4;8	GCA_020875875.1	20	480,026
Lactuca saligna	#N/A	GCA_902860255.1		2165.76
Oryza sativa f. spontanea	#N/A	GCA_000576065.1	12	337.95

Organism/Name	Ploidy	Assembly Accession	chromosom	Size (Mb)	GC
Gossypium arboreum	2	GCA_013265605.1	13	94.64	35.30
Arabidopsis thaliana	2	GCA_024498515.1	5	120.09	36.19
Oryza brachyantha	2	GCA_000710545.1	12	144.40	40.90
Boechera stricta	2	GCA_018361405.1	7	189.46	35.91
Oryza nivara	2	GCA_000710535.2	12	194.24	43.30
Bruguiera parviflora	2	GCA_019804595.1	18	210.39	35.80
Camelina neglecta	2	GCA_023864065.1	6	210.45	36.48
Camelina laxa	2	GCA_024034495.1	6	213.07	36.66
Typha latifolia	2	GCA_019914945.1	15	214.33	37.76
Cucurbita argyrosperma subsp. arg	2	GCA_004115005.2	20	228.92	36.54
Prunus mira	2	GCA_020226265.1	8	239.89	37.51
Cucumis sativus	2	GCA_016161875.1	7	242.88	33.53
Prunus persica	2	GCA_024337555.1	8	243.54	37.94
Prunus davidiana	2	GCA_020226225.1	8	243.91	37.56
Prunus armeniaca	2	GCA_020424065.1	8	251.33	37.55
Prunus dulcis	2	GCA_021292205.2	8	257.66	37.99
Arabis montbretiana	2	GCA_001484125.2	8	257.69	36.58
Oryza officinalis	2	GCA_000717455.1	12	261.89	44.70
Prunus salicina	2	GCA_020226455.1	8	267.14	37.58
Eutrema salsugineum	2	GCA_016617915.1	7	295.49	37.61
Gillenia trifoliata	2	GCA_018257905.1	9	296.28	38.28
Citrus sinensis	2	GCA_022201045.1	9	298.98	34.21
Amphicarpaea edgeworthii	2	GCA_014843725.1	11	299.06	32.06
Citrus trifoliata	2	GCA_018350135.1	9	303.07	34.13
Linum usitatissimum	2	GCA_010665275.2	15	306.38	39.01
Morella rubra	2	GCA_003952965.2	8	313.01	37.83
Ricinus communis	2	GCA_019578655.1	10	316.11	33.06
Salvia hispanica	2	GCA_023119035.1	6	321.47	36.17
Bauhinia variegata	2	GCA_022379115.2	14	326.37	34.94
Salix dunnii	2	GCA_015731905.1	19	328.09	33.09
Morus alba	2	GCA_012066045.3	14	336.46	34.40
Camelina hispida	2	GCA_023864115.1	7	339.78	36.78
Prunus avium	2	GCA_014155035.1	8	344.34	38.44
Oryza glaberrima	2	GCA_000147395.3	12	347.32	42.87
Oryza barthii	2	GCA_000182155.4	12	347.72	42.87
Carica papaya	2	GCA_022788785.1	10	349.88	36.04
Salix suchowensis	2	GCA_017552425.1	19	355.72	34.95
Lemna minuta	2	GCA_024174645.1	21	360.44	39.26
Cucumis melo subsp. agrestis	2	GCA_014525375.1	12	366.17	33.67
Citrullus lanatus subsp. cordophan	2	GCA_018142915.1	11	367.12	33.60
Cucumis melo	2	GCA_020920055.1	12	367.30	33.75
Mangifera indica	2	GCA_021014955.1	20	368.78	32.82
Corylus avellana	2	GCA_901000735.2	11	369.78	35.92
Syzygium aromaticum	2	GCA_024500025.1	11	370.26	39.91
Corylus heterophylla	2	GCA_016403345.1	11	370.75	35.83
Brassica rapa	2	GCA_016163755.1	10	370.90	37.19
Amaranthus cruentus	2	GCA_019425755.1	17	370.91	33.09
Aeschynomene evenia	2	GCA_013621005.1	10	375.94	34.93
Oryza sativa Indica Group	2	GCA_019338905.1	12	378.29	43.44
Oryza sativa Japonica Group	2	GCA_014526345.1	12	378.86	43.30
Oryza glumipatula	2	GCA_000576495.2	12	388.59	44.08
Populus trichocarpa	2	GCA_024362645.1	19	392.26	33.75
Oryza meridionalis	2	GCA_000338895.3	12	393.64	43.39
Setaria viridis	2	GCA_012934335.1	9	397.00	45.99
Citrullus lanatus	2	GCA_004801215.2	11	397.83	34.24

Eucalyptus tenuipes	2	GCA_014182365.1	11	397.94	39.62
Oryza sativa	2	GCA_019137765.1	12	407.50	43.94
Nephelium lappaceum	2	GCA_021234005.1	16	409.26	32.84
Trifolium pratense	2	GCA_020283565.1	7	414.03	33.72
Amaranthus hypochondriacus	2	GCA_000753965.2	16	417.46	34.51
Coffea humblotiana	2	GCA_023065735.1	11	420.72	36.03
Phaseolus vulgaris	2	GCA_015708805.1	11	423.74	33.99
Centella asiatica	2	GCA_014636745.1	9	430.22	34.62
Eucalyptus curtisii	2	GCA_017140595.1	11	435.42	39.67
Stellera chamaejasme	2	GCA_024586325.1	9	439.66	41.33
Eucalyptus microcorys	2	GCA_014182515.1	11	441.07	39.38
Psidium guajava	2	GCA_016432845.1	11	443.76	39.47
Vigna angularis	2	GCA_016808095.1	11	447.81	33.61
Litchi chinensis	2	GCA_020101635.1	15	450.29	34.51
Dimocarpus longan	2	GCA_022984855.1	15	454.30	33.95
Vigna mungo	2	GCA_023940565.1	11	454.43	33.53
Avicennia marina subsp. marina	2	GCA_013168755.1	32	457.00	36.72
Avicennia marina	2	GCA_019155195.1	31	457.34	33.46
Raphanus sativus	2	GCA_019703475.1	9	459.82	37.40
Oryza rufipogon	2	GCA_023541355.1	12	462.58	44.16
Raphanus sativus var. niger	2	GCA_019705885.1	9	465.11	36.77
Mentha longifolia	2	GCA_001642375.2	12	468.95	36.86
Xanthoceras sorbifolium	2	GCA_020796215.1	15	470.00	35.09
Primulina huaijiensis	2	GCA_012295235.1	18	470.00	35.89
Eucalyptus guilfoylei	2	GCA_016097605.1	11	472.52	39.50
Raphanus sativus var. caudatus	2	GCA_019705895.1	9	473.76	36.18
Eucalyptus cloeziana	2	GCA_014182715.1	11	480.24	39.39
Vaccinium darrowii	2	GCA_020921045.1	12	480.50	38.50
Vaccinium macrocarpon	2	GCA_022606695.1	12	484.92	37.94
Raphanus sativus var. raphanistroi	2	GCA_019705965.1	9	486.12	36.92
Ensete glaucum	2	GCA_021527575.1	9	494.94	38.24
Eucalyptus regnans	2	GCA_014182855.1	11	495.13	39.40
Eucalyptus brandiana	2	GCA_014182725.1	11	507.24	39.47
Eucalyptus salubris	2	GCA_014182395.1	11	508.09	39.46
Paulownia fortunei	2	GCA_019321725.1	20	511.77	32.56
Eucalyptus marginata	2	GCA_014182565.1	11	513.05	39.54
Raphanus sativus var. oleiformis	2	GCA_019705865.1	9	514.66	37.26
Rosa chinensis	2	GCA_002994745.2	7	515.12	38.84
Vaccinium myrtillus	2	GCA_016920895.1	12	524.29	38.64
Juglans regia	2	GCA_002916465.2	16	525.08	36.20
Thlaspi arvense	2	GCA_911865555.2	7	525.56	38.38
Senna tora	2	GCA_014851425.1	13	526.36	35.45
Juglans mandshurica	2	GCA_022457165.1	16	528.15	36.50
Rhododendron simsii	2	GCA_014282245.1	13	528.64	38.90
Eucalyptus pumila	2	GCA_016097595.1	11	529.92	39.38
Eucalyptus virginea	2	GCA_014182375.1	11	532.95	39.47
Brassica nigra	2	GCA_016432835.1	8	534.24	38.43
Gardenia jasminoides	2	GCA_013103745.1	11	536.00	35.92
Eucalyptus erythrocorys	2	GCA_014182555.1	11	539.36	39.63
Fagus sylvatica	2	GCA_907173295.1	12	540.34	35.63
Paspalum notatum	2	GCA_022530915.1	10	540.95	45.67
Eucalyptus cladocalyx	2	GCA_017140615.1	11	544.25	39.32
Eucalyptus globulus	2	GCA_014182545.1	11	545.19	39.38
Phaseolus lunatus	2	GCA_013389735.1	11	546.42	35.90
Rhododendron ovatum	2	GCA_019656835.1	13	549.69	38.84
Eucalyptus victrix	2	GCA_016097545.1	11	557.32	39.20

Eucalyptus camaldulensis	2	GCA_014182705.1	11	558.61	39.28
Eucalyptus viminalis	2	GCA_014182385.1	11	558.87	39.36
Eucalyptus leucophloia subsp. eurc	2	GCA_017140325.1	11	568.64	39.15
Musa beccarii	2	GCA_024322285.1	9	569.62	38.73
Catharanthus roseus	2	GCA_024505715.1	8	572.92	34.17
Platycodon grandiflorus	2	GCA_016624345.1	9	574.71	36.85
Gynostemma pentaphyllum	2	GCA_020536105.1	11	582.95	32.87
Eucalyptus paniculata subsp. matu	2	GCA_017140255.1	11	589.01	39.26
Eucalyptus caleyi	2	GCA_014182885.2	11	589.48	39.23
Eucalyptus fibrosa	2	GCA_017140475.1	11	590.07	39.17
Cajanus cajan	2	GCA_000340665.2	11	590.52	33.71
Eucalyptus decipiens	2	GCA_014182575.1	11	591.12	39.40
Eucalyptus sideroxylon	2	GCA_014182405.1	11	592.32	39.21
Eucalyptus shirleyi	2	GCA_017140165.1	11	597.34	39.21
Vigna unguiculata	2	GCA_003958685.2	11	597.52	35.00
Eucalyptus polyanthemos subsp. p	2	GCA_017140185.1	11	603.44	39.18
Vitis amurensis	2	GCA_016071775.1	19	603.56	34.30
Eucalyptus sideroxylon x Eucalyptu	2	GCA_016097485.1	11	603.74	39.30
Urochloa ruziziensis	2	GCA_015476505.1	9	604.56	46.58
Eucalyptus coolabah	2	GCA_014182585.1	11	606.47	39.25
Eucalyptus albens	2	GCA_014182695.1	11	607.09	39.18
Gossypium raimondii	2	GCA_013467475.1	13	615.03	33.99
Eucalyptus grandis	2	GCA_016545825.1	11	616.53	39.39
Diospyros lotus	2	GCA_014633365.1	15	630.10	36.57
Eucalyptus lansdowneana	2	GCA_017140395.1	11	633.71	39.26
Rhododendron henanense subsp. l	2	GCA_020567845.1	13	634.29	40.86
Malus sylvestris	2	GCA_916048215.2	17	640.97	38.06
Akebia trifoliata	2	GCA_017979445.1	16	652.80	35.00
Impatiens glandulifera	2	GCA_907164915.1	9	653.88	32.12
Gossypium trilobum	2	GCA_013467465.1	13	655.38	34.53
Luffa aegyptiaca	2	GCA_017139565.1	13	656.03	35.88
Carya illinoinensis	2	GCA_018689175.1	16	656.69	36.25
Actinidia eriantha	2	GCA_019202715.1	29	657.10	35.70
Gossypium gossypioides	2	GCA_013467495.1	13	664.72	34.33
Vitellaria paradoxa	2	GCA_019916065.1	12	667.21	32.94
Gossypium klotzschianum	2	GCA_013677235.1	13	670.96	34.57
Rhododendron griersonianum	2	GCA_018127125.1	13	674.98	40.76
Malus sieversii	2	GCA_020795835.1	17	683.35	38.02
Gossypium davidsonii	2	GCA_013677245.1	13	704.22	33.75
Eucalyptus dawsonii	2	GCA_016097615.1	11	707.06	39.26
Jacaranda mimosifolia	2	GCA_018894105.1	18	707.41	33.87
Luffa acutangula	2	GCA_012295215.1	13	710.00	38.53
Digitaria exilis	2	GCA_902859565.1	19	716.47	45.53
Sorghum bicolor	2	GCA_015952705.1	10	729.38	44.01
Gossypium schwendimanii	2	GCA_013677275.1	13	729.43	34.12
Solanum commersonii	2	GCA_018258275.1	12	731.62	34.41
Gossypium harknessii	2	GCA_013677255.1	13	732.16	34.17
Osmanthus fragrans	2	GCA_019395295.1	23	733.26	34.30
Asparagus setaceus	2	GCA_012295165.1	10	735.53	36.10
Gossypium aridum	2	GCA_013487665.1	13	739.12	34.26
Populus tomentosa	2	GCA_018804465.1	38	739.80	33.64
Gossypium lobatum	2	GCA_013467485.1	13	744.54	34.24
Macadamia integrifolia	2	GCA_013358625.1	14	744.94	39.29
Macadamia tetraphylla	2	GCA_022985045.1	14	750.87	39.28
Fraxinus pennsylvanica	2	GCA_912172775.1	23	756.79	36.02
Manihot esculenta	2	GCA_020916445.1	18	762.40	38.60

Solanum tuberosum	2	GCA_020169555.1	12	774.06	35.21
Gossypium armourianum	2	GCA_013677265.1	13	780.95	36.70
Quercus robur	2	GCA_932294415.1	12	789.74	35.68
Solanum lycopersicum	2	GCA_022405115.1	12	797.15	34.59
Fraxinus excelsior	2	GCA_019097785.1	23	807.61	34.42
Primulina eburnea	2	GCA_022965805.1	18	812.38	37.09
Nelumbo nucifera	2	GCA_014319735.1	8	821.29	38.97
Telopea speciosissima	2	GCA_018873765.1	11	823.06	40.12
Gossypium laxum	2	GCA_013511315.1	13	833.90	35.99
Quercus lobata	2	GCA_001633185.5	12	845.95	35.43
Solanum stenotomum	2	GCA_019186545.1	12	846.25	36.57
Buddleja alternifolia	2	GCA_019426215.1	19	853.76	36.10
Quercus gilva	2	GCA_023621385.1	12	889.84	35.79
Spinacia oleracea	2	GCA_020520425.1	6	894.26	37.91
Quercus glauca	2	GCA_023736055.1	12	903.13	35.99
Medicago ruthenica	2	GCA_018208015.1	8	904.13	35.91
Cannabis sativa	2	GCA_016165845.1	10	914.40	33.78
Quercus aquifolioides	2	GCA_019022515.1	12	926.49	36.60
Coptis chinensis	2	GCA_015680905.1	9	935.66	37.34
Glycine latifolia	2	GCA_013407115.1	20	939.49	33.89
Eucommia ulmoides	2	GCA_016647705.1	17	947.85	35.17
Oxytropis ochrocephala	2	GCA_020916435.1	8	958.91	37.46
Glycine soja	2	GCA_014282345.1	20	975.92	34.79
Pharus latifolius	2	GCA_019359835.1	12	1002.92	44.31
Glycine max	2	GCA_022114995.1	20	1011.40	34.89
Arachis duranensis	2	GCA_018207795.1	10	1099.87	35.92
Gossypium anomalum	2	GCA_019455425.1	13	1193.34	34.28
Olea europaea subsp. cuspidata	2	GCA_023089605.1	23	1197.68	35.39
Dendrobium nobile	2	GCA_022539455.1	19	1199.12	35.31
Elaeis guineensis	2	GCA_015461965.1	16	1209.42	40.46
Dendrobium officinale	2	GCA_019514585.1	19	1228.67	35.32
Arachis cardenasii	2	GCA_018493915.1	10	1238.08	37.25
Dendrobium huoshanense	2	GCA_016618105.1	19	1284.29	35.75
Olea europaea subsp. europaea	2	GCA_902713445.1	23	1316.68	34.56
Litsea cubeba	2	GCA_012931725.1	12	1325.68	40.51
Arachis stenosperma	2	GCA_014773155.1	10	1328.99	37.55
Dendrobium chrysotoxum	2	GCA_019925795.1	19	1368.17	35.54
Arachis ipaensis	2	GCA_013265535.1	10	1438.65	36.57
Gossypium stocksii	2	GCA_020496765.1	13	1448.11	34.95
Vicia sativa	2	GCA_021764765.1	6	1653.55	35.63
Lycium barbarum	2	GCA_019175385.1	12	1669.72	38.10
Arctium lappa	2	GCA_023525745.1	18	1727.36	36.65
Cenchrus americanus	2	GCA_021560375.1	7	1908.26	49.06
Pogostemon cablin	2	GCA_023678885.1	63	1940.59	34.37
Panax stipuleanatus	2	GCA_020205555.1	12	1965.48	35.22
Chloranthus sessilifolius	2	GCA_021018995.1	15	2168.75	39.32
Zea mays	2	GCA_024505845.1	10	2345.81	46.97
Lolium rigidum	2	GCA_022539505.1	7	2438.47	44.77
Papaver somniferum	2	GCA_010119995.1	11	2637.75	38.54
Panax notoginseng	2	GCA_016801055.1	12	2660.68	34.45
Chrysanthemum lavandulifolium	2	GCA_022545495.1	9	2670.47	36.03
Hibiscus mutabilis	2	GCA_019671005.1	46	2675.93	35.35
Miscanthus floridulus	2	GCA_019320115.1	19	2684.45	45.29
Camellia oleifera	2	GCA_022316695.1	15	2889.51	34.52
Limonium bicolor	2	GCA_023374045.1	8	2925.44	39.22
Helianthus annuus	2	GCA_002127325.2	17	3010.05	38.82

Camellia sinensis var. sinensis	2	GCA_020536495.1	15	3062.74	38.53
Camellia sinensis var. lasiocalyx	2	GCA_020536555.1	15	3062.77	38.53
Camellia sinensis var. assamica	2	GCA_020536565.1	15	3062.80	38.52
Camellia sinensis	2	GCA_020536515.1	15	3062.86	38.53
Capsicum annuum	2	GCA_021292125.1	12	3077.74	34.91
Zingiber officinale	2	GCA_018446385.1	22	3090.43	39.16
Avena longiglumis	2	GCA_023614385.1	7	3736.64	44.37
Hemerocallis citrina	2	GCA_017893485.1	11	3775.58	39.87
Pisum sativum	2	GCA_024323335.1	7	3796.64	38.12
Aegilops speltoides	2	GCA_021437245.1	7	4110.19	46.38
Aegilops tauschii subsp. strangulat	2	GCA_002575655.2	7	4218.18	46.42
Hordeum vulgare subsp. vulgare	2	GCA_907165075.1	7	4439.13	44.64
Hordeum vulgare subsp. spontaneu	2	GCA_907165085.1	7	4498.60	44.66
Thinopyrum elongatum	2	GCA_011799875.1	7	4634.14	45.88
Triticum urartu	2	GCA_003073215.2	7	4849.19	45.94
Hordeum vulgare	2	GCA_024137805.1	7	5111.00	44.00
Aegilops speltoides subsp. speltoid	2	GCA_944222845.1	7	5116.92	47.32
Aegilops searsii	2	GCA_021605185.1	7	5336.42	46.21
Amorphophallus konjac	2	GCA_022559845.1	13	5598.56	45.56
Aegilops longissima	2	GCA_021605205.1	7	5796.09	46.40
Aegilops sharonensis	2	GCA_021641835.1	7	5892.84	46.35
Aegilops bicornis	2	GCA_021605145.1	7	5902.72	46.40
Secale cereale	2	GCA_902687465.1	8	6735.23	46.05
Ceratopteris richardii	2	GCA_020310875.1	39	7462.46	38.71
Allium cepa	2	GCA_905187595.1	8	14937.40	33.67
Allium sativum	2	GCA_014155895.2	9	16243.20	35.73
Coffea eugenioides	2	GCA_003713205.1	11	1094.45	36.96
Olea europaea var. sylvestris	2	GCA_002742605.1	23	1141.15	36.77
Asparagus officinalis	2	GCA_001876935.1	10	1187.54	39.36
Gossypium longicalyx	2	GCA_010883175.1	13	1190.21	34.19
Schrenkiella parvula	2	GCA_000218505.1	7	137.07	35.74
Hevea brasiliensis	2	GCA_010458925.1	18	1473.45	33.90
Coix aquatica	2	GCA_009725075.1	10	1615.47	46.76
Coix lacryma-jobi var. lacryma-jobi	2	GCA_009763385.1	10	1731.46	46.76
Gossypium australe	2	GCA_005393395.2	13	1743.39	36.48
Erysimum cheiranthoides	2	GCA_011420285.1	8	177.18	36.29
Fragaria vesca subsp. vesca	2	GCA_000184155.1	7	214.37	38.98
Zea mays subsp. mays	2	GCA_002682915.2	10	2197.97	46.74
Cocos nucifera	2	GCA_008124465.1	16	2202.46	37.83
Prunus mume	2	GCA_000346735.1	8	234.03	38.33
Brassica rapa subsp. pekinensis	2	GCA_008629595.1	10	234.69	35.53
Nicotiana attenuata	2	GCA_001879085.1	12	2365.68	41.33
Fragaria iinumae	2	GCA_009720345.1	7	240.58	39.71
Cucurbita pepo subsp. pepo	2	GCA_002806865.2	20	261.36	37.26
Leersia perrieri	2	GCA_000325765.3	12	266.69	42.60
Andrographis paniculata	2	GCA_004354405.1	24	269.41	33.34
Brachypodium distachyon	2	GCA_000005505.4	5	271.30	46.42
Sesamum indicum	2	GCA_000512975.1	16	275.06	35.22
Solanum pennellii	2	GCA_001406875.2	12	2768.13	35.74
Lagenaria siceraria	2	GCA_002890555.2	11	297.88	31.95
Capsicum chinense	2	GCA_002271895.2	12	3070.91	34.86
Arabis alpina	2	GCA_900128785.1	8	311.64	37.40
Punica granatum	2	GCA_007655135.2	8	320.49	40.37
Capsicum baccatum	2	GCA_002271885.2	12	3215.61	35.43
Theobroma cacao	2	GCA_000208745.2	10	324.88	34.99
Apium graveolens	2	GCA_009905375.1	11	3332.58	35.70

Salix brachista	2	GCA_009078335.1	19	339.59	34.16
Citrus maxima	2	GCA_002006925.1	9	345.76	34.99
Cicer arietinum	2	GCA_006345785.1	8	347.25	32.77
Ziziphus jujuba	2	GCA_001835785.2	12	362.58	32.96
Oryza longistaminata	2	GCA_009805545.1	12	371.35	42.07
Carex littledalei	2	GCA_011114355.1	29	373.85	35.45
Oryza sativa aus subgroup	2	GCA_009831335.1	12	383.24	43.79
Oryza sativa tropical japonica subg	2	GCA_009831315.1	12	384.20	43.50
Cucumis melo var. inodorus	2	GCA_009760825.1	12	386.50	33.88
Scutellaria baicalensis	2	GCA_005771605.1	9	386.67	34.34
Oryza sativa aromatic subgroup	2	GCA_009831255.1	12	391.87	43.61
Nymphaea colorata	2	GCA_008831285.1	14	409.01	38.59
Cicer reticulatum	2	GCA_003689015.2	8	416.90	31.52
Daucus carota subsp. sativus	2	GCA_001625215.1	9	421.54	36.04
Erigeron canadensis	2	GCA_010389155.1	9	426.38	34.07
Medicago truncatula	2	GCA_003473485.2	8	429.61	33.44
Aegilops tauschii	2	GCA_000347335.2	7	4310.35	46.40
Populus simonii	2	GCA_007827005.2	19	441.41	33.66
Vigna angularis var. angularis	2	GCA_001723775.1	11	444.44	31.77
Ipomoea trifida	2	GCA_004706985.1	15	460.93	35.83
Ipomoea triloba	2	GCA_003576645.1	16	461.83	36.36
Vigna radiata var. radiata	2	GCA_000741045.2	11	463.64	34.40
Setaria italica	2	GCA_001652605.1	9	477.54	46.28
Vitis vinifera	2	GCA_000003745.2	19	486.20	35.03
Panicum hallii var. hallii	2	GCA_003061485.1	9	487.47	46.95
Brassica oleracea var. oleracea	2	GCA_000695525.1	9	488.95	37.33
Musa balbisiana	2	GCA_004837865.1	11	492.78	37.99
Vitis riparia	2	GCA_004353265.1	19	500.11	34.43
Fagopyrum tataricum	2	GCA_002319775.1	8	505.88	38.76
Rhododendron williamsianum	2	GCA_009746105.1	13	532.29	34.22
Pyrus betulifolia	2	GCA_007844245.1	17	532.75	37.55
Panicum hallii	2	GCA_002211085.2	9	535.89	46.91
Gossypioides kirkii	2	GCA_005610355.1	12	538.06	33.26
Beta vulgaris subsp. vulgaris	2	GCA_002917755.1	9	540.53	35.85
Actinidia chinensis var. chinensis	2	GCA_003024255.1	29	553.84	35.80
Brassica oleracea	2	GCA_900416815.2	9	554.98	36.46
Lupinus angustifolius	2	GCA_002285895.2	20	557.91	33.39
Lupinus albus	2	GCA_010261695.1	25	558.90	36.82
Coffea canephora	2	GCA_900059795.1	11	568.61	38.89
Gossypium thurberi	2	GCA_004027125.1	13	582.01	34.24
Actinidia chinensis	2	GCA_009663005.1	29	653.93	35.64
Acer yangbiense	2	GCA_008009225.1	13	665.89	35.97
Solanum pinnatisectum	2	GCA_009887355.1	12	724.56	38.98
Alloteropsis semialata	2	GCA_004135705.1	9	747.77	46.29
Gossypium turneri	2	GCA_008044935.1	13	755.20	33.21
Fragaria nilgerrensis	2	GCA_010134655.1	7	772.25	37.20
Phoenix dactylifera	2	GCA_009389715.1	18	772.47	40.31
Spatholobus suberectus	2	GCA_004329165.1	9	798.47	31.77
Quercus mongolica	2	GCA_011696235.1	12	809.99	35.85
Sequoiadendron giganteum	2	GCA_007115665.2	11	8125.60	35.45
Benincasa hispida	2	GCA_009727055.1	12	912.95	34.99
Ananas comosus	3	GCA_902162155.2	25	381.91	38.24
Oryza punctata	4	GCA_000710525.1	24	224.65	43.60
Potentilla anserina	4	GCA_933775445.1	7	237.42	38.03
Arabidopsis suecica	4	GCA_019202805.1	13	272.25	36.00
Polygonum aviculare	4	GCA_934048045.1	10	352.07	39.33

Prunus fruticosa	4	GCA_018703695.1	8	375.29	37.92
Oryza minuta	4	GCA_000632695.1	24	451.66	43.81
Trifolium occidentale	4	GCA_012979555.1	8	501.00	37.47
Eragrostis tef	4	GCA_024500355.1	20	575.08	45.60
Triadica sebifera	4	GCA_023653625.1	22	739.40	32.31
Salvia splendens	4	GCA_004379255.2	22	806.49	38.85
Brassica juncea	4	GCA_020002515.1	18	904.41	38.11
Brassica napus	4	GCA_020379485.1	19	1001.50	37.12
Brassica carinata	4	GCA_016771965.1	17	1086.99	37.23
Cenchrus purpureus	4	GCA_022644695.1	14	2018.00	46.97
Panax japonicus	4	GCA_020205505.1	24	2028.86	33.96
Gossypium barbadense	4	GCA_018997955.1	26	2210.13	34.18
Gossypium tomentosum	4	GCA_018144435.1	26	2226.97	34.19
Gossypium mustelinum	4	GCA_017165895.1	26	2297.22	34.39
Gossypium hirsutum	4	GCA_024600755.1	26	2331.22	34.40
Arachis hypogaea	4	GCA_016103905.1	20	2490.84	36.29
Arachis hypogaea subsp. fastigiata	4	GCA_022829005.1	20	2535.29	36.41
Panax ginseng	4	GCA_020205605.1	24	3355.15	34.25
Panax quinquefolius	4	GCA_020205615.1	24	3565.32	34.12
Avena insularis	4	GCA_023614405.1	14	7519.20	43.51
Triticum dicoccoides	4	GCA_900184675.1	14	10495.00	45.96
Spirodela polyrhiza	4	GCA_001981405.1	20	136.67	42.72
Mikania micrantha	4	GCA_009363875.1	19	1790.64	36.20
Miscanthus sacchariflorus	4	GCA_002993905.1	19	2074.92	46.48
Gossypium darwinii	4	GCA_007990325.1	26	2182.96	34.14
Arachis monticola	4	GCA_003063285.2	20	2618.65	37.41
Isatis tinctoria	4	GCA_010577795.1	7	293.87	38.18
Coffea arabica	4	GCA_003713225.1	22	699.90	37.15
Panicum miliaceum	4	GCA_003046395.2	18	854.79	46.90
Trifolium repens	4	GCA_005869975.1	16	919.83	36.14
Brassica juncea var. tumida	4	GCA_001687265.1	18	954.86	37.34
Triticum turgidum subsp. durum	4	GCA_900231445.1	14	9964.34	46.00
Actinidia deliciosa	6	GCA_024454175.1	29	621.99	35.44
Echinochloa crus-galli	6	GCA_020466025.1	27	1340.74	45.88
Dendrocalamus latiflorus	6	GCA_017311315.1	70	2748.73	44.66
Avena sativa	6	GCA_023646675.1	21	10757.50	43.76
Triticum aestivum	6	GCA_918797515.1	21	14679.40	46.17
Camelina sativa	6	GCA_000633955.1	20	641.36	37.49
Ipomoea batatas	6	GCA_002525835.2	15	837.01	35.87
Fragaria x ananassa	8	GCA_019022445.1	28	805.68	39.35
Saccharum spontaneum	8	GCA_022457205.1	40	2761.16	44.57
Saccharum officinarum	8	GCA_020631735.1	80	6804.89	44.58
Utricularia gibba	16	GCA_002189035.1	14	100.69	41.14
Spirodela intermedia	-	GCA_902729315.2	18	136.68	42.01
Kandelia obovata	-	GCA_021464305.1	18	190.32	35.21
Acanthochlamys bracteata	-	GCA_019914995.1	20	197.97	35.06
Aristolochia contorta	-	GCA_022405105.1	7	210.54	39.41
Ceriops tagal	-	GCA_021533255.1	18	231.92	36.61
Cucurbita argyrosperma subsp. sor	-	GCA_018691285.1	20	255.12	36.64
Melastoma candidum	-	GCA_023653495.1	12	256.22	42.95
Aristolochia fimbriata	-	GCA_019845555.1	7	257.86	40.85
Boehmeria nivea	-	GCA_021020685.1	14	266.60	35.19
Arabidopsis thaliana x Arabidopsis	-	GCA_019202795.1	13	269.00	35.97
Boehmeria nivea var. tenacissima	-	GCA_018132145.1	14	270.21	35.22
Azadirachta indica	-	GCA_022749755.1	14	281.70	32.22
Aquilegia kansuensis	-	GCA_020826895.1	7	293.21	36.82

Begonia peltatifolia	-	GCA_022433055.1	15	310.00	37.96
Salix arbutifolia	-	GCA_021905355.1	19	324.00	33.98
Syntrichia caninervis	-	GCA_016097705.1	13	329.82	43.75
Tripterygium wilfordii	-	GCA_016880815.1	23	342.58	37.29
Brassica rapa subsp. trilocularis	-	GCA_017639395.1	10	346.51	36.54
Rhodamnia argentea	-	GCA_020921035.1	11	346.71	40.38
Ceratodon purpureus	-	GCA_014871845.1	13	349.46	42.50
Abrus pulchellus subsp. cantoniens	-	GCA_024086825.1	11	381.27	32.00
Angophora floribunda	-	GCA_014182895.1	11	388.37	39.13
Vitis rotundifolia	-	GCA_022557335.1	20	393.82	33.54
Corymbia calophylla	-	GCA_014182845.1	11	394.90	39.47
Sphagnum fallax	-	GCA_021442195.1	20	395.14	36.48
Corymbia maculata	-	GCA_014182735.1	11	403.98	39.21
Ziziphus jujuba var. spinosa	-	GCA_020796205.1	12	406.16	33.09
Raphanus raphanistrum subsp. lan	-	GCA_019706005.1	9	418.30	36.11
Raphanus raphanistrum subsp. rap	-	GCA_019706035.1	9	421.54	36.56
Populus deltoides	-	GCA_015852605.2	19	424.59	33.39
Sphagnum magellanicum	-	GCA_021904315.1	20	439.01	36.28
Gynochthodes officinalis	-	GCA_020080225.1	11	484.87	33.82
Raphanus raphanistrum x Raphanı	-	GCA_019705995.1	9	488.88	36.92
Leptodermis oblonga	-	GCA_016801395.1	11	497.30	36.09
Taxillus chinensis	-	GCA_023512835.1	9	521.91	40.20
Thymus quinquecostatus	-	GCA_024222315.1	13	528.68	40.40
Juglans nigra	-	GCA_002916485.2	16	532.00	36.40
Corymbia citriodora subsp. variega	-	GCA_014858505.1	11	544.19	39.66
Pugionium cornutum	-	GCA_018901935.1	11	550.39	36.06
Brassica oleracea var. capitata	-	GCA_018177695.1	9	565.47	36.77
Musa troglodytarum	-	GCA_023547065.1	10	603.59	39.58
Eucalyptus melliodora	-	GCA_004368105.3	11	639.60	39.27
Begonia loranthoides	-	GCA_022433065.1	19	671.67	36.63
Pohlia nutans	-	GCA_022496805.1	23	698.20	38.95
Cynara cardunculus var. scolymus	-	GCA_001531365.2	17	724.96	36.04
Chimonanthus praecox	-	GCA_022113865.1	11	737.03	36.61
Forsythia suspensa	-	GCA_023638005.1	14	737.53	33.64
Begonia darthvaderiana	-	GCA_022432945.1	15	785.39	38.29
Begonia masoniana	-	GCA_022432975.1	15	799.69	38.45
Ilex asprella	-	GCA_023539305.1	19	804.07	37.66
Chimonanthus salicifolius	-	GCA_013350335.1	11	853.43	36.87
Cichorium endivia	-	GCA_023376185.1	9	886.98	34.83
Saccharum hybrid cultivar	-	GCA_020102875.1	10	903.62	45.06
Gastrodia elata	-	GCA_016760335.1	18	1046.14	34.26
Sindora glabra	-	GCA_020226215.1	12	1113.92	28.02
Microstegium vimineum	-	GCA_022036555.1	23	1118.67	45.01
Tetracentron sinense	-	GCA_015143295.1	24	1161.36	38.45
Bretschneidera sinensis	-	GCA_023935145.1	9	1213.74	35.81
Codonopsis lanceolata	-	GCA_013146195.2	8	1273.26	37.21
Cichorium intybus	-	GCA_023525715.1	9	1278.75	35.51
Solanum lycopersicoides	-	GCA_022817965.1	12	1287.24	35.14
Gentiana dahurica var. dahurica	-	GCA_024500145.1	13	1416.54	37.77
Flaveria linearis	-	GCA_024085815.1	18	1654.55	37.00
Lactuca sativa	-	GCA_002870075.3	9	2388.97	38.72
Papilionanthe hookeriana x Papilioı	-	GCA_022702705.1	19	2570.13	35.27
Smallanthus sonchifolius	-	GCA_023525975.1	29	2716.52	37.44
Areca catechu	-	GCA_021397845.1	16	2823.08	41.33
Cymbidium sinense	-	GCA_021442155.1	20	3525.77	32.59
Artemisia tridentata subsp. tridenta	-	GCA_023558565.1	9	4198.57	36.17

Taxus chinensis	-	GCA_019776745.2	12	10238.00	36.78
Cycas panzhihuaensis	-	GCA_023213395.1	11	10482.70	35.89
Taxus wallichiana var. yunnanensis	-	GCA_018340775.1	12	10738.30	36.90
Triticum aestivum subsp. tibeticum	-	GCA_014338645.1	21	14708.20	46.24
Nyssa sinensis	-	GCA_008638375.1	22	1001.45	35.98
Rhamnella rubrinervis	-	GCA_007844105.2	12	245.35	34.05
Echium plantagineum	-	GCA_003412495.2	8	349.03	34.72
Carpinus fangiana	-	GCA_006937295.1	8	381.95	37.26
Pyrus ussuriensis x Pyrus commun	-	GCA_008932095.1	17	510.64	37.37
Juglans microcarpa x Juglans regia	-	GCA_004785595.1	16	527.90	36.40
Eragrostis curvula	-	GCA_007726485.1	7	603.07	45.64
Capsicum annuum var. glabriuscul	-	GCA_000950795.1	12	760.07	34.76
Physcomitrella patens	1;2	GCA_000002425.2	27	472.08	33.89
Malus domestica	2;3	GCA_022606005.1	17	754.68	38.16
Colocasia esculenta	2;3	GCA_014218235.1	14	2405.85	42.24
Dioscorea rotundata	2;3	GCA_002240015.2	21	2155.82	46.62
Ficus carica	2;3	GCA_009761775.1	13	333.44	34.41
Musa acuminata subsp. malaccens	2;3	GCA_000313855.2	11	472.23	40.72
Vanilla planifolia	2;3;4	GCA_023853775.1	14	1967.63	31.74
Arabidopsis arenosa	2;4	GCA_905216605.1	8	149.66	35.95
Dianthus caryophyllus	2;4	GCA_023091065.1	15	636.30	37.35
Solanum pimpinellifolium	2;4	GCA_014964335.1	12	808.10	34.50
Lonicera japonica	2;4	GCA_021464415.1	9	886.13	34.33
Bothriochloa decipiens	2;4	GCA_023333625.1	20	1213.27	45.99
Lolium perenne	2;4	GCA_019359855.1	7	2277.55	44.31
Hordeum marinum	2;4	GCA_022496015.1	7	3815.97	44.51
Adiantum capillus-veneris	2;4	GCA_014529385.2	30	4822.57	41.50
Mercurialis annua	2;4;6	GCA_937616625.1	8	453.17	33.92
Chenopodium formosanum	2;4;6;10	GCA_024500155.1	27	1629.91	36.04
Dioscorea cayenensis subsp. rotun	4;6;8	GCA_009730915.2	28	584.31	36.36
Panicum virgatum	4;6;8	GCA_016808335.1	18	1130.00	46.80
Dioscorea alata	4;8	GCA_020875875.1	20	480.03	36.41
Lactuca saligna	#N/A	GCA_902860255.1		2165.76	38.44
Oryza sativa f. spontanea	#N/A	GCA_000576065.1	12	337.95	42.94

Journal Title Here, 2022, pp. 1–9

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Paper

SUPPLEMENTARY MATERIAL

A Review and Benchmark of assembling nuclear genomes of plants.

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1], Gisele Nunes[1], Bent Petersen[3,4], Thomas Sicheritz-Pontén[3,4], and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, [3] Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Denmark, [4] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia.

*Corresponding author. renato.renison@gmail.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Advancements in sequencing technologies have allowed exponential growth of genomes deposited in public databases, with data generation outpacing the decrease in sequencing costs. Assembling large eukaryotes genomes, particularly plants, remains complex and expensive due to inherent characteristics like polyploidy, genome size, and repetitive regions. Many assembly software have been developed to address these issues, but a fair comparison of their effectiveness in assembling plant nuclear genomes is lacking in the literature. To address this gap, we collected information on 856 complete plant nuclear genomes deposited in the NCBI database by August 2022, along with associated data such as ploidy, genome size, chromosome number, GC content, sequencing technologies, and assembly software. We sequenced by simulation two diploid plant species (*Setaria italica* and *Oryza sativa*), generating short and long reads. Three pipelines used in recent publications of complete plant nuclear genomes were compared to identify optimal strategies and areas for improvement. WTDBG2 and SMARTdenovo generated assemblies with larger contig sizes and higher N50 values but with many assembly errors compared to the CANU and SOAPdenovo. SOAPdenovo generated fragmented assemblies, even when combined with long-read assemblers. CANU had fewer assembly errors, with good contig length and N50 values. Quickmerge joined different assemblies, increasing N50 values without introducing many errors. PurgeHaplotigs identified syntenic contigs from highly heterozygous regions, increasing the final assembly's N50 values. SSPACE and GapCloser formed new scaffolds and filled over 90% gaps. Our results highlight areas for improvement in existing pipelines and suggest opportunities for developing new assembly strategies.

Keywords: Plant genome, benchmark, pipeline comparison, genome assembly.

Table S2. Assembly and completeness metrics of the results obtained for *Setaria italica* in the three pipelines executed. Under each pipeline is shown in hours the total execution time of the pipeline.

Setaria italica	***N50 (bp)***	*54,898,046*					
	Min contig (bp)	39,546,705					
	Max contig (bp)	65,039,919					
	Maxn (bp)	2,617,739					
	Total contigs	9					
	Total bases (bp)	458,457,535					
	Total gaps	13,592,106					
	C BUSCOs (%)	97.8					
	S BUSCOs (%)	87.4					
	D BUSCOs (%)	10.4					
	F BUSCOs (%)	0.2					
	M BUSCOs (%)	2.0					
Pipeline 1 (7,84 h)	**Ferramenta**	***WTDBG2 >***	***Smartdenovo >***	***Quickmerge >***	***Pilon***		
	N50 (bp)	134,427	153,059	165,240	164,919		
	Min contig (bp)	4,648	14,949	4,648	4,649		
	Max contig (bp)	1,800,461	1,133,783	1,840,191	1,830,816		
	Total contigs	4,228	3,343	3,500	3,500		
	Total bases (bp)	324,647,463	341,757,627	344,166,036	343,926,106		
	C BUSCOs (%)	87.9	87.7	87.5	87.8		
	S BUSCOs (%)	86.0	85.3	84.4	84.7		
	D BUSCOs (%)	1.9	2.4	3.1	3.1		
	F BUSCOs (%)	1.1	1.0	1.1	1.0		
	M BUSCOs (%)	11.0	11.3	11.4	11.2		
	Tempo exec (h)	0.28	7.13	0.001	0.43		
Pipeline 2 (30,06 h)	**Ferramenta**	***CANU >***	***>***	***>***	***Pilon >***	***PurgeHaplotigs***	
	N50 (bp)	115,387			115,387	143,918	
	Min contig (bp)	8,165			8,165	9,783	
	Max contig (bp)	876,462			876,488	876,488	
	Total contigs	7,462			7,462	4,000	
	Total bases (bp)	417,151,309			416,978,699	347,883,128	
	C BUSCOs (%)	88.7			88.8	88.7	
	S BUSCOs (%)	80.6			80.8	86.1	
	D BUSCOs (%)	8.1			8.0	2.6	
	F BUSCOs (%)	1.0			1.0	1.0	
	M BUSCOs (%)	10.3			10.2	10.3	
	Tempo exec (h)	20.45			4.3	5.31	
Pipeline 3 (56,12 h)	**Ferramenta**	***CANU >***	***SOAPdenovo2 >***	***Quickmerge >***	***SSPACE >***	***GapCloser***	**PurgeHaplotigs**
	N50 (bp)	115,387	9,634	77,810	77,810	77,802	153,767
	Min contig (bp)	8,165	100	100	100	100	100
	Max contig (bp)	876,462	200,646	927,130	927,130	927,144	927,144
	Maxn (bp)		4,060	4,060	4,060	2,417	78
	Total contigs	7,462	4,477,581				
	Total scaffolds		848,077	657,240	657,150	657,150	5,461
	Scaffolds with N's		67,652	41,171	41,193	17,979	987
	Total bases (bp)	417,151,309	437,953,995	556,928,331	556,926,614	557,131,920	368,906,870
	Total gaps		10,964,747	6,334,089	6,334,089	35,826	1,474
	C BUSCOs (%)	88.7	91.0	96.5	96.5	96.8	95.2
	S BUSCOs (%)	80.6	89.2	89.4	89.4	89.4	92.3
	D BUSCOs (%)	8.1	1.8	7.1	7.1	7.4	2.9
	F BUSCOs (%)	1.0	2.7	0.8	0.8	0.7	0.7
	M BUSCOs (%)	10.3	6.3	2.7	2.7	2.5	4.1
	Tempo exec (h)	20.45	10.18	0.08	0.28	0.43	24.7

Maxn indicates the longest sequence of N's found in a contig or scaffold;
C = Complete; S = Complete and single-copy; D = Complete and duplicated; F = Fragmented; M = Missing.

Journal Title Here, 2022, pp. 1–9

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Paper

SUPPLEMENTARY MATERIAL

A Review and Benchmark of assembling nuclear genomes of plants.

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1], Gisele Nunes[1], Bent Petersen[3,4], Thomas Sicheritz-Pontén[3,4], and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, [3] Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Denmark, [4] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia.

*Corresponding author. renato.renison@gmail.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Advancements in sequencing technologies have allowed exponential growth of genomes deposited in public databases, with data generation outpacing the decrease in sequencing costs. Assembling large eukaryotes genomes, particularly plants, remains complex and expensive due to inherent characteristics like polyploidy, genome size, and repetitive regions. Many assembly software have been developed to address these issues, but a fair comparison of their effectiveness in assembling plant nuclear genomes is lacking in the literature. To address this gap, we collected information on 856 complete plant nuclear genomes deposited in the NCBI database by August 2022, along with associated data such as ploidy, genome size, chromosome number, GC content, sequencing technologies, and assembly software. We sequenced by simulation two diploid plant species (*Setaria italica* and *Oryza sativa*), generating short and long reads. Three pipelines used in recent publications of complete plant nuclear genomes were compared to identify optimal strategies and areas for improvement. WTDBG2 and SMARTdenovo generated assemblies with larger contig sizes and higher N50 values but with many assembly errors compared to the CANU and SOAPdenovo. SOAPdenovo generated fragmented assemblies, even when combined with long-read assemblers. CANU had fewer assembly errors, with good contig length and N50 values. Quickmerge joined different assemblies, increasing N50 values without introducing many errors. PurgeHaplotigs identified syntenic contigs from highly heterozygous regions, increasing the final assembly’s N50 values. SSPACE and GapCloser formed new scaffolds and filled over 90% gaps. Our results highlight areas for improvement in existing pipelines and suggest opportunities for developing new assembly strategies.

Keywords: Plant genome, benchmark, pipeline comparison, genome assembly.

Table S3. Assembly and completeness metrics of the results obtained for Oryza sativa in the three pipelines executed.

Oryza sativa	***N50 (bp)***	*32,301,089*					
	Min contig (bp)	24,008,571					
	Max contig (bp)	44,746,243					
	Maxn (bp)	200					
	Total contigs	12					
	Total bases (bp)	392,033,263					
	Total gaps	500					
	C BUSCOs (%)	97.7					
	S BUSCOs (%)	95.3					
	D BUSCOs (%)	2.4					
	F BUSCOs (%)	0.5					
	M BUSCOs (%)	1.8					
Pipeline 1	**Ferramenta**	***WTDBG2 >***	***Smartdenovo >***	***Quickmerge >***	***Pilon***		
	N50 (bp)	683,889	1,431,093	2,254,883	2,254,788		
	Min contig (bp)	3,314	15,282	3,314	3,313		
	Max contig (bp)	4,784,173	5,464,793	7,816,664	7,816,446		
	Total contigs	1,241	511	403	403		
	Total bases (bp)	367,761,831	392,878,033	395,718,844	395,628,824		
	C BUSCOs (%)	97.0	97.7	95.3	95.6		
	S BUSCOs (%)	94.7	95.2	88.0	88.3		
	D BUSCOs (%)	2.3	2.5	7.3	7.3		
	F BUSCOs (%)	0.5	0.5	0.6	0.5		
	M BUSCOs (%)	2.5	1.8	4.1	3.9		
	Tempo exec (h)						
Pipeline 2	**Ferramenta**	***CANU >***	***>***	***>***	***Pilon >***	***PurgeHaplotigs***	
	N50 (bp)	370,565			370,570	404,735	
	Min contig (bp)	10,700			10,705	10,705	
	Max contig (bp)	2,807,783			2,807,774	2,807,774	
	Total contigs	3,465			3,465	1,887	
	Total bases (bp)	417,723,799			417,719,727	387,157,479	
	C BUSCOs (%)	97.4			97.6	97.6	
	S BUSCOs (%)	92.7			92.9	95.0	
	D BUSCOs (%)	4.7			4.7	2.6	
	F BUSCOs (%)	0.6			0.6	0.6	
	M BUSCOs (%)	2.0			1.8	1.8	
Pipeline 3	**Ferramenta**	***CANU >***	***SOAPdenovo2 >***	***Quickmerge >***	***SSPACE >***	***GapCloser >***	***PurgeHaplotigs***
	N50 (bp)	370,565	13,640	272,847	272,847	272,847	417,057
	Min contig (bp)	10,700	100	100	100	100	100
	Max contig (bp)	2,807,783	214,251	3,102,429	3,102,429	3,102,488	3,102,488
	Maxn (bp)		2,906	2,184	2,184	106	75
	Total contigs	3,465					
	Total scaffolds		878,433	643,032	643,029	643,029	2,248
	Scaffolds with N's		51,905	25,804	25,805	8,247	61
	Total bases (bp)	417,723,799	416,758,941	511,858,208	511,858,115	511,915,326	387,089,085
	Total gaps		10,082,933	4,253,772	4,253,772	10,087	144
	C BUSCOs (%)	97.4	94.4	97.2	97.2	97.2	97.2
	S BUSCOs (%)	92.7	92.4	93.2	93.2	93.2	94.7
	D BUSCOs (%)	4.7	2.0	4.0	4.0	4.0	2.5
	F BUSCOs (%)	0.6	1.9	0.5	0.5	0.5	0.5
	M BUSCOs (%)	2.0	3.7	2.3	2.3	2.3	2.3

Maxn indicates the longest sequence of N's found in a contig or scaffold;
C = Complete; S = Complete and single-copy; D = Complete and duplicated; F = Fragmented; M = Missing.

Journal Title Here, 2022, pp. 1–9

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Paper

SUPPLEMENTARY MATERIAL

A Review and Benchmark of assembling nuclear genomes of plants.

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1], Gisele Nunes[1], Bent Petersen[3,4], Thomas Sicheritz-Pontén[3,4], and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, [3] Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Denmark, [4] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia.

*Corresponding author. renato.renison@gmail.com

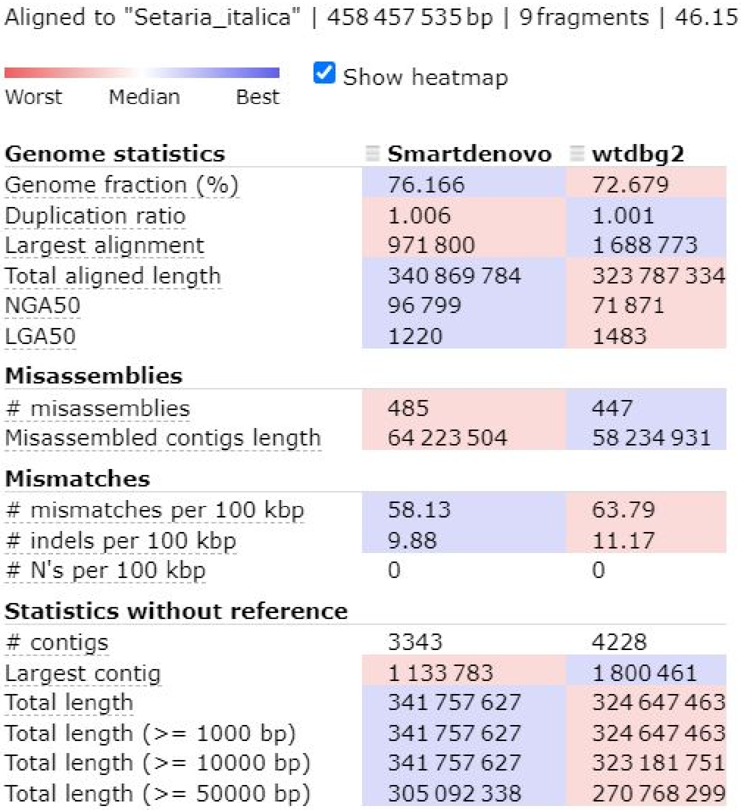
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Advancements in sequencing technologies have allowed exponential growth of genomes deposited in public databases, with data generation outpacing the decrease in sequencing costs. Assembling large eukaryotes genomes, particularly plants, remains complex and expensive due to inherent characteristics like polyploidy, genome size, and repetitive regions. Many assembly software have been developed to address these issues, but a fair comparison of their effectiveness in assembling plant nuclear genomes is lacking in the literature. To address this gap, we collected information on 856 complete plant nuclear genomes deposited in the NCBI database by August 2022, along with associated data such as ploidy, genome size, chromosome number, GC content, sequencing technologies, and assembly software. We sequenced by simulation two diploid plant species (*Setaria italica* and *Oryza sativa*), generating short and long reads. Three pipelines used in recent publications of complete plant nuclear genomes were compared to identify optimal strategies and areas for improvement. WTDBG2 and SMARTdenovo generated assemblies with larger contig sizes and higher N50 values but with many assembly errors compared to the CANU and SOAPdenovo. SOAPdenovo generated fragmented assemblies, even when combined with long-read assemblers. CANU had fewer assembly errors, with good contig length and N50 values. Quickmerge joined different assemblies, increasing N50 values without introducing many errors. PurgeHaplotigs identified syntenic contigs from highly heterozygous regions, increasing the final assembly's N50 values. SSPACE and GapCloser formed new scaffolds and filled over 90% gaps. Our results highlight areas for improvement in existing pipelines and suggest opportunities for developing new assembly strategies.

Keywords: Plant genome, benchmark, pipeline comparison, genome assembly.

Table S4. Assembly metrics found by QUAST for the *Setaria italica* results of Smartdenovo and Wtdbg2 assemblers. In red are the worst values and in blue are the best metric values.



Aligned to "Setaria_italica" | 458 457 535 bp | 9 fragments | 46.15 % G+C

☑ Show heatmap

Genome statistics	Smartdenovo	wtdbg2
Genome fraction (%)	76.166	72.679
Duplication ratio	1.006	1.001
Largest alignment	971 800	1 688 773
Total aligned length	340 869 784	323 787 334
NGA50	96 799	71 871
LGA50	1220	1483
Misassemblies		
# misassemblies	485	447
Misassembled contigs length	64 223 504	58 234 931
Mismatches		
# mismatches per 100 kbp	58.13	63.79
# indels per 100 kbp	9.88	11.17
# N's per 100 kbp	0	0
Statistics without reference		
# contigs	3343	4228
Largest contig	1 133 783	1 800 461
Total length	341 757 627	324 647 463
Total length (>= 1000 bp)	341 757 627	324 647 463
Total length (>= 10000 bp)	341 757 627	323 181 751
Total length (>= 50000 bp)	305 092 338	270 768 299

Journal Title Here, 2022, pp. 1–9

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Paper

SUPPLEMENTARY MATERIAL

A Review and Benchmark of assembling nuclear genomes of plants.

Renato R. M. Oliveira[1,2]*, Santelmo Vasconcelos[1], Gisele Nunes[1], Bent Petersen[3,4], Thomas Sicheritz-Pontén[3,4], and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale, Belém, Pará, Brazil, [2]Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, [3] Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Denmark, [4] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia.

*Corresponding author. renato.renison@gmail.com

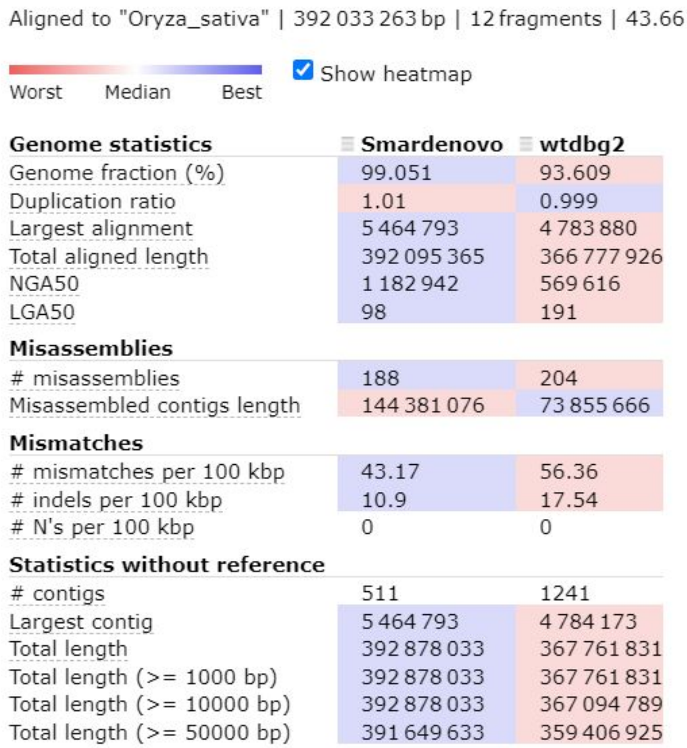
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Advancements in sequencing technologies have allowed exponential growth of genomes deposited in public databases, with data generation outpacing the decrease in sequencing costs. Assembling large eukaryotes genomes, particularly plants, remains complex and expensive due to inherent characteristics like polyploidy, genome size, and repetitive regions. Many assembly software have been developed to address these issues, but a fair comparison of their effectiveness in assembling plant nuclear genomes is lacking in the literature. To address this gap, we collected information on 856 complete plant nuclear genomes deposited in the NCBI database by August 2022, along with associated data such as ploidy, genome size, chromosome number, GC content, sequencing technologies, and assembly software. We sequenced by simulation two diploid plant species (*Setaria italica* and *Oryza sativa*), generating short and long reads. Three pipelines used in recent publications of complete plant nuclear genomes were compared to identify optimal strategies and areas for improvement. WTDBG2 and SMARTdenovo generated assemblies with larger contig sizes and higher N50 values but with many assembly errors compared to the CANU and SOAPdenovo. SOAPdenovo generated fragmented assemblies, even when combined with long-read assemblers. CANU had fewer assembly errors, with good contig length and N50 values. Quickmerge joined different assemblies, increasing N50 values without introducing many errors. PurgeHaplotigs identified syntenic contigs from highly heterozygous regions, increasing the final assembly's N50 values. SSPACE and GapCloser formed new scaffolds and filled over 90% gaps. Our results highlight areas for improvement in existing pipelines and suggest opportunities for developing new assembly strategies.

Keywords: Plant genome, benchmark, pipeline comparison, genome assembly.

Table S5. Assembly metrics found by QUAST for the *Oryza sativa* results of Smartdenovo and Wtdbg2 assemblers. In red are the worst values and in blue are the best metric values.



Aligned to "Oryza_sativa" | 392 033 263 bp | 12 fragments | 43.66 % G+C

Worst Median Best

☑ Show heatmap

Genome statistics	Smardenovo	wtdbg2
Genome fraction (%)	99.051	93.609
Duplication ratio	1.01	0.999
Largest alignment	5 464 793	4 783 880
Total aligned length	392 095 365	366 777 926
NGA50	1 182 942	569 616
LGA50	98	191
Misassemblies		
# misassemblies	188	204
Misassembled contigs length	144 381 076	73 855 666
Mismatches		
# mismatches per 100 kbp	43.17	56.36
# indels per 100 kbp	10.9	17.54
# N's per 100 kbp	0	0
Statistics without reference		
# contigs	511	1241
Largest contig	5 464 793	4 784 173
Total length	392 878 033	367 761 831
Total length (>= 1000 bp)	392 878 033	367 761 831
Total length (>= 10000 bp)	392 878 033	367 094 789
Total length (>= 50000 bp)	391 649 633	359 406 925