UNIVERSIDADE FEDERAL DE SÃO CARLOS
CAMPUS SOROCABA
CENTRO DE CIÊNCIAS E TECNOLOGIAS PARA A SUSTENTABILIDADE
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA E MONITORAMENTO AMBIENTAL

LAÍS ROSSETTO FERRAZ DE BARROS

COMPARAÇÃO DO DESEMPENHO DE DIFERENTES ESTRATÉGIAS DE MONTAGEM DO GENOMA DE *Bertholletia excelsa* Bonpl. (Lecythidaceae)

Sorocaba
2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CAMPUS SOROCABA
CENTRO DE CIÊNCIAS E TECNOLOGIAS PARA A SUSTENTABILIDADE
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA E MONITORAMENTO AMBIENTAL

LAÍS ROSSETTO FERRAZ DE BARROS

COMPARAÇÃO DO DESEMPENHO DE DIFERENTES ESTRATÉGIAS DE MONTAGEM DO GENOMA DE *Bertholletia excelsa* Bonpl. (Lecythidaceae)

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia e Monitoramento Ambiental da Universidade Federal de São Carlos, *campus* Sorocaba, para obtenção do título de mestre em Biotecnologia e Monitoramento Ambiental.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Orientação: Prof. Dra. Karina Martins

Sorocaba
2020

DEDICATÓRIA

Dedico aos meus pais e ao meu irmão pelo apoio e incentivo.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por essa oportunidade e por me guiar durante todo o processo.

À Profª Dra. Karina Martins pela oportunidade, pela idealização do projeto, pelo apoio e dedicação.

Aos meus pais, Ednir e Dourival, e ao meu irmão Regis pelo apoio incondicional.

Aos meus tios Edna e Francis por acreditarem no meu potencial, pela amizade e carinho.

Ao meu tio Eduardo por sempre torcer pelo meu sucesso.

À minha prima Yedda pela amizade, consolo e por me ajudar com a revisão de textos.

Ao meu primo Braulio pela atenção e incentivo.

Ao meu namorado Augusto pela paciência e companheirismo.

À Juliane Gouvêa pela amizade e pelo trabalho de coaching.

Aos demais amigos e familiares que colaboraram de alguma forma para meu desenvolvimento.

Ao Rodrigo Theodoro Rocha, da Universidade de Brasília (UnB), por realizar a montagem do genoma com o algoritmo Canu.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida.

Ao Programa de Pós-Graduação em Biotecnologia e Monitoramento Ambiental da Universidade Federal de São Carlos *campus* Sorocaba pela oportunidade.

À Universidade Federal de São Carlos a que devo a minha formação superior.

“Sempre faço o que não consigo fazer para aprender o que não sei.”

PABLO PICASSO

RESUMO

BARROS, L. R. F. Comparação do desempenho de diferentes estratégias de montagem do genoma de *Bertholletia excelsa* Bonpl. (Lecythidaceae). 2020. 68 f. Dissertação (Mestrado em Biotecnologia e Monitoramento Ambiental) – Universidade Federal de São Carlos, *campus* Sorocaba, Sorocaba, 2020.

A Castanheira-do-Brasil (*Bertholletia excelsa)* é uma espécie arbórea nativa da Amazônia pela qual existe preocupação referente à sua sustentabilidade ecológica. Desse modo, sequenciar e montar seu genoma pode contribuir para definição de estratégias para a conservação da diversidade genética, manutenção das populações naturais, além de auxiliar na compreensão do sucesso de populações futuras. Portanto, os objetivos do projeto consistem em montar o genoma de *Bertholletia excelsa*, comparar diferentes metodologias de montagem, e definir a metodologia que obtenha a melhor qualidade. A metodologia consistiu em sequenciar amostras de DNA de um indivíduo adulto de *Bertholletia excelsa* com a tecnologia *Single Molecule Real Time* da Pacific Biosciences. Essa etapa, resultou em uma cobertura genômica de 187× e um N50 de *reads* brutas de 14,02 kb. Em seguida, procedeu-se com a estimativa do tamanho do genoma haploide a partir da distribuição *k-mer*, resultando em ~596 Mpb. Posteriormente, foram testados cinco *assemblers* (wtdbg2, MECAT2, SMARTdenovo, Flye e Canu) com seis coberturas genômicas (47×, 63×, 97×, 126×, 187× e 60× com *reads* corrigidas por Canu). QUAST foi utilizado para comparar a eficiência desses métodos através dos parâmetros: *contig* N50, número de *contigs*, L50 e tamanho da montagem em relação a estimativa do tamanho do genoma. BUSCO foi usado para determinar a completude gênica. LAI *score* validou a qualidade considerando a contiguidade e a integridade de sequências repetitivas. Dessa forma, conclui-se que a montagem com a melhor qualidade foi SMARTdenovo com cobertura genômica de 187×, resultando em um *contig* N50 de ~2,6 Mb, completude gênica de 95,1%, LAI *score* igual a 10,53; totalizando a montagem em 649.349.366 pb.

Palavras-chave: MECAT2. Wtdbg2. SMARTdenovo. Canu. Flye.

ABSTRACT

BARROS, L. R. F. Performance comparison of different strategies for assembling the *Bertholletia excelsa* Bonpl. genome (Lecythidaceae). 2020. 68 p. Dissertation (Master's degree in Biotechnology and Environmental Monitoring) – Universidade Federal de São Carlos, *campus* Sorocaba, Sorocaba, 2020.

The Brazil Nut (*Bertholletia excelsa*) is an arboreal species, native to the Amazon forest, by which there's a concern referring to its ecological sustainability. Therefore, sequencing and assembling the genome might contribute to the conservation of genetic diversity and maintenance of natural populations, besides assisting the comprehension of the success of future populations. Therefore, the objectives are to assemble the genome of *Bertholletia excelsa*, compare different assembly methodologies, and define the methodology that results in the best quality. The methodology consisted in sequencing DNA samples from an adult individual of *Bertholletia excelsa* with Pacific Biosciences is Single Molecule Real Time. This step resulted in a genomic coverage of 187× and a N50 of raw reads of 14,02 kb. Following up, the haploid genome size was estimated from k-mer distribution (k=22), resulting in ~596Mpb. Subsequently, there were tested five assemblers (wtdbg2, MECAT2, SMARTdenovo, Flye and Canu) with six genomic coverages (47×, 63×, 97×, 126×, 187× and 60× with corrected reads by Canu). QUAST was used to compare the efficiency of these methods through the following parameters: contig N50, number of contigs, L50 and assembly size in relation to genome size estimative. BUSCO was used to determine the genomic completeness. The LAI score validated its quality, considering the contiguity and integrity of repetitive sequences. In conclusion, the best assembly was SMARTdenovo with genomic coverage of 187×, resulting in a contig N50 of ~2,6 Mb, genic completeness of 95,1% and LAI score of 10,53; totaling an assembly of 649.349.366 pb.

Keywords: MECAT2. Wtdbg2. SMARTdenovo. Canu. Flye.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BUSCO Benchmarking Universal Single-Copy Orthologs

CTAB Brometo de cetiltrimetilamônio

DNA Ácido desoxirribonucleico

FBG fuzzy-Bruijn graph

FDDs Fatores de distância

Gb Giga bases

GB Gigabyte

GC Guanina-Citosina

gDNA DNA genômico

Hi-C Captura de Conformação Cromossômica

kb Kilo bases

LAI Long-terminal repeat (LTR) assembly index

Mb Mega bases

MB Megabyte

NCBI National Center for Biotechnology Information

NCGAS National Center for Genome Analysis Support

NGS Next Generation Sequencing

OLC Overlap-layout-consensus

PacBio Pacific Biosciences

pb Par de bases

PFGE Eletroforese em gel de campo pulsado

RAM Random Access Memory

RO Rondônia

SMRT Single Molecule Real Time

STRs Simple Tandem Repeats

TB Terabyte

TEs Elementos Transponíveis

UnB Universidade de Brasília

USD Dólar Americano

vCPU CPUvirtual

SUMÁRIO

1. INTRODUÇÃO

Bertholletia excelsa Bonpl., também conhecida como Castanheira-do-Brasil (em inglês *Brazil nut tree*), castanha-do-Brasil ou castanha-do-Pará é uma espécie arbórea nativa da Floresta Amazônica pertencente à família Lecythidaceae (BALDONI, WADT, PEDROZO, 2020; RIBEIRO et al., 1999). Suas sementes são um produto comercial importante; em 2016, por exemplo, foram destinados ao mercado interno no Brasil, USD $ 30 milhões obtidos com a comercialização (IBGE, 2016). A castanha é utilizada principalmente como alimento *in natura* ou industrializado, também é matéria-prima para cosméticos, além de ser objeto de diversas pesquisas: desde sua ação anti-inflamatória em pacientes submetidos à hemodiálise, como fonte de gene para transgenia com objetivo de aumentar o valor nutricional de outros alimentos (CAMARGO et al., 2010; PINTO et al., 2014; ARAGÃO et al., 1999). Por outro lado, a intensa atividade extrativista ao longo de décadas e o desflorestamento são uma ameaça à regeneração natural dessa espécie; a exemplo disso, ações de conservação no âmbito estadual, nacional e internacional classificam a castanha-do-brasil como espécie vulnerável (CAMARGO et al., 2010; CNCFlora, 2012; COEMA-PA, 2007; MINISTÉRIO DO MEIO AMBIENTE, 2008; INTERNACIONAL UNION FOR CONSERVATION OF NATURE, 2011).

Os avanços da tecnologia de sequenciamento (*next generation sequencing* – NGS) que no decorrer da década de 2000 possibilitaram a transição da abordagem por genética clássica – em que apenas algumas regiões do genoma são analisadas - para análise em escala genômica (EKBLOM e WOLF, 2014). Assim, a capacidade de conhecer a sequência total do genoma dos organismos permite, por exemplo, a realização de estudos comparativos, como a compreensão de diferenças genéticas que provocam doenças; e estudos evolutivos – como potencial adaptativo das populações e respostas às mudanças ambientais (METZKER, 2010; EKBLOM e WOLF, 2014; ALLENDORF et al. 2010; OUBORG et al. 2010). É possível também guiar o melhoramento de plantas pela seleção de genes relacionados a produtividade, resistência ao estresse ambiental e aos patógenos (LI et al., 2017; YANO et al, 2016; REIG-VALIENTE, et al., 2018; HAZZOURI et al.,2015). Em outras áreas de estudo pode suscitar questões acerca de depressão por endogamia e de expressão gênica (EKBLOM e WOLF, 2014; KARDOS et al., 2016; BÉRÉNOS et al., 2016; HOFFMAN et al., 2014). Dessa forma, sequenciar e montar o genoma de

Bertholletia excelsa pode contribuir para que a comunidade científica aumente sua compreensão sobre a biodiversidade, a conservação de espécies e ecossistemas ameaçados, podendo até resultar na descoberta de genes úteis, proteínas e novas vias metabólicas (LEWIN et al., 2018).

Contudo, a tecnologia de sequenciamento não é capaz de criar uma sequência única com a totalidade do genoma, pois a técnica fragmenta todo o DNA (MARTINS, 2013). Essa limitação tecnológica é intensificada conforme a complexidade do genoma estudado: poliploidia, genoma grande, alta taxa de heterozigose, alto conteúdo GC, elementos transponíveis, e elementos repetitivos (constituem em aproximadamente 50% do genoma humano, por exemplo), que são características de genomas complexos, e particularmente de plantas (SOHN e NAM, 2018; TÜRKTAS et al., 2014). Assim, essas problemáticas podem ser específicas do projeto de sequenciamento e montagem do genoma da Castanha-do-Brasil, visto que *Bertholletia excelsa* é uma espécie alógama, ou seja, possui fecundação cruzada, uma condição favorável para heterozigosidade e conteúdos repetitivos (MAUÉS, 2002; BALDONI, WADT, PEDROZO, 2020). A possibilidade de alta taxa de heterozigose na espécie é reforçada com a conclusão do estudo de Cabral et al. (2017), que demonstrou alta diversidade genética e ausência de consanguinidade em uma população de *Bertholleita excelsa*. Além disso, Buckley et al. (1988) concluíram a espécie é diploide; e Prance e Mori (1979) identificaram que o número cromossômico é 17 (n=17).

Portanto, possivelmente quanto maior a complexidade do genoma, maior será a dificuldade na determinação da posição dos fragmentos e menor será a contiguidade da montagem. Levando em consideração que as características do genoma podem se apresentar de forma diversa e específica em diferentes organismos, não há como pressupor uma metodologia única, correta e executável em todos os projetos de montagem de genoma (*genome assembly*). Entretanto, mesmo sendo indispensável utilizar e/ou testar diferentes abordagens para o genoma em questão, os últimos estudos convergem para uma tendência de metodologias que vem trazendo os melhores resultados de qualidade da montagem em diversos organismos, como o tipo de sequenciamento e o algoritmo de montagem, por exemplo. Para isso, é preciso compreender os desafios acerca do

tema para auxiliar no planejamento e execução de um projeto de sequenciamento e montagem do genoma.

2. OBJETIVOS

O objetivo principal do projeto consiste em sequenciar e montar o genoma de *Bertholletia excelsa* trabalhando com diferentes metodologias de montagem *de novo* do genoma; comparar essas metodologias de montagem, e definir a metodologia que obtenha a melhor qualidade da montagem.

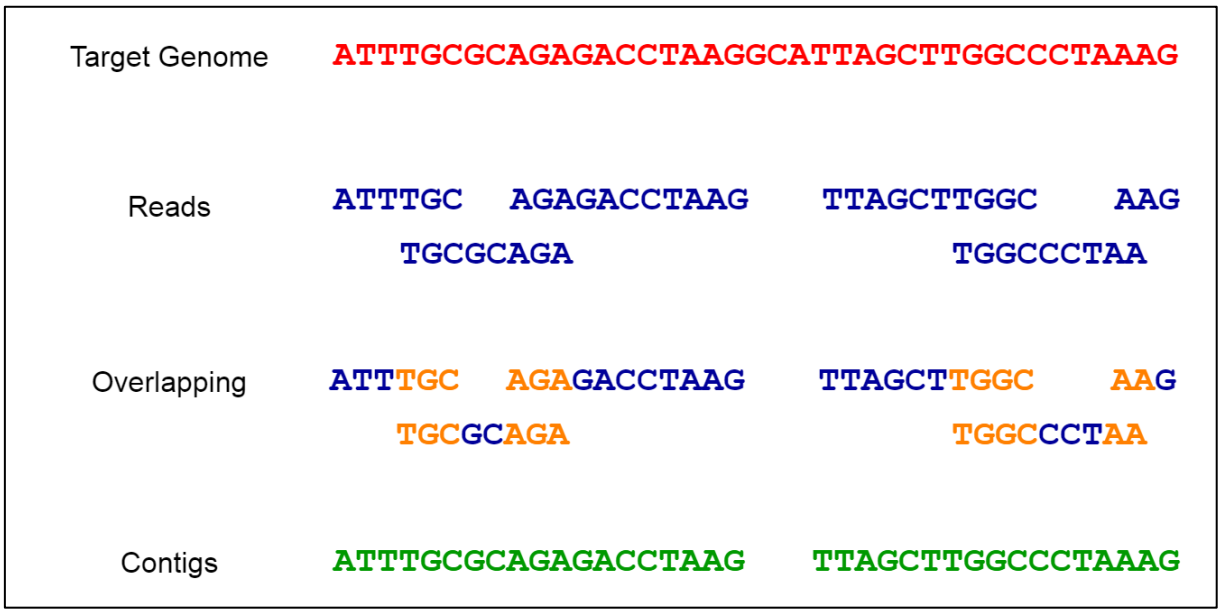
3. REVISÃO DE LITERATURA

3.1. O QUE SIGNIFICA MONTAR UM GENOMA?

As atuais tecnologias de sequenciamento de DNA não são capazes de sequenciar todo o genoma de uma espécie, apenas pequenos pedaços do genoma podem ser sequenciados, e essas peças devem ser montadas para determinar o genoma das espécies estudadas. Logo, a montagem do genoma consiste em procedimentos que buscam organizar um grande número de sequências de DNA de maneira linear, com o objetivo final de representar a molécula de DNA que compõe cada cromossomo da espécie estudada (MARTINS, 2013). Em projetos de sequenciamento, todo o DNA de uma fonte (geralmente um único organismo) é primeiro cortado em milhões de pequenos fragmentos aleatórios. Esses fragmentos são sequenciados em máquinas automatizadas; o produto do sequenciamento é chamado de *reads* (leituras, em português) (EL-METWALLY et al., 2013). Em seguida, essas leituras são unidas para formar leituras contíguas mais longas − conhecidas como *contigs* − por um programa de computador conhecido como *assembler* (montador). Para montagem correta é preciso obter sobreposição (*overlap*) entre as *reads,* o que requer uma alta profundidade de leitura (*read depth ou coverage*). Assim, *contigs* são juntados para formar trechos mais longos – os *scaffolds* (Figura 1) (EL-METWALLY et al., 2013; EKBLOM e WOLF, 2014). Contudo, haverá lacunas entre os trechos, os quais são chamados de *gaps* (intervalos), que também podem ser preenchidos com o caractere não informativo “N” (EKBLOM e WOLF, 2014). No último passo, *scaffolds* são unidos em grupos de ligação ou em cromossomos; entretanto, há uma grande dificuldade de obter esses

dados, então a alternativa é utilizar cromossomos putativos (EKBLOM e WOLF, 2014).

Figura 1 – Esquema simplificado da montagem do genoma.



Fonte: Mallawaarachchi, 2020.

Entre os passos descritos anteriormente, também há passos de avaliação da qualidade da montagem. Em geral, independentemente da tecnologia de sequenciamento, o fluxo de trabalho é o mesmo. Primeiramente, as leituras são examinadas quanto à qualidade geral, presença de adaptadores e a presença de possíveis contaminantes (DEL ANGEL, 2018).

Ademais, em projetos de sequenciamento e montagem, *assemblers* são testados em paralelo com a finalidade de comparar os resultados e corrigir erros. Com isso, muitos *assemblers* são executados novamente com novos parâmetros para a validação da montagem (DEL ANGEL, 2018).

Existem também duas abordagens para a montagem do genoma: a abordagem comparativa e a abordagem *de novo*. Durante a montagem comparativa, também conhecida como montagem baseada em referência, um genoma de referência do mesmo organismo ou de uma espécie intimamente relacionada é usado como um mapa para guiar o processo de montagem, alinhando os fragmentos montados (EL-METWALLY et al., 2013; KYRIAKIDOU et al., 2018). Esse tipo de abordagem é comum quando se pretende re-sequenciar genomas para corrigir ou

estender montagens (KYRIAKIDOU et al., 2018). Já, a montagem *de novo* (termo *de novo* vem do latim e significa "desde o princípio") nenhum mapa ou orientação está disponível para a montagem do genoma, portanto, a montagem *de novo* é usada para reconstruir genomas que não são semelhantes aos genomas previamente sequenciados (MARTIN e WANG, 2011). Dessa forma, a montagem de sequências genômicas de espécies não modelo − que não possuem genoma de referência − é mais complexa (CALI, 2018). Assim, a decisão para usar a estratégia de montagem *de novo* ou baseada na referência, baseia-se na aplicação biológica, no custo, no esforço necessário para atingir a acurácia necessária e considerações de tempo de montagem (METZKER, 2010).

A montagem *de novo* pode ser dada por diferentes abordagens, seja pelo uso de apenas um tipo de sequenciador e um algoritmo, ou até mesmo uso de diferentes tecnologias de sequenciamento com diferentes *assemblers*, introduzindo também ferramentas para aumento da qualidade a fim de melhorar a qualidade do genoma montado; esse é o caso de estudos com *Arabidopsis thaliana, Arachis hypogaea, Broussonetia papyrifera, Capsicum annuum* e rosa (MICHAEL et al., 2018; CHEN et al., 2019; PENG et al., 2019; HULSE-KEMP et al., 2018; RAYMOND et al., 2018).

Assim, é importante ressaltar que não há uma metodologia única e correta para projetos de montagem de genoma. Entretanto, a qualidade pode divergir entre uma técnica e outra. Em suma, ainda é válido ressaltar que o tipo de método empregado depende da qualidade que o pesquisador pretende alcançar, dos recursos financeiros e computacionais disponíveis e da especificidade do genoma estudado (BRADNAM et al., 2013; EKBLOM e WOLF, 2014).

A título de conhecimento, podemos verificar o número total de publicações (até 14 de janeiro de 2020) com montagens de genomas através do banco de dados “Genome” do NCBI (National Center for Biotechnology Information) (Tabela 1). Esse banco de dados fornece informações sobre genomas, incluindo sequências, mapas, cromossomos, montagens e anotações.

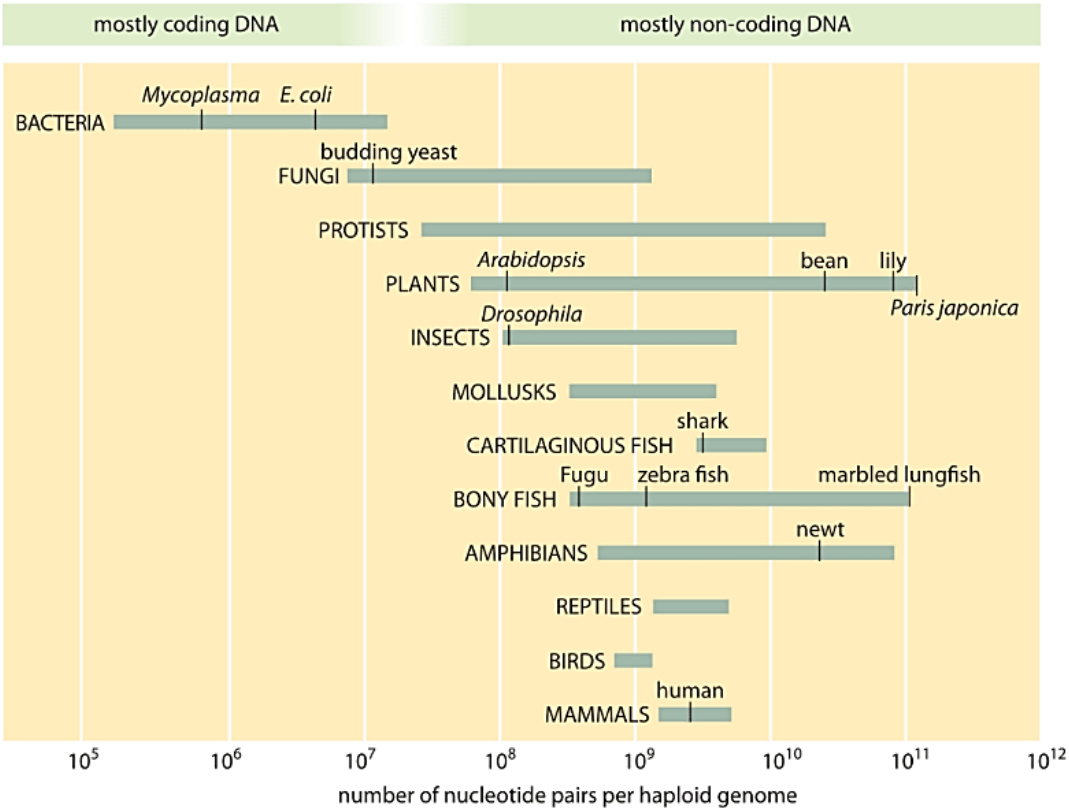
Tabela 1 – Número total de organismos, por grupo, com montagem de genoma obtidos no banco de dados “Genome” do NCBI, até 14 de janeiro de 2020.

Grupos	Número de organismos
Animais	1683
Plantas (terrestres)	467
Fungos	2073
Protistas	325
Bactérias (Reino)	25984
Vírus (Reino)	16730

Fonte: Elaborado pelo autor a partir do NCBI - Banco de Dados de Genoma, 2020.

Observando a tabela é possível inferir que, em geral, quanto maior o genoma e sua complexidade, menor é a quantidade de montagens disponíveis, visto que bactérias possuem entre 10^5 e 10^7 pb (pares de bases), e plantas podem chegar a 10^{11} pb, por exemplo (Figura 2, MILO e PHILLIPS, 2016). Além disso, o número expressivo de genomas bacterianos pode ser explicado pela abundância de organismos e o interesse médico sobre doenças humanas relacionadas a esses tipos de patógenos (JUDGE et al. 2016; SU, SANTOLA, READ, 2019; DECANO et al., 2019). Um dos motivos que justificam o número pequeno de montagens de genomas de plantas é porque plantas podem apresentar genomas poliploides − fator que provoca dificuldades na montagem (MEYERS e LEVIN, 2006; WOOD et al. 2009). Richards (2015) constatou que existiam aproximadamente o mesmo número de genomas de plantas e mamíferos, apesar de plantas serem um Reino e mamífero uma Ordem.

Figura 2 – Comparação do tamanho do genoma entre diferentes grupos de organismos.



Fonte: MILO e PHILLIPS, 2016.

3.2. QUALIDADE DA MONTAGEM

Finalizada a montagem *de novo* é preciso validar o genoma montado, pois na ausência do genoma referência, o resultado gerado é um “esboço” do genoma (*draft genome*), ou seja, uma hipótese da sequência original (EKBLOM e WOLF, 2014; XIAO et al., 2016). Dessa forma, segundo Shendure e Ji (2008) é fundamental ter métricas que reflitam confiabilidade, reprodutibilidade e relevância biológica do genoma montado. Assim, podemos indicar as métricas amplamente utilizadas com base no estado-da-arte como número de *contigs,* e o L50 (uma variação dessa métrica) que mede o número de *contigs* que representam 50% da montagem; assim os valores mais baixos desses parâmetros indicam montagens menos fragmentadas (KREMER; MCBRIDE; PINTO, 2017; JAYAKUMAR e SAKAKIBARA, 2017).

Já a estatística N50 é definida como o comprimento mínimo do *contig* necessário para cobrir 50% do genoma, ou seja, um N50 de 10.000 pb significa que 50% das bases montadas estão contidas em *contigs* de pelo menos 10 000 pb de

comprimento (XIAO et al., 2016; LI et al., 2017; JAYAKUMAR e SAKAKIBARA, 2017; SOHN e NAM, 2018; AYLING; CLARK; LEGGETT, 2019). Assim, uma boa montagem segundo esse parâmetro é aquela com o valor mais alto de N50 para *contig* ou *scaffold* (LI et al., 2017). Há também as estatísticas NG50, NA50 ou NGA50 que são resultantes da modificação do N50, e incorporam o tamanho esperado do genoma, dessa forma, resultam em um N50 normalizado (BRADNAM et al.,2013; SOHN e NAM, 2018). É importante ressaltar que N50 e suas variações são métricas padrões para avaliar contiguidade e não a precisão da montagem (EKBLOM e WOLF, 2014; KREMER; MCBRIDE; PINTO, 2017; AYLING; CLARK; LEGGETT, 2019).

Por fim, outro aspecto que também indica a qualidade da montagem é a escolha da cobertura de leitura (refere-se à quantidade de vezes que o genoma foi sequenciado) (METZKER, 2010; MILLER; KOREN; SUTTON, 2010; SCHATZ; DELCHER; SALZBERG, 2010). De acordo com a experiência de Ekblom e Wolf (2014), a amostragem de 100x a 50x de cobertura em uma biblioteca de tamanho de inserção curta pode melhorar significativamente algumas etapas do processo de montagem. Já, Li et al. (2017) afirmam que coberturas de genomas de 90x a 95x são boas frente a presença de conteúdos repetitivos.

Mesmo compreendendo os principais parâmetros, é notório que, a avaliação da qualidade é um desafio, porque não há uma consolidação sobre os valores estatísticos (EKBLOM e WOLF, 2014; KYRIAKIDOU et al., 2018). Pensando nisso o VGP (Vertebrate Genome Project) estabeleceu um padrão de qualidade para todos os genomas liberados que consiste em: tamanho mínimo de N50 de 1 Mb para *contigs* e 10 Mb para *scaffolds*, que a frequência de erro não seja superior a 1 a cada 10.000 bases, que as variantes sejam confirmadas por várias tecnologias e que pelo menos 90% da sequência esteja atribuída aos cromossomos e ao *haplotype phased* (NATURE BIOTECHNOLOGY, 2018).

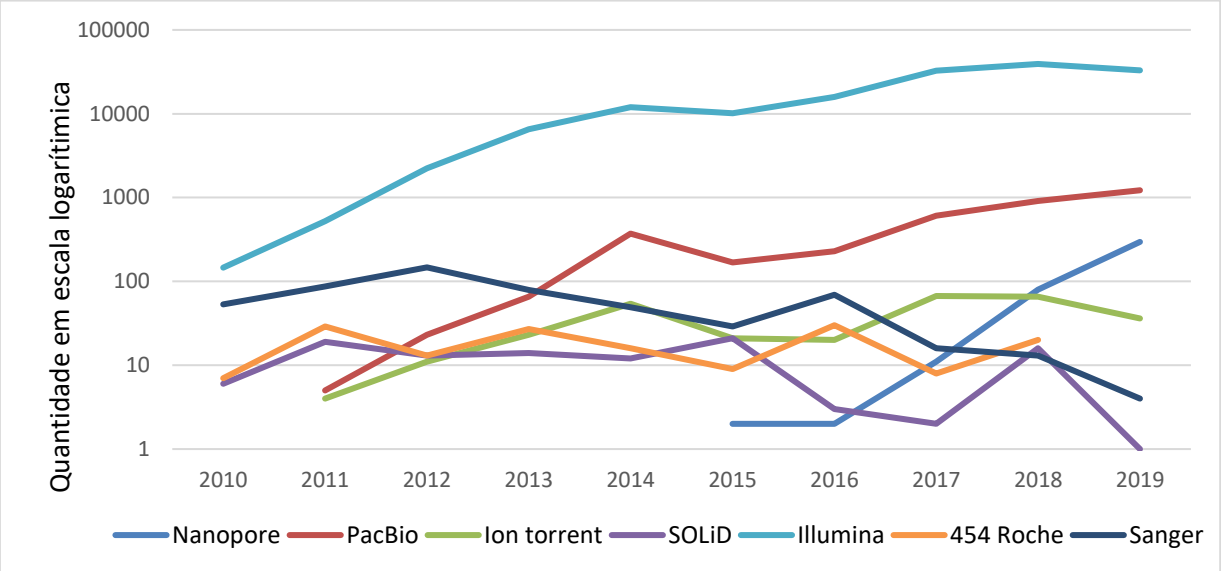
Assim, a análise de qualidade pode ser obtida, por exemplo, pelas ferramentas QUAST (Gurevich et al., 2013), REAPR (Hunt et al., 2013), ALE (Clark et al., 2013) e GMvalue (Kosugi et al., 2015) (XIAO et al., 2016; SOHN e NAM, 2018). Além disso, para análise de completude da montagem em nível genético BUSCO (SIMÃO et al., 2015) e CEGMA (PARRA; BRADNAM; KORF, 2007) são

exemplos de ferramentas usadas para anotação genética (JAYAKUMAR e SAKAKIBARA, 2017; PAAJANEN et al., 2019; SOHN e NAM, 2018).

3.3. DESAFIOS DA MONTAGEM DO GENOMA INERENTES À TÉCNICA: *LONG E SHORT READS*

Primeiramente, antes de entendermos os desafios relacionados as técnicas *long e short reads,* é importante considerar quais plataformas de sequenciamento têm sido utilizadas nos últimos anos; assim foi realizada uma pesquisa no banco de dados “Assembly” do NCBI (fornece informações sobre a estrutura de genomas montados, nomes das montagens, outros metadados, relatórios estatísticos e links para dados de sequência) com as palavras-chave de cada tipo de sequenciamento (“Sanger”, “454 Roche”, “Illumina”, “SOLiD”, “Ion torrent”, “PacBio”, “Nanopore”) e computados em número de projetos que envolvam tais sequenciamentos por ano, nos últimos nove anos (2010 a 2019) (Figura 3).

Figura 3 – Número de projetos de montagem de genoma por tipo de plataforma de sequenciamento.



Fonte: Elaborado pelo autor a partir do NCBI - Banco de Dados de Montagem, 2020.

Neste gráfico (Figura 3) é possível ver que Illumina é o sequenciamento predominantemente mais utilizado entre 2010 e 2019. Illumina foi desenvolvido em 2006, e é definido como sequenciamento por síntese – SBS (HEATHER e CHAIN, 2016; SHENDURE e JI, 2008). Mesmo sendo um tipo se sequenciamento

considerado de segunda geração, Illumina possui tecnologia mais barata do que as de terceira geração (PacBio e Nanopore) possibilitando um acesso maior a esse tipo de recurso (ver Tabela 2; NCGAS, 2019). Além disso, possui a vantagem de ser uma técnica precisa na determinação das bases: a taxa de erro é inferior a 1% (RICE E GREEN, 2019). Em contrapartida, Illumina produz, em geral, leituras curtas (*short reads*) entre 75–300 bp (RICE E GREEN, 2019); esse comprimento de leitura é um grande desafio na montagem de genomas, pois gera montagens fragmentadas, causadas por regiões repetitivas maiores do que o comprimento de leitura (MAGI, 2017).

Tabela 2 – Comparação dos custos entre as plataformas de sequenciamento.

	Illumina	**PacBio**	**Nanopore**
Custo por Gb	NextSeq (2x150, 120Gb): US$ 42/Gb; MiSeq (2x300, 15Gb): US$ 133/Gb	Sequel (CLR de 20Gb, HiFi de 50Gb): US$ 30-50/Gb	PromethION: US$ 45-130/Gb; MinION: US$ 16-48/Gb
Outros Custos	Preparação da biblioteca abaixo de US$ 100/bilioteca, sem incluir a seleção de tamanho, etc.	Preparação da biblioteca: aproximadamente US$ 4-500; aproximadamente US$ 1-1400 cada célula.	Kit de sequenciamento: US$ 600; flow cells: US$ 7-900/cada

Fonte: Adaptado de National Center for Genome Analysis Support - NCGAS, 2019.

Pode-se observar também o sequenciamento PacBio em destaque a partir de 2014. PacBio é uma plataforma de sequenciamento de terceira geração, lançada em 2010, chamada de sequenciamento em tempo real de moléculas únicas - *Single Molecule Real Time* - SMRT (HEATHER e CHAIN, 2016). Esse destaque nos últimos anos não é somente pelo fato de PacBio ser umas das tecnologias mais atuais de sequenciamento, mas pela sua eficiência em resolver ambiguidades e reduzir substancialmente os *gaps*. Como é capaz de produzir leituras longas (*long reads*), acima de 10 kb; é útil para enfrentar e superar os desafios da montagem de genomas grandes, repetitivos e complexos, como é o caso dos genomas de plantas (RICE E GREEN, 2019; KYRIAKIDOU, 2018).

Através da Figura 3, é possível inferir inclusive que o método Sanger se destaca principalmente nos primeiros anos, pois foi a primeira plataforma de

sequenciamento; porém, com a evolução das tecnologias há um decréscimo do seu uso. Já, as demais tecnologias possuem valores e flutuações ao longo do período muito semelhantes; exceto Nanopore, que aparece apenas em 2015 e apresenta progressão em 2017. Nanopore é o tipo de sequenciamento mais recente, comercialmente disponível em 2015 (OXFORD NANOPORE TECHNOLOGIES, 2018). Essa tecnologia detecta uma corrente iônica que é passada através do nanoporo (membrana de polímero eletricamente resistente). A corrente é alterada à medida que as bases passam pelo poro em diferentes combinações (OXFORD NANOPORE TECHNOLOGIES, 2018; CLARKE et al., 2009). Nanopore foi recentemente usada para gerar conjuntos de sequências altamente contíguos para espécies de *Brassica* e banana (BELSER et al., 2018; KERSEY, 2019). Outro aspecto importante sobre as atuais tecnologias de sequenciamento de leitura longa, como as fornecidas pela Oxford Nanopore Technologies (Nanopore) ou Pacific Biosciences (PacBio), é a produção de leituras de qualidade inferior, com taxas de erro em torno de 10% (RHOADS e AU, 2015; JAIN, 2016). Entretanto, essa realidade está se modificando, pois a plataforma mais recente da Pacific Biosciences, a Sequel II promete uma precisão maior que 99,9% (Pacific Biosciences, 2019).

Por fim, nessa pesquisa no banco de dados “Assembly” do NCBI também se observou que diversos projetos utilizaram mais de um tipo de sequenciamento; geralmente uma tecnologia de *short reads* (leituras curtas) com uma de *long reads* (leituras longas), em especial, muitos deles utilizam Illumina e posteriormente PacBio. Isso, possivelmente deve-se ao fato de que tecnologias de leituras curtas são mais baratas e possuem maior exatidão na identificação de nucleotídeos, e as tecnologias de leituras longas podem orientar melhor a disposição de *reads* e cobrir *gaps*.

3.4. DESAFIOS DA MONTAGEM DO GENOMA RELACIONADOS ÀS PROPRIEDADES DO GENOMA

3.4.1. Repetições

Repetições ou elementos repetitivos são fragmentos de DNA que ocorrem em múltiplas cópias no genoma, perto de centrômeros, telômeros ou satélites nos cromossomos (DEL ANGEL et al.,2018; EKBLOM e WOLF, 2014; SOHN e NAM,

2018). As sequências repetitivas podem ser encontradas em tandem (chamadas de STRs - *Simple Tandem Repeats*) ou como repetições intercaladas dispersas por todo o genoma (CORDAUX e BATZER, 2009; KOITO e IKEDA, 2013; MEHROTRA et al., 2014; EKBLOM e WOLF, 2014). As repetições em tandem são classificadas de acordo com o tamanho dos motivos de repetição em microssatélites (motivos de 1 a 6 nucleotídeos), minissatélites (motivos de 7 a 100 nucleotídeos) e satélites (motivos com mais de 100 nucleotídeos) (FOULONGNE-ORIOL et al., 2013). As repetições intercaladas compreendem os elementos transponíveis (TEs) que são fragmentos de DNA capazes de mudar de localização no genoma por movimentação física (FESCHOTTE; JIANG; WESSLER, 2002).

Microssatélites e TEs são abundantes nos genomas eucariotos; em humanos isso corresponde a 50%, e em plantas de 50% a 90%; por exemplo, em uva, cerca de 41% do genoma é formado por elementos repetitivos; já em milho, representam mais de 80% do genoma (CORDAUX e BATZER, 2009; KOITO e IKEDA, 2013; SCHNABLE et al., 2009; JAILLON et al., 2007; MEHROTRA e GOYAL, 2014).

A quantidade e a distribuição de repetições representam um grande desafio na montagem de genomas de plantas, na maioria das vezes, sequências altamente repetitivas levam a caminhos ambíguos e a uma montagem fragmentada com lacunas no genoma final (DEL ANGEL et al., 2018; DESCHAMPS e LLACA, 2016; SOHN e NAM, 2016). Isso ocorre devido ao montador (*assembler*) possuir dificuldade em determinar corretamente o local e o comprimento dessas sequências que são altamente semelhantes. Para resolver essa problemática é preciso utilizar novas tecnologias de sequenciamento que produzam leituras longas o suficiente para incluírem essas repetições e as sequências que as flanqueiam. Isso também pode ser resolvido com o aumento da profundidade de leitura (*coverage*), principalmente se o tipo de sequenciamento for baseado em leituras curtas (*short reads*) (DEL ANGEL et al., 2018; SOHN e NAM, 2018).

3.4.2. Conteúdo GC

Sequências repetitivas podem ser ricas em AT, por exemplo, como em CG (HESLOP-HARRISON e SCHWARZACHER, 2011). Projetos que envolvam sequenciamento de genomas devem avaliar se o genoma em questão possui alto conteúdo repetitivo em GC, pois a plataforma Illumina possui dificuldades em

identificar tais nucleotídeos em sequência (DEL ANGEL et al.,2018; EKBLOM e WOLF, 2014; CHEN et al., 2013). Nesse caso, o sequenciamento dessas regiões pode resultar em baixíssimas coberturas; por isso, recomenda-se aumentar a cobertura de leitura (*coverage*) ou utilizar novas tecnologias como PacBio e Nanopore, pois essas plataformas não apresentam esse entrave (DEL ANGEL et al.,2018; EKBLOM e WOLF, 2014; CHEN et al., 2013).

3.4.3. Tamanho do Genoma

O conhecimento do tamanho do genoma é importante para a triagem de abordagens técnicas em projetos genômicos (uso de leituras curtas para genomas pequenos e leituras longas para genomas grandes, por exemplo), avaliar as diferenças entre as contiguidades das montagens, e avaliar o tamanho da montagem final (avaliar se a montagem representa o genoma em sua totalidade) (HESLOP-HARRISON e SCHWARZACHER, 2011; EKBLOM e WOLF, 2014).

Se o conhecimento do tamanho do genoma não estiver disponível para a espécie de interesse é necessário obter uma estimativa do tamanho do genoma antes de prosseguir com os dados de sequência. Para isso, a citometria de fluxo é geralmente utilizada para estimar o tamanho do genoma (EKBLOM e WOLF, 2014; HESLOP-HARRISON e SCHWARZACHER, 2011; DEL ANGEL et al., 2018). Alternativamente a essa técnica, há uma abordagem baseada em *k-mers* (fragmentos de leituras com quantidade fixa de nucleotídeos), o tamanho do genoma pode ser estimado utilizando o número total de *k-mers* dividido pela cobertura do sequenciamento (EKBLOM e WOLF, 2014; LI e HARKESS, 2018).

Outras formas de estimar o tamanho do genoma é através de bancos de dados online para plantas (http://data.kew.org/cvalues), animais (http://www.genomesize.com) e fungos (http://www.zbi.ee/fungal-genomesize) (EKBLOM e WOLF, 2014; HESLOP-HARRISON e SCHWARZACHER, 2011; DEL ANGEL et al., 2018). Dentre as plantas, o tamanho médio do genoma (haploide) de angiospermas é de 5800 Mb, de angiospermas basais a média é de 2300 Mb e eudicotiledôneas é de 2800 Mb. Quanto às gimnospermas o tamanho médio dos genomas é de 18.200 Mb (para mais informações ver Tabela 3; HESLOP-HARRISON e SCHWARZACHER, 2011).

Tabela 3 – Valores mínimos e máximos de tamanho de genoma de diferentes grupos de plantas.

	Valor mínimo		Valor máximo	
Grupo de Plantas	**Espécie**	**Tamanho (Mb)**	**Espécie**	**Tamanho (Mb)**
Angisperma	*Cardamine amara*	49	*Paris japonica*	149185
Angsperma basal	*Aristolochia pallida*	279,3	*Illicium henryi*	14357
Eudicotilledônea	*Genlisea tuberosa*	65	*Viscum album*	100845
Gimnosperma	*Gnetum ula*	2205	*Pinus ayacahuite*	35280

Fonte: Elaborado pelo autor, com base no banco de dados de plantas http://data.kew.org/cvalues/cvalOrigReference.html, 2020.

3.4.4. Heterozigose e Ploidia

Outro desafio para a montagem *de novo* do genoma de plantas é a questão da poliploidia (COMAI, 2005). A poliploidia é um evento preponderante na evolução das plantas, são responsáveis desde alterações moleculares até alterações ecológicas. Estima-se que aproximadamente 80% das plantas sejam poliploides (MEYERS e LEVIN, 2006). Outros estudos apontam que 30% a 50% das espécies de angiospermas sejam poliploides (CUI et al., 2006; SOLTIS et al., 2014).

Sabe-se que um genoma altamente heterozigoto cria dificuldades para as leituras de sequências de alelos homólogos, pois os algoritmos de montagem possuem a tendência de reduzir as diferenças alélicas para uma sequência consenso, ou seja, eles geram uma montagem haploide (DEL ANGEL et al.,2018; EKBLOM e WOLF, 2014). Na prática, as diferenças alélicas ao invés de serem montadas juntas, serão montadas separadamente; de modo geral, no resultado final as regiões heterozigotas serão colocadas duas vezes, no caso de organismos diploides, ou seja, o tamanho da montagem será maior que o tamanho do genoma (DEL ANGEL et al., 2018; PRYSZCZ e GABALDÓN, 2016). Por isso, recomenda-se utilizar organismos homozigotos para a montagem do genoma ou recorrer às tecnologias de terceira geração, pois as leituras longas ajudam a distinguir homólogos individuais (JIAO e SCHNEEBERGER, 2017).

3.5. DESAFIOS DA MONTAGEM RELACIONADOS ÀS ESTRATÉGIAS DE MONTAGEM

Os algoritmos de montagem do genoma utilizam todas as leituras ao mesmo tempo para alinhá-las umas às outras para identificar as regiões onde pelo menos dois segmentos de leitura se sobrepõem. Estas sobreposições têm o propósito de formar um resultado linear e contínuo. Quanto mais curtas as sequências, maior a quantidade de sobreposições necessárias para executar esta tarefa.

Em geral, existem duas classes de montadores (*assemblers*): montadores baseados em abordagem *overlap-layout-consensus* (OLC) e montadores *De Bruijn graph* (grafos De Bruijn) (NAGARAJAN e POP, 2013). Grafos De Bruijn é um algoritmo que quebra os segmentos de leitura em *k-mers* antes de montá-los em *contigs.* A abordagem de grafos forma *contigs* ligando dois fragmentos *(k-mers*) com k ou mais nucleotídeos sobrepostos (COMPEAU; PEVZNER; TESLER 2011). Já OLC identifica regiões de sobreposição das leituras. Graficamente cada leitura é representada por um nó e as sobreposições são as arestas que unem os nós (COMMINS; TOFT; FARES, 2009). Assim, o algoritmo define o melhor caminho para uma sequência consenso através do gráfico; o *draft genome* (“esboço” do genoma) é então representado por essa sequência consenso que representa o genoma (COMMINS; TOFT; FARES, 2009). Montadores (*assemblers*) OLC são bem adaptados à maioria das tecnologias de leituras longas (*long reads*), e a abordagem mais utilizada para a montagem de dados de leitura curta (*short reads*) é baseada nos grafos De Bruijn, (EKBLOM e WOLF, 2014).

Desse modo, softwares como o SOAPdenovo, ALLPATHS-LG, ABySS e o Velvet são programas de montagem comuns para leituras curtas, que se baseiam em grafos De Bruijn (LUO et al., 2012; GNERRE et al., 2011; SIMPSON et al., 2009; ZERBINO e BIRNEY, 2008). Em contrapartida, softwares como Celera, Arachne e PCAP foram importantes, principalmente, para dados do sequenciamento Sanger; e usam a abordagem OLC (BATZOGLOU et al., 2002; HUANG et al., 2003; DENISOV et al., 2008). No princípio, Celera teve uma estratégia razoável para um *draft genome* (“esboço” do genoma) devido à abrangência de sequências repetitivas, resultando em uma de alta qualidade (SHENDURE, 2017).

No entanto, nos últimos anos com o desenvolvimento de novas tecnologias de sequenciamento de leituras longas (*long reads*), houve um rápido desenvolvimento

de novos algoritmos para lidar com os novos tipos de dados genômicos. Assim, o sequenciamento PacBio possui um algoritmo de montagem *de novo* específico, o FALCON, também baseado na abordagem OLC (XIAO et al., 2016). Outros exemplos de algoritmos que lidam com leituras longas (*long reads*) são Canu e Miniasm (KOREN et al. 2017; LI, 2016). Canu corrige os erros de sequenciamento na sua etapa inicial, com objetivo de melhorar a precisão das bases nas leituras, além de eliminar leituras (*reads)* que não apresentam sobreposição (KOREN et al. 2017). Em seguida, segue com as sobreposições entre as leituras corrigidas. Por outro lado, Miniasm não possui a etapa de correção de erros, por isso depende da precisão das leituras iniciais; e pode ser necessária uma etapa de polimento (*polishing*) (CALI, 2018).

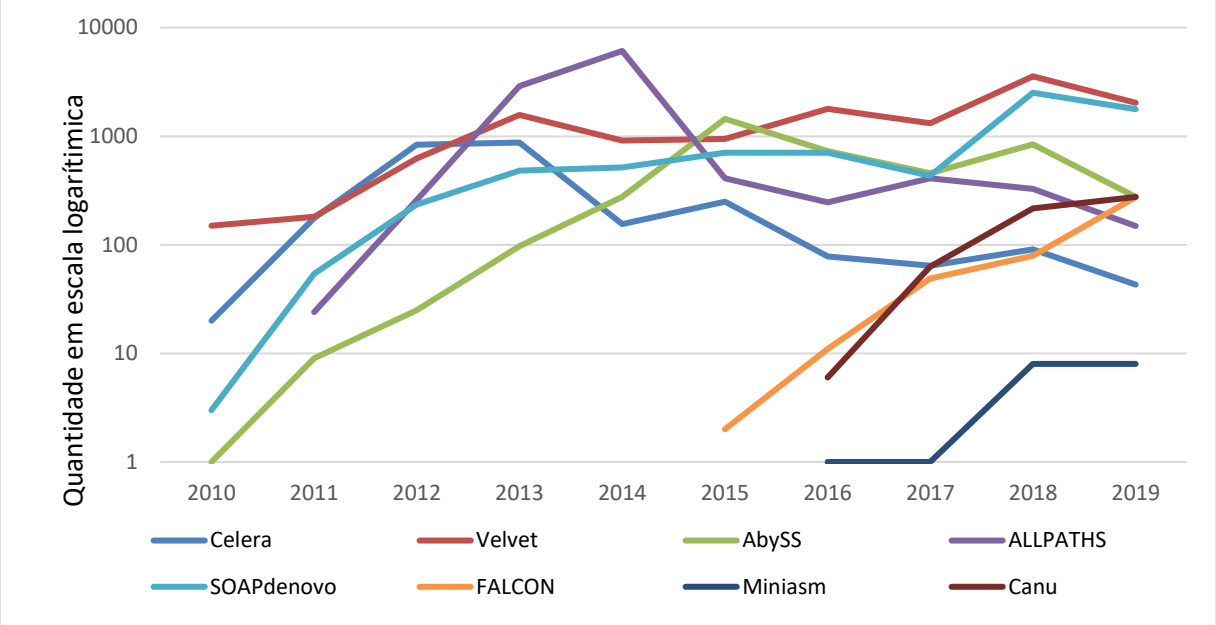
Para avaliar o estado da arte em tecnologias de montagem foram publicados em 2011 o *Assemblathon e o Genome Assembly Gold-Standard Evaluation* (GAGE). São projetos que avaliam o desempenho de vários *assemblers* em uma estrutura competitiva com conjuntos de dados simulados e reais (SCHATZ; WITKOWSKI; MCCOMBIE, 2012).

Numa perspectiva similar, a fim de demonstrar como os algoritmos têm sido utilizados ao longo do tempo, foram plotados dados do banco de dados “Assembly” do NCBI, com as palavras-chave dos montadores (*assemblers)* mais citados na literatura no período de 2010 a 2019 (Figura 4). Assim, analisando, nota-se que os resultados de Celera confirmam destaque no seu uso logo nos primeiros anos devido a metodologia de Sanger, decaindo posteriormente devido às tecnologias de sequenciamento de segunda geração.

Consequentemente, Velvet, AbySS, ALLPATHS e SOAPdenovo são os montadores de leituras curtas (*short reads)* mais utilizados e estão em evidência. Nessa pesquisa é possível observar que esses algoritmos são responsáveis por analisar, principalmente, dados gerados pelo sequenciamento Illumina, já demonstrado anteriormente.

Contudo, o grande número de trabalhos que usaram o ALLPATHS no ano de 2014 deve-se em especial a uma pesquisa que ao invés de agrupar todas as sequências obtidas e lançar como um arquivo único, acabou lançando separadamente os milhares de trechos no banco de dados; mas, sua importância foi nítida também no ano de 2013.

Figura 4 - Número de projetos por tipo de algoritmo de montagem.



Fonte: Elaborado pelo autor a partir do NCBI - Banco de Dados de Montagem, 2020.

No que diz respeito a regularidade de Velvet ao longo do tempo deve-se primeiramente por ser um dos primeiros algoritmos para dados de tecnologia de nova geração. A Tabela 4 apresenta informações sobre a data de criação de montadores (*assemblers)* e seu uso em relação ao tipo de leitura. Em segundo lugar, Velvet é um programa adaptado a leituras longas (*long reads)*, ou seja, adaptado aos dados de tecnologias atuais de sequenciamento; assim, essa característica pode explicar o aumento expressivo em 2018.

Outro algoritmo que teve destaque em 2018 é SOAPdenovo. Todavia, a relevância do número de projetos pode ser explicada pelo seu uso em ferramentas atuais de *polishing* baseadas em sequenciamento Illumina. Além disso, SOAPdenovo é um dos montadores (*assemblers)* mais recentes para leituras curtas (*short reads)* (Tabela 3), por isso foi menos frequente nos anos anteriores.

Por fim, FALCON, Canu e Miniasm aparecem somente nos últimos anos (2015 a 2019), pois, como mencionado anteriormente, foram criados para receber dados gerados pelos sequenciamentos de terceira geração. Assim, sua utilização ainda está em evolução.

Tabela 4 – Algoritmos de montagem, tipos de leituras processadas e data de criação.

Nome	Tipo de leitura	Referência
SUTTA	longa e curta	(Narzisi e Mishra, 2010)
ARACHNE	longa	(Batzoglou et al., 2002)
CABOG	longa e curta	(Miler et al., 2008)
Celera	longa	(Myers et al., 2000)
Edena	curta	(Hernandez et al., 2008)
Minimus (AMOS)	longa	(Sommer et al., 2007)
Newbler	longa	454/Roche
CAP3	longa	(Huang e Madan, 1999)
PCAP	longa	(Huang et al., 2003)
Phrap	longa	(Green, 1999)
Phusion	longa	(Mullikin e Ning, 2003)
TIGR	longa	(Sutton et al., 1995)
ABySS	curta	(Simpson et al., 2009)
ALLPATHS	curta	(Butler et al., 2008/2011)
Euler	longa	(Pevzner et al., 2001)
Euler-SR	curta	(Chaisson e Pevzner, 2008)
Ray	longa e curta	(Boisvert et al., 2010)
SOAPdenovo	curta	(Li et al., 2010)
Velvet	longa e curta	(Zerbino e Birney, 2008/2009)
PE-Assembler	curta	(Ariyaratne e Sung, 2011)
QSRA	curta	(Bryant et al., 2009)
SHARCGS	curta	(Dohm et al., 2007)
SHORTY	curta	(Hossain et al., 2009)
SSAKE	curta	(Warren et al., 2007)
Taipan	curta	(Schmidt et al., 2009)
VCAKE	curta	(Jeck et al., 2007)
FALCON	longa	(Chin et al.,2016)
Canu	longa	(Koren et al., 2017)
Miniasm	longa	(Li, 2016)

Fonte: Adaptado de NARZISI e MISHRA, 2011.

Softwares como, GapCloser, GapFiller e iMAGE são usados para solucionar *gaps* (TÜRKTAS et al., 2014; EKBLOM e WOLF, 2014). Outra ferramenta interessante é o RepeatMasker, que detecta homologias e está entre os programas mais comuns para detectar elementos repetitivos. Um exemplo disso é o trabalho com *Glycyrrhiza uralensis,* que demonstrou que RepeatMasker foi capaz de identificar 161 Mb (36,48% do “esboço do genoma”) como elementos transponíveis e repetições (MOCHIDA, et al., 2017).

RepeatMasker (http://www.repeatmasker.org) é um dos programas utilizados para anotação genômica. O processo de anotação pode ser dividido em duas etapas: a etapa computacional onde dados de outros genomas e transcriptoma da

espécie são usados para previsões iniciais; e a segunda etapa é de fato a anotação (EKBLOM e WOLF, 2014). RepeatMasker é usado antes da primeira etapa para “mascarar” sequências repetitivas e elementos transponíveis, posteriormente algoritmos de previsão da sequência de codificação (*coding sequence* − CDS), como por exemplo, AUGUSTUS (STANKE et al., 2006) são utilizados para melhorar a precisão de suas previsões (YANDELL e ENCE, 2012). Outro recurso que pode melhorar a precisão é o alinhamento de proteínas como o tblastx, por exemplo (EKBLOM e WOLF, 2014). A etapa final, a de anotação propriamente dita, pode ser dada de forma automatizada usando JIGSAW (ALLEN e SALZBERG, 2005), EVidenceModeler (HAAS et al., 2008), GLEAN (ELSIK et al., 2007), MAKER (CANTAREL et al. 2008) ou PASA (HAAS et al. 2003) que incorporam e pesam as evidências de várias fontes (YANDELL e ENCE, 2012; EKBLOM e WOLF, 2014).

3.6. ESTRATÉGIAS PARA AUMENTAR A QUALIDADE

Devido à dificuldade de montar um genoma completo a partir dos dados de sequenciamento, principalmente pela quantidade de repetições e pela complexidade do genoma das plantas; ainda são necessárias técnicas adicionais para alcançar a contiguidade no nível dos cromossomos, como mapas genéticos ou físicos/ópticos (SHENDURE, 2017; LI et al., 2017; KYRIAKIDOU et al., 2018).

Uma das soluções para desafios da montagem em escala cromossômica é a captura de conformação cromossômica - Hi-C (LIEBERMAN-AIDEN, 2009). O método usa uma biblioteca Illumina e é baseada na ligação de fragmentos de DNA que estão fisicamente próximos; sendo então possível mapear e ordenar *contigs* e, então montar e determinar a posição de cada *scaffold (*RICE e GREEN, 2019; LI e HARKESS, 2018). Um protocolo Hi-C modificado interessante é fornecido como um serviço pela *Dovetail Genomics* desde 2014 (PUTNAM, 2016; LI et al., 2017). Em um trabalho com cevada, Hi-C foi utilizada para ordenar a sequência referência em uma região altamente repetitiva (MASCHER et al., 2017). Outro exemplo com Hi-C, é o estudo com rosa que demonstrou alta congruência entre a sequência e o mapa genético, 97,7% da montagem (503Mb) foi orientada com coeficientes de correlação de Pearson (entre 0,986 a 0,996) (RAYMOND et al., 2018).

O mapeamento BioNano também pode ajudar usando informações de ligação a partir da localização física obtida da digestão por enzimas de restrição e

fluorescência. É uma tecnologia de alto rendimento, sem sequenciamento, que pode identificar erros de montagem, ligar *scaffolds* e melhorar a contiguidade (RICE e GREEN, 2019; LI et al., 2017). Assim, uma montagem de genoma referência de milho usando leituras longas (*long reads*) foi reduzida a 625 *scaffolds* (a partir de 2958 *contigs* inicialmente) com o uso de mapeamento BioNano (JIAO et al., 2017).

Por último, o sistema chamado Chromium, introduzido pela 10X Genomics em 2015, utiliza um grupo de “códigos de barras” de leituras curtas ligadas que se originam da mesma molécula de DNA individual (JIAO e SCHNEEBERGER, 2017; PHILLIPY, 2017; ZHENG et al., 2016; LI et al., 2017). Dessa forma, a técnica pode até determinar locos heterozigotos associando-os a cada “código de barra” (KYRIAKIDOU et al., 2018). Para exemplificar sua utilização, essa abordagem foi capaz de “ancorar” mais de 83% da sequência total à montagem final do genoma de *Capsicum annuum* (HULSE-KEMP et al., 2018).

Vale ressaltar que as tecnologias de alto rendimento de mapeamento físico, como BioNano, Captura de Conformação Cromossômica (Hi-C) e 10X Genomics surgiram como alternativas aos métodos tradicionais de mapeamento genético ou físico (LI et al., 2017).

Desse modo, com base em todas as informações apresentadas, o objetivo desse trabalho é sequenciar e montar o genoma de *Bertholletia excelsa* usando diferentes metodologias de montagem, e definir a metodologia que obtenha a melhor qualidade da montagem.

4. METODOLOGIA

4.1. MATERIAL VEGETAL E EXTRAÇÃO DE DNA GENÔMICO

A castanheira utilizada nesse projeto é um indivíduo adulto nativo (Figura 5) situado na EMBRAPA Rondônia em Porto Velho/RO (localização: 8°48'19.5"S 63°51'12.8"W). Essa castanheira não está isolada de outros indivíduos nativos da mesma espécie; e há também um plantio próximo ao local. No entanto, não há dados e estudos sobre essa população e sequer sabe-se a origem das árvores estabelecidas no plantio.

Para a extração de DNA foram coletadas e enviadas folhas frescas; o DNA foi extraído usando o método brometo de cetiltrimetilamônio (CTAB) baseado no protocolo descrito por Doyle e Doyle (1987) (descrito na íntegra no Anexo A) pela

empresa Reasearch and Testing Laboratory (RTL Genomics) − Lubbock, Texas, EUA.

Figura 5 – Árvore de *Bertholletia excelsa* utilizada no projeto.

Fonte: Karina Martins.

4.2. PREPARAÇÃO DE BIBLIOTECA E SEQUENCIAMENTO

A preparação da biblioteca e sequenciamento foram realizadas também pela empresa Reasearch and Testing Laboratory (RTL Genomics) − Lubbock, Texas, EUA. A metodologia seguiu a descrição do “Procedimento & Lista de Verificação - Preparando Bibliotecas SMRTbell®> 30 kb usando Megaruptor® Shearing e Seleção de Tamanho BluePippin™ para Sistemas PacBio RS II e Sequel®” disponibilizado pelo fabricante.

Antes de iniciar a preparação da biblioteca, a primeira etapa consistiu em avaliar o tamanho e a integridade do DNA genômico (gDNA) por eletroforese em gel de campo pulsado (PFGE). Em seguida, o DNA genômico não passou pelo processo de fragmentação, pois após a extração os fragmentos de DNA já eram superiores a 30 kb (entre 50 – 70 kb).

A construção da biblioteca de sequenciamento deu-se a partir do SMRTBell Template Prep Kit 1.0 e SMRTbell Damage Repair (Pacific Biosciences) de acordo com as instruções do fabricante. A construção da biblioteca incluiu o pré-tratamento

do DNA com Exo VII (com a finalidade de remover as extremidades de fita simples), reparo do DNA, sendo que na etapa “Repair Ends” houve uma modificação em relação as instruções do fabricante, onde o tempo foi estendido para 30 minutos (tempo determinado pelo fabricante de 5-10min).

Na sequência, para a construção da biblioteca, o DNA foi purificado com 0.45X AMPure® PB Beads (Pacific Biosciences) e os adaptadores *hairpin* foram ligados aos fragmentos (incubados a 25 ºC *overnight*). Após a ligação, foi realizada a digestão com Exo III/VII para remover os produtos de ligação que estariam falhados. Dessa forma, a biblioteca SMRTbell foi purificada novamente com 0.45X AMPure® PB Beads (Pacific Biosciences), seguida da etapa de seleção de tamanho com BluePippin Software v6.20 (Sage Science) para 10 kb (removendo todo fragmento inferior a 10 kb.

Para finalizar a construção da biblioteca, o DNA foi purificado com 1X AMPure® PB Beads (Pacific Biosciences), reparado com SMRTbell Damage Repair (Pacific Biosciences) e novamente purificado com 1X AMPure® PB Beads (Pacific Biosciences). Enfim, a concentração final da biblioteca SMRTbell foram avaliados usando o Qubit 3.0 Fluorometer e o kit dsDNA HS (Life Technologies). A biblioteca final SMRTbell resultou em 38 kb (antes do processo de seleção a biblioteca SMRTbell era de aproximadamente 27 kb).

Por fim, para o processo de sequenciamento, houve a ligação de 10X polimerase Sequel à biblioteca SMRTbell e iniciou-se o processo de sequenciamento no equipamento Sequel da Pacific Biosciences com um total de 10 células, sendo estimado um total de 70 Gb de dados (7Gb por célula). Mas, a quantidade de dados gerados foi de 112,66 Gb (aproximadamente 11 Gb, em média, por célula).

4.3. SISTEMA COMPUTACIONAL

As montagens do genoma foram produzidas em quatro máquinas com 40 vCPUs pela computação em nuvem da UFSCar (Cloud@UFSCar, Plataforma Ubuntu 16.04.1 LTS), sendo três máquinas com 370GB de memória RAM e 1TB de memória em HD, e uma com 370GB de memória RAM e 700 MB de memória em HD. Os arquivos resultantes do sequenciamento PacBio foram transferidos para as

máquinas e convertidos do formato “bam” para “fastq” com os comandos “sort –n” e “bam2fq” do software SAMtools v. 0.1.19 (LI et al., 2009).

4.4. AVALIAÇÃO DA QUALIDADE INICIAL DAS LEITURAS E ESTIMATIVA DO TAMANHO DO GENOMA

A avaliação da qualidade dos dados brutos resultantes do sequenciamento deu-se a partir do *assembler* Canu v 1.8 (KOREN et al., 2017) em sua primeira etapa: a correção – correção de erros do sequenciamento. A partir dele foram avaliados o número e tamanho de *reads* brutos e corrigidos; o N50 e a mediana desses dados foram calculados com base no histograma gerado pelo programa. As *reads* corrigidas foram produzidas com o parâmetro “correctedErrorRate=0.035” (*default* = 0.045), recomendado pelo Canu devido à alta cobertura genômica e, também para melhoria da eficiência computacional. Esse parâmetro representa a diferença entre duas sequências corrigidas sobrepostas.

As *reads* corrigidas foram utilizadas para gerar uma estimativa do tamanho do genoma haploide. Primeiro foi gerada uma distribuição de 22-mers através do Jellyfish v.2.2.7 (MARCAIS e KINGSFORD, 2011). Em seguida, utilizou-se a ferramenta GenomeScope versão 1.0 (VURTURE et al, 2017) para determinar a estimativa do tamanho do genoma haploide, conteúdo repetitivo e nível de heterozigose. Para essas estimativas, foram testados valores de cobertura de kmers (*max kmer coverage*) de 1.000, 10.000 e 100.000.

As *reads* brutas e as corrigidas foram geradas para servirem como *inputs* para algoritmos de montagem com a finalidade de comparar os resultados entre diferentes *assemblers*; porém, as *reads* corrigidas tem como objetivo principal comparar os resultados dos algoritmos que não realizam a etapa de correção como o Canu.

4.5. MONTAGEM *DE NOVO* E AVALIAÇÃO DA QUALIDADE FINAL

Para a montagem foram testados cinco *assemblers* específicos para dados de sequenciamento de terceira geração: Canu v 1.8 (KOREN et al., 2017), wtdbg2 v 2.3 (RUAN e LI, 2019), SMARTdenovo (https://github.com/ruanjue/smartdeno), MECAT2 (2019.2) (XIAO et al., 2017) e Flye v 2.5 (KOLMOGOROV et al., 2019). Dentre eles, wtdbg2 é um montador rápido baseado no grafo *fuzzy-Bruijn* (*fuzzy-Bruijn graph* -

FBG) que é análogo ao grafo *De Bruijn*, mas permite *mismatches* e *gaps* (KONO e ARAKAWA, 2019). Já o SMARTdenovo utiliza o sistema OLC (*overlap-layout-consensus*), caracterizando uma montagem rápida e razoavelmente precisa; entretanto, também não possui etapa de correção (KONO e ARAKAWA, 2019).

Flye usa o grafo de repetição (*repeat graph)* o qual é criado a partir de correspondências aproximadas, ou seja, são mais toleráveis aos altos erros do sequenciamento de *long reads;* contrapondo-se ao grafo *De Bruijn* que exige correspondências exatas de *k-mer* (KOLMOGOROV et al., 2019).

MECAT2 é um algoritmo que permite o alinhamento de sequências a partir de uma pontuação resultante das diferenças entre fatores de distância (FDDs) (XIAO et al., 2017). MECAT2 é capaz de produzir alta qualidade da montagem *de novo,* especialmente de genomas grandes com baixo custo computacional (XIAO et al., 2017; JAYAKUMAR e SAKAKIBAR, 2017).

Por fim, Canu é um algoritmo de montagem hierárquico, pois é composto por três etapas: correção, *trimming* (filtragem e edição das sequências) e *assembly* (construção do BOG – *best overlap graph*) (KOREN et al., 2017).

Assim, todos os *assemblers* foram rodados com os seus respectivos parâmetros-padrão (*default*). Contudo, a montagem com o uso de Canu foi realizada por Rodrigo Theodoro Rocha, colaborador da Universidade de Brasília (UnB), utilizando o supercomputador Santos Dumont. Isso porque os recursos computacionais disponíveis não foram suficientes para esse programa.

Além dos diferentes montadores, também foram testados para cada um deles (exceto Canu) diferentes coberturas genômicas dos dados brutos iniciais, com a finalidade de entender como elas influenciam na qualidade da montagem. Para a obtenção de diferentes coberturas genômicas utilizou-se os dados brutos originais, com cobertura 187×, com os parâmetros *default* do programa Filtlong v 0.2.0 (https://github.com/rrwick/Filtlong) para gerar arquivos com as seguintes coberturas genômicas: 47×, 63×, 97× e 126×. Testou-se ainda o arquivo com cobertura de 60× com *reads* corrigidas com Canu (60×_cor).

O programa QUAST v 5.0.2 (GUREVICH et al., 2013) foi utilizado para avaliação da qualidade da montagem no quesito contiguidade. Assim, para inferir sobre o grau de qualidade de cada estratégia de montagem, os parâmetros foram ranqueados considerando as melhores montagens aquelas que obtiverem menor

número de *contigs*, maior valor de N50, menor valor de L50, menor número de *gaps* e maior porcentagem do genoma montado. Primeiramente, as estratégias de montagens foram ordenadas em relação a qualidade perante a esses os cinco parâmetros; ou seja, resultaram em cinco ordenações, uma para cada parâmetro. Dessa forma, cada estratégia de montagem recebeu uma pontuação por colocação em cada ordenação. A partir daí dois *rankings* foram criados com critérios diferentes. No *ranking* denominado “Sem sistema de pontuação diferenciada por parâmetro” o resultado se dá pela somatória de pontos de cada estratégia de montagem conquistada nos cinco parâmetros descritos acima. Já o *ranking* denominado “Com sistema de pontuação diferenciada por parâmetro”, as pontuações por parâmetros recebem pesos diferentes: para N50 e número de *contigs* a pontuação é multiplicada por três, para L50 a pontuação é multiplicada por dois, número de gaps e proporção do genoma montado são multiplicadas por um. Por fim, também se prossegue com a somatória desses pontos para revelar o *ranking* das melhores montagens. Esses pesos são dados com base no estado-da-arte de diversos estudos, levando em consideração o grau de importância dos parâmetros em relação a escolha da melhor montagem.

Por fim, BUSCO (*Benchmarking Universal Single-Copy Orthologs*) (SIMÃO et al., 2015) foi usado para avaliar a integridade das montagens a nível gênico; BUSCO procura um conjunto de genes de cópia única altamente conservados em um determinado grupo taxônomico (WANG et al., 2020). Para esse trabalho a base de referência utilizada foi de Embriófitas. E LAI *score* (OU; CHEN; JIANG, 2018) − *long-terminal repeat (LTR) assembly index* – foi usada para avaliar a qualidade considerando a contiguidade e a integridade de sequências repetitivas do tipo LTR. Para essas duas avalições foram utilizadas só as abordagens com maior cobertura genômica inicial, ou seja, apenas uma montagem obtida de cada *assembler*.

5. RESULTADOS E DISCUSSÃO

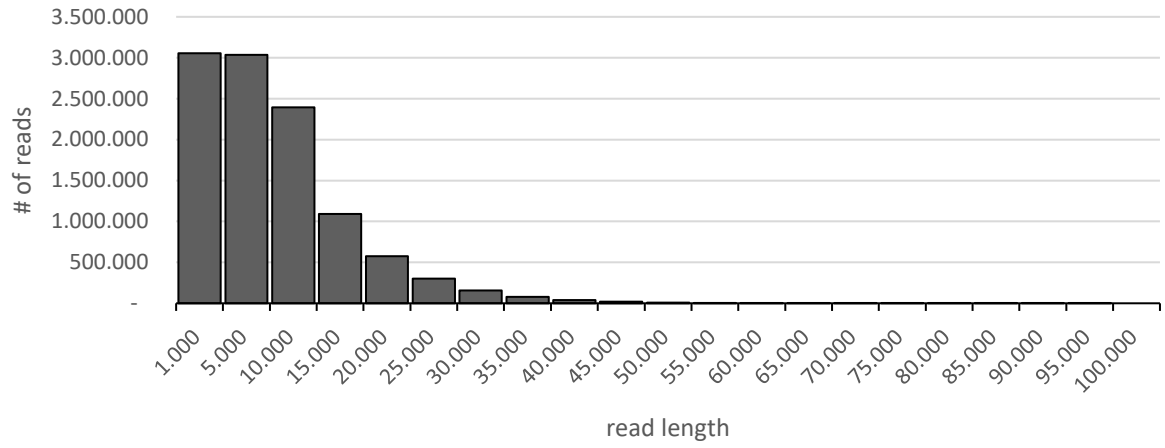
5.1. DESCRIÇÃO E AVALIAÇÃO DA QUALIDADE INICIAL DAS LEITURAS BRUTAS E CORRIGIDAS

Com o uso de Canu foram identificadas 11.711.241 *reads* resultantes do sequenciamento (leituras brutas), entretanto 947.803 *reads* foram excluídas pelo

programa, pois eram menores que 1 kb, devido ao limite do *default* (parâmetro padrão); essas *reads* representam 8,09% do total de *reads* brutas.

Dessa forma, 10.763.438 *reads* foram processados pelo programa como dados brutos. Essas *reads* representam 112.204.232.021 bases. Dessas r*eads*, 78,79% (8.480.169) são *reads* com até 15 kb, ou seja, apenas 21,21% das *reads* são maiores que 15 kb; sendo apenas seis leituras acima de 90 kb, e uma única *read* de ~100 kb (Figura 6). Assim, de acordo com a literatura, PacBio realmente produz leituras longas (acima de 15 kb), entretanto essas *reads* não são a maioria dos dados, o que ainda pode gerar dificuldades na montagem de genomas grandes e repetitivos.

Figura 6 – Distribuição de *reads* brutas por tamanho.



Fonte: Elaborado pelo autor, a partir do programa Canu.

Além disso, a partir do histograma gerado pelo Canu foi possível calcular o N50 e a mediana das leituras brutas, resultando em um N50 de 14,02 kb e uma mediana de aproximadamente 50 kb. Esses resultados estão na média em relação a literatura: no trabalho de Xia et al. (2018) o N50 das *reads* brutas foi de 15,7 kb, já no trabalho Yang et al. (2019), com *Acer yangbiense* (árvore sudoeste da China), o N50 foi de 16,8 kb.

No que concerne ao estágio de correção, o algoritmo detectou 1.255.073 *reads* (11,66%), das 10.763.438 *reads,* que não apresentaram sobreposições e não são candidatas a correção (Tabela 4). Assim, das *reads* com sobreposições (9.508.365 *reads*), 84,11% das leituras (8.045.118 *reads*) foram candidatas à

correção; porém foram utilizadas apenas as leituras mais longas (entre 14 kb a 99 kb). Essa seleção de leituras longas também deve ser justificada pela redução da cobertura, ou seja, se há redução da cobertura genômica é importante que permaneçam *long reads* para que a montagem não apresente aumento da fragmentação e, também, para que não ocorram sobreposições equivocadas que resultem na queda da qualidade da montagem.

Consequentemente, foram selecionadas 1.772.515 leituras longas candidatas à correção. Canu também estipulou que 181.718 *reads* fossem resgatadas (Tabela 5); essas *reads* possuem tamanho mínimo de 1 kb, e provavelmente são importantes para o processo de montagem do genoma.

Tabela 5 – *Reads* brutas, candidatas à correção e resgatadas.

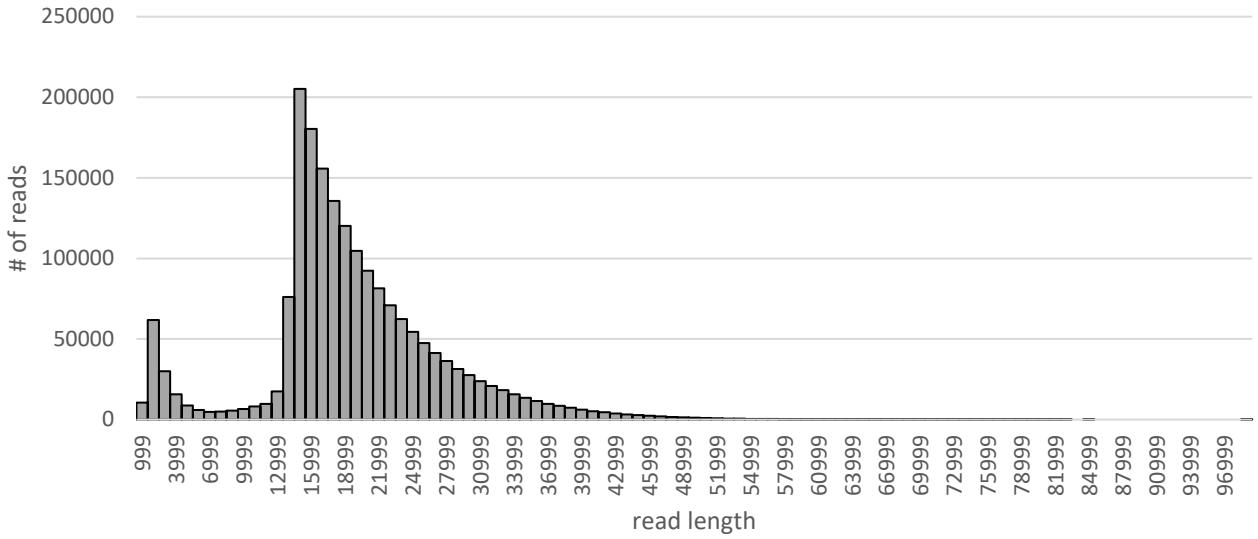
Categorias	Leituras com sobreposições	Leituras sem sobreposições	Leituras candidatas à correção	*Reads* brutas a serem corrigidas	*Reads* brutas a serem resgatadas
Número de *reads*	9.508.365	1.255.073	8.045.118	1.772.515	181.718
Número de bases	101.479.724.711	4.619.008.885	96.185.539.008	39.873.639.577	780.218.569
Cobertura	107,957	4,914	102,3	42,4	0,83
Mediana	9.159	0	10.320	20.265	3.129
N50	14.399	11.159	14.921	22.800	5.235
Mínimo – máximo	1.000 – 10.4359	0 – 89.019	1.000 – 104.359	13.944 – 99.707	1.003 – 56.632

Fonte: Elaborado pelo autor, a partir do programa Canu.

Por fim, a etapa de correção resultou em 1.882.527 *reads*, totalizando 36.076.143.157 pb (Figura 7); ou seja, os dados corrigidos representam 32,15% do total bases das *reads* brutas descritas anteriormente (112.204.232.021 bases), entretanto essa queda é resultado da redução da cobertura genômica gerada pelo programa.

Com base no histograma gerado pelo Canu também foi possível calcular a mediana e o N50 das leituras corrigidas, resultando em uma mediana de aproximadamente 50 kb, o mesmo valor da mediana das leituras brutas, mas um N50 com valor entre 20 a 21 kb (valor superior ao N50 dos dados brutos), demonstrando que a etapa de correção do algoritmo Canu melhora a contiguidade dos dados.

Figura 7 – Distribuição *reads* corrigidos pelo *assembler* Canu.



Fonte: Elaborado pelo autor, a partir do programa Canu.

5.2. ESTIMATIVA DO TAMANHO DO GENOMA VIA K-MER

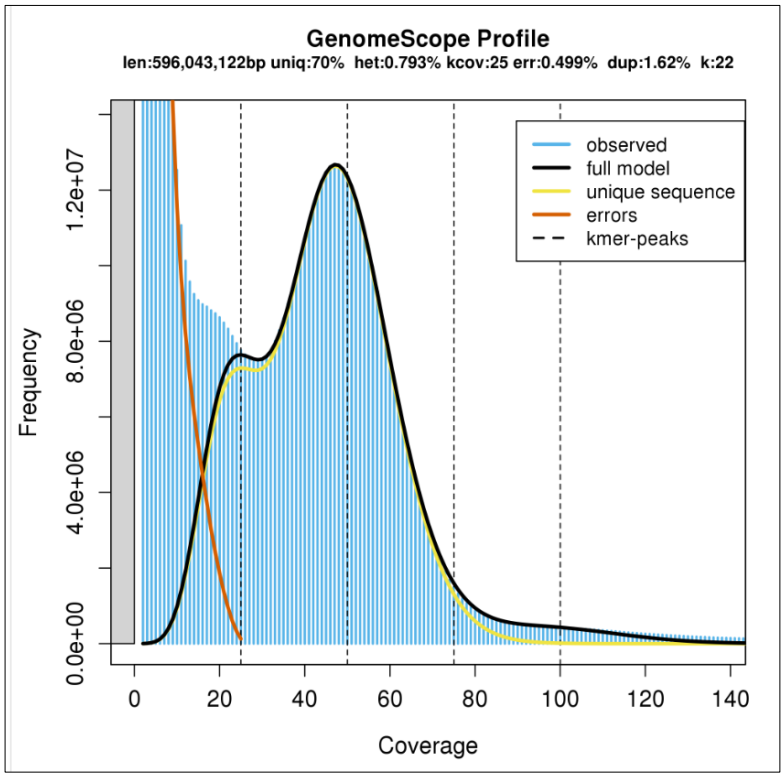
Através do GenomeScope (Vurture et al. 2017) foi possível estimar o tamanho do genoma com uma distribuição de 22 *mers*. Essa estimativa utilizou as *reads* corrigidas de Canu porque, segundo Wang et al. (2020), é necessário utilizar a correção para que a estimativa não seja subestimada por repetições genômicas logo, resulta em uma estimativa mais precisa. Um exemplo disso é o trabalho de Paajanen et al (2019), eles demonstraram que ao utilizar leituras resultantes do sequenciamento PacBio, após um processo de *polishing,* obtiveram em um conjunto gênico muito semelhante ao de uma biblioteca Illumina.

Logo, após testar diferentes valores para “max k-mer coverage” (Anexo B), conclui-se que a estimativa do tamanho do genoma haploide é de aproximadamente 596 Mpb (Figura 8 e Tabela 5); dessa forma, a cobertura genômica para os dados brutos (~112,66 Gb) é de 187×.

Na figura 8, o pico à esquerda demonstra baixa heterozigosidade, em torno de 0,8% (informação detalhada na Tabela 6). Já, o conteúdo repetitivo representa cerca de 30% do genoma (179.004.411 bp); um valor bem menor se comparado a *Malania oleífera* e *Eucalyptus grandis* (~50% de conteúdo repetitivo) – valores

obtidos com estimativa via distribuição *k-mer* e anotação gênica, respectivamente (XU et al., 2019; MYBURG et al., 2014).

Figura 8 – Distribuição 22mers obtida com as *reads* corrigidas por Canu.



Fonte: Gráfico gerado por pelo programa GenomeScope.

Tabela 6 – Propriedades do genoma obtidos pela distribuição de *k-mers*.

Propriedade	Mínimo	Máximo
Heterozigose	0.79%	0.79%
Tamanho do genoma haploide	595.904.308 bp	596.043.122 bp
Tamanho do genoma repetitivo	178.962.722 bp	179.004.411 bp
Tamanho de genoma único	416.941.586 bp	417.038.711 bp
Ajuste do modelo	95.47%	98.70%
Taxa de erro nas *reads*	0.50%	0.50%

Fonte: Tabela gerada por pelo programa GenomeScope.

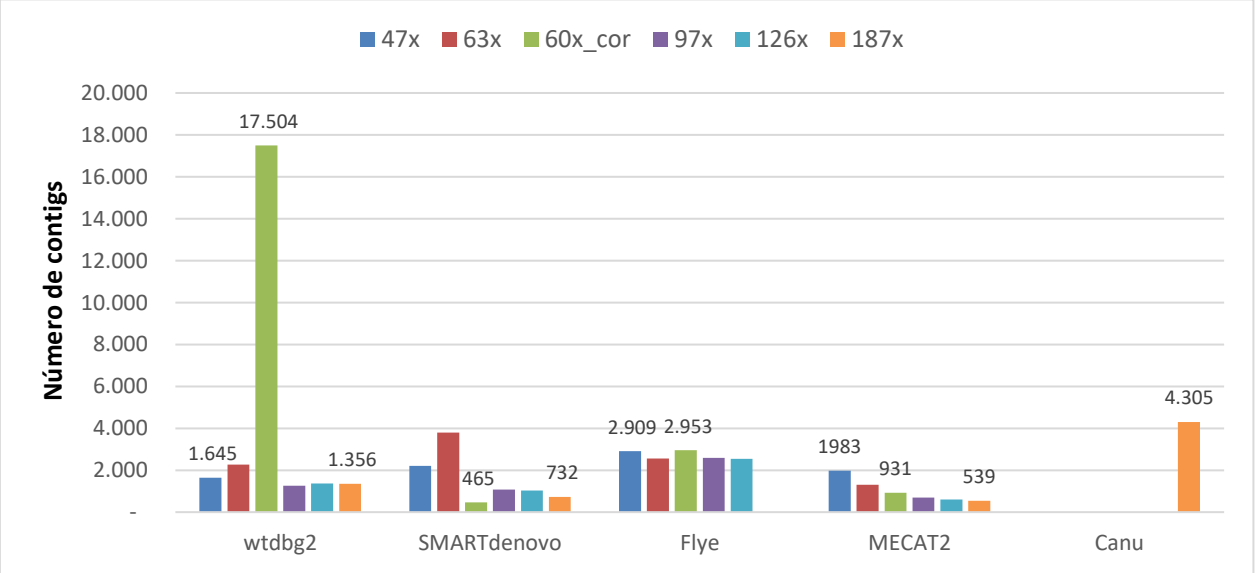
5.3. COMPARAÇÃO DAS DIFERENTES ESTRATÉGIAS DE MONTAGENS *DE NOVO*

Para comparar as montagens entre diferentes *assemblers* e diferentes coberturas genômicas foram criados os gráficos com os principais parâmetros que

medem a contiguidade da montagem: número de *contigs, contig* N50, L50 e tamanho da montagem (Figuras 9, 10, 11 e 12). As tabelas que dão origem aos gráficos estão disponíveis no Anexo C. A montagem de Canu foi realizada apenas com as suas próprias *reads* corrigidas, portanto, há uma única montagem resultante da abordagem de 187× de cobertura genômica A única proposta de montagem não sucedida foi Flye com 187× de cobertura por falta de memória computacional.

Dessa forma, observando os resultados em relação ao número de *contigs* e L50 (Figuras 9 e 10), a montagem wtdbg2 com *reads* corrigidas por Canu ("60×_cor wtdbg2") foi a estratégia de montagem mais fragmentada, seguida da montagem resultante de Canu. Em contrapartida, a montagem com menor número de *contigs* entre as montagens corrigidas, e também entre as demais abordagens, foi SMARTdenovo com *reads* corrigidas por Canu ("60×_cor SMARTdenovo"), apresentando melhor qualidade em termos de fragmentação. Desse modo, se compararmos "60×_cor SMARTdenovo" e "60×_cor wtdbg2" vemos que não existe uma associação da utilização de *reads* corrigidas, mas que a qualidade da montagem está mais relacionada com o tipo de algoritmo de montagem.

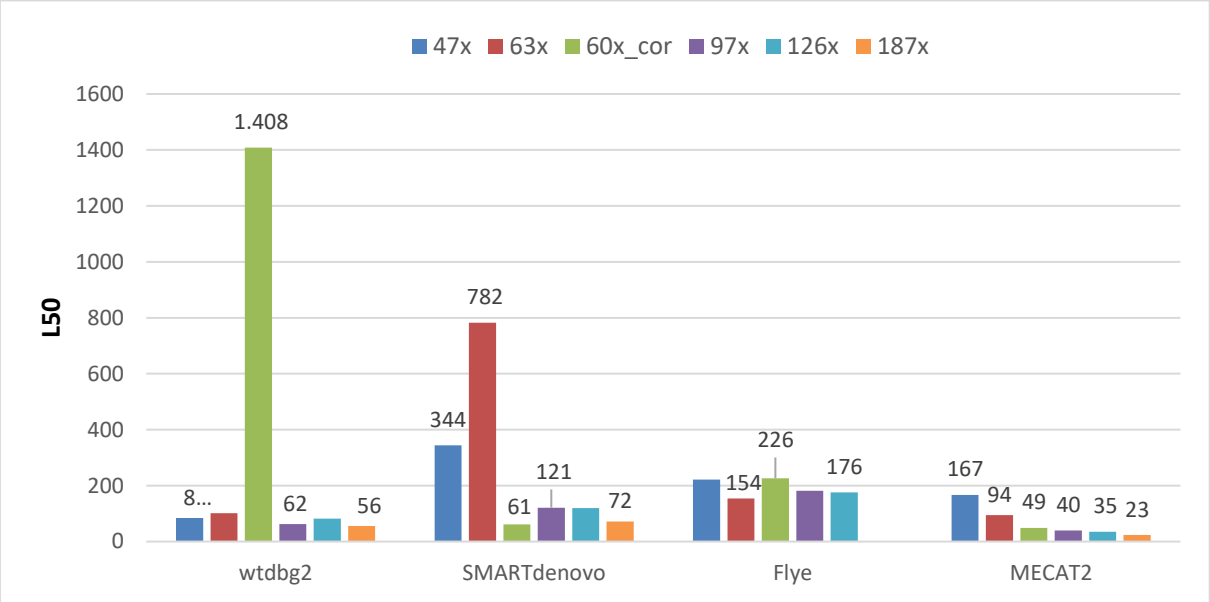
Figura 9 – Número de *Contigs* em montagens realizadas por cinco *assemblers* com diferentes coberturas genômicas iniciais.



Fonte: Elaborado pelo autor, a partir do programa QUAST.

Além disso, é possível inferir que MECAT2 é o único *assembler* que produz uma tendência dos dados; mostrando que quanto maior a cobertura genômica, menor é a quantidade de *contigs* e L50, ou seja, menos fragmentada é a montagem. De modo geral, MECAT2 destaca-se como melhor algoritmo nesses quesitos.

Figura 10 – Valor de L50 em montagens realizadas por quatro *assemblers* com diferentes coberturas genômicas iniciais.

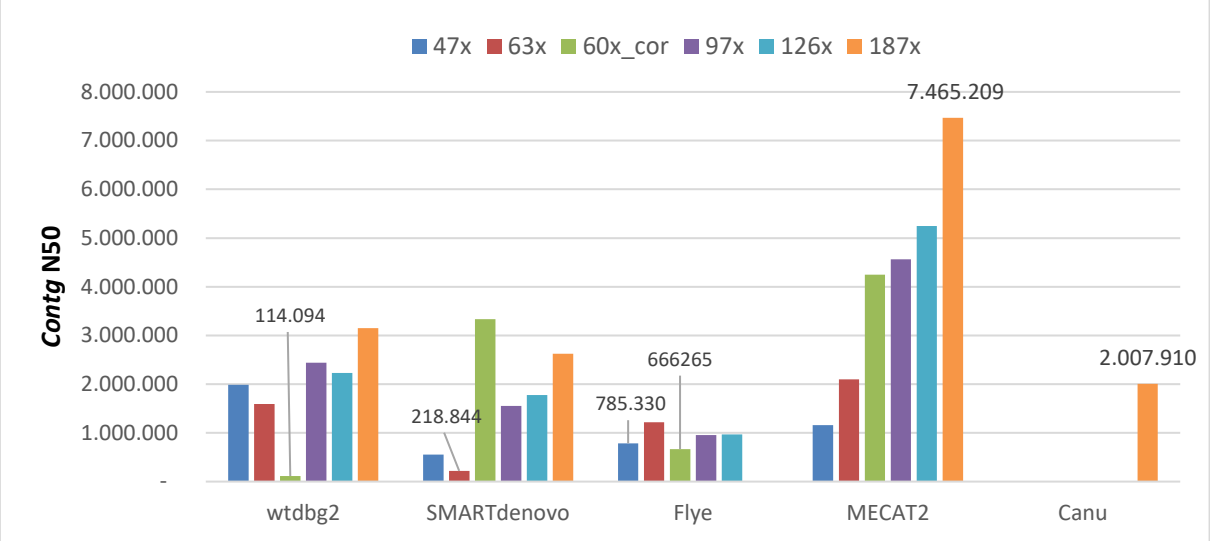


Fonte: Elaborado pelo autor, a partir do programa QUAST.

De forma análoga o melhor algoritmo no critério "*Contig* N50" (principal critério para escolha das melhores montagens conforme a literatura atual) é o MECAT2 a partir de uma cobertura de 97× ou 60× com correção ("60×_cor MECAT2"). Desse modo, MECAT2 com 187× de cobertura genômica obteve o melhor valor de N50, aproximadamente 7,47 Mb (Figura 11). Esse resultado alinha-se à ideia amplamente discutida na literatura, que altas coberturas genômicas garantem melhor qualidade de contiguidade.

Portanto, MECAT2 é o que apresenta melhor performance em relação as demais abordagens. E Flye foi o pior *assembler* nesse quesito, apresentado em todas as suas montagens valores baixíssimos de N50.

Figura 11 – Tamanho do *Contig* N50 em montagens realizadas por cinco *assemblers* com diferentes coberturas genômicas iniciais.



Fonte: Elaborado pelo autor, a partir do programa QUAST.

Para o tamanho total da montagem (Figura 12) a melhor estratégia nesse quesito foi “187× wtdbg2” com tamanho total 597.523.347 pb. wtdbg2 (com exceção de “60×_cor wtdbg2”) foi o *assembler* que apresentou os resultados mais próximos da estimativa do tamanho do genoma. A segunda abordagem com o tamanho da montagem mais próxima da estimativa foi “187× MECAT2” com 593.634.560 pb.

As piores abordagens nesse quesito foram “60×_cor wtdbg2” com tamanho total de 901.650.871 pb, seguido de Canu com 831.215.698 pb.

> Canu divide os haplótipos em contigs separados sempre que a divergência alélica for maior do que a taxa de erro de sobreposição pós-correção. Esse limite geralmente é de 1,5% para dados recentes do PacBio. Essa divisão resulta em um tamanho de montagem maior que o tamanho do genoma haploide (KOREN et al., 2017, p. 730).

Isso também pode justificar o porquê da montagem de wtdbg2 superestimar o tamanho do genoma quando utiliza as *reads* corrigidas de Canu.

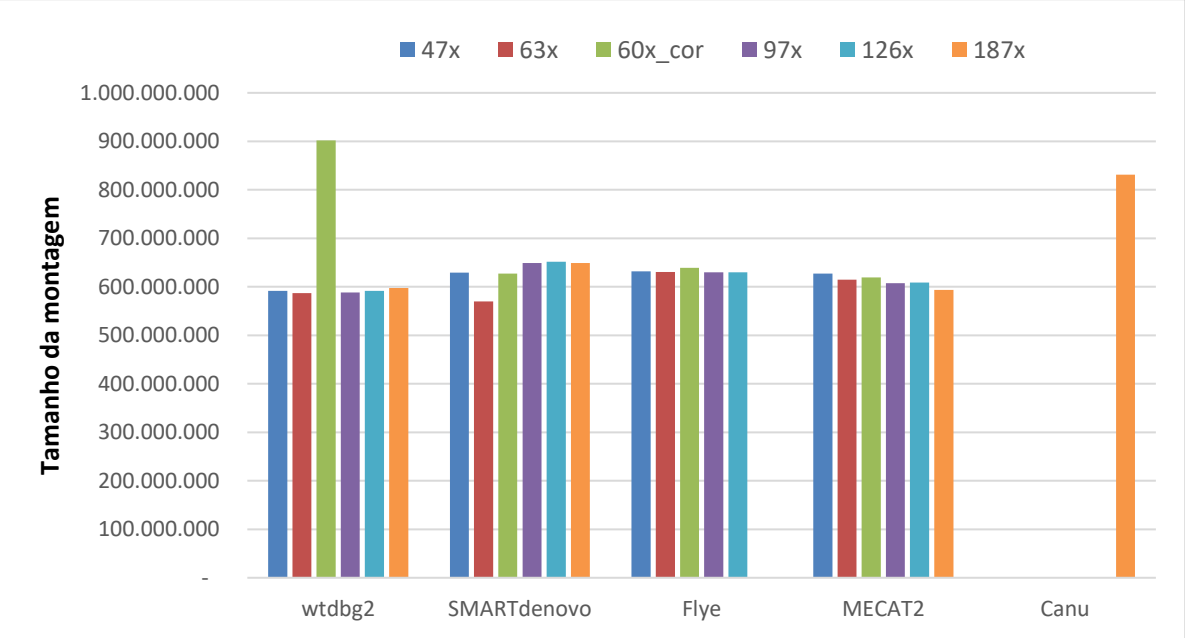
Flye apesar de apresentar-se constante na determinação do tamanho da montagem, acabou sendo o *assembler* que mais superestimou esses resultados (valores acima de 620Mpb).

Em geral, praticamente todas as estratégias geraram resultados próximos, em média montaram 105,51% do tamanho do genoma estimado. Comparando os resultados com o trabalho de Yang et al. (2019), por exemplo, os valores em relação ao tamanho estimado de *Acer yangbiense* também estão superestimados tanto para

a abordagem de correção com Canu e montagem com SMARTdenovo, como a de correção com Canu e montagem com wtdbg.

Por fim, é válido destacar que a porcentagem de heterozigose/conteúdos repetitivos dos genomas interferem no tamanho da montagem, como justifica Xia et al. (2018) em sua montagem com MECAT. Nesse trabalho o conteúdo de GC variou de 34,28% a 34,92% (Anexo C) aproximando-se de outros estudos com plantas que resultaram em torno de 33,57% a 36,29% de conteúdo CG (XIA et al., 2019; MORRISSEY et al., 2019; SCHMIDT et al., 2017).

Figura 12 – Tamanho das montagens obtidas por cinco *assemblers* com diferentes coberturas genômicas iniciais.



Fonte: Elaborado pelo autor, a partir do programa QUAST.

5.4. AVALIAÇÃO DA QUALIDADE GLOBAL DAS MONTAGENS *DE NOVO*

Por fim, as diferentes estratégias foram ranqueadas para comparar a qualidade global das montagens (Tabela 7). Assim, é possível observar que nos dois métodos de pontuação estão presentes as mesmas abordagens de montagem, sendo MECAT2 com 187×, 126× e 97× de cobertura as três melhores montagens nos dois processos perante os parâmetros analisados. Isso, deu-se principalmente porque MECAT2 apresenta-se nas três primeiras colocações para maiores valores de N50 e menores valores de L50 nas coberturas de 187×, 126× e 97×,

respectivamente. Dessa forma, para esse trabalho MECAT2 foi o algoritmo com maior regularidade, além de proporcionar as melhores montagens para *Bertholletia excelsa.*

Tabela 7 – Rankings das melhores montagens do genoma de *Bertholletia excelsa* com base em parâmetros de contiguidade.

Ranking	Sem sistema de pontuação diferenciada por parâmetro	Com sistema de pontuação diferenciada por parâmetro
1	187× MECAT2	187× MECAT2
2	126× MECAT2	126× MECAT2
3	97× MECAT2	97× MECAT2
4	187× wtdbg2	60× _cor SMARTdenovo
5	60× _cor SMARTdenovo	60× _cor MECAT2
6	97× wtdbg2	187× wtdbg2

Fonte: Elaborado pelo autor.

Porém, em termos de números de *contigs* “187× MECAT2” e “126× MECAT2” perdem apenas para “60×_cor SMARTdenovo”. No ranking com sistema de pontuação diferenciada, MECAT2 com *reads* corrigidas “60× _cor MECAT2” aparece devido ao valor de seu N50.

A superioridade de MECAT é afirmada, por exemplo, no trabalho de Zhang et al. (2019), e reforçada pelos resultados de Jayakumar e Sakakibara (2017) para *Ipomoea nil*; dentre os *assemblers* em comum a este estudo, MECAT2 produziu a montagem com a melhor qualidade (3º lugar), seguido por SMARTdenovo (5º lugar) e wtdbg2 foi o pior (8º lugar/penúltima colocação).

Essa estratégia de utilizar *reads* corrigidas por Canu seguida pela montagem com SMARTdenovo foi escolhida em alguns estudos como a melhor montagem para prosseguir com *polishing* e/ou mapas ópticos ou físicos, a fim de aumentar ainda mais a qualidade da montagem (XU et al., 2019; SCHMIDT et al., 2017; DESCHAMPS et al., 2018; YANG et al, 2019). De modo geral, a escolha da melhor montagem é determinada principalmente pelo N50 de *contigs*; dentre esses estudos citados, o maior N50 foi de 3.005.621 pb (DESCHAMPS et al., 2018), valor inferior a mesma abordagem presente nesse trabalho, o qual o N50 foi de 3.337.434 pb para “60×_cor SMARTdenovo”.

Em relação à montagem “187× wtdbg2”, ela aparece no ranking sem sistema de pontuação, pois é a melhor abordagem em termos de tamanho da montagem; mas, perde colocação no *ranking* com sistema de pontuação, porque seu o número de *contigs* é maior do que o das abordagens “60× _cor MECAT2” e “60× _cor SMARTdenovo”

O algoritmo Flye foi o único a apresentar *gaps* a cada 100kb (Anexo C4), além disso, foi o algoritmo com pior desempenho; mesma conclusão do trabalho de Wang et al. (2020) com *Eucalyptus pauciflora*. Por outro lado, a montagem com o *assembler* Canu mostra-se altamente fragmentada, com baixo valore de N50 e alto valor para número de *contigs* e tamanho da montagem.

De modo geral, é possível observar que a melhoria da qualidade da montagem apenas com os dados de leitura longa, depende do aumento da cobertura genômica (GHURYE e POP, 2019).

5.5. AVALIAÇÃO DA INTEGRIDADE GÊNICA

Foram escolhidas para essa avaliação apenas as montagens com maior cobertura genômica (187×), uma de cada *assembler* (wtdbg2, MECAT2, SMARTdenovo) pois, foram as melhores montagens dentro de cada montador. Flye não foi utilizado porque não houve a montagem com 187×. Para essa avaliação também foi utilizada a montagem de Canu.

A partir da Tabela 8, é possível observar que MECAT2 foi o pior *assembler* para BUSCO. O valor de completude gênica foi de 70,7%, 14,2% são genes fragmentados e 15,1% são genes ausentes. Xia et al. (2018) apresenta em seu estudo que MECAT (mesmo passando pela etapa de *polishing* com Quiver e Pilon) só conseguiu 89,2% de genes completos com a avaliação de BUSCO. O viés da montagem obtida com MECAT também é apontada pela baixo valor de LAI *score*, igual a 8,59 – segundo Yang et al. (2019), o intervalo do valor de referência está entre 10 e 20. LAI representa regiões LTR (*long-terminal repeat*), portanto mesmo MECAT2 (187×) se apresentando como o melhor *assembler* em termos de contiguidade (*contig* N50, número de *contigs*), ele não conseguiu representar o genoma de *Bertholletia excelsa*, por falhar na determinação de regiões repetitivas. Como LTR também flanqueiam regiões gênicas, MECAT2 acabou suprimindo essas regiões também; por isso, a porcentagem de genes ausentes e fragmentados é alto

− 15,1% e 14,2%, respectivamente (MURRAY; ROSENTHAL; PFAÜER, 2006). Outra montagem que conduz a uma discussão semelhante é wtdbg2 (187×); até porque é um montador que permite *mismatches (*relatado na página 21).

Apesar de apresentar uma montagem excessivamente maior do que a estimativa do genoma, Canu resultou em 95,8% de completude gênica e um LAI *score* de 10,56. Aparentemente Canu realizou uma boa montagem nesse sentido, porém separou haplótipos, resultando em montagem maior (ver informação na página 30), por isso há uma porcentagem alta de genes duplicados (22,1%); dessa forma, realizar a etapa de *polishing* é necessária para remover a redundância dessa montagem (KOREN et al., 2017).

Por fim, a melhor montagem foi SMARTdenovo (187×), pois obteve os melhores valores para integridade gênica (95,1%), fragmentação gênica (1,6%), genes ausentes (3,3%) e LAI *score* = 10,53; portanto, essa é a montagem de maior qualidade, e que está mais próxima de representar o genoma da Castanheira-do-Brasil. Na montagem do genoma de *Acer yangbiense,* SMARTdenovo associada à outras técnicas que visam aumentar a qualidade da montagem − como Hi-C, por exemplo – resultou numa completude gênica de 95,5% e LAI *score* de 12,21 (YANG et al., 2019). Isso sugere uma excelência da montagem de *Bertholletia excelsa* com SMARTdenovo (187×), pois os resultados estão próximos dos obtidos por Yang et al. (2019), visto que a montagem de *Bertholletia excelsa* não passou por etapas de *polishing* ou foram associadas a mapas genéticos ou físicos/ópticos.

Em relação a abordagem utilizada pelo *assembler* SMARTdenovo, ele é o único dentre os cinco montadores testados que se baseia em OLC (*overlap-layout-consensus*) − metodologia adequada para lidar com *long noisy reads* (KONO e ARAKAWA, 2019; https://github.com/ruanjue/smartdenovo).

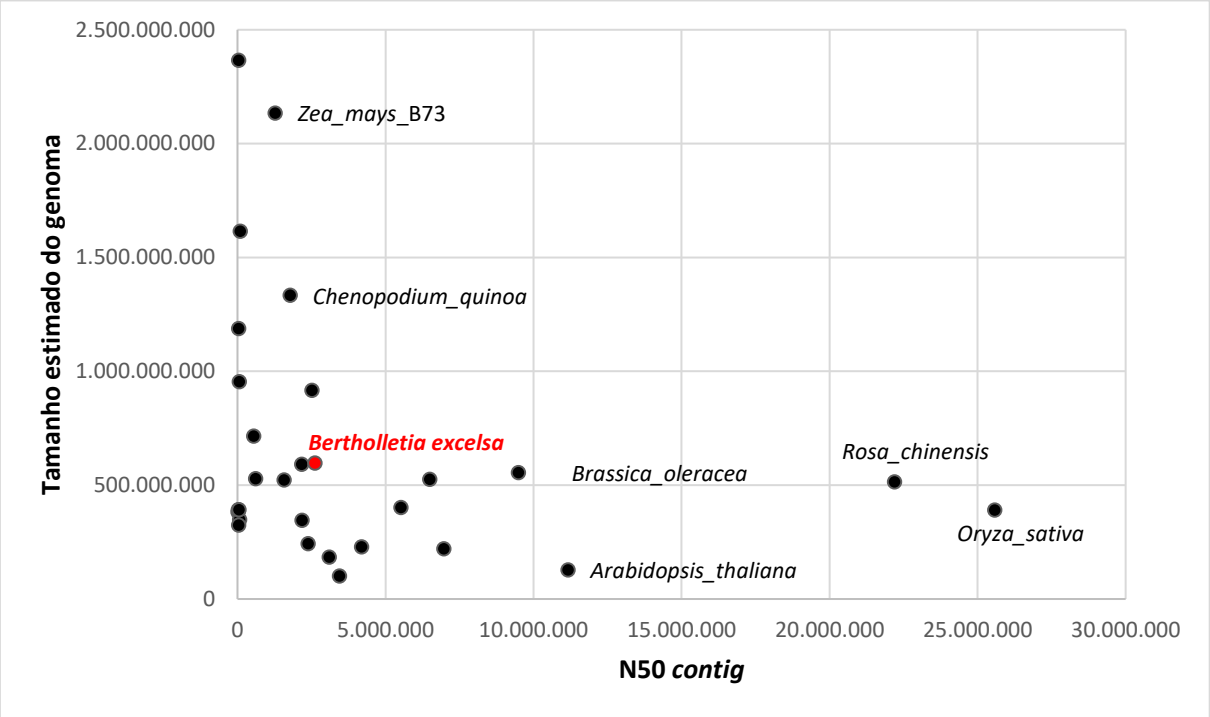
Tabela 8 – Estatísticas de contiguidade e de integridade gênica das montagens.

Assembly	Coverage	Assembly Statistics				BUSCO scores								LAI scores
						Complete genes		*Duplicated genes*		*Fragmented genes*		*Missing genes*		
		Length (pb)	*Number of contigs*	*Largest contig (bp)*	*contig N50 (pb)*	Number	%	Number	%	Number	%	Number	%	
Canu	187×	831.215.698	4.305	11.129.176	2.007.910	412	95,8	95	22,1	3	0,7	15	3,5	10,56
MECAT2	187×	593.696.041	573	30.393.383	7.465.209	304	70,7	8	1,9	61	14,2	65	15,1	8,59
wtdbg2	187×	597.529.250	1.359	13.061.757	3.149.022	364	84,6	10	2,3	31	7,2	35	8,2	8,99
SMARTdenovo	187×	649.349.366	732	14.715.192	2.623.932	409	95,1	19	4,4	7	1,6	14	3,3	10,53

Fonte: Elaborado pelo autor através do QUAST, BUSCO e LAI.

Por fim, ao comparar a melhor estratégia obtida nesse trabalho com montagens citadas por Belser et al. (2018) a partir de *long reads* (Figura 13), conclui-se que a montagem de *Bertholletia excelsa*, a partir do algoritmo SMARTdenovo com 187× de cobertura genômica, produziu um N50 (*contig*) de 2.623.932pb, e portanto, é a décima primeira (11º) montagem mais contígua.

Figura 13 – Tamanho do genoma *versus* tamanho do N50 das montagens de plantas a partir de sequenciamento de *long reads*.



Fonte: Elaborado pelo autor, adaptado de Belser et al., 2018.

6. CONCLUSÕES

- Acredita-se que o tamanho do genoma seja de 596.043.122 bp, como foi dado pela estimativa por *k-mer*.
- A montagem que resultou num tamanho de genoma mais próximo da estimativa obtida por distribuição de *k-mers* foi wtdbg2 com 187× de cobertura.
- MECAT2 (187×) apesar de apresentar excelentes resultados para contiguidade, demonstrou ser o pior *assembler* frente aos resultados de integridade gênica. Logo, MECAT2 suprimiu regiões gênicas conservadas, possivelmente pela sua dificuldade em determinar regiões repetitivas (que principalmente flanqueiam regiões gênicas), resultando num colapso da montagem do genoma.
- O *assembler* Canu não obteve boas estatísticas de contiguidade em comparação aos demais montadores, entretanto no que diz respeito à completude gênica, foi o segundo melhor *assembler*. Ademais, o alto valor de genes duplicados, sugere que Canu não obteve êxito em representar um genoma haploide; mas, é preciso considerar o uso de outras ferramentas e métodos, como *polishing* e Hi-C, para refinar a qualidade da montagem.
- A melhor montagem obtida é a abordagem SMARTdenovo com 187× de cobertura genômica, pois apresentou melhores valores de completude gênica e LAI *score* com um N50 de aproximadamente 2,6 Mpb.
- As altas coberturas de leituras (*coverage*) e o uso de de *reads* corrigidas ajudam a reduzir a taxa de erros resultantes do sequenciamento PacBio, consequentemente melhoram a contiguidade da montagem.
- No geral, os resultados reforçam as hipóteses de que em projetos de sequenciamento e montagem *de novo* de genomas se faz necessário testar diversos algoritmos de montagem; e acrescentar a montagem ferramentas que realizem o *polishing* ou que associem mapas genéticos ou físicos/ópticos, pois são importantes para obter uma montagem robusta.

7. REFERÊNCIAS

ALLEN, J. E.; SALZBERG, S. L. JIGSAW: integration of multiple sources of evidence for gene prediction. **Bioinformatics** 21, 3596–3603, 2005.

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

ARAGÃO, F. J. L. et al. Expression of a methionine-rich storage albumin from the Brazil nut (*Bertholletia excelsa* H.B.K., Lecythidaceae) in transgenic bean plants (*Phaseolus vulgaris* L., Fabaceae). **Genetics and Molecular Biology**. v.22, n. 3, p. 445-449, set. 1999.

ARIYARATNE, P. N.; SUNG, W. K. PE-Assembler: de novo assembler using short paired-end reads. **Bioinformatics**. 27: 167–174, jan. 2011.

AYLING, M.; CLARK, M. D.; LEGGETT, R. M. New approaches for metagenome assembly with short reads**. Briefings in Bioinformatics**, bbz020, fev. 2019.

BALDONI, A.B.; WADT, L. H. O.; PEDROZO, C. A. Brazil nut (Bertholletia excelsa Bonpl.) Breeding. In: AL-KHAYRI, J. M.; JAIN, S. M.; JOHNSON, D. V. **Advances in Plant Breeding Strategies: Nut and Beverage Crops, Volume 4.** Springer International Publishing, 2020. Cap. 3, p. 57-76.

BATZOGLOU, S. et al. ARACHNE: a whole-genome shotgun assembler. **Genome Research.** v.12, p. 177–189, jan. 2002.

BELSER, C. et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. **Nat Plants**, v. 4, p. 879-887, nov.2018.

BÉRÉNOS, C. Genomic analysis reveals depression due to both individual and maternal inbreeding in a free-living mammal population. **Molecular Ecology**. v. 25, n. 13, p. 3152– 3168, jul. 2016.

BOISVERT, S.; LAVIOLETTE, F.; CORBEIL, J. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. **Journal of Computational Biology**. 17: 1519–1533, nov. 2010.

BRADNAM, K. R. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. **GigaScience** 2:10, jul. 2013.

BRYANT, D; WONG, W.K.; MOCKLER, T. Qsra - a quality-value guided de novo short read assembler. **BMC Bioinformatics**. 10: 69, fev. 2009.

BUCKLEY, D. P. Genetics of Brazil nut (Bertholletia excelsa Humb. & Bonpl.: Lecythidaceae). **Theoretical and Applied Genetics**. 76(6), 923–928, dez. 1988.

BUTLER, J. et al. ALLPATHS: De novo assembly of whole-genome shotgun microreads. **Genome Research**. 18: 810–820, mai. 2008.

CABRAL, J. C. et al. Diversity and genetic structure of the native Brazil nut tree (*Bertholletia excelsa* Bonpl.) population. **Genet Mol Res**. V.16, n. 3, gmr16039702, jul. 2017.

CALI, D. S. et al. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. **Briefings in Bioinformatics**. bby017, abr. 2018.

CAMARGO, F. F. et al. Genetic variability for morphometric characteristics in brazilian nut parent trees from northern Mato Grosso, Amazon rain forest. **Acta Amazonica.** v.40, n. 4, p.705-710, dez. 2010.

CANTAREL, B. L. et. al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**. Genome Research**. 18:188–196, 2008.

CHAISSON, M. J.; PEVZNER, P. A. Short read fragment assembly of bacterial genomes. Genome Research. 18: 324–330, fev. 2008.

CHEN, X. et al. Sequencing of Cultivated Peanut,*Arachis hypogaea*, Yields Insights into Genome Evolution and Oil Improvement. **Molecular Plant.** Mar 2019.

CHEN, Y.C. et al. Effects of GC Bias in Next-Generation-Sequencing Data on *De Novo* Genome Assembly**. PLoS One**. v. 8, n. 4, p. e62856, abr. 2013.

CHIN, C-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. **Nature Methods**. V.13, n. 12, dez. 2016.

CLARK, S. C. et al. ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. **Bioinformatics**, v. 29, n.4, p. 435-443, fev. 2013.

CLARKE, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. **Nature Nanotechnology**. v. 4, p. 265–270, fev. 2009.

CNCFlora. ***Bertholletia excelsa*** **in Lista Vermelha da flora brasileira versão 2012.2 Centro Nacional de Conservação da Flora**. Disponível em <http://cncflora.jbrj.gov.br/portal/pt-br/profile/Bertholletia excelsa>. Acesso em 6 janeiro 2020.

COMAI, L. The advantages and disadvantages of being polyploid. **Nature Reviews Genetics**. v. 6, p.836–846, out. 2005.

COMMINS, J.; TOFT, C.; FARES, M. A. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. **Biological Procedures Online.** V. 11, n. 1, p. 52-78, abr. 2009.

COMPEAU, P.E.C.; PEVZNER, P. A.; TESLER, G. How to apply de Bruijn graphs to genome assembly. **Nature Biotechnology**. v. 29, p. 987-991, nov. 2011.

CONSELHO ESTADUAL DE MEIO AMBIENTE, PARÁ. **Resolução COEMA nº 54 de 24 de outubro de 2007**. Homologa a lista de espécies da flora e da fauna ameaçadas no Estado do Pará., Belém, PA, 2007.

CORDAUX, R.; BATZER, M. A. The impact of retrotransposons on human genome evolution. **Nature Reviews Genetics.** v.10, p. 691–703, out. 2009.

CUI, L. et al. Widespread genome duplications throughout the history of flowering plants. **Genome Research**, v. 16, n. 6, p. 738-749, jun. 2006.

DACCORD, N. et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. **Nature Genetics**. v. 49, n. 7, p. 1099–1106, jun. 2017.

DECANO, A. G. et al. Complete Assembly of *Escherichia coli* Sequence Type 131 Genomes Using Long Reads Demonstrates Antibiotic Resistance Gene Variation within Diverse Plasmid and Chromosomal Contexts. **Clinical Science and Epidemiology.** V. 4, n. 3, e00130-19, mai. 2019.

DEL ANGEL, V.D. et al. Ten steps to get started in Genome Assembly and Annotation [version 1; referees: 2 approved]. **F1000Research,** fev. 2018.

DENISOV, G. Consensus generation and variant detection by Celera Assembler. **Bioinformatics**. v.24, n. 8, p.1035–1040, abr. 2008.

DESCHAMPS, S. et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. **Nat. Commun**. 9: 4844, nov. 2018.

DESCHAMPS, S.; LLACA, V. Strategies for Sequence Assembly of Plant Genomes. In: ABDURAKHMONOV, I. Y. **Plant Genomics.** InTech, 2016.cap.3, p.46-66.

DOHM, J. C. et al.. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. **Genome Research**. 17: 1697–1706, nov. 2007.

EKBLOM, R.; WOLF, J. B. A field guide to whole-genome sequencing, assembly and annotation. **Evolutionary Applications**. v.7, n.9, p.1026-1042, jun. 2014.

EL-METWALLY, S. et al. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. **PLOS Computational Biology.** v. 9, n. 12, p. e1003345, dez. 2013.

ELSIK, C. G. et al. Creating a honey bee consensus gene set. **Genome Biol.** 8, R13, 2007.

FESCHOTTE, C.; JIANG, N.; WESSLER, S.R. Plant transposable elements: where genetics meets genomics. **Nature Reviews Genetics**. v.3. p.329–341, mai. 2002.

FOULONGNE-ORIOL, M. et al. Genome wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus.* **Fungal Genetics and Biology**. v. 55, p. 6–21, jun 2013.

GHURYE, J.; POP, M. Modern technologies and algorithms for scaffolding assembled genomes. **PLoS Comput Biol.** 15(6): e1006994, jun. 2019.

GNERRE, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. **Proceedings of the National Academy of Sciences**. v.108, n. 4, p.1513–1518, jan. 2011.

GOODWIN, S.; MCPHERSON, J.D.; MCCOMBIE, W.R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics.** v**.** 17, p. 333-351, jun. 2016.

GREEN, P. **Phrap documentation**. 1996. Disponível em: <http://www.phrap.org/phredphrap/phrap.html>. Acesso em: 18 de outubro de 2018.

GUREVICH, A. et al. QUAST: Quality assessment tool for genome assemblies. **Bioinformatics**.29(8), p.1072-1075, abr 2013.

HAAS, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. **Genome Biol**. 9, R7, 2008.

HAAS, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic **Acids Research**. 31:5654–5666, 2003.

HAZZOURI, K. M. et al. Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. **Nature Communications**. 6:8824, nov. 2015.

HEATHER, J. M.; CHAIN, B. The sequence of sequencers: The history of sequencing DNA**. Elsevier Reviews Genomics** v. 10, n. 1, p. 1-8, jan. 2016.

HENSON, J.; TISCHLER, G.; NING, Z. Next-generation sequencing and large genome assemblies. **Pharmacogenomics**. v.13, n.8, p.901-915, jun. 2012.

HERNANDEZ D. et al. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. **Genome Research**. 18: 802–809, mai. 2008.

HESLOP-HARRISON, J.S.; SCHWARZACHER, T. Organisation of the plant genome in chromosomes. **The Plant Journal**. v. 66, n.1, p. 18-33, mar. 2011.

HOFFMAN, J. I. et al. High-throughput sequencing reveals inbreeding depression in a natural population. **Proceedings of the National Academy of Sciences USA.** v.111, n.10, p. 3775– 3780, mar. 2014.

HOSSAIN, M.; AZIMI, N.; SKIENA, S. Crystallizing short-read assemblies around seeds. **BMC Bioinformatics**. 10: S16, jan. 2009.

HUANG, X. et al. PCAP: a whole-genome assembly program. **Genome Research**. v.13, p. 2164–2170, set. 2003.

HUANG, X; MADAN, A. CAP3: A DNA Sequence Assembly Program. **Genome Research**. 9: 868–877, set. 1999.

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. 4, jan. 2018.

HUNT, M et al. REAPR: a universal tool for genome assembly evaluation. **Genome Biology**.14:R47, mai 2013.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção da Extração Vegetal e da Silvicultura**. Rio de Janeiro, v. 31, p.1-54, 2016.
INTERNACIONAL UNION FOR CONSERVATION OF NATURE. AMERICAS REGIONAL WORKSHOP (CONSERVATION & SUSTAINABLE MANAGEMENT OF TREES, COSTA RICA). Bertholletia excelsa in IUCN Red List of Threatened Species. Version 2011.2, IUCN. IUCN.

JAILLON, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. **Nature**, v. 449, n. 7161, p. 463–467, set. 2007.

JAIN, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. **Genome Biology**. v.17, p. 239, 250, nov. 2016.

JAYAKUMAR, V.; SAKAKIBARA, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. **Briefings in Bioinformatics.** bbx147, nov. 2017.

JECK, W. R. et al. Extending assembly of short DNA sequences to handle error. **Bioinformatics**. v. 23, n. 21, p. 2942–2944, nov. 2007.

JIAO, W.-B.; SCHENEEBERGER, K. The impact of third generation genomic technologies on plant genome assembly. **Current Opinion in Plant Biology**. v. 36, p. 64–70, fev. 2017.

JIAO, Y. et al. Improved maize reference genome with single-molecule technologies. **Nature**, 546, p. 524–527, 2017.

JUDGE, K. et al. Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. **Microbial Genomics**. V. 2, n. 9, e000085, set. 2016.

KARDOS, M. et al. Genomics advances the study of inbreeding depression in the wild. **Evolutionary Applications.** v. 9, n.10, p. 1205-1218, dez. 2016.

KERSEY, P. J. Plant genome sequences_ past, present, future. **Current Opinion in Plant Biology,** v. 48, p. 1–8, abr. 2019.

KOITO, A.; IKEDA, T. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. **Frontiers in Microbiology**. v. 4, p. 1-28, fev. 2013.

KOLMOGOROV, M. et al. Assembly of Long Error-Prone Reads Using Repeat Graphs. **Nature Biotechnology**. v. 37, p. 540-546, abr. 2019.

KONO, N.; ARAKAWA, K. Nanopore sequencing: Review of potential applications in functional genomics. **Development, Growth & Differentiation**, 61(5), 316–326, 2019.

KOREN, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. **Genome Research**. v. 27, p. 722–736, mar. 2017.

KOSUGI, S.; HIRAKAWA, H.; TABATA, S. GMcloser: Closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. **Bioinformatics,** v. 31, n. 23, p. 3733-3741, dez. 2015.

KREMER, F. S.; MCBRIDE, A. J. A.; PINTO, L.S. Approaches for in silico finishing of microbial genome sequences. **Genetics and Molecular Biology**. 40(3), p. 553-576, jul-set. 2017.

KYRIAKIDOU, M. et al. Current Strategies of Polyploid Plant Genome Sequence Assembly. Frontiers in Plant Science. v. 9, n. 1660, nov. 2018.

LEWIN, H. A. et al. Earth BioGenome Project: Sequencing life for the future of life. **Proc Natl Acad Sci U S A**. v.115, n.17, p. 4325–4333, abr. 2018.

LI, C. et al. Genome Sequencing and Assembly by Long Reads in Plants. **Genes.** v. 9, n. 1, p. 6-14, dez. 2017.

LI, F.W.; HARKESS, A. A guide to sequence your favorite plant genomes. **Appl Plant Sci**., 6(3), e1030, mar 2018.

LI, H. et al. The Sequence alignment/map (SAM) format and SAMtools. **Bioinformatics**. 25, 2078-9, 2009.

LI, H. Minimap and Miniasm: fast mapping and de novo assembly for noisy long sequences. **Bioinformatics.** v. 32, n. 14, p. 2103–2110, jul. 2016.

LI, R. De novo assembly of human genomes with massively parallel short read sequencing. **Genome Research**. 20: 265–272, fev. 2010.

LIEBERMAN-AIDEN, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. **Science.** v.326, n. 5950, p. 289-293, out. 2009.

LUO, R.et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. **GigaScience**. v.1, n. 1, p. 2047-217X-1-18, dez. 2012.

MAGI, A. et al. Nanopore sequencing data analysis: state of the art, applications and challenges. **Briefings in Bioinformatics**. bbx062, jun. 2017.

MALLAWAARACHCHI, V. **Genome Assembly — The Holy Grail of Genome Analysis: Assembling the 2019 novel coronavirus genome**. 2020. Disponível em: <https://towardsdatascience.com/genome-assembly-the-holy-grail-of-genome-analysis-fae8fc9ef09c>

MARCAIS, G.; KINGSFORD, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. **Bioinformatics**, 27(6):764–70, mar. 2011.

MARTIN, J.A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics.** v. 12, p. 671-682, set. 2011.

MARTINS, A. M. **Sequenciamento de DNA, montagem de novo do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis***. Tese (Doutorado) - Curso de Biologia Molecular, Ciências Biológicas, Universidade de Brasília, Brasília, p. 198. 2013.

MASCHER, M. et al. A chromosome conformation capture ordered sequence of the barley genome. **Nature**. 544, p. 427–433. 2017.

MAUÉS, M. M. Reproductive phenology and pollination of the brazil nut tree (*Bertholletia excelsa* Humb. & Bonpl. Lecythidaceae) in Eastern Amazonia**.** IN: Kevan P & Imperatriz Fonseca VL (eds) - **Pollinating Bees - The Conservation Link Between Agriculture and Nature** – Brasília; Ministério do Meio Ambiente. p.245-254, 2002.

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, ago. 2014.

METZKER, M.L. Sequencing technologies – the next generation. **Nature Reviews Genetics.** v. 11, p. 31–46, jan. 2010.

MEYERS, L.A.; LEVIN, D.A. On the abundance of polyploids in flowering plants. **Evolution.** v. 60, n. 6, p. 1198–1206, jun. 2006.

MICHAEL, T. P. et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. **Nature Communications**. V. 9, n. 541, fev. 2018.

MILLER J. R. et al. Aggressive assembly of pyrosequencing reads with mates. **Bioinformatics**. 24: 2818–2824, dez. 2008.

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics.** v. 95, n. 6, p. 315–327, jun. 2010.

MILO, R.; PHILLIPS, R. **Cell biology by the numbers**. New York, NY: Garland Science, Taylor & Francis Group. 2016.

MINISTÉRIO DO MEIO AMBIENTE. **Instrução Normativa n. 6, de 23 de setembro de 2008. Espécies da flora brasileira ameaçadas de extinção e com deficiência de dados**, Diário Oficial [da] República Federativa do Brasil, Poder Executivo, Brasília, DF, 24 set. 2008. Seção 1, p.75-83, 2008.

MOCHIDA, K. et al. Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. **The Plant Journal.** v.89, n. 2, p.181-194, jan. 2017.

MORRISSEY, J. et al. Low-cost assembly of a cacao crop genome is able to resolve complex heterozygous bubbles. **Hortic. Res**. 6:44, abr. 2019.

MULLIKIN, J. C.; NING, Z. The Phusion Assembler. **Genome Research**. 13: 81–90, jan. 2003.

MURRAY, P. R.; ROSENTHAL, K. S.; PFAÜER, M. A. **Microbiologia Médica**. 5. ed. Madrid: Elsiver España, 2006. 976 p.

MYBURG, A. A. et al. The genome of *Eucalyptus grandis*. **Nature**. 510, 7505:356-62, jun. 2014.

MYERS, E.W. et al. A Whole-Genome Assembly of *Drosophila*. **Science**. 287: 2196–2204, mar. 2000.

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

NARZISI, G.; MISHRA, B. Comparing De Novo Genome Assembly: The Long and Short read. **PLoS ONE.** v.6, n.4, p. e19175v, abr. 2011.

NARZISI, G; MISHRA, B. Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons. **Bioinformatics**. v. 27, n.2, p. 153–160, jan. 2011.

NATURE BIOTECHNOLOGY. A reference standard for genome biology. **Nat. Biotechno**l. 36, 1121, dez. 2018. Disponível em:<https://doi.org/10.1038/nbt.4318>. Acesso em: 14 mar. 2020.

NCGAS - National Center for Genome Analysis Support. **Sequencing technology comparisons update**. Disponível em < https://ncgas.org/Blog_Posts/Blog%20Images/sequencing%20technology%20comparisons%20update%202019.htm>. Acesso em 14 jan. 2020.

OU, S.; CHEN, J.; JIANG, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). **Nucleic Acids Res**. 46(21):e126–e26. Nov. 2018.

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

PAAJANEN, P. et al. A critical comparison of technologies for a plant genome sequencing Project. **Gigascience,** v.8, n. 3, giy163, mar. 2019.

PACIFIC BIOSCIENCES. **Pacific Biosciences Launches New Sequel II system, Featuring ~8 Times the DNA Sequencing Data Output**. Abr. 2019. Disponível em:<https://www.pacb.com/press_releases/pacific-biosciences-launches-new-sequel-ii-system-featuring-8-times-the-dna-sequencing-data-output/>. Acesso em: 12 mar. 2020.

PARRA, G.; BRADNAM, K.; KORF, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. **Bioinformatics**, v. 23, n.9, p. 1061-1067, mai 2007.

PENG, X. et al. A Chromosome-Scale Genome Assembly of Paper Mulberry (*Broussonetia papyrifera*) Provides New Insights into Its Forage and Papermaking Usage. **Molecular Plant.** Fev. 2019.

PEVZNER, P.A.; TANG, H.; WATERMAN, M. S. An Eulerian path approach to DNA fragmente assembly**. Proceedings of the National Academy of Sciences of the United States of America.** 98: 9748–9753, ago. 2001.

PINTO, M. B. S. et al. Brazil Nut (Bertholletia excelsa, H.B.K.) Improves Oxidative Stress and Inflammation Biomarkers in Hemodialysis Patients. **Biological Trace Element Research**. v. 158, n. 1, p. 105-112, fev. 2014.

PRANCE, G.T.; MORI, A. S. Lecythidaceae. **Flora Neotropica**. 21:1-270, 1979.

PRYSZCZ, L.P.; GABALDÓN, T. Redundans: an assembly pipeline for highly heterozygous genomes. **Nucleic Acids Research**. v. 44, n. 12, p. e113, jul. 2016.

PUTNAM, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. **Genome Research**. v. 26, p. 342-350, fev. 2016.

RAYMOND, O. et al. The *Rosa* genome provides new insights into the domestication of modern roses. **Nature Genetics***.* v.50, p. 772–777, abr. 2018.

REIG-VALIENTE, J. L. et al. Genome-wide association study of agronomic traits in rice cultivated in temperate regions. **BMC Genomics.** 19:706, set. 2018.

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

RIBEIRO, J.E.L.S.; HOPKINS, M.J.G.; VICENTINI, A. ET AL. **Flora da Reserva Ducke: guia de identificação das plantas vasculares de uma floresta de terra-firme na Amazônia Central**. Manaus, AM: Instituto Nacional de Pesquisas da Amazônia, INPA, 1999.

RICE, E. S.; GREEN, R. E. New Approaches for Genome Assembly and Scaffolding. **Annual Review of Animal Biosciences**. V. 7, p. 17-40, fev. 2019.

RICHARDS, S. It's more than stamp collecting: how genome sequencing can unify biological research. **Trends Genet**. 31(7):411-21, jul. 2015.

RUAN, J.; LI, H. Fast and accurate long-read assembly with wtdbg2. **BioRxiv**. 2019.

SCHATZ, M. C.; DELCHER, A.L.; SALZBERG, S.L. Assembly of large genomes using second generation sequencing. **Genome Research.** v.20, p. 1165–1173, mai. 2010.

SCHATZ, M.; WITKOWSKI J, MCCOMBIE, W. R. Current challenges in *de novo* plant genome sequencing and assembly. **Genome Biology**. v. 13, p. 243, abr. 2012.

SCHMIDT, B. A fast hybrid short read fragment assembly algorithm. **Bioinformatics**. 25: 2279–2280, 2009.

SCHMIDT, M. H-W. De Novo Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing. **Plant Cell**. 29(10), p. 2336–2348, out. 2017.

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

SHENDURE, J. et al. DNA sequencing at 40: past, present and future. **Nature.** v. 550, p. 345–353, out. 2017.

SHENDURE, J.; JI, H. Next-generation DNA sequencing**. Nature Biotechnology**. v. 26, n.10, p. 1135-1145, out. 2008.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v.31, n.19, p. 3210-3212, out. 2015.

SIMPSON, J. T. ABySS: a parallel assembler for short read sequence data. **Genome Research**. v.19, p. 1117–1123, fev. 2009.

SOHN, J.; NAM, JW. The present and future of *de novo* whole-genome assembly. **Briefings in Bioinformatics**. v.19, n.1, p.23-40, jan. 2018.

SOLTIS, D. E.; VISGER, C. J.; SOLTIS, P. S. The polyploidy revolution then...and now: Stebbins revisited. **American Journal of Botany**, v. 101, n. 7, p. 1057–1078, jul. 2014.

SOMMER D. et al. Minimus: a fast, lightweight genome assembler. **BMC Bioinformatics**. 8: 64, fev. 2007.

STANKE, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. **Nucleic Acids Research**. 34(Suppl 2):W435–W439, 2006.

SU, M.; SANTOLA, S. W.; READ, T. D. Genome-Based Prediction of Bacterial Antibiotic Resistance. **Journal of Clinical Microbiology**. V. 57, n.3, e01405-18, mar. 2019.

SUTTON, G. G. et al. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. **Genome Science and Technology**. 1: 9–19, abr. 1995.

TÜRKTAS, M. et al. Sequencing of plant genomes – a review. **Turkish Journal of Agriculture and Forestry.** 2014.

VURTURE, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. **Bioinformatics.** V. 33, n. 14, p. 2202-2204, jul. 2017.

WANG, W. et al. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. **GigaScience.** 9(1), giz160, jan. 2020.

WARREN, R. L. et al. Assembling millions of short DNA sequences using SSAKE. **Bioinformatics**. 23: 500–501, fev. 2007.

WOOD et al. The frequency of polyploid speciation in vascular plants. **PNAS**. V. 106, n. 33, p. 13875-13879, ago. 2009.

XIA, M. et al. Improved de novo genome assembly and analysis of the Chinese cucurbit *Siraitia grosvenorii*, also known as monk fruit or luo-han-guo. **GigaScience**. v. 7, n. 6, giy067, jun. 2018.

XIAO, C. L. et al. MECAT: an ultra-fast mapping, error correction and de novo assembly tool for singlemolecule sequencing reads. **Nat Methods**. 14 (11): 1072-1074, nov. 2017.

XIAO, W. et al. Challenges, Solutions, and Quality Metrics of Personal Genome Assembly in Advancing Precision Medicine. **Pharmaceutics**. v. 8, n.2, abr. 2016.

XU, C-Q. et al. Genome sequence of *Malania oleifera*, a tree with great value for nervonic acid production. **Gigascience**. 8(2): giy164, jan. 2019.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**. v. 13, p- 329-341, mai. 2012.

YANG, J. et al. *De novo* genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. **GigaScience**. v. 8, n. 7, giz085, jul. 2019.

YANO, K. et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. **Nature Genetics**. v. 48, n. 8, p. 927–934. Jun. 2016.

YIN, D. et al. Genome of an allotetraploid wild peanut *Arachis monticola*: a de novo assembly. **GigaScience**. v. 7, n. 6, giy066, jun. 2018.

ZERBINO, D. R. et al. Pebble and rock band: Heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. **PLoS ONE**. 4: e8407, dez. 2009.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**. v. 18, p. 821–829, mar. 2008.

ZHANG, W. et al. A Sequence-Based Novel Approach for Quality Evaluation of Third-Generation Sequencing Reads. **Genes.** 10(1): 44, jan. 2019.

ANEXOS

ANEXO A − Protocolo da extração de DNA

As amostras da planta foram trituradas em nitrogênio líquido até o ponto de pó fino com auxílio de almofariz e pistilo pré-refrigerados. O pó resultante da trituração foi transferido para microtubos de 2 mL (não ultrapassando 100 mg por tubo), sendo adicionados 500 µL de tampão CTAB e 5 µL de solução RNase A. Em seguida os microtubos foram invertidos para mistura, incubados a 60 ºC por 1 hora e centrifugados por 10 minutos a 13.000 g.

Com o uso de ponteiras largas, o sobrenadante foi transferido para um novo tubo ao qual foi adicionado em igual volume a solução de Fenol/Clorofórmio/Álcool isoamílico. Novamente, os tubos foram invertidos vigorosamente até completa mistura, seguido de centrifugação por 5 minutos a 10.000 g. Assim, cuidadosamente a camada superior aquosa foi removida e inserida em um novo tubo. Esse processo é repetido até que a camada superior aquosa esteja limpa; após atingir esse resultado, se prossegue novamente com a adição em igual volume da solução de Fenol/Clorofórmio/Álcool isoamílico, inversão vigorosa dos tubos e centrifugação por 5 minutos a 10.000 g, sendo este processo repetido por mais duas vezes.

Dessa forma, camada superior aquosa foi transferida para um novo tubo, e o DNA foi precipitado com a adição de 0,7 de volume de isopropanol, assim os tubos foram invertidos gentilmente entre 15 a 20 vezes, incubados a -20 ºC por 15 minutos e centrifugados por 5 minutos a 15.000 g (para formação do *pellet*). O *pellet* então é lavado duas vezes com 250 µL de etanol (70%) gelado, brevemente seco, ressuspenso em 50 µL de água ultra purificada e dissolvido com agitação suave a 60 ºC por 1 hora.

Anexo B − Estimativas do tamanho genoma obtidas com k=22

read length 1000 max kmer coverage 1000 genome size: 534.6 Mb http://genomescope.org/analysis.php?code=v2IrQNAIyCtUxtAmtFRy
max kmer coverage 1,000 genome size: 534.6 Mb http://genomescope.org/analysis.php?code=7cMZEINX6M6gEtpO554r
read length 1000 max kmer coverage 10,000 genome size: 586.9 Mb http://genomescope.org/analysis.php?code=Z4k9wf99jwuK98uryzCH
read length 10,000 max kmer coverage 10,000 genome size: 586.9 Mb http://genomescope.org/analysis.php?code=cJgjcTW2gd9flYdfzRuC
read length 15,000 max kmer coverage 10,000 genome size: 586.9 Mb http://genomescope.org/analysis.php?code=W30yaNxuAHUKfmQgtYtS
read length 10,000 max kmer coverage 100,000 genome size: 596 Mb http://genomescope.org/analysis.php?code=WTBOEnMI5a7TkU99KfWR
read length 15000 max kmer coverage 1,000,000 genome size: 597 Mb http://genomescope.org/analysis.php?code=OJX8FbQE65iCq60FMtUj

ANEXO C − Tabelas resumidas das montagens de novo de *Bertholletia excelsa*

ANEXO C1 − Resumo dos dados da montagem com wtdbg2 obtidas através do QUAST.

wtdbg2						
Coverage	**47x**	**63x**	**60x_cor**	**97x**	**126x**	**187x**
# contigs	1645	2266	17504	1266	1374	1356
Largest contig (pb)	10210582	13088739	4345111	19268925	11850723	13061757
Total length (pb)	591411051	586787077	901650871	588112684	591422550	597523347
GC (%)	34,66	34,62	34,28	34,7	34,72	34,71
N50	1983893	1589793	114094	2437419	2229496	3149022
N75	922074	727954	40611	1170540	1187848	1425982
L50	84	101	1408	62	82	56
L75	195	240	4928	147	171	133
# N's per 100 kbp	0	0	0	0	0	0
% do genoma montado em relação à estimativa por citometria de fluxo	99,22	98,45	151,27	98,67	99,225	100,25

ANEXO C2 − Resumo dos dados da montagem com SMARTdenovo obtidas através do QUAST.

SMARTdenovo						
Coverage	**47x**	**63x**	**60x_cor**	**97x**	**126x**	**187x**
# contigs	2205	3803	465	1075	1027	732
Largest contig (pb)	2965487	1435613	12691500	6731163	5633454	14715192
Total length (pb)	629563127	569562240	627062221	648922074	651460819	6,49E+08
GC (%)	34,79	34,92	34,68	34,69	34,69	34,69
N50	551387	218844	3337434	1550750	1777024	2623932
N75	284142	114216	1759658	840966	989604	1478945
L50	344	782	61	121	120	72
L75	737	1678	126	262	240	155
# N's per 100 kbp	0	0	0	0	0	0
% do genoma montado em relação à estimativa por citometria de fluxo	105,62	95,56	105,20	108,87	109,30	108,94

ANEXO C3 − Resumo dos dados da montagem com MECAT2 obtidas através do QUAST.

MECAT2						
Coverage	**47x**	**63x**	**60x_cor**	**97x**	**126x**	**187x**
# contigs	1983	1309	931	688	607	539
Largest contig (pb)	6027679	6919104	12514354	18867244	24391902	30393383
Total length (pb)	627488129	614501610	619240864	607659065	608916729	5,94E+08
GC (%)	34,83	34,84	34,74	34,81	34,78	34,83
N50	1158703	2095483	4251011	4564686	5244646	7465209
N75	577835	1064280	2516717	2558651	3216654	3996619
L50	167	94	49	40	35	23
L75	353	196	96	83	72	50
# N's per 100 kbp	0	0	0	0	0	0
% do genoma montado em relação à estimativa por citometria de fluxo	105,28	103,10	103,89	101,95	102,16	99,60

ANEXO C4 − Resumo dos dados da montagem com Flye obtidas através do QUAST.

Flye						
Coverage	**47x**	**63x**	**60x_cor**	**97x**	**126x**	**187x**
# contigs	2909	2556	2953	2594	2539	
Largest contig (pb)	4504065	6581720	4913329	5925866	6510658	
Total length (pb)	631608319	630290332	638886281	629905677	629617182	
GC (%)	34,75	34,74	34,73	34,73	34,75	
N50	785330	1218812	666265	953380	965035	Não foi possível por falta de memória.
N75	296288	481899	290486	354009	373278	
L50	221	154	226	182	176	
L75	544	350	584	449	434	
# N's per 100 kbp	0,4	1,44	0,59	0,57	0,51	
% do genoma montado em relação à estimativa por citometria de fluxo	105,97	105,75	107,19	105,68	105,63	

ANEXO C5 − Resumo dos dados da montagem com Canu obtidas por Rodrigo Theodoro Rocha.

Canu	
Coverage	**60x**
# contigs	4305
Largest contig (pb)	11129176
Total length (pb)	831215698
N50	2007910
% do genoma montado em relação à estimativa por citometria de fluxo	139,46