Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Montagem *De novo* de genomas contrastantes de soja para estudo da resistência ao complexo de percevejos

Marcos Antonio de Godoy Filho

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração Genética e Melhoramento de Plantas

Piracicaba
2020

Marcos Antonio de Godoy Filho
Bacharel em Biotecnologia

Montagem *De novo* de genomas contrastantes de soja para estudo da resistência ao complexo de percevejos

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **JOSÉ BALDIN PINHEIRO**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração Genética e Melhoramento de Plantas

Piracicaba
2020

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Godoy Filho, Marcos Antonio de

Montagem *De novo* de genomas contrastantes de soja para estudo da resistência ao complexo de percevejos / Marcos Antonio de Godoy Filho. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - - Piracicaba, 2020.

71 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Soja 2. Genoma 3. Variantes 4. 10x Chromium 5. Percevejos I. Título

Dedico este trabalho aos meus pais e meus avós, com amor e muita admiração pelo exemplo de determinação e força.

Aos meus amigos e familiares que sempre estiveram presentes, gratidão.

AGRADECIMENTOS

A Deus pela vida e todas as graças concedidas que me permitiram chegar até aqui.

Aos meus pais Gislaine e Marcos por todo amor, carinho, apoio incondicional e principalmente pelo exemplo de superação que me proporcionaram, gratidão

Aos meus avós paternos Irene (*in memorian*) e José (*in memorian*), por todo amor e carinho dedicados em minha infância e aos meus avós maternos Helena e Antonio (*in memorian*), que sempre me apoiaram, amaram e foram meus maiores exemplos na vida, gratidão.

Ao Prof. Dr. José Baldin Pinheiro, por toda a preocupação com o trabalho e com minha pessoa, pela orientação e por todas as oportunidades durante esses dois anos, gratidão.

A Escola superior de agricultura "Luiz de Queiroz', ao Departamento de Genética e o CNPq, pelo apoio para a realização desta pesquisa.

Ao Dr. Steven J. Clough da Universidade de Illinois pela oportunidade de parceria que me permitiu a realização desta pesquisa.

Ao Lucas Borges da Universidade de Campinas pela ajuda com o *assembly* e em especial ao Dr. João Paulo Gomes Viana da Universidade de Illinois, por toda a ajuda com alguns *softwares* e principalmente pelas discussões e orientações que me ampliaram os horizontes.

Aos meus colegas do Laboratório de Diversidade e Melhoramento e do programa de Genética e Melhoramento da ESALQ, pelos dias e bons momentos compartilhados assim como os ensinamentos.

A Dr. Maria Imaculada Zucchi, pelo incentivo a ciência e amizade.

Ao Dr. Gabriel Rodrigues Alves Margarido, pela administração do servidor dayhoff do departamento de genética e por sua ajuda com o mesmo.

A Superintendência da Tecnologia da Informação da universidade de São Paulo, pelo apoio com o servidor águia.

A todos os funcionários do departamento de genética da ESALQ, em especial aos funcionários Marcio, Amaral e Claudinei.

A todos os professores do departamento de genética e também a todos os meus professores, do jardim de infância até a universidade, pelo inestimável conhecimento passado aos alunos.

As minhas duas orientadoras da graduação, Monalisa Sampaio Carneiro e Alessandra Alves de Souza, pelas oportunidades e ensinamentos, assim como meus colegas dos respectivos grupos de pesquisa.

A minha segunda e terceira família, as repúblicas Só-K-Bota da UFSCar de araras e Pikreta da ESALQ, pelos bons momentos compartilhados.

A meus amigos, por todo apoio e amizade, em especial a Alisson Mello, Antonio Beraldo, Bruna Lopes, Felipe Barreto, Lucas Girotto, Lucas Messa, Lucas Rabelo, Marina Oliveira e Thais Gonsales.

> “Se a seleção consistisse apenas em separar algumas variedades e raças muito distintas, usando-as depois para procriarem, o princípio de seleção seria tão óbvio que podia ser digno de menção, mas não de discussão. A sua importância reside no grande efeito produzido pela acumulação, num determinado sentido e ao longo de gerações sucessivas, de diferenças absolutamente imperceptíveis ao olho humano, a menos que muito treinado – diferenças que em vão tentei apreciar. São menos de um em mil, os homens que têm precisão no olhar e capacidade de discernimento suficientes para serem criadores de renome. Se alguém tiver estas qualidades, estudar o assunto durante alguns anos, e dedicar toda a sua vida à criação com uma perseverança invencível, então será bem sucedido, e poderá conseguir melhoramentos extraordinários nas suas crias”
>
> Charles Darwin

SUMÁRIO

RESUMO

Montagem *De novo* de genomas contrastantes de soja para estudo da resistência ao complexo de percevejos

A soja é uma importante *commodity* no Brasil, em 2019, a soja e seus derivados responderam por US$ 32,63 bilhões em exportações. No entanto, existem alguns obstáculos à obtenção de maior produtividade, como ataque de insetos, principalmente pelo complexo de percevejos, representados principalmente por *Euschistus heros*, *Nezara viridula* e *Piezodorus guildinii*, que diminuem o rendimento da cultura e podem transmitir fitopatógenos. Uma alternativa ao uso de inseticidas é o desenvolvimento de cultivares resistentes. Para um melhor entendimento da arquitetura genética envolvida na resistência da soja ao complexo do percevejo, foram sequenciados os genomas da cultivar de soja IAC-100 (resistente) e da cultivar CD-215 (suscetível), foram usados os dados de genotipagem por sequenciamento de uma população de 236 de linhagens endogâmicas recombinantes oriundas do cruzamento entre as cultivares estudadas, visando encontrar variantes comuns as mais resistentes e presentes apenas na cultivar IAC-100. Para a montagem dos genomas, utilizamos a tecnologia Chromium Genome (10x Genomics, San Francisco, EUA). A montagem do genoma foi realizada com os *softwares* Supernova 2.1.1, Ragoo 1.1 e REAPR 1.0.17, gerando montagens com aproximadamente 1 GB e *scaffold* N50 próximos de 54 MB, para ambas as cultivares nossos genomas apresentam de alta sintênia com os cromossomos da Williams 82 com número de genes e elementos repetitivos muito próximos (aproximadamente 58 mil genes e 44% de elementos repetitivos).A partir dos dados do sequenciamento do genoma, o *pipeline* do GATK foi usado para a descoberta de variantes e usando dados de genotyping-by-sequencing (GBS), de 236 indivíduos de uma população de linhagens endogâmicas recombinantes, oriundas do cruzamento entre a IAC-100 e CD-215, o *pipeline* do GATK foi usado para a chamada de SNPs. Foram selecionadas, a partir de dados históricos, as 25 linhagens mais resistentes ao complexo de percevejos, e polimorfismos compartilhados entre as mesmas e IAC-100 foram identificados. Foram identificados um total de 290.189 mil SNPs únicos na IAC-100 quando comparados com CD-215, além de 335.229 mil Indels únicos. A partir dos dados de GBS foram identificados 852 SNPs, compartilhados entre as 25 linhagens mais resistente da população e IAC-100, com esses SNPs muito próximos da posição de QTLs relatados para a resistência. Variantes estruturais foram visualizadas com o *software* IGV na região de QTLs relatados para a resistência, onde grandes variantes únicas na IAC-100, foram encontradas próximas de genes considerados candidatos a resistência, envolvidos com a biossíntese de fenilpropanóides/lignina, terpenos, número de vagens e desenvolvimento de sementes.Além de disponibilizar dois genomas de soja brasileiros com os melhores *scaffolds* N50 até hoje relatados, esses resultados abrem as portas para vários outros estudos e aplicações no melhoramento, como mapeamento fino, seleção assistida, expressão de genes candidatos e modificações genéticas nos mesmos usando CRISPR ou outros métodos.

Palavras-chave: Soja, Genoma, Variantes, 10x Chromium, Genotipagem por sequenciamento, Percevejos

ABSTRACT

De novo assembly of contrasting soybean genomes for the study of resistance to the stink bug complex

Soybean is an important commodity in Brazil in 2019, the soybean and its derivatives accounted for $ 32,63billion in exports. However, there are some obstacles to achieving greater productivity as insect attacks, particularly by stink bug complex that reduce crop yield and can transmit pathogens. An alternative to the use of insecticides is the development of resistant cultivars. For a better understanding of the genetic architecture involved in soybean resistance to the bed bug complex, the genomes of soybean cultivar IAC-100 (resistant) and cultivar CD-215 (susceptible) were sequenced. Genotyping data by sequencing a population of 236 recombinant inbred lines from the crossing of the two cultivars studied were also used, to find common variants, the most resistant and present only in the cultivar IAC-100. The genome assemblies, we use the Chromium Genome technology (10x Genomics, San Francisco, USA). The genome assemblies was performed with the *software* Supernova 2.1.1, 1.0.17 Ragoo 1.1, and REAPR generating assemblies with about 1 GB and N50 *scaffold* next 54 MB for both genomes cultivars. Our genomes have high synteny with the Williams82 chromosomes with several genes and repetitive elements very close (approximately 58 thousand genes and 44% repetitive elements). From the genome sequencing data, the GATK pipeline was used to discover variants and using data from genotyping-by-sequencing (GBS), of 236 individuals from a population of recombinant inbred lines, originated from the crossing between the IAC -100 and CD-215, the GATK pipeline was used for calling SNPs. The 25 most resistant genotypes to the bed bug complex were selected from historical data, and polymorphisms shared between them and IAC-100 were identified. A total of 290.189 thousand unique SNPs were identified in the IAC-100 when compared to CD-215, in addition to 335.229 thousand unique Indels. From the GBS data, 852 SNPs were identified, shared among the 25 most resistant genotypes in the population and IAC-100, with these SNPs very close to the position of QTLs reported for resistance. Structural variants were visualized with the IGV software in the region of QTLs reported for resistance, where large unique variants in IAC-100, were found close to genes considered candidates for resistance, involved with the biosynthesis of phenylpropanoids / lignin, terpenes, number of pods and seed development. In addition to providing two Brazilian soybean genomes with the best N50 scaffolds reported to date, these results open the door to several other studies and applications in breeding, such as fine mapping, assisted selection, expression of candidate genes and genetic modifications in them using CRISPR or other methods.

Keywords: Soybean, Genome, Variants, 10x Chromium, Genotyping by sequencing, Stink bugs

1 INTRODUÇÃO

A soja (*Glycine max* [L.] Merrill) é um dos grãos mais importantes no mundo. As sementes são uma excelente fonte de proteína, óleo e micronutrientes, podendo ser utilizada na alimentação animal, produção de bicombustíveis e plásticos. (Costa Neto; Rossi.,2000).

O Departamento de Agricultura dos Estados Unidos (UNITED STATE DEPARTMENT OF AGRICULTURE – USDA) estimou que a Produção Global de Soja em 2019/2020 foi de 335,35 milhões de toneladas. Segundo a CONAB (COMPANHIA NACIONAL DE ABASTECIMENTO), a safra 2019/2020 de soja no Brasil teve uma produtividade de aproximadamente 120 milhões de toneladas, ranqueando o país como o maior produtor de soja do mundo, superando os Estados Unidos. No ano de 2018, as exportações de soja, óleo de soja e farelo de soja somaram aproximadamente US$ 32,63 bilhões, sendo a principal *commodity* do Brasil (CONAB.,2019). Entretanto, existem entraves para se lograr elevada produtividade, como a incidência de doenças e insetos-praga que afetam a cultura.

Segundo Pimentel (1997), fitopatógenos, insetos-praga e plantas daninhas são responsáveis pela perda de mais de 40% da produção de alimentos mesmo com o uso em larga escala de agroquímicos, que pode gerar diversos problemas ambientais e de saúde. De acordo com a Organização Mundial de Saúde (OMS, 1999), anualmente, três milhões de pessoas sofrem intoxicações por agroquímicos, destes, mais de 2 milhões de casos acontecem nos países em desenvolvimento. Desse modo, alternativas ao uso de produtos químicos na agricultura são muito importantes, como o desenvolvimento dos métodos de controle biológico de pragas e plantas resistentes.

Dentre os principais insetos-praga da soja, encontra-se o complexo de percevejos, esses insetos podem variar em cor e tamanho, mas apresentam características comuns, como corpos arredondados ou ovalados, aparelho bucal do tipo sugador, antenas que apresentam um total de cinco segmentos e escutelo em forma de triângulo (Panizzi et al.,2000). Os hábitos alimentares desses insetos são variados, eles podem ser herbívoros, predadores e as vezes até mesmo onívoros, apresentando hábitos generalistas ou especialistas em preferências alimentares (Koch et al.,2019).

Os percevejos se alimentam de quase todas as partes da planta, caules, pecíolos, folhas, flores, frutos e sementes, sua alimentação ocorre inserindo seu aparelho bucal nos tecidos das plantas, injetando enzimas digestivas e sugando nutrientes disponíveis no local. Na soja, a preferência alimentar desses insetos-praga, são vagens e sementes em

desenvolvimento, onde ao realizar a inserção do aparelho bucal, causam lesões mecânicas e químicas, ocasionadas pelas enzimas injetadas (Goodwin; McPherson, 2000).

Dentre as fases de desenvolvimento do percevejo, o quinto ínstar e os adultos causam danos mais graves do que os primeiros instares, sendo que os danos mecânicos causados pelas diferentes espécies podem variar devido a duração da alimentação e da profundidade das lesões nas sementes, devido a variação do aparelho bucal. (Corrêa-Ferreira; De Azevedo.,2002).

Em geral, os danos diretos da alimentação durante o desenvolvimento das vagens e sementes pode resultar em perda de vagens, aborto do enchimento dos grãos e vagens vazias, a alimentação direta das sementes durante o seu desenvolvimento resulta em sementes murchas, deformadas e pequenas; além de causar uma ligeira deformação das sementes e resultar em marcas de punção, afetando assim o rendimento, qualidade das sementes e também as taxas de germinação (Koch et al.,2019).

Figura 1: Danos nas sementes de soja ocasionados pela alimentação de percevejos, da esquerda para a direita temos o aumento do dano devido a maior alimentação dos percevejos. Fonte: Koch et al., 2019

Dentre as espécies de percevejos que mais se destacam no Brasil, são três pentatomídeos fitófagos: *Euschistus heros* (Fabricius) ou percevejo marrom, que no Brasil que ataca a soja geralmente de Novembro a Abril, permanecendo o restante do ano em estado de dormência, atualmente o mesmo é o mais encontrado, podendo chegar a 84% da população em certas lavouras (Corrêa-Ferreira et al.,2010a). Também encontramos o *Nezara viridula* (Linnaeus) que, ao contrário do percevejo marrom, pode ser encontrado o ano todo, utilizando outras plantas como hospedeiras na entressafra da soja, e por fim, temos *Piezodorus guildinii* (Westwood) ou percevejo verde-pequeno, que também utiliza outras plantas como hospedeiras na entressafra da soja, sendo considerado o mais prejudicial, causando maiores

danos à qualidade dos grãos e maior retenção foliar (permanência da planta na fase vegetativa, ocasionando a maturação desuniforme dos grãos) (Sosa-Gómez; Moscardi.,1995); (Corrêa-Ferreira; Panizzi.,1999). Entretanto, podemos encontrar outras espécies dos gêneros *Dichelops*, *Acrosternum*, *Edessa* e *Thyanta*, que também afetam negativamente a produtividade e qualidade da soja (Hoffmann-Campo et al.,2000).

Além dos danos diretos ocasionados pela alimentação, os percevejos podem transmitir o fungo *Eremothecium coryli*, que afeta a qualidade das sementes e seu valor comercial, além de outros fitopatógenos, causando mais prejuízos aos produtores (Kimura et al.,2008).

Uma estimativa realizada no estado do Mato Grosso, demonstrou uma diminuição da produtividade de até 30% devido à ocorrência de percevejos (Vivan; Degrande, 2011), além de perdas que chegam a 600 milhões de dólares por ano no país. (Cepea/Esalq; Andef, 2017).

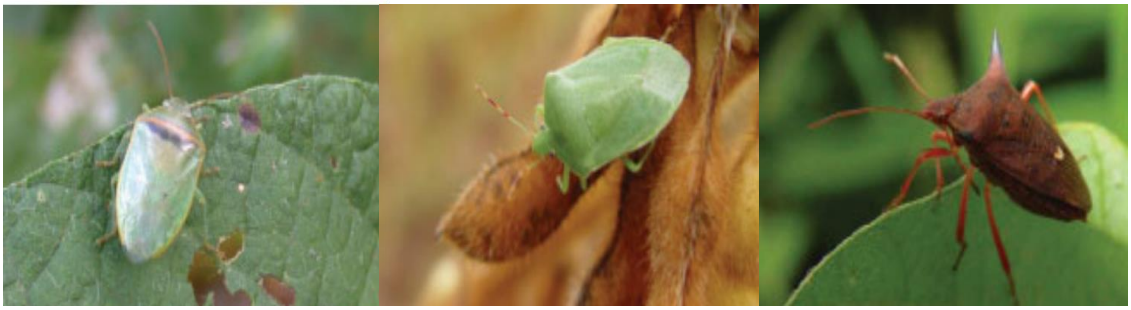


Figura 2: Principais percevejos da soja, da esquerda para a direita, *Piezodorus guildinii, Nezara viridula, Euschistus heros.* Fonte: Acervo fundação MT.

Dentre os estádios fenológicos da cultura, a fase reprodutiva da soja vai de R1, correspondente ao início da floração, até a fase R8, onde temos a maturação, e posterior colheita. A fase R5, que corresponde ao início do desenvolvimento dos grãos, é dividida em mais cinco subfases, cada uma correspondendo a certa porcentagem do desenvolvimento dos grãos, até R7, onde os grãos cessam seu desenvolvimento. (Fer; Caviness.,1977). Nas fases de R5 temos os maiores danos causados pelos percevejos, o que coincide com a fase de crescimento de sua população, atingindo um pico populacional ao fim desta fase (Corrêa-Ferreira; Krzyzanowski; Minami.,2009).

Para controlar a população de percevejos, são utilizados principalmente inseticidas, o que gera altas despesas aos produtores. Segundo Corrêa-Ferreira; Krzyzanowski; Minami, (2009), o controle de percevejos através de inseticidas tem sido pouco eficiente, devido ao aumento da população resistente aos produtos. Diversas populações de insetos resistentes já

são encontradas e, futuramente, podem se tornar um grande problema aos produtores (Gomez; Roggia et al., 2012).

Uma alternativa ao uso de inseticidas, com menor custo aos produtores e um menor impacto ambiental é o uso de cultivares resistentes a insetos. Essa característica não tem sido uma prioridade dos programas de melhoramento genético, entretanto, existem algumas variedades de soja consideradas resistentes, mesmo não se conhecendo em muitos detalhes os fatores genéticos e bioquímicos da resistência (Pinheiro et al., 2016).

No Brasil, o Instituto Agronômico de Campinas (IAC) foi o primeiro a introduzir linhagens resistentes a insetos, com a incorporação das linhagens PI 227687, PI 229358, PI 171451, PI 171444 e linhagens derivadas da PI 274454. O instituto já liberou 5 cultivares resistentes IAC 100, IAC 17, IAC 19, IAC 23 e IAC 24, sendo as três primeiras consideradas resistentes ou tolerantes ao complexo de percevejo (Pinheiro et al., 2016).

Desde 1990, a Escola Superior de Agricultura Luiz de Queiroz ESALQ-USP vem trabalhando num programa de melhoramento genético de soja visando o desenvolvimento de cultivares resistentes a insetos (Pinheiro et al.,2016). Além destas instituições, a Embrapa Soja lançou cultivares resistentes aos percevejos, que devem estar disponíveis no mercado nas próximas safras (Embrapa, 2018).

Apesar de sua grande importância, ainda são escassos os conhecimentos sobre a resistência de percevejos, sendo poucos os QTLs ou alelos identificados até o momento. Como essa é uma herança poligênica e com alta influência ambiental (Godoi; Pinheiro; 2009), a montagem do genoma de cultivares contrastantes para a identificação de variantes e de genes candidatos, que podem estar associados à resistência ao complexo de percevejos, pode ser uma poderosa ferramenta.

Desse modo, um dos objetivos desse trabalho, foi montar os genomas das cultivares brasileiras IAC-100 e CD-215, a primeira resistente ao complexo de percevejos e a segunda suscetível. Essas cultivares foram utilizadas em diversos estudos de resistência e são os genitores de uma população de linhagens endogâmicas recombinantes (RILs), utilizadas para o mapeamento de QTLs para resistência ao complexo de percevejos, além de serem alvo de estudos de expressão gênica diferencial. Outro objetivo desse trabalho, é utilizar dados de genotyping-by-sequencing (GBS) da população de RILs, para identificar SNPs compartilhados entre os genótipos mais resistentes dessa população e o genitor IAC-100, portador dos alelos de resistência.

Além disso, a partir dos dados de sequenciamento das cultivares IAC-100 e CD-215, iremos selecionar grandes variantes estruturais (> 1 Kb) presentes apenas na IAC-100, na

região de QTLs já mapeados para a resistência ao complexo de percevejos e identificar genes candidatos, próximos a essas grandes variantes.

Com a identificação de variantes e genes candidatos, abrem-se as portas para novos estudos, como a utilização de CRISPR e RNA de interferência em trabalhos de silenciamento gênico, visando comprovar genes e vias envolvidas na resistência, além da possibilidade do desenvolvimento de plantas geneticamente modificadas e o desenvolvimento de marcadores moleculares para mapeamento fino e seleção assistida, que facilita e acelera o trabalho de lançar novas cultivares resistentes, algo importante, pois atualmente as cultivares resistentes são pouco competitivas em termos de produtividade e adaptabilidade quando comparadas com as cultivares comerciais, não sendo, portanto, utilizadas pelos produtores.

De maneira geral, conseguiremos fornecer genomas e variantes de qualidade que ampliando a possibilidade de futuros estudos sobre a resistência ao complexo de percevejos. Além disso, teremos disponíveis dois genomas de cultivares de soja brasileiras, o que não temos disponível na literatura até o momento, representando não apenas um grande avanço no estudo da resistência ao complexo de percevejos, mas também, um grande avanço para a genômica de soja.

1.1 Mecanismos de defesa contra Insetos

Inicialmente, a planta precisa reconhecer a interação com os insetos, para então poder desencadear uma resposta de defesa, para isso ela usa padrões moleculares associados a insetos herbívoros (HAMPs), que são moléculas conservadas que apresentam funções essenciais no inseto (Mithöfer; Boland, 2008); (Medzhitov; Janeway, 1997). Os HAMPs são identificados por receptores de reconhecimento de padrões moleculares (PRRs), esses receptores são proteínas transmembranas que, ao reconhecer HAMPs, dão início a uma cascata de sinais bioquímicos que irá regular positiva ou negativamente diversos genes envolvidos na defesa da planta (Dalio et al., 2014).

As plantas, de modo geral, apresentam diversos mecanismos de defesa contra insetos além do anteriormente exemplificado para insetos herbívoros. De forma primária, os mecanismos envolvem a presença de barreiras físicas, que incluem tricomas, espinhos e ceras, que podem impedir que insetos se liguem a superfície da planta, se alimentem, e até mesmo ovipositem. Um exemplo desse mecanismo físico é o caso do ácaro *Tetranychus urticae*, que teve sua oviposição reduzida significativamente em genótipos de framboesa com altas densidades de tricomas foliares (White; Eigenbrode,2000); (Karley et al., 2016).

Outro mecanismo de defesa é a liberação de voláteis, induzidos por herbívoros (HIPVs), estes compostos são capazes de atrair predadores, como demonstrado por Bruce et al. (2015), que transformou uma cultivar de trigo para produzir um feromônio de defesa, que conseguiu repelir pulgões e atrair predadores em condições controladas. Esses compostos, induzidos pelos insetos, são capazes ainda de induzir respostas de defesa em plantas vizinhas (Erb et al., 2015).

Muitos compostos químicos do metabolismo secundário das plantas podem ser induzidos a partir da percepção dos insetos, como terpenóides, alcalóides, fenóis, isoflavonas, flavonóides e fitoalexinas (Hanley et al., 2007).

Os terpenóides ou isoterpenóides apresentam uma alta gama de estruturas, com mais de 25 mil tipos de moléculas (Chen et al., 2011). Essas moléculas podem ser voláteis ou não, podendo atuar contra insetos como repelentes, afetando o crescimento ou sendo tóxicos, aumentando a mortalidade dos mesmos, as mesmas já foram relatadas em inúmeras culturas afetando diversas espécies de insetos como *Manduca sexta, Myzus persicae, Bemisia tabaci, Tetranychus urticae, Ips typographus, Diaphorina citri, Ips typographus, Pissodes strobe, Tetranychus urticae, Phyllotreta nemorum* (Kortbeek et al., 2018)

Flavonóides e isoflavonóides (Figura 3) foram relatados como dissuasores de alimentação de insetos, sendo citotóxicos, protegendo assim as plantas contra insetos, influenciando seu comportamento, crescimento e desenvolvimento (Simmonds et al., 2001); (Simmonds et al., 2003).

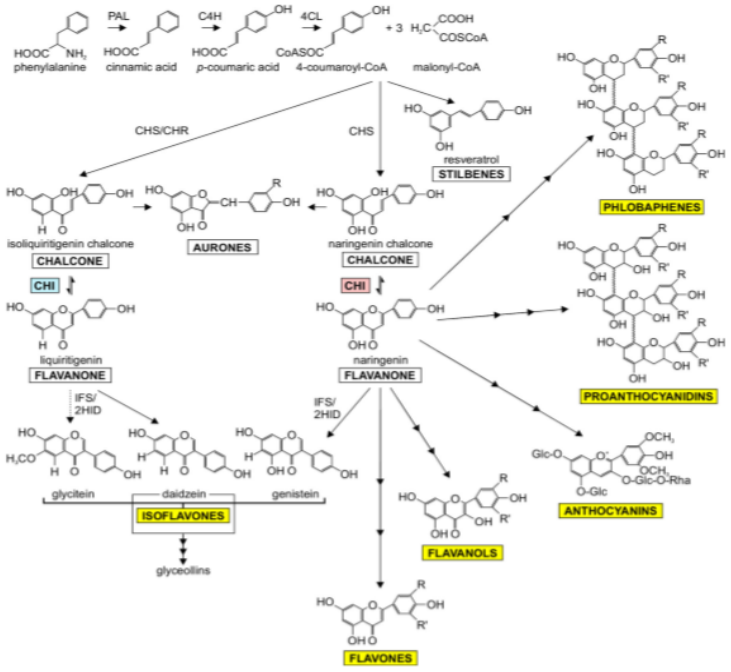


Figura 3: Rota metabólica Fenilpropanóide na soja, fonte: Dastmalchi & Dhaubhadel, 2014

A presença de isoflavonóides vem sendo relatada em diversos estudos sobre resistência de plantas a insetos. A genisteína e a daidzeína (isoflavonas) são conhecidas por serem dissuasoras do nematóide do solo *Radopholus similis*, além de ser responsável pela inibição do crescimento do pulgão *Aphis craccivora* (Harborne; Wuyts, 2000).

Piubelli et al. (2003b) demonstraram que sementes imaturas de cultivares de soja consideradas resistentes, após serem usadas como alimento pelo percevejo *Nezara viridula*, apresentaram maior teor de genisteína e daidzeína do que o controle. Como a linhagem PI 227687, apresentou os maiores teores de genisteína e daidzeína, antes e depois da infestação por insetos, Piubelli et al. (2003b) utilizaram um extrato da PI 227687, para conduzir um experimento de preferência alimentar, tratando vagens de uma variedade suscetível (BR- 16) com água ou o extrato da linhagem PI 227687 após a alimentação de *Nezara viridula*. As vagens de BR-16 tratadas com água, foram 4 vezes mais procuradas para a alimentação do que as tratadas com o extrato, evidenciando um efeito de dissuasão do inseto, provavelmente por baixa palatabilidade.

A espécie *Nezara viridula*, apresentou alta mortalidade quando alimentada com a linhagem PI 227687, e IAC-100 (66.2 %, e 48.8% respectivamente), ambas resistentes, apresentando altos valores de genisteína e a daidzeína quando comparadas com linhagens suscetíveis após alimentação do percevejo. Os percevejos alimentados com a variedade BR-16 apresentam uma taxa de mortalidade de apenas 27.5%. O peso médio de indivíduos adultos também foi menor quando os insetos foram alimentados com PI 227687 e IAC-100, em comparação com BR-16. Fêmeas do percevejo também acumularam menos lipídio quando se alimentaram de ‘IAC-100’ (4,8 mg) e PI 227687 (4,3 mg), o que pode influenciar negativamente a reprodução do inseto (Piubelli et al., 2003a).

Segundo Silva et al. (2013), PI 227687 e "IAC-100," causaram alta mortalidade de ninfas de *P. guildinii* (maior que 90%), o que corrobora os resultados obtidos por Piubelli et al. (2003a). No trabalho conduzido por Souza et al. (2013), os autores demonstraram que *N. viridula* possui menor preferência alimentar por PI 227687, corroborando o efeito de dissuasão demonstrado também por Piubelli et al. (2003b).

A cultivar chinesa 'Zhongdou 27', apresenta altos níveis de isoflavonóides sendo considerada resistente ao pulgão de soja (*Aphis glycines* Matsumura), um inseto sugador, assim como os percevejos. Um estudo de mapeamento utilizando essa cultivar, mapeou dois QTLs associados à resistência ao pulgão e aos altos níveis de isoflavonóides, o estudo ainda conclui que os dois QTLs são fortes candidatos a serem utilizados em MAS (seleção assistida

por marcadores), visando resistência ao pulgão e aumento do conteúdo de isoflavonóides (Meng et al., 2011).

A partir de uma população de 228 plantas F2, oriunda do cruzamento entre as cultivares IAC-100 e CD-215, Moller (2010) avaliou diversas características associadas à tolerância ao complexo de percevejos, período de granação (PEG), retenção foliar (RF), número de vagens por planta (NVP), índice percentual de dano nas vagens (IPDV), número de sementes (NS), peso de cem sementes (PCS), peso de sementes boas (PSB) e peso de sementes manchadas (PSM). Através de um modelo restrito de múltiplos QTLs e análise de Kruskal-Wallis, foram identificados oito QTLs para PSM, três QTLs para PSB, um QTL para PG e um QTL para VA, sendo esse o primeiro registro na literatura de QTLs associados à resistência ao complexo de percevejos.

A partir da população F2 utilizada por Moller (2010), foi originada uma população F2:3 de 15 indivíduos, divididos em 3 repetições para avaliação fenotípica de caracteres agronômicos e caracteres associados à resistência ao complexo de percevejos. Um total de 29 QTLs foram identificados para características associadas a resistência a percevejos, para isso, foi utilizado o método de mapeamento de intervalos múltiplos univariado (MIM), as características avaliadas foram as mesmas do trabalho anterior. (Moller, 2017)

Um estudo comparando a expressão gênica entre as variedades IAC-100 e CD-215 (resistente e suscetível respectivamente) ao percevejo *Piezodorus guildinii*, demonstrou diferenças constitutivas na expressão gênica das duas variedades, o gene de uma proteína de ligação a oxysterol, presente no metabolismo de terpenos, foi identificada 20 vezes mais expressa em IAC-100, o mesmo gene foi identificado também dentro de um QTL associado à resistência ao percevejo, o gene de uma proteína semelhante a aliinase, enzima que sintetiza aliina, um sulfóxido conhecido por ser tóxico para muitos insetos, também foi relatada diferencialmente expressa, 40 vezes mais em IAC-100, e seu gene também está localizado dentro de um QTL associado à resistência aos percevejos. Também foram identificados dois fatores de transcrição homólogos ao do WRKY, envolvido na via do ácido jasmônico, além de proteínas de repetição rica em leucina (LRR) (Santos, 2012). O fator de transcrição WRKY60 é encontrado com maior expressão em soja tolerante à pulgão, podendo ser responsável por regular vias envolvidas na tolerância, além disso esses fatores de transcrição são relatados como indutores de proteínas PR no arroz e no trigo, proteínas que contribuem para resistência a patógenos no arroz e ao pulgão do trigo (Chapman et al., 2018).

Outro trabalho de expressão gênica utilizando também as variedades IAC-100 e CD-215 demonstrou que as vias ligadas ao metabolismo de lipídios, que está relacionada com as

vias do ácido jasmônico, carotenóides, terpenos e isoflavonóides são diferencialmente expressas após 24 horas de infestação de *Piezodorus guildinii*, entretanto, na cultivar IAC 100, essas rotas aparecem com maiores níveis de transcrição do que na variedade suscetível CD-215 (Silva, 2014).

1.2 Genomas e tecnologias de Sequenciamento

Em 1977 tivemos a introdução da primeira tecnologia de sequenciamento, a de Sanger, isso permitiu que genomas fossem sequenciados, gerando um grande salto, para o entendimento dos seres vivos e abrindo portas para inúmeros estudos e novos questionamentos (Mardis, 2017). O primeiro genoma completo foi o do bacteriófago phi-x174 com apenas 5375 bases (Sanger et al., 1977), algum tempo depois o método *Shotgun sequencing* usando a tecnologia de Sanger foi desenvolvida, gerando o sequenciamento do bacteriófago lambda (Sanger et al., 1982).

O sequenciamento de Sanger foi a base da genômica, responsável, por exemplo, pelo genoma humano e também o genoma da *Arabidopsis thaliana*. Em meados dos anos 2000 tivemos o surgimento das plataformas de segunda geração de sequenciamento, fazendo com que o custo para o sequenciamento de genomas caísse drasticamente, sendo que o genoma humano poderia ser agora sequenciado por um custo 50.000 vezes menor do que o original, montado a partir da tecnologia Sanger (Goodwin et al., 2016).

A principal tecnologia dentre as plataformas de segunda geração de sequenciamento, também chamadas de nova geração de sequenciamento (NGS), é a Illumina, que consiste inicialmente na fragmentação do DNA, geralmente, em sequências de 250-250 pb, seguido da ligação de *primers* nas extremidades das moléculas fragmentadas. O DNA é então desnaturado e fica como fita simples, onde um dos seus primers se ligam a oligonucleotídeos que estão fixos em uma *flow cell*, posteriormente a sequência passa por uma amplificação em fase sólida, também chamada de amplificação por ponte, nessa etapa temos a criação de *clusters* que contém a mesma sequência de DNA. O processo chega então ao sequenciamento, onde uma DNA polimerase adiciona bases que emitem um sinal luminoso detectado por um *laser*, os *clusters* são formados exatamente para amplificar o sinal, já que todas as sequências ali presentes são iguais, assim o *laser* pode detectar com uma maior confiabilidade a base correta (Metzker, 2010).

Dentre os equipamentos Illumina, temos um custo por gigabase sequenciado variando entre 7 dólares nos equipamentos mais modernos e 996 dólares nos equipamentos mais

antigos, com uma variação de 1.6 GB de dados de saída, como em um Illumina MiniSeq High output, com um tempo total de 7 horas, até 900GB por *flowcell* em um Illumina HiSeq X, com um tempo total de aproximadamente 3 dias. Desse modo pode-se ter o sequenciamento de um genoma em apenas algumas horas ou dias (Goodwin et al., 2016).

A partir de 2010 temos a terceira geração de sequenciadores, estes conseguem gerar moléculas de 8 kb a 40 kb, muito maiores do que as *reads* do sequenciamento Illumina que geralmente tem entre 150 e 250 pb, esse tamanho de molécula auxilia a montagem de genomas complexos e aumenta a continuidade dos genomas, além de melhorar a identificação de variantes estruturais, entretanto esse tipo de sequenciamento apresenta uma taxa de erro de até 13% enquanto a tecnologia Illumina tem em torno de 1% de erro. O custo em sequenciamentos de terceira geração é bem superior ao da segunda geração quando comparado com os equipamentos mais modernos da Illumina, onde temos um custo médio próximo dos 50 dólares contra custos que chegam a 1000 dólares (Goodwin et al., 2016); (Jung et al., 2019); (Mardis, 2017)

O sequenciamento de terceira geração da Pacific Biosciences é muito utilizado, ele acontece a partir de longos fragmentos de DNA que receberam adaptadores circulares, criando uma molécula circular de DNA. Essa molécula circular é sequenciada em uma célula especializada com milhares de poços com fundos transparentes, chamadas de guias de onda de modo zero, no fundo desse poço temos uma polimerase fixa, esta vai adicionando nucleotídeos fluorescentes, que devido as guias de modo zero podem ser lidas por um *laser* que em tempo real identifica as bases adicionadas. (Mardis, 2017)

Dentre a terceira geração de sequenciamento podemos citar também o 10X Chromium Genome technology, que simula o sequenciamento de *reads* longas, onde teremos grandes fragmentos de DNA de até 100 Kb particionados em pequenas esferas chamadas de GEMs, cada uma das GEMs contém um *barcode* único , assim as grandes sequências de DNA são fragmentadas e recebem esse *barcode* único, temos então o sequenciamento Illumina gerando pequenas *reads*, estas contém os *barcodes* que identificam de qual grande molécula ela foi originada, assim na hora da montagem conseguimos reconstruir essas grandes moléculas através das *linked reads*. Esse método apresenta algumas vantagens, como um custo menor em relação às outras tecnologias de terceira geração e a qualidade do sequenciamento Illumina com uma baixa taxa de erros (Mardis, 2017); (Jiao et al., 2017).

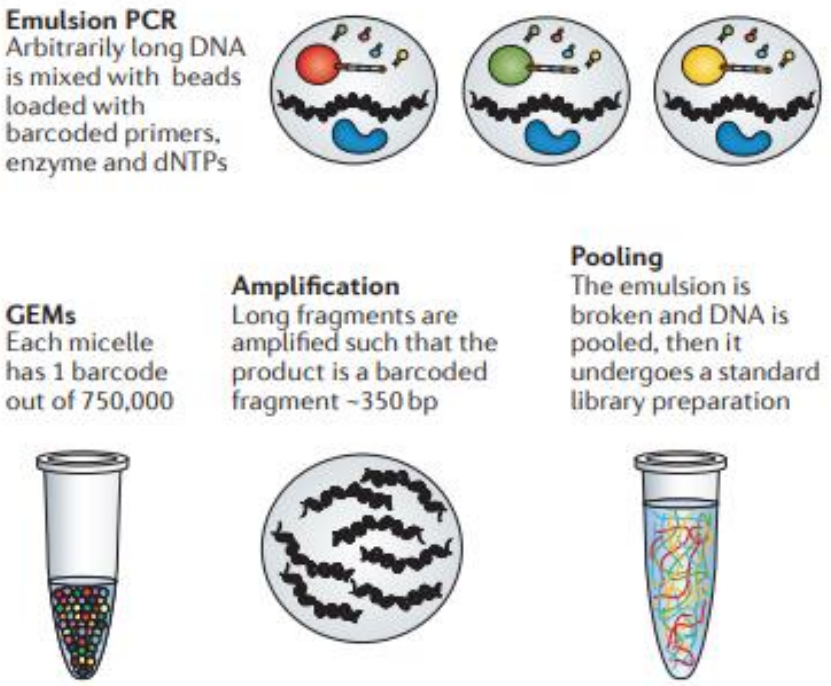


Figura 4: 10X Chromium Genome technology, fonte: Goodwin et al.,2016

Com o avanço das tecnologias de sequenciamento e a diminuição dos custos, temos mais de 14.000 genomas depositados no US National Center for Biotechnology Information (NCBI) e não temos apenas genomas de referência de espécies, com a diminuição dos custos, já possuímos o genoma de diversos indivíduos da mesma espécie (Goodwin et al., 2016).

Apesar do alto número de genomas disponíveis no NCBI, para a maioria das espécies temos genomas com milhares de sequências que não estão mapeadas em seus cromossomos, isso ocorre, pois muitos deles foram realizados a partir de tecnologia de segunda geração, mas com o advento da terceira geração esse cenário pode ficar para trás por permitir montagens mais contínuas, além disso o 10X Chromium Genome technology pode permitir o início de estudos populacionais usando montagens de genoma de toda uma população por um custo um pouco maior do que bibliotecas de representação reduzida (Jiao et al.,2017).

1.3 Estratégias para Montagem *De Novo* de Genomas

A montagem da sequência genômica é um grande quebra-cabeça, temos milhões de *reads*, curtas ou longas, dependendo da tecnologia de sequenciamento escolhida. As *reads* precisam ser organizadas, encaixadas umas sobre as outras para formar *contigs*, estes podem ser unidos formando *scaffolds* que contém dois ou mais *contigs*, formando sequências ainda maiores, mas que ainda não representam um cromossomo inteiro no caso de eucariotos ou um

genoma inteiro no caso de procariotos. Os *scaffolds* no caso de eucariotos podem ainda ser ancorados em cromossomos.

A montagem *De novo* de genomas é computacionalmente muito complexa, considerada um problema NP-Difícil, também conhecido como problema do ‘caixeiro viajante’, que precisa lidar com a presença de diferentes haplótipos, erros de sequenciamento e regiões repetitivas (Narzisi et al., 2011). Diversas abordagens são utilizadas pelos montadores, dentre as mais comuns e utilizadas temos os montadores baseados em Grafos de Bruijn, *Overlap Layout-Consensus* (OLC) e os algoritmos *Greedy*, sendo que os dois primeiros métodos são os mais utilizados por apresentarem melhores resultados (Kwon et al., 2019).

O método de montagem baseado em *Overlap Layout-Consensus*, consiste em sobrepor todas as leituras obtidas no sequenciamento, para isso pode ser usado a transformação Burrows–Wheeler, criando um *index* com todas as *reads*, permitindo o encontro de sobreposição das mesmas, após isso é construído um *layout* a partir das sobreposições, criando um gráfico, para então chegarmos a sequência consenso utilizando um caminho hamiltoniano, foi desenvolvido inicialmente por Staden em 1980 (Li et al., 2011). Essa abordagem foi muito utilizada na época do sequenciamento de Sanger, entretanto voltou a ser utilizada recentemente para a montagem de genomas utilizando as tecnologias de sequenciamento de terceira geração (*long reads*), pois vem apresentando melhores resultados quando comparado a outras abordagens (Lannoy et al., 2017).

Os algoritmos *Greedy* são de fácil implementação computacional, se baseiam na sobreposição de *reads*, assim como o *Overlap Layout-Consensus*, entretanto eles assumem diferentes pontuações para diferentes sobreposições, mesclando as sobreposições com altas pontuações, isso se repete até que nenhuma sequência possa ser mesclada, além disso montadores baseados em *Greedy* usam heurísticas para auxiliar a montagem, muito devido a dificuldades em montar regiões repetitivas, muitas vezes colapsando essas sequências. (Pop et al.,2002).

Os grafos de Bruijn se baseiam em dividir as *reads* em sequências menores (*reads*), de tamanho K, chamadas de *K-mers*, onde os *K-mers* serão sobrepostos por K-1 criando vértices, onde dois vértices X e Y podem se conectar caso se sobreponham por K-2 entre o sufixo de X e o prefixo de Y. As arestas representam a montagem das sequências, sendo o caminho euleriano aquele que percorre apenas uma vez cada *K-mer*, formando a sequência correta (Compeau et al., 2011).

Para que o grafo fique completo, necessita-se de todos os *K-mers* presentes no genoma (ou em uma região específica montada), estes vértices precisam estar conectados a um número par de arestas, sendo assim todos os vértices balanceados, com exceção de dois vértices, semi-balanceados, nas extremidades do grafo, permitindo que possamos encontrar um caminho euleriano, percorrendo todas as arestas uma única vez (o que representa passar por todos os *K-mers*), (Compeau et al., 2011).

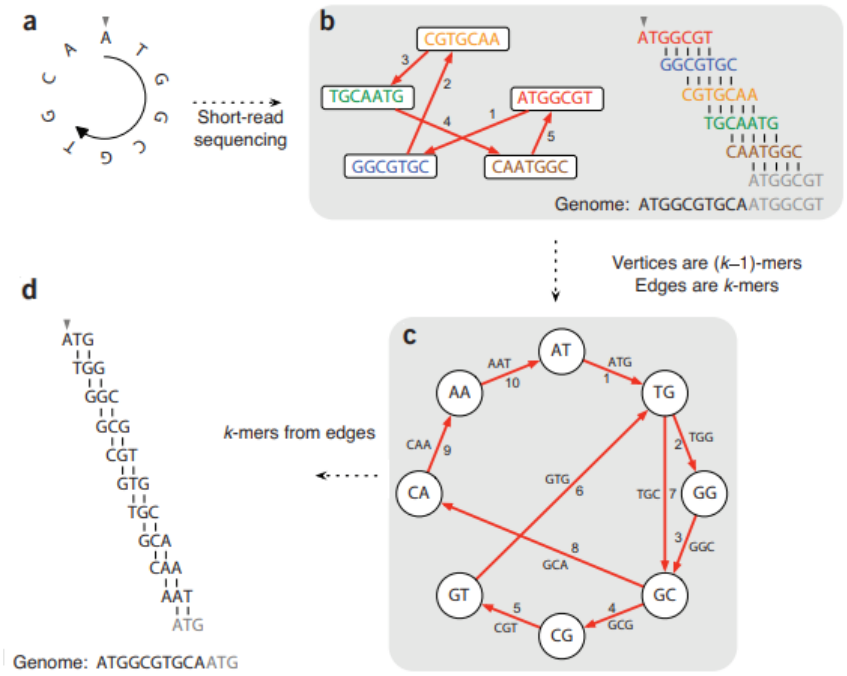


Figura 5: a: Sequência original a ser sequenciada; b *reads* e montagem por sobreposição, é possível notar que temos sempre duas bases livres, o que impede o uso dos grafos de Bruijn; c fragmentação das *reads* em K=3, gerando apenas uma base livre na sobreposição e permitindo a criação do grafo e a descoberta da sequência correta a partir de um caminho euleriano que passa pelas arestas; d *k-mers* e sequência genômica final da montagem. Fonte: adaptado de Compeau et al., 2011.

1.4 Ordenação e Anotação de Genomas

Mesmo que um genoma esteja fragmentado em centenas ou milhares de *contigs*, são muito úteis para diversos propósitos dentro da genética, permitindo inúmeros estudos antes impossíveis, entretanto, a organização desses *scaffolds* em escala cromossômica permite inúmeros outros estudos, principalmente referente a evolução do cariótipo da espécie e identificação de grandes variantes estruturais que podem ser importantes para determinadas características (Rice et al., 2019).

Dentre os métodos utilizados para ordenar e orientar os *contigs*, formando *scaffolds* que formarão os pseudo cromossomos, temos os mapas genéticos. Para construir um mapa, precisamos de uma população obtida através de recombinação entre dois parentais, a progênie é então genotipada e o mapa é obtido através das frequências de recombinação entre os marcadores, utilizando diversas ferramentas estatísticas para se determinar a distância e correta entre os marcadores e a ordem dos mesmos dentro de um grupo de ligação (Fierst et al., 2015).

Os marcadores usados na construção do mapa de genético podem então ser alinhados contra as sequências do genoma, assim os *contigs* podem ser unidos, fragmentados, orientados além de serem corretamente alocados em seus devidos grupos de ligação, dentre os *softwares* usados nessa prática temos o Chromonomer e o ALLMAPS, o segundo é baseado ainda no uso de múltiplos mapas genéticos (Tang et al., 2015).

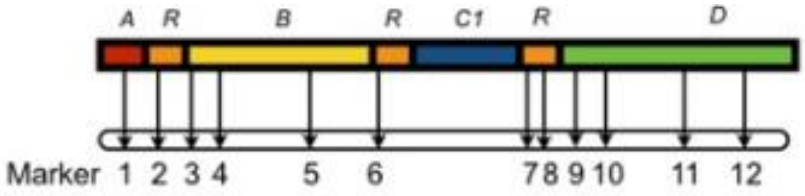


Figura 6: Ordenação e orientação de *contigs* a partir de marcadores de um mapa genético, formando um pseudo cromossomo , podemos observar a presença de *contigs* oriundos de duplicações (laranjas), que apresentam diferentes variantes, sendo corretamente ordenados dentro grupo de ligação. Fonte: Fierst et al., 2015.

Um método que vem sendo muito utilizado para conseguir *scaffolds* em escala cromossômica é o de ligação por proximidade chamado de HI-C, essa técnica foi desenvolvida para estudar a interação entre os cromossomos dentro do núcleo e a sua estrutura tridimensional, mas vem sendo utilizada na montagem de genomas, o método se baseia no sequenciamento *paired-end*, onde os pares de leitura representam regiões genômicas próximas, sendo que a probabilidade das regiões próximas interagirem é maior do que existir a interação entre regiões distantes, sendo a maior probabilidade a de interação intracromossômica (Ghurye et al., 2017). Desse modo, as frequências de contatos entre as sequências podem determinar as distâncias entre os loci, assim podemos ordenar e orientar as sequências para formar *scaffolds* em escala cromossômica (Ghurye et al., 2017); (Rice et al., 2019).

Pode ser utilizado também dados de outros genomas disponíveis da mesma espécie, assim os *contigs* gerados podem ser alinhados contra esse genoma a fim de se agrupar, quebrar *contigs* quiméricos e orientá-los, o *software* Ragoo, que implementa essa metodologia, foi até mesmo comparado com a metodologia de HI-C, apresentando melhores resultados quanto erros de variantes estruturais e continuidade dos *scaffolds*, mas por outro lado pode deixar de encontrar essas variantes devido ao viés do uso da referência (Alonge et al., 2019).

A partir do genoma completo, chegamos a etapa de anotação dos genes, algo crucial para conseguir retirar informações biológicas da sequência montada, para isso existem inúmeras ferramentas disponíveis, a primeira etapa de qualquer *pipeline* de anotação é a identificação dos elementos repetitivos e o seu mascaramento (Angel et al., 2018). Esses elementos são a maior parte dos genomas de plantas, podendo chegar a 90% de todo conteúdo genômico como no caso do trigo (Choulet et al., 2010).

Dentre os elementos repetitivos temos a regiões microssatélites ou repetições simples repetidas (SSR), que consistem em repetições uma certa sequência chamada de motivo, essas sequências são muito comuns em procariotos e eucariotos e apresentam alto grau de variante alélico, sendo que em plantas o motivo A/T é o mais comum (Bhargava et al., 2010).

Temos também os chamados elementos de transposição, a capacidade de movimentação desses elementos dentro do genoma os tornaram um elemento muito importante na evolução das espécies, pois acabam gerando variabilidade, podendo afetar a organização dos genomas, plasticidade e a expressão e função de genes (Angel et al., 2018).

Os elementos de transposição foram descritos por Barbara McClintock, no milho, em meados dos anos 40, podendo ser divididos em duas classes, a primeira (tipo I) é a dos retrotransposons, que usam um RNA intermediário para gerar uma nova fita de DNA por uma transcriptase reversa, esse DNA é integrado ao genoma por enzimas, desse modo, cada vez que um desses elementos fica ativo no genoma, temos pelo menos uma nova cópia do mesmo; a segunda classe (tipo II) é conhecida por apresentar o mecanismo de ‘recorta e cola’, onde uma transposase corta o elemento, inserindo-o em outro local do genoma (Bennetzen et al., 2014). Existem cinco ordens de elementos transponíveis LTR, LINEs, DIRs-like, PLEs e SINEs, os LTR (*Long Terminal Repeat*) são os mais comuns em plantas (Wicker et al., 2007).

A identificação dessas repetições é complicada pois são pouco conservadas, geralmente se faz o uso de bibliotecas específicas para cada espécie que podem ser construídas a partir de alguns *softwares*, entretanto para muitas espécies já existem bibliotecas prontas para o uso, e a falha no mascaramento dessas regiões interfere nas próximas etapas da

anotação do genoma, podendo causar a identificação errada de muitos genes (Yandell et al., 2012).

Após o mascaramento, um dos métodos usados para a anotação de genes é o alinhamento de sequências EST, proteínas e dados de RNA-seq contra o genoma montado e mascarado, desse modo é possível identificar os *exons* e *introns*, determinando a posição e estrutura dos genes, geralmente são usados dados do mesmo organismo, mas dados de organismos próximos também podem ser usados, principalmente dados de proteínas, já que estas sequências são mais conservadas entre as espécies, entretanto os dados de RNA-seq são os mais informativos para determinar a estrutura de um gene, fornecendo mais evidências para delimitações de *exons*, sites de *splicing* e até dados de *splicings* alternativos caso a profundidade do sequenciamento seja alta (Angel et al., 2018).

Outro método muito utilizado, são os preditores *Ab initio*, eles utilizam *machine learning* para predizer a localização e estrutura dos genes, uma das vantagens é que não são necessários outros dados para realizar a anotação , onde os parâmetros dos modelos de muitos organismos já estão determinados, por outro lado o melhor método para usar esses preditores é treiná-los com um conjunto de genes do próprio organismo, pois mesmo organismos próximos apresentam diferenças quanto a estrutura gênica do seu genoma, com bons dados de treinamento podemos chegar até em 100% dos genes previstos, com até 70% das estruturas gênicas corretas (Yandell et al., 2012).

Alguns *softwares* como o MAKER acabam por unir diversos outros *softwares* num *pipeline* para anotação, onde temos a anotação dos genes a partir de evidências de RNAs e proteínas que são usados no treinamento de preditores *Ab initio*, aumentando assim a eficácia do processo de anotação de um genoma. Os preditores *Ab initio*, podem ser treinados usando até mesmo as evidências das anotações já existentes de genomas da mesma espécie (Holt et al., 2011).

1.5 Genomas de plantas e sua evolução

Existem aproximadamente 399 mil espécies de plantas e algas verdes no mundo, com uma enorme variabilidade no tamanho dos seus genomas, que pode ser de 63 MB na *Genlisea margaretae carnívora* (Greilhuber et al., 2006) até aproximadamente 150 Gb como a *Paris japonica* (Pellicer et al., 2010).

Geralmente o tamanho dos genomas estimados via citometria não abrange o tamanho das montagens dos genomas publicados, ficando estes em torno de 85% do tamanho estimado,

isso ocorre devido a fração do genoma altamente repetitiva como repetições ribossômicas, regiões teloméricas e centroméricas além dos elementos transponíveis (Michael et al., 2013).

Quanto ao conteúdo gênico, as plantas geralmente apresentam entre 20 e 124 mil genes (Wendel et al., 2016). Esses números acontecem devido às diferenças biológicas e também devido às ferramentas usadas para a anotação, onde até mesmo o tamanho dos *scaffolds* tem influência no número de genes anotados, além disso os elementos transponíveis podem ser anotados como genes, o genoma do arroz já foi anotado tendo aproximadamente 55 mil genes, entretanto entre 10 e 15 mil foram identificados como na verdade sendo elementos transponíveis (Michael et al., 2013).

Desde a primeira publicação do genoma da *Arabidopsis thaliana*, centenas de outros genomas de plantas foram publicados e a tendência é de que esse número aumente em número de espécies e também em indivíduos de uma mesma espécie, hoje o genoma da *Arabidopsis thaliana*, que apresenta um tamanho de 140 MB e 5 cromossomos, está sequenciado quase que continuamente, apresentando poucos gaps, devido aos avanços nas tecnologias de sequenciamento de longas sequências e nos algoritmos de montagem de genomas (Kersey et al., 2019).

A união das tecnologias de sequenciamento, principalmente as de segunda geração e terceira geração, junto com métodos de ordenação baseados principalmente em HI-C, vem obtendo ótimos resultados, gerando genomas grandes e complexos com alta qualidade, pois facilitam a montagem de regiões repetitivas e a sua continuidade facilita a correta anotação dos genes e suas estruturas, entretanto falando de genomas poliplóides, ainda faltam ferramentas de bioinformática que lidem com esse tipo de dado, sendo que hoje genomas poliplóides geralmente são sequenciados e disponibilizados como genomas haplóides (Jung et al., 2019).

Diversos estudos apontam que as angiospermas passaram por diversos eventos de poliploidização, a *Arabidopsis thaliana* passou por 3 eventos de poliploidização durante sua história evolutiva, mas mesmo assim apresenta um genoma compacto, isso ocorre devido a perda diferencial de genes, onde as plantas perdem a maioria do material genético duplicado, apenas nos últimos 10 milhões de anos a *Arabidopsis thaliana* perdeu quase metade do seu genoma devido a pequenas deleções (Proost et al., 2011).

Essa perda de genes também ocorreu com o milho, que perdeu cerca de metade dos genes duplicados no último evento de ploidização que gerou o progenitor do milho, apesar da perda, alguns pares de genes se mantém, entretanto esse não é um processo aleatório, sendo preferencial a pares de genes onde ao menos um adquire nova função ou que a maior

expressão devido a duplicação seja benéfica e a mutação leve a perda de função (Adans et al., 2005).

O genoma ancestral das angiospermas provavelmente tinha um número de genes entre 12 e 14 mil , resultado obtido a partir de modelos que simulam o nascimento e morte de genes em pequena e larga escala, baseado em genomas de diversas espécies, esse genoma passou por um primeiro evento de duplicação gerando um tetraploide, que hibridou com um diplóide gerando um triploide, este triploide passou por mais um evento de duplicação do genoma inteiro, gerando um hexaplóide com 21 cromossomos, que seria o ancestral das dicotiledôneas, essa teoria é suportada a partir de dados dos genomas da videira, mamão, café, soja e outros (Proost et al., 2011).

Diversos estudos demonstram a importância da poliploidização na história evolutiva das plantas, mas a proliferação de elementos transponíveis e também a sua remoção estão muito presentes na história evolutiva das plantas, o *Oryza australiensis* possui um genoma com mais do que o dobro do tamanho do *Oryza sativa*, sendo predominantemente devido ao aumento de aproximadamente 400 MB de três famílias de elementos retrotransponíveis, no algodão também podemos ver a influência desses elementos na evolução das plantas, em um clado australiano as espécies têm um genoma até três vezes maior do que o clado americano devido ao aumento e da deleção de diferentes famílias de elementos transponíveis (Ammiraju et al., 2007); (Hawkins et al., 2006).

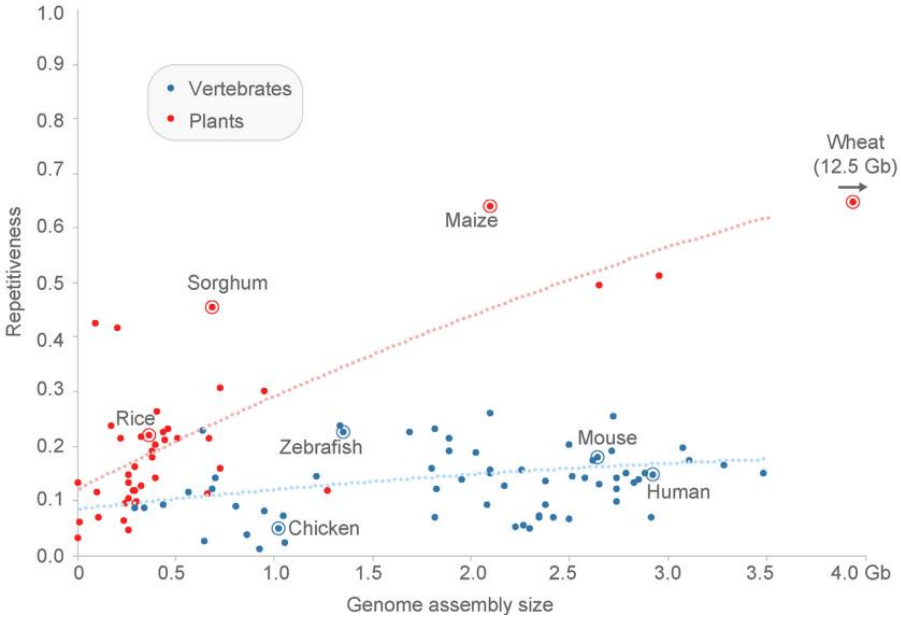


Figura 7: Comparação entre o tamanho do genoma de plantas e vertebrados e sua relação com regiões repetitivas, pode-se notar que geralmente em plantas tem-se genomas com maior quantidade de elementos repetitivos e estes apresentam correlação positiva com o tamanho do genoma. Fonte: Jiao et al. (2017).

O genoma da soja, além das duplicações comuns a angiospermas e dicotiledôneas passou por um evento de triplicação do genoma, comum a muitas dicotiledôneas, esse evento ocorreu a aproximadamente 117 milhões de anos atrás (Bowers et al., 2003). Após esse evento a soja ainda passou por mais duas rodadas de duplicações de genoma, uma há 55 milhões de anos atrás e outra entre 13 e 5 milhões de anos atrás (Gill et al., 2009). O genoma da soja apresenta 57% dos seus genes em regiões de heterocromatina, ricas em repetição e distantes dos centrômeros, que são ricos em elementos repetitivos, na região dos centrômeros o número de recombinações é menor quando comparado ao restante do cromossomo. Regiões de eucromatina tem uma relação de distância genética com distância física de bases de aproximadamente 1 cM por 197 kb, enquanto regiões de heterocromatina 1 cM por 3,5 MB (Schmutz et al.,2010). A estrutura dos genes, *exons* e *introns* na soja, se mostra bastante conservada, onde 99,45% dos *introns* compartilhados entre o álamo e a videira, estão presentes na soja, essas duas espécies apresentam uma taxa lenta de evolução, indicando que essa estrutura é conservada na soja (Schmutz et al., 2010).

Hoje temos 4 genomas de soja disponíveis no NCBI, com um tamanho variando de 927 MB até 1.017 MB, conteúdo GC entre 34.6% e 35.34%, com montagens apresentando entre 108601 *scaffolds* e 475 *scaffolds*, o genoma de referência mais utilizado é o da cultivar Williams 82 que está na versão 2.1, com um *scaffold* N50 de 48.57MB. O genoma de maior *scaffold* N50 disponível no NCBI é o da cultivar Zhonghuang 13, com N50 de 51 MB e 1310 sequências, montado a partir de leituras originadas de Illumina e PacBio.

O genoma da cultivar Zhonghuang 13 foi recentemente atualizada para sua segunda versão, chamada de referência de ouro, onde novas sequências foram usadas para melhorar a montagem do genoma, assim houve uma melhora dos parâmetros de qualidade, com o *contig* N50 indo de aproximadamente 3 MB para 22.6 MB, e um *scaffold* N50 de 51.98 MB, a anotação do genoma indicou 55,443 genes codificadores de proteínas e um total de 53.06% de sequências repetitivas (Shen et al., 2019).

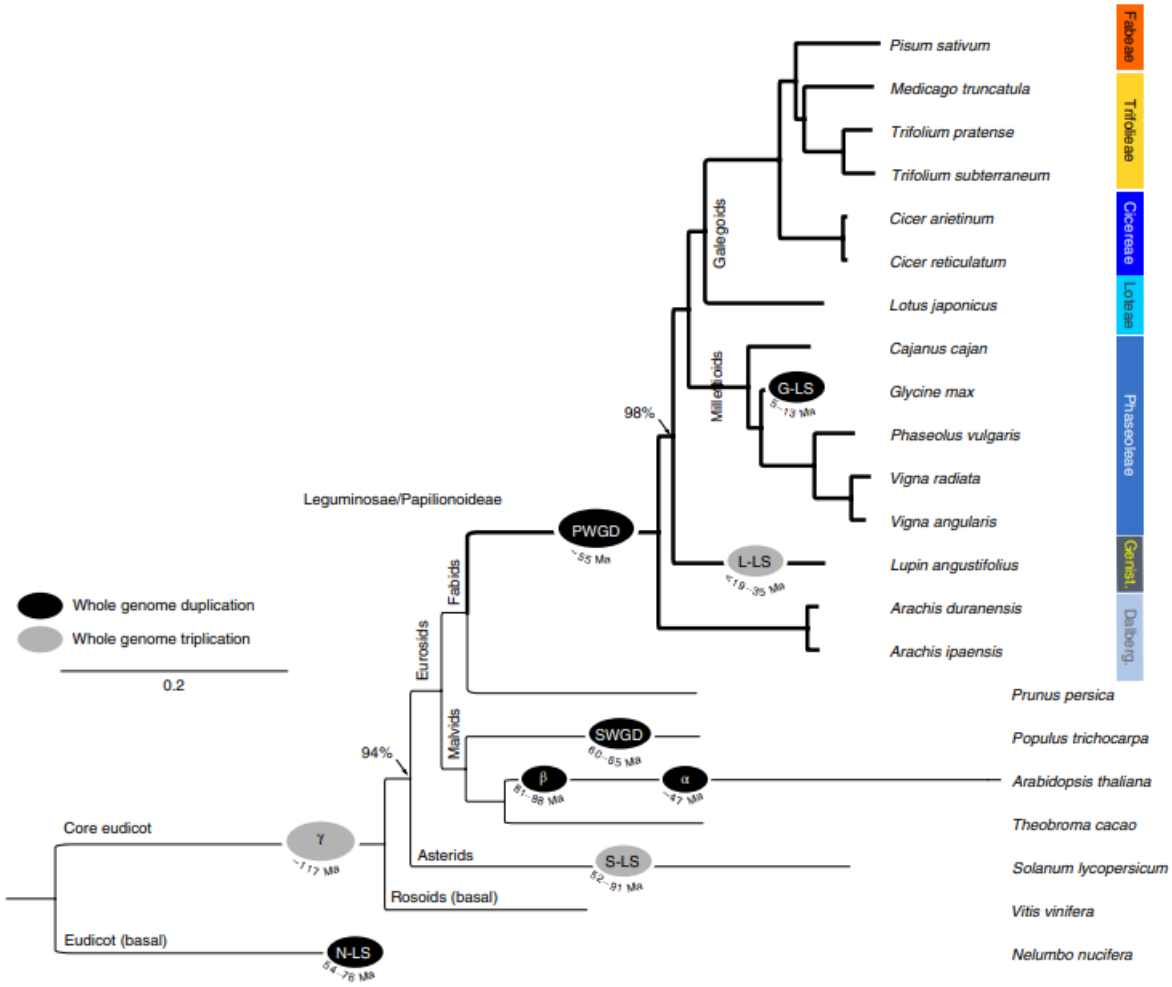


Figura 8: Eventos de triplicação e duplicação de eudicotiledoneas, é possível observar que a soja passou por um evento de triplicação e dois eventos de duplicação do genoma. Fonte: Kreplak et al. (2019).

Também temos 26 montagens de genomas de soja representativos da variabilidade de mais de 2 mil variedades e acesos, principalmente dos Estados Unidos e China. Essas 26 montagens variam entre 992 MB até 1039 MB, esses 26 genomas, juntos com outros três genomas disponíveis, foram usados para um pan genoma baseado em grafos, permitindo o acesso a 124,222 variantes estruturais (Liu et al., 2020). Essas variantes estruturais são muito importantes e podem ser usadas no melhoramento genético de plantas, pois podem estar associadas a características agronômicas de interesse (Lye; Purugganan, 2019); (Patil et al., 2019); (Liu et al., 2020).

1.6 Descoberta de variantes

Os avanços nas tecnologias de sequenciamento, que a cada dia se tornam mais eficientes em gerar dados a um menor custo, facilitaram o sequenciamento *De novo* de

genomas de muitas culturas, abrindo as portas para diversos outros estudos importantes para o melhoramento genético (Xu; Bai, 2015).

O resequenciamento de genoma total (WGR) de plantas permite a descoberta de diversos tipos de variantes de DNA, como SNPs, Indels, variações no número de cópias (do inglês, CNV) e variações presença/ausência (do inglês, PAV) como discutido por Xu & Bai (2015), além da descoberta de genes importantes para certas características presentes no indivíduo ou população estudada.

Esses estudos expandiram o conhecimento das variações genéticas nas culturas e forneceram uma gigantesca quantidade de dados genômicos para estudos genéticos (Xu; Bai, 2015).

O alinhamento completo entre genomas é muito usado para identificação de variantes estruturais e estudos evolutivos, se tornando um problema a partir do início da era genômica, pois os alinhadores disponíveis não eram apropriados para tal trabalho, já que a principal inferência desse tipo de alinhamento é a história evolutiva, precisando ser levado em consideração a ortologia dos genes, desse modo o alinhamento entre genomas, na maioria das vezes tem segmentos correspondentes, alinhados apenas no caso serem ortólogos, tendo como dificuldades extras grandes variantes como inversões, translocações e também duplicações, esse último, principalmente se tivermos eventos de duplicações de todo o genoma (Dewey, 2019).

Os alinhamentos completos de genomas são uma poderosa ferramenta na identificação de variantes estruturais, geralmente eles trabalham gerando um grande número de alinhamentos locais com uma posterior filtragem de qualidade para remover falsos positivos, posteriormente temos um refinamento, onde sequências homólogas passam por outro alinhamento a partir de um alinhador colinear (Armstrong et al., 2018).

A chamada de SNPs e indels a partir de dados de resequenciamento de genoma total é muito comum na genética humana, sendo responsável pela identificação de inúmeras mutações com importância médica, sendo o *software* GATK muito utilizado nessas análises (Kishikawa et al., 2019). Na agricultura são menos comuns esses tipos de trabalhos, entretanto diversos trabalhos foram recentemente publicados, com resultados promissores para a aplicação no melhoramento genético de plantas (Jiang et al., 2017); (Ahn et al., 2018); (Chen et al., 2017); (Yamamoto et al., 2018); (Qi et al., 2014).

Recentemente a cultivar arroz japonica Longdao 24 e seus pais (Longdao 5 e Jigeng 83) foram resequenciados, sendo encontradas variantes de nucleotídeo único (SNPs) e indels em Longdao24 e em seus pais. Ao analisar genes envolvidos com a produtividade e qualidade

de amido foram encontradas variantes que geram alterações de códons em 10 genes, dando vantagens Longdao24 (Jiang et al., 2017).

Visando a identificação de SNPs, foram resequenciados os genomas de duas cultivares de pimenta das espécies *Capsicum baccatum* e *Capsicum annuum*, a primeira resistente a Oídio e a segunda suscetível ao fungo. Foram encontrados 6281 SNPs associados a 46 genes candidatos de resistência, a partir disso foram desenvolvidos 36 primers HRM, utilizados para validar os SNPs em uma progênie do cruzamento das duas cultivares, ao todo 19 primers foram capazes de distinguir a população resistente e suscetível a partir da correlação entre os mesmos com os escores de avaliação da doença fenotípica para cada indivíduo, podendo ser utilizados como marcadores de seleção assistida (MAS) (Ahn et al., 2018).

Uma variedade de arroz adaptada à água do mar, Sea Rice 86 (SR86), que apresenta características únicas, foi sequenciada a fim de se identificar variantes envolvidas com a característica, após a comparação com 3 mil acessos de arroz foram identificados 47 indels e 7223 SNPs com impacto funcional em proteínas. Foram então desenvolvidos vinte e quatro marcadores a partir de indels maiores que 28 bp, selecionados a partir do meio de cada braço de cada cromossomo, podendo serem utilizados em futuros estudos em progênies visando a resistência a salinidade (Chen et al., 2017).

Usando o genoma de referência do milho (B73) e o resequenciamento de genoma total de 15 linhagens de milho tolerantes à seca, foram buscados SNPs não-sinônimos (nsSNPs) para a tolerância a seca através da Ferramenta ANNOVAR e pelo método de análise de clusters baseada em SNP. Um total de 524 SNPs foram encontrados pelos métodos, gerando 261 genes candidatos. A partir de dados de QTLs para a característica e a anotação GO dos genes, foi mostrado que os genes candidatos estavam envolvidos na tolerância à seca, além disso 70% deles se mostraram diferencialmente expressos sob estresse (Xu et al., 2014).

Na África existem as cultivares de arroz ‘New Rice for África’ (NERICA), híbridos interespecíficos entre variedades de arroz asiáticas (*Oryza sativa*) e de arroz africano (*Oryza glaberrima Steud*), adaptadas ao clima e modo de produção na África ocidental. Foi realizado então o resequenciamento de quatro linhagens NU e os parentais *Oryza sativa* (WAB56-104) e *Oryza glaberrima* (CG14), possibilitando a identificação de possíveis genes importantes para características agronômicas e genes únicos candidatos a estarem envolvidos na resistência a estresses bióticos e abióticos (Yamamoto et al., 2018).

Um acesso silvestre de soja (W05), com alta tolerância a salinidade e distante geneticamente das cultivares modernas foi resequenciado, também foi obtida uma população de mapeamento a partir do cruzamento do acesso com uma cultivar com baixa tolerância à

salinidade, que também foi sequenciada, entretanto com uma profundidade de apenas 1X. Foi encontrado um QTL para a resistência a salinidade, a partir dele, uma análise de ponto de ruptura recombinante de linhagens resistentes e suscetíveis da população levou a uma região de 388-Kb no Chr03, com 43 genes candidatos, um dos genes (GmCHX1) , apresenta uma inserção em um dos seus *exons* no genoma de referência, essa inserção não está presente no acesso W05. Uma comparação entre as linhagens tolerantes e suscetíveis demonstrou que esse gene é conservado entre as tolerantes, além disso, análises via PCR em tempo real mostraram que o gene é mais expresso nas linhagens resistentes (Qi et al., 2014).

2MATERIAL E MÉTODOS

2.1 Extração de DNA e Sequenciamento

O DNA das cultivares IAC-100 e CD-215 foi extraído a partir do protocolo Doyle e Doyle (1987). As amostras foram sequenciadas no Roy J. Carver Biotechnology Center da Univeristy of Illinois Urbana-Champaign pela plataforma NovaSeq S4 da Illumina. As bibliotecas foram construídas utilizando o equipamento 10X Chromium Controller com o kit Genome & Gel Bead, que utiliza a tecnologia 10X Chromium Genome, permitindo a geração de *linked reads*. O sequenciamento das bibliotecas foi paired-end 2 x 150nt em um sequenciador Illumina NovaSeq 6000, utilizando uma lane da *flow cell* NovaSeq S4

2.2 Montagem dos genomas

As leituras fastq obtidas a partir do sequenciamento foram verificadas visualmente quanto à qualidade média das *reads* com o programa FASTQC. A montagem foi realizada através do *software* Supernova, proposto por Weisenfeld et al. (2017), ao qual se baseia no uso do algoritmo gráficos de Bruijn (de Bruijin., 1946), considerando também as linked-reads para melhorar a montagem, criando um gráfico composto por diferentes ramificações, onde cada caminho (*path*) corresponde a variações do haplótipo em decorrência da presença de SNPs ou indels, mas com a possibilidade de terem sido geradas por erros no sequenciamento ou alinhamento.

A partir do gráfico, um arquivo FASTA foi gerado, utilizando-se o comando supernova *mkoutput* e que, através da opção *pseudohap*, seguiu os *paths* de apenas um dos haplótipos a fim de gerar sequências que representam fragmentos do genoma montado. Em

decorrência dos altos níveis de homozigosidade em *G. max*, o uso de apenas um dos haplótipos não interfere significativamente nas análises posteriores.

A fim de se conseguir alocar os *contigs* em *scaffolds*, quebrar *contigs* quiméricos e ordenar os *contigs* criando pseudocromossomos, foram testadas duas abordagens, a primeira baseada em mapas genéticos usando o *software* Chromonomer, para isso foi utilizado o mapa genético consenso da soja baseado em SNPs disponível no Soybase, o mapa continha um total de 3626 SNPs, ainda usando o Soybase, a partir do nome de cada SNP suas posições no genoma foram identificadas, e usando o *software* bedtools V2.2.29.0 uma sequência com 150 pb que flanqueiam as posições foram salvas num arquivo .fasta a partir do genoma da Williams82 v2.1.

As sequências que flanqueiam os SNPs foram então alinhadas contra os genomas da IAC-100 e CD-215 usando o *software* BWA com configurações padrão gerando um arquivo .SAM que foi convertido em .BAM, esses alinhamentos serviram de entrada para o *software* Chromonomer que foi usado para realizar a ordenação.

A outra abordagem para a ordenação foi utilizando o *software* Ragoo, que usa o minimap2 para alinhar as sequências do draft contra uma referência, para isso usamos de referência do genoma da Williams82 v2.1. com um tamanho máximo da gap de 200 pb, com a opção de correção de *contigs* quiméricos.

Após o ordenamento foi realizada a correção de erros pelo *software* REAPR, usando como *input* um alinhamento entre os genomas montados e o conjunto total de *reads* filtradas pelo Trimmomatic, para o alinhamento foi usado o *software* BWA com configurações padrão. Após a correção foi realizado outro ordenamento pelo *software* Ragoo com tamanho máximo de gap 200.

2.3 Anotação dos genomas

A anotação das sequências repetitivas foi realizada a partir do *software* RepeatMasker, para isso foi usada a biblioteca de elementos repetitivos disponível no soybase, a SoyTEdb (Du et al., 2011) e a biblioteca Dfam V.3.1

A partir dos genomas mascarados obtidos com o repeatmasker foi usado o preditor *Ab initio* AUGUSTUS para realizar a anotação dos genes, como não existem dados treinados para *Glycine max*, foi utilizado o conjunto treinado para *Arabidopsis thaliana*

2.4 Avaliação da qualidade

O *software* BUSCO (Simão et al., 2015) forneceu um *score* para a montagem, utilizando-se de um *dataset* composto de genes anotados para o grupo Embryophyta (v10.0). *Scripts* em Perl também foram usados para gerar parâmetros quantitativos de qualidade.

Para se verificar a sintênia entre os cromossomos da Williams82 e o das nossas montagens, nossos genomas foram filtrados a partir de um *Script* em perl, selecionando os 20 maiores *scaffolds*, desse modo, foi realizado um alinhamento contra os 20 cromossomos da Williams82, usando o *software* NUCcmer com a opção -mum, usando o delta filter com 99% de identidade entre os alinhamentos gerados.

2.5 Busca por variantes

Para a chamada de SNPs e Indels, os dados pré processados com Trimmomatic para retirada dos *barcodes*, primers e seleção de qualidade das cultivares IAC100 e CD-215, foram alinhados ao genoma da Williams82 v2.1, que também é suscetível ao complexo de percevejos. Para o alinhamento foi utilizado o *software* Bowtie2 usando a opção --very-sensitive-local, gerando um arquivo .SAM, que foi transformado pelo Samtools para .BAM.

Para a chamada dos SNPs e indels foi usado o *software* Genome Analysis Toolkit 4.0.11.0 (GATK), foram marcadas as *reads* duplicadas com Picard, seguido da chamada das variantes com a opção HaplotipeCaller e filtragem usando o hard filter. Foram selecionados os SNPs e indels únicos na cultivar IAC-100 através do VCFtools usando a opção isec.

Foram usados dados de GBS disponíveis internamente no Laboratório de Diversidade e Melhoramento ESALQ-USP para realizar a chamada de variantes de uma população de 236 RILs, obtida do cruzamento entre IAC-100 e CD-215. O *pipeline* do GATK usado para as variantes da IAC-100 e CD-215 foram utilizados novamente para essa chamada, com a inclusão da etapa de recalibração de bases usando o conjunto total de SNPs da IAC e a etapa GenotypeGVCFs, entretanto sem a etapa de marcação de duplicatas.

O VCFtools isec com a opção -n 25 foi usado para selecionar variantes únicas da IAC-100 que estavam presentes nas 25 linhagens mais resistentes da população, os dados de resistência foram obtidos de dados históricos do programa de melhoramento do Laboratório de diversidade e melhoramento ESALQ-USP, sendo a seleção das 25 linhagens realizada pela maior porcentagem de sementes boas após experimentos em diversos anos e locais sem controle de percevejos.

O *software* IGV (Robinson et al.,2011) foi utilizado para visualizar os arquivos .bam gerados pelo bowtie2 na chamada de variantes da IAC-100 e CD-215, permitindo a visualização de grandes eventos deleções nas regiões que apresentam QTLs já mapeados para a resistência a percevejos, foram escolhidos os três QTLs mais significativos para essa etapa. Para a exploração de genes próximas a variantes foram usados os bancos de dados NCBI e uniprot.

3 RESULTADOS E DISCUSSÃO

3.1 Sequenciamento e montagem

O sequenciamento dos genomas utilizando a metologia 10 X da Chromium, gerou um total de aproximadamente 915 milhões de reads para a cultivar IAC-100, enquanto para a cultivar CD-215 foram geradas aproximadamente 801 milhões de reads. Os resultados obtidos a partir da função pseudohap do software Supernova, para os dois genomas podem ser visualizados na tabela 1, abaixo.

Tabela 1: Parâmetros de qualidade dos genomas após montagem com supernova

Parâmetros	IAC-100	CD-215
Número de *reads* usadas	460.02 M	460.02 M
Tamanho médio de *reads* após filtragem	138.50 pb	138.50 pb
Cobertura total	58.91 X	58.65 X
Cobertura efetiva	44.25 X	43.54 X
Tamanho estimado do genoma	1.17 Gb	1.18 Gb
Tamanho dos fragmentos de DNA	65.74 Kb	74.82 Kb
Número de *scaffolds*	31518	31455
Número de *scaffolds* > 10.000 pb	5822	5730
Scaffold N50	3.68 MB	3.94 MB
Scaffold N50 > 10.000 pb	4.17 MB	4.58 MB
Conteúdo GC	34.56 %	34.52 %
Tamanho da montagem	1.051 Gb	1.055 Gb
Tamanho da montagem > 10.000 pb	958 MB	962 MB
Número de *Scaffolds* com 75% da montagem >10.000 pb	159	144

Os tamanhos dos nossos genomas e o conteúdo de GC estão próximos dos quatro conjuntos de genoma de soja disponíveis no NCBI até o momento, com conteúdo de GC igual a 35,12%, 35,34%, 35,27% e 34,60% e tamanhos dos *assemblies* variando entre 927 MB a 1.017MB. Além disso, os *scaffolds* N50 dos nossos *assemblies*, considerando sequências maiores que 10.000 pb foram de 4.17 MB para a IAC-100 e 4.68 MB para a CD-215, a diferença deste parâmetro de qualidade entre as duas cultivares, provavelmente se deu pela maior qualidade de extração de DNA obtido com a cultivar CD-215, que teve um tamanho de molécula estimado em 74.82 Kb, enquanto a cultivar IAC-100 teve um tamanho estimado de 65.74 Kb, essa discussão sobre o tamanho da molécula de DNA também é discutida por Armstrong. (2019), que também observou essa correlação com o tamanho do *scaffold* N50.

Esses resultados para *scaffold* N50 são melhores ou semelhantes a vários genomas de referência disponíveis no NCBI, como *Citrus sinensis*, *Hevea brasiliensis*, *Lactuca sativa* e *Solanum tuberosum* com *scaffolds* N50 entre 1,28 MB e 1,77 MB.

Os resultados das nossas montagens são também parecidos com os dados obtidos na montagem de referência de *Capsicum annuum* utilizando apenas a tecnologia 10x da Chromium, que conseguiu obter um *scaffold* N50 de 3.96 MB em um genoma estimado de 3.21 Gb, entretanto os autores discutem que apesar da qualidade da montagem, regiões repetitivas ainda são difíceis de serem montadas devido ao sequenciamento de leituras curtas (Hulse-Kemp et al., 2018).

As regiões repetitivas que são muito presentes em todos os genomas de plantas provavelmente dificultam a montagem, resultando em um *scaffold* N50 reduzido, isso não ocorre em outros organismos , a montagem do genoma do peixe *Perca fluviatilis* , gerou uma montagem de 1 Gb com um *scaffold* N50 de 6.3 MB, enquanto a montagem de três indivíduos de cachorro selvagem africano *Lycaon pictus*, com um genoma de aproximadamente 2.3 Gb, apresentaram um *scaffold* N50 de 7MB, 15MB e 21 MB (Ozerov et al., 2018); (Armstrong et al., 2019).

Analisando apenas as sequências maiores que 10.000 pares de base, temos dois conjuntos altamente representativos dos genomas das duas cultivares, com mais de 950 MB de tamanho, maiores do que o menor genoma de soja disponível no NCBI. Além disso os dois genomas apresentam uma alta continuidade, onde 75% da montagem do genoma da cultivar IAC-100 está em apenas 159 *scaffolds*, enquanto a cultivar CD-215 apresenta 75% da montagem em 144 *scaffolds*. Considerando a montagem toda, podemos perceber que o número de *scaffolds* passa dos 30 mil, com pouco mais de 5 mil *scaffolds* maiores que 10.000

pares de base, demonstrando que apesar da maioria dos dados estarem alocados em grandes *scaffolds*, o *software* supernova gera uma grande quantidade de pequenos *scaffolds*.

A análise usando o *software* BUSCO, com o banco de dados de genes ortólogos de cópia única do grupo das embriófitas, ao qual a soja pertence, revelou um total 97.1% dos genes do grupo presentes no genoma da CD-215, enquanto o genoma da IAC-100 apresentou 97.6% dos genes presentes, a maioria dos genes estão presentes em cópias duplas e não simples, aproximadamente 53% deles, como pode ser observado no gráfico abaixo, isso ocorre devido às diversas duplicações genômicas ocorridas durante a história evolutiva das plantas. Pode ser observado que a montagem do genoma da cultivar Williams 82, apresenta o mesmo padrão de cópias duplas encontrado em nossos genomas.

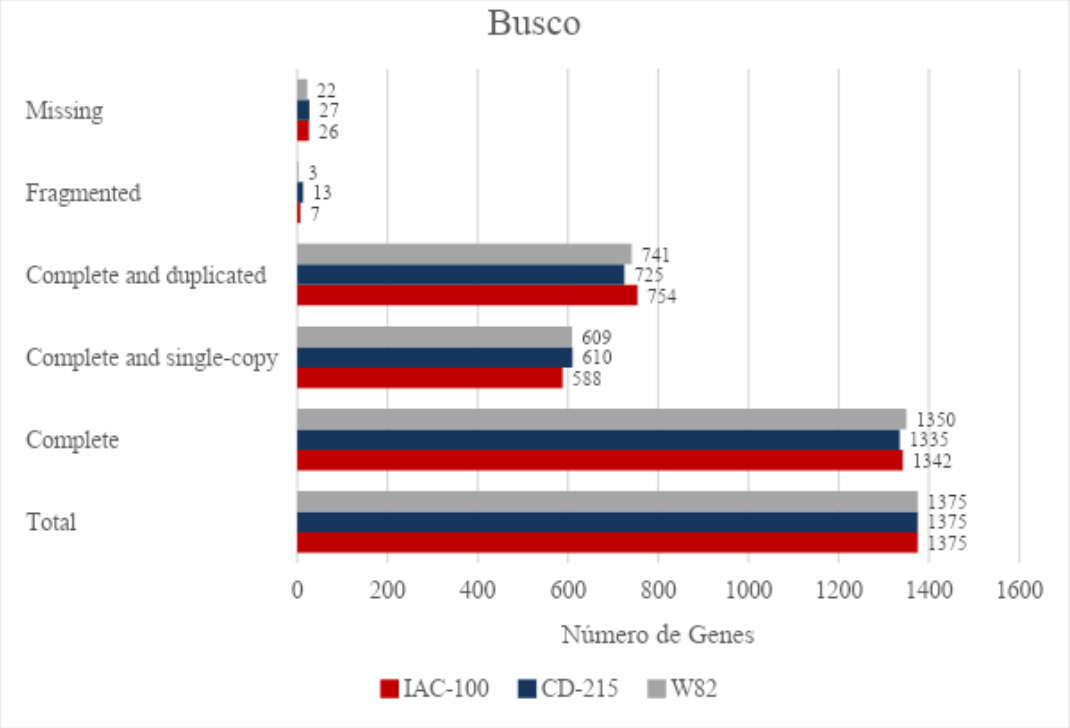


Figura 9: Avaliação de genes ortólogos de cópia única do grupo das embriófitas nos *assemblies* contrastantes para resistência ao complexo de percevejos, IAC-100 e CD-215 gerados com o supernova; pode ser observado uma alta completude dos genes avaliados em nossas montagens (~ 97%), assim como no genoma de referência da Williams82.

3.2 Ancoragem e ordenamento

Ancoragem e ordenamento de *contigs* é uma etapa fundamental na montagem de um bom genoma, sendo ainda a etapa onde conseguimos identificar e corrigir erros do montador,

como *contigs* quiméricos, que apresentam sequências erroneamente unidas, que pertencem a loci distantes dentro de um cromossomo ou até mesmo de diferentes cromossomos.

A abordagem utilizando o mapa genético consensus da soja identificou um total de 2763 SNPs na montagem da IAC-100 e um total de 2782 SNPs na montagem da CD-215, de um total de 3627 SNPs do mapa utilizado, isso provavelmente ocorreu devido a variações entre os genomas, impedindo o alinhamento dessas regiões, além disso essas regiões podem estar ausentes nas montagens ou até mesmo com problemas na montagem.

As ordenações baseadas no mapa genético, resultaram em genomas ordenados com tamanhos aproximados de 750 MB e uma média de 17,5 *Scaffolds* quiméricos identificados no processo e corrigidos, além disso tivemos um aumento do *Scaffold* N50 em aproximadamente 10x. Resultados de outras montagens que utilizaram o chromonomer atingiram 87% e 86% do tamanho da montagem inicial, com um aumento de até 11x do *scaffold* N50, enquanto conseguimos apenas 75% de ancoragem (Cm et al., 2016); (Conte et al., 2017).

Tabela 2: Parâmetros de qualidade dos genomas após a ordenação baseada no mapa consensus da soja

Genoma / **Resultados**	**IAC-100**	**CD-215**
Scaffold N50	35.7 MB	37.5 MB
Número de *Scaffolds* com 75% da montagem	44	34
SNPs usados	2763	2782
Scaffolds quiméricos quebrados	16	19
Scaffolds ancorados e ordenados	342	331
Tamanho da montagem ancorada e ordenada	746 MB	758 MB

Esses resultados poderiam ser melhorados de duas formas, primeiro com o uso de um número maior de marcadores, que provavelmente aumentaria o tamanho da montagem ordenada, outros trabalhos utilizaram mais de 4 mil marcadores para realizar o ordenamento, apresentando melhores resultados (Cm et al., 2016); (Conte et al., 2017). Um segundo modo de melhorar esses resultados, seria a partir do uso de um mapa genético de alta densidade

obtido a partir de uma população oriunda do cruzamento entre as duas cultivares, gerando linhagens endogâmicas recombinantes (RILs), permitindo uma melhora quantitativa e qualitativa do ordenamento, pois desse modo teríamos um mapa genético mais próximo da realidade dos nossos genomas, que podem conter variantes em relação ao mapa consenso, como inversões e translocações, por exemplo.

Outro método de ancoragem e ordenamento é através do uso de outro genoma da espécie como referência, para isso foi utilizado o *software* Ragoo, que usa o minimap2 para obter montagens em escala cromossômica em poucos minutos (Alonge et al., 2019).

O uso do Ragoo foi mais eficiente do que o chromonomer em todos os parâmetros avaliados, aumentando o tamanho do genoma ordenado para aproximadamente 1GB, que é o tamanho esperado para um genoma de soja, o *Scaffold* N50 do genoma da cultivar IAC-100 aumentou quase 15 vezes quando comparado com a montagem inicial do supernova, indo para 53.5 MB, enquanto o *scaffold* N50 da cultivar CD-215 aumentou quase 14X, indo para 53.8 MB.

Tabela 3: Parâmetros de qualidade dos genomas após a ordenação baseada em alinhamentos contra referência

Genoma / **Resultados**	**IAC-100**	**CD-215**
Scaffold N50	53.5 MB	53.8 MB
Número de *Scaffolds* com 75% da montagem	15	15
Scaffolds quiméricos quebrados	61	49
Scaffolds ancorados e ordenados	396	380
Tamanho da montagem ancorada e ordenada	1.057 Gb	1.061 Gb
Tamanho da montagem com os 20 maiores *scaffolds*	1.033 Gb	1.042 Gb

O *software* também foi mais eficiente em identificar *Scaffolds* quiméricos, identificando uma média de 55 *scaffolds* quiméricos, o *software* utiliza alinhamentos discordantes para realizar essa quebra, no caso de *contigs* quiméricos Inter cromossômicos, o

parâmetro utilizado para essa essa definição é que 5% do total do alinhamento, cubra ao menos 100kbp de outro cromossomo.

O uso de mapas genéticos é amplamente difundido e altamente recomendado, entretanto é necessária a disponibilidade de bons mapas, com alta densidade de marcadores, a obtenção desses dados leva tempo e podem ter custos elevados, sendo uma alternativa para espécies que ainda não tem um genoma de referência e apresentam bons mapas genéticos disponíveis.

Para espécies com bons genomas de referência, a utilização do *software* Ragoo pode ser uma boa opção, gerando ótimos resultados e sem custos adicionais, além de ser um *software* rápido e com pouca demanda computacional, com resultados que podem ser melhores do que utilizando dados de captura de conformação cromossômica, como foi discutido por Alonge et al., 2019, desse modo, os resultados desse *software* foram escolhidos para o prosseguimento da montagem.

O *software* REAPR identifica erros na montagem de genomas a partir de alinhamentos, dando pontuações de qualidade para todas as bases do genoma, os erros nos *contigs* são substituídas por Ns e *scaffolds* são quebrados. O *software* identificou com precisão um total de 34787 erros no genoma da CD-215 e 33387 erros no genoma da IAC-100, aproximadamente 63.5% dos erros foram devido a regiões sem cobertura de *reads*, causadas pelo *software* Ragoo que insere gaps entre *scaffolds* e também a erros de montagem do Supernova. Entretanto o *software* acaba fragmentando o genoma pela quebra de *scaffolds*, o *software* Ragoo foi novamente usado para a ordenação do genoma, gerando nossos *assemblies* finais.

Recentemente tivemos a publicação da chamada ‘referência de ouro’ da soja, da cultivar chinesa Zhonghuang 13, para a montagem dessa ‘referência de ouro’, foram utilizadas 5 diferentes técnicas, entre elas, *short reads* Hi-seq, *long reads* PacBio, mapas óticos Bionano marcados com BspQI e BssSI e captura de conformação cromossômica (Hi-C), esses dados geraram uma montagem com aproximadamente 1GB e um *Scaffold* N50 de 51.95 MB (Shen et al., 2019). Os resultados obtidos com REAPR e Ragoo, alavancaram a qualidade dos nossos *assemblies*, gerando os dois genomas de soja com melhor continuidade até a data presente publicados, superando o *scaffold* N50 da Zhonghuang 13 e também o genoma da cultivar Williams 82, que apresenta um *scaffold* N50 de 48.57MB que utilizou diversas técnicas de sequenciamento, nossos N50 também superam os N50 dos 26 genomas publicados por Liu et al (2020), onde o melhor tem um N50 de 52.3MB.

Tabela 4: Comparação dos parâmetros de qualidade dos *assemblies* gerados e dos *assemblies* das cultivares Williams 82 e Zhonghuang 13.

Genoma / Resultados	IAC-100	CD-215	Williams 82	Zhonghuang 13
Scaffold N50	53.6 MB	54.2 MB	48.57 MB	51.95 MB
Número de *Scaffolds*	396	350	1192	58
Tamanho da montagem	1.058 MB	1.061 MB	979 MB	997MB

A utilização de genomas de referência para a ordenação de *assemblies* pode ser uma ótima estratégia para a montagem de outros genomas, sendo que até a presente data, este é o único trabalho utilizando esse método a partir de uma montagem gerada unicamente com *linked- reads* da 10X Chromium Genome Technology.

Mesmo após a utilização do *pipeline* para correção de erros ancoragem e ordenamento, a qualidade do genoma avaliada através do banco de dados de genes ortólogos de cópia única do grupo das embriófitas, permaneceu semelhante a montagem do supernova, como pode ser visto na figura 10.

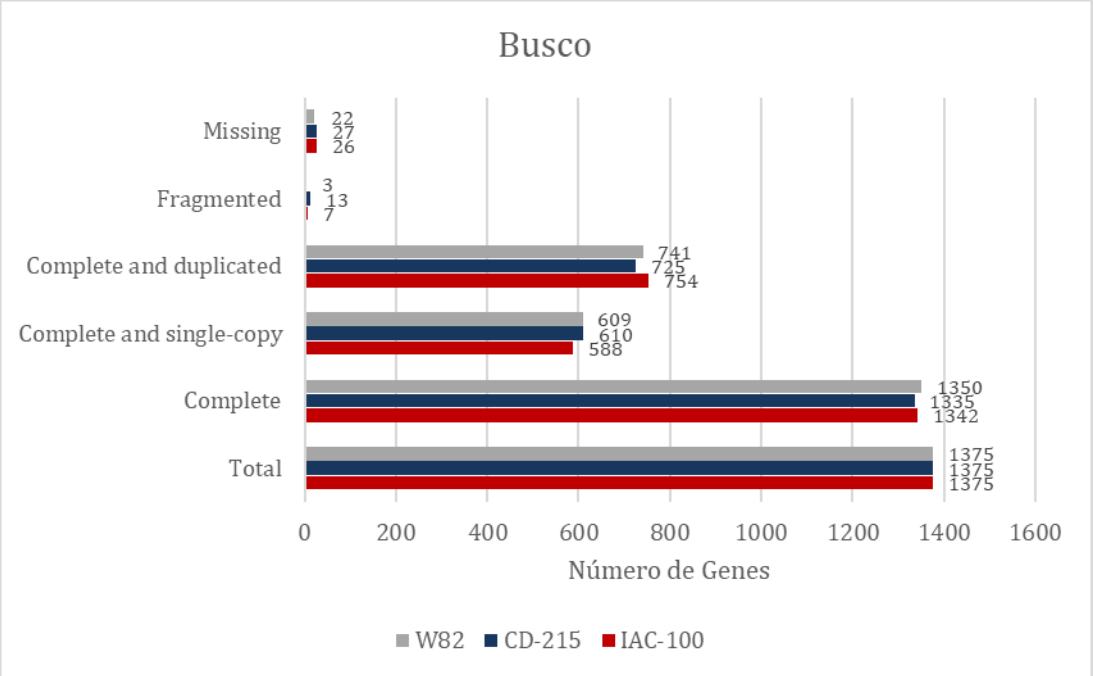


Figura 10: Figura 9: Avaliação de genes ortólogos de cópia única do grupo das embriófitas nos *assemblies* contrastantes para resistência ao complexo de percevejos, IAC-100 e CD-215 nas montagens finais; pode ser observado uma alta completude dos genes avaliados em nossas montagens (~ 97%), assim como no genoma de referência da Williams82.

A partir do alinhamento realizado com o NUCmer é possível observar uma alta sintenia entre todos os cromossomos de nossos *assemblies* e os cromossomos da Williams82, mostrando uma boa qualidade dos *assemblies* montados, como pode ser observado nas figuras 11 e 12. Apesar disso, existem algumas variantes estruturais (em azul) entre as mesmas e a Williams82, podendo ser observado que essas variantes nem sempre são comuns às duas, indicando variantes entre as mesmas, resultado esperado devido a divergência genética das mesmas.

Alinhamentos em vermelho, fora da diagonal principal, são resultados de alinhamentos entre posições diferentes dos cromossomos, algo comum em soja devido à alta incidência de genes duplicados devido sua história evolutiva e também elementos transponíveis, que constituem grande parte dos genomas de plantas (Gill et al., 2009).

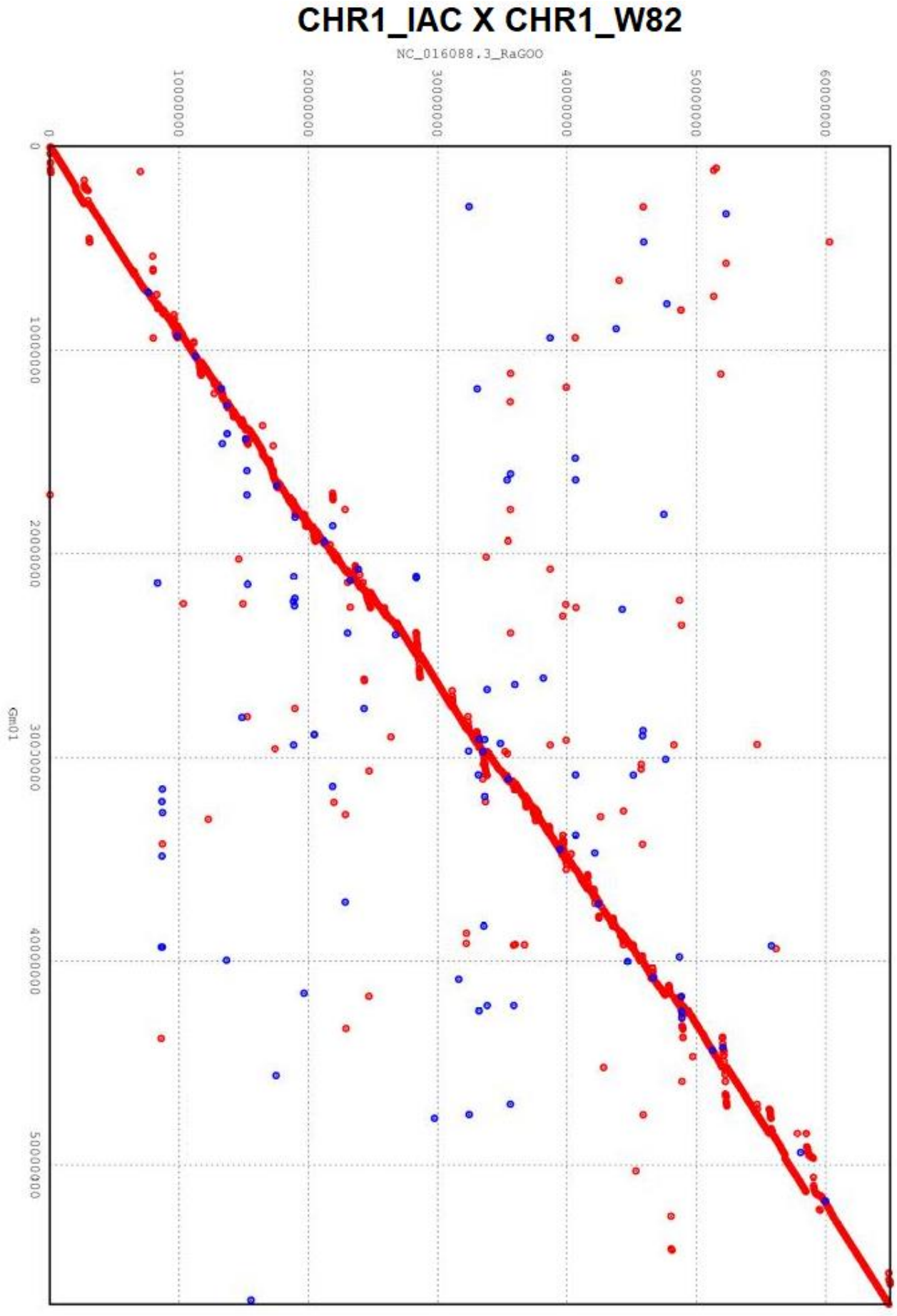


Figura 11: Plot do alinhamento entre o cromossomo 1 da IAC-100 e Williams 82.

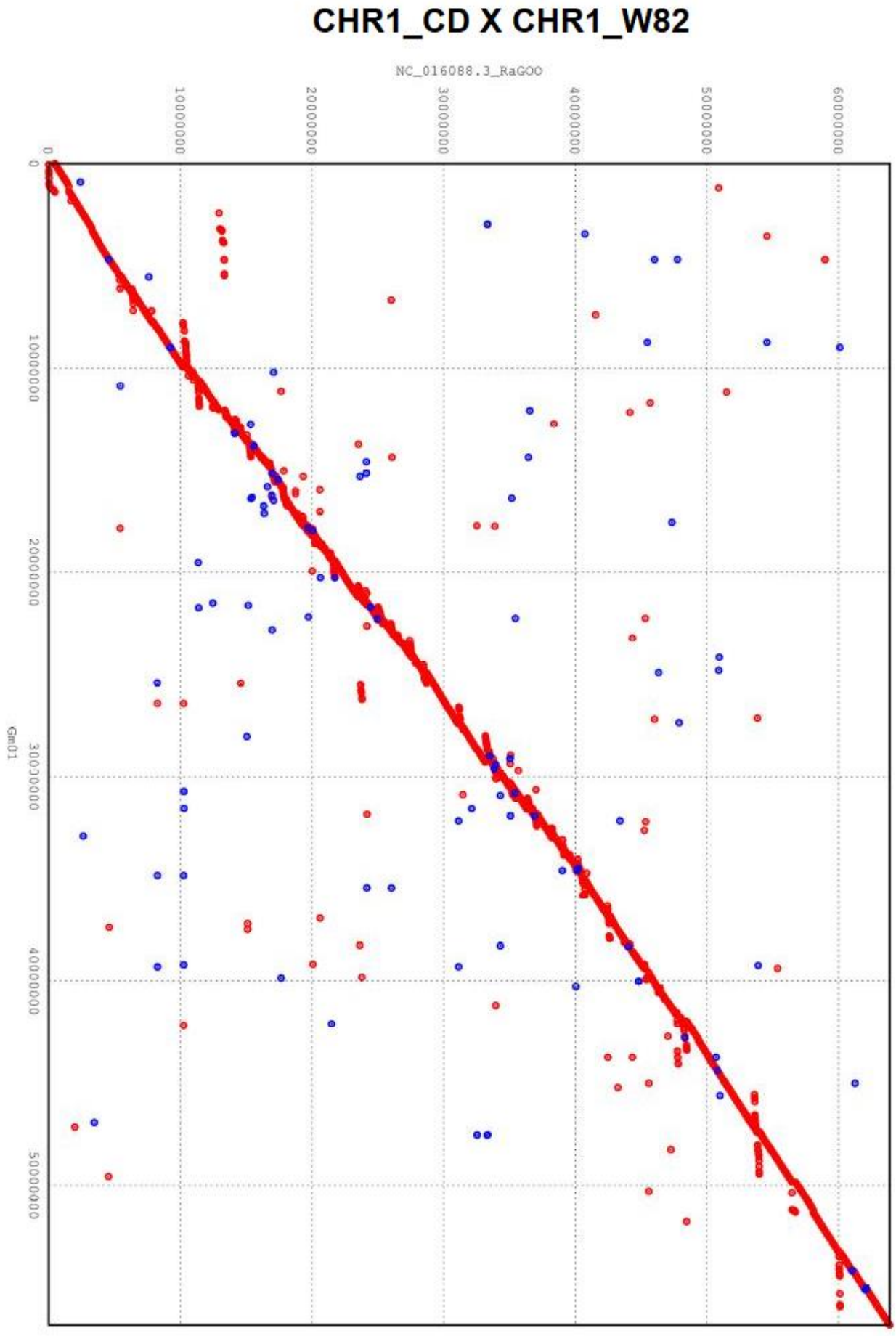


Figura 12: Plot do alinhamento entre o cromossomo 1 da CD-215 e Williams82

3.3 Anotação

O *software* repeatmasker identificou aproximadamente 44% de elementos transponíveis não identificados nos genomas montados, essa categorização sem identificação de classes, acontece devido ao banco de dados utilizado não ter distinção das classes desses elementos. Essa porcentagem de elementos presentes nos genomas é compatível com a identificada no genoma da cultivar Williams82 (43%) (Du et al., 2011). Apesar da identificação dos elementos transponíveis, não foram identificados DNAs satélites, repetições simples ou de baixa complexidade.

A anotação dos genes a partir das previsões com o *software* AUGUSTUS, possibilitou a identificação de 58.444 genes na montagem do genoma da cultivar IAC-100 e 57.946 genes n montagem da cultivar CD-215. O *software* se mostra eficiente na previsão de genes *Ab initio* de soja, pois esses números estão entre a quantidade de genes identificados na Zhonghuang 13 e na Williams 82, que são 55.443 e 59.847 genes (Shen et al.,2019). Esses números também estão próximos da média de 56.522 genes dos 26 genomas de soja montados por Liu et al. (2020).

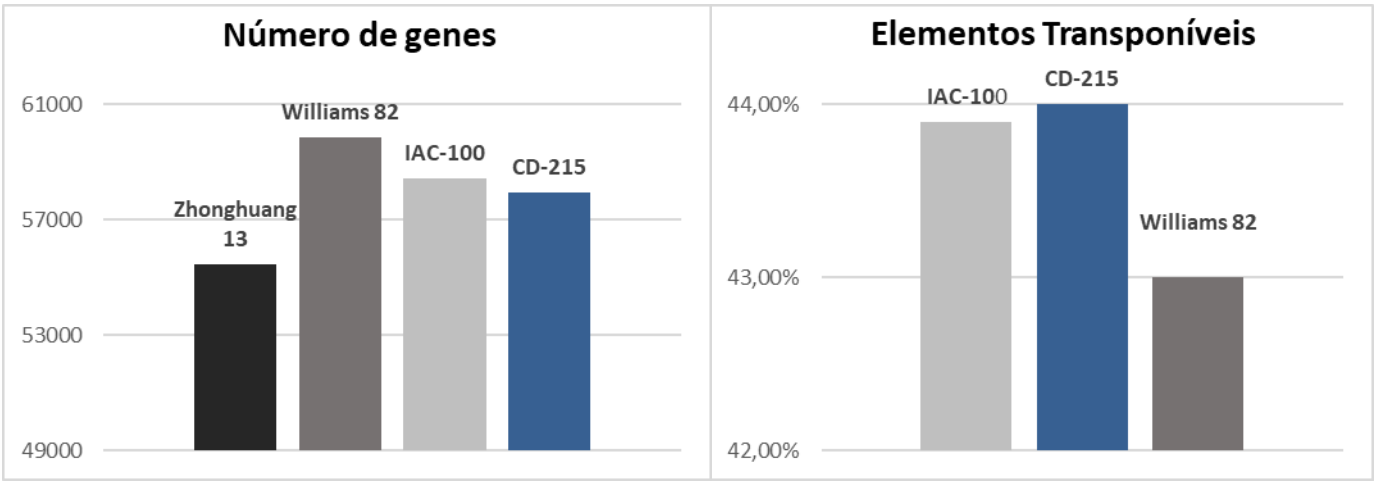


Figura 13: Vemos no gráfico da direita, uma comparação da porcentagem elementos repetitivos identificados nas montagens dos genomas da IAC-100 e CD-215 com Williams 82, enquanto a esquerda temos um comparativo de número de genes das montagens com Williams 82 e Zhonghuang 13.

3.4 Variantes

Com a crescente disponibilidade de dados genéticos e fenotípicos, melhoristas têm em mãos valiosos dados que podem ser aplicados ao melhoramento genético através de estudos de associação genômica ampla, seleção genômica e edição genética (Hu et al., 2018). Quando genomas anotados de alta qualidade estão disponíveis, os genes e variantes genéticas que contribuem para as características agronômicas de interesse podem ser identificados e as alterações feitas durante os processos de melhoramento podem ser avaliadas em nível molecular, entretanto para que essas associações entre genótipo e fenótipo possam ocorrer, é primordial a identificação de variantes de alta qualidade.

Uma média de 850 milhões de *reads* sequenciadas para as cultivares IAC-100 e CD-215 permitiu alinhamentos com altas profundidades de *reads*, aumentando a confiabilidade de SNPs, indels encontrados a partir do *pipeline* do GATK , segundo Kishikawa.,(2019), uma profundidade de aproximadamente 15 vezes, é suficiente para uma chamada de SNPs com 99% de acurácia e Indels necessitam de mais de 300 *reads* de profundidade para 95% de acurácia, entretanto profundidades entre 50 e 100 *reads* permitem uma acurácia de aproximadamente 80% para Indels.

A partir do *pipeline* do GATK foi obtido um total de 480.163 SNPs na cultivar IAC-100 e 442.754 SNPs na cultivar CD-215, sendo que 189.799 SNPs são compartilhados e 290.189 SNPs são únicos na IAC-100. Esses números são dentro do esperado considerando trabalhos que encontraram aproximadamente 9 e 10 milhões de SNPs em populações de soja genotipadas com alta profundidade (Happ et al., 2019); (Valliyodan et al., 2016). Foram obtidos um total de 335.229 indels únicos na IAC-100 e 248.725 indels compartilhados com CD-215.

Variantes são relatadas como importantes para diversas características agronômicas em muitas espécies, como tamanho e número de grãos em arroz, resistência a doenças, tolerância a deficiência de fósforo, arquitetura de planta, sensibilidade ao fotoperíodo, tolerância a alumínio, tolerância a frio, tempo de florescimento, resistência a nematóides, entre outras (Tao et al., 2019).

Especificamente em soja, já conhecemos que variantes estruturais são responsáveis por tolerância a salinidade e resistência ao nematoide do cisto, entretanto, provavelmente estão envolvidos em inúmeras outras características (Tao et al., 2019).

Essas variantes únicas encontradas na IAC-100 podem ser utilizadas em futuros estudos, sendo usadas em *chips* ou como marcadores tradicionais para a genotipagem de

populações visando mapeamento de QTLs, mapeamento fino, estudos de associação genômica ampla e seleção genômica, pois representam haplótipos derivados da IAC-100, onde dentro desse conjunto de variantes, podem haver algumas importantes para a resistência ao complexo de percevejos.

Figura 14: Diagramas de Venn das duas cultivares usadas nesse estudo, mostrando o número de SNPs e indels compartilhados pelas mesmas, usando como referência o genoma da Williams82.

A busca por variantes estruturais, como grandes deleções, na região dos QTLs mapeados por Moller (2010), para a resistência ao complexo de percevejos, trouxe uma série de variantes presentes unicamente na cultivar IAC-100, a posição dos que estão próximos a genes candidatos estão na tabela suplementar 1. A identificação dessas grandes variantes estruturais do genoma da IAC-100 pode ser usada para genotipar uma população de RILs oriundas do cruzamento entre IAC-100 e CD-215 ou Williams82, visando realizar mapeamento fino nessas regiões e encontrar associação entre as variantes e a resistência, permitindo diminuir o número de genes candidatos e encontrando marcadores fortemente associados resistência para uso em seleção assistida.

A partir das variantes observadas e os genes presentes na região dos QTLs, foi possível notar um grande número de genes ainda não caracterizados, com funções desconhecidas, mas ao mesmo tempo alguns genes se mostraram bons candidatos a serem explorados em futuros estudos, por estarem relacionados biologicamente com a resistência a insetos e/ou fitopatógenos segundo a literatura.

No cromossomo I/GM20, é relatado um QTL na posição de 40 cM, correspondente aproximadamente a posição 34580130 de bases na Williams82. Próximo de variantes estruturais únicos na IAC-100 foram encontrados genes como cinnamate 4-hydroxylase (Gene

ID: 100812188), TGA26 bZIP transcription factor (Gene ID: 100775280), trichome birefringence-like 36 (Gene ID: 100803650), germin-like protein subfamily 3 member 2-like (100803113) , NBS-LRR (Gene ID: 100305369), GEM-like protein 8 (Gene ID: 100778858) e Diphosphomevalonate decarboxylase MVD2 (Gene ID: 100800990)

O gene cinnamate 4-hydroxylase, é uma enzima da rota de biossíntese de fenilpropanóide/lignina, ela dá origem a coumaric acid (COA), que pode dar origem a flavonoides, isoflavonas, fitoalexinas e ligninas. O aumento da expressão desse gene pode ocasionar acúmulo de lignina, uma resposta chave para as plantas, aumentando a resistência a fitopatógenos, como a *Phytophthora sojae*, *Verticillium dahlia* e *Fusarium culmorum* (Kang et al., 2000); (Gayoso et al., 2010); (Yan et al., 2019). Já o acúmulo de flavonóides e isoflavonóides são conhecidos por seus efeitos contra insetos, como pulgões e percevejos, sendo encontradas diferenças entre suas concentrações em cultivares de soja suscetíveis e resistentes, com a cultivar IAC-100 apresentando altas concentrações desses compostos quando comparada com cultivares suscetíveis a percevejos (Harborne; Wuyts, 2000); (Piubelli et al., 2003a); (Meng et al., 2011)

O gene TGA26 bZIP é de uma família de fatores de transcrição, conhecidos por estarem envolvidos com desenvolvimento de sementes, flores e resistência a fitopatógenos e estresses abióticos em Arabidopsis (JAKOBY et al., 2002). Em soja, um dos membros da família está relacionado a indução de uma chalcona sintase, se ligando ao seu promotor, sabemos que a chalcona sintase faz parte da via metabólica da biossíntese de Fenilpropanóide/lignina, sendo conhecido por induzir isoflavonóides na interação com fitopatógenos (DRÖGE et al., 1997). Entretanto, segundo Ullah (2019), outros membros da família, em soja, estão relacionados a nodulação simbiótica e em resposta à disponibilidade de nitrogênio na soja, dessa maneira uma caracterização do gene TGA26 se faz necessária para determinar em que processos está envolvido.

O gene trichome birefringence-like 36 pode estar envolvido na deposição de celulose na parede secundária de tricomas (Bischoff et al., 2010), enquanto os genes da família do germin-like protein subfamily 3 member 2-like estão envolvidos na resposta de defesa contra diversos fitopatógenos (Pei et al., 2019). Outro gene próximo de uma variante estrutural foi o de uma proteína NBS-LRR, conhecidas pelo seu papel na resistência a doenças de plantas e insetos (Belkhadir et al., 2004). Também temos o gene putative GEM-like protein 8 de uma família de proteínas envolvidas na sinalização do ácido abscísico (Mauri et al.,2016). Já o gene Diphosphomevalonate decarboxylase MVD2 está envolvido na síntese de terpenóides,

os compostos mais estudados para resistência a insetos, podendo ser voláteis ou não, causando a dissuasão de insetos em diversas espécies vegetais (Kortbeek et al., 2019).

No cromossomo C2/GM6 temos dois QTLs, nas posições aproximadas de 16029425 e 16985392 (90 cM e 99 cM) de bases, usando como referência a Williams 82. Uma região com muitos genes aparentemente sem correlação com resistência ou sem função conhecida, entretanto foram encontrados dois genes promissores para futuros estudos, próximos de variantes estruturais únicos da IAC-100, como o gene adenosylhomocysteinase-like (GeneID: 100779609), que realiza metilações. Uma menor expressão desse gene ocasiona uma maior resposta de defesa contra fitopatógenos, observada em diversas culturas como tomate, Arabidopsis, tabaco e batata, entretanto a perda parcial de função desse gene altera o desenvolvimento das plantas, com plantas atrofiadas, raízes pequenas e sem dominância apical (Hui et al., 2015).

Outro gene encontrado na região foi o GASA11 (Gene ID: 100306354), membros dessa família gênica respondem a hormônios como a giberilina e atuam no desenvolvimento de plantas, sempre sendo relacionados a tecidos jovens e órgãos em crescimento ativo, tempo de floração, tamanho e o peso das sementes, desenvolvimento de flores e siliques (estruturas muito parecidas com vagens) (Zhang et al., 2008); (Nahirñak et al., 2012); (Roxrud et al., 2007). A giberilina que induz os genes GASA é conhecida por aumentar o número de vagens e produtividade da soja, sem aumentar a sua produção de parte aérea, essas características são relatadas como características associadas a resistência ao complexo de percevejos, sendo utilizadas nas fenotipagens para estudos de mapeamentos de QTLs (Bertolin et al., 2010); (Pinheiro et al., 2016). Além disso a Gibberellin 2-oxidase e Gibberellin 3-oxidase estão presentes entre os transcritos diferencialmente expressos nas cultivares IAC-100 e CD215 na interação com *Piezodorus guildinii*, entretanto apenas a IAC expressa a Gibberellin 3-oxidase (Glyma15g01500) e a Gibberellin 2-oxidase (Glyma13g28970) está em uma posição diferente do genoma (Silva, 2014).

Além dos genes envolvidos com as rotas da biossíntese de fenilpropanóide/lignina e terpenos encontrados nos QTLs, diversos genes correlacionados com essas vias foram encontrados diferencialmente expressos por Silva (2014), apenas na cultivar IAC-100 e não na CD-215, juntamente com algumas proteínas NBS-LRR.

Os dados de literatura sobre resistência a insetos em plantas, assim como as variantes estruturais encontradas nas regiões com QTLs, próximos a genes das rotas de biossíntese de fenilpropanóide/lignina e terpenos, além do trabalho de Silva (2014), corroboram que essas vias metabólicas são importantes para a resistência ao complexo de percevejos. Desse modo

análises químicas, determinando a quantidade desses compostos em progênies podem ser usadas durante a seleção de genótipos superiores para a resistência ao complexo de percevejos.

A partir dos dados de GBS, foram encontradas variantes nas 25 linhagens mais resistentes de uma população de RILs. Estas variantes, podem indicar regiões genômicas candidatas a serem melhor estudadas. Foram encontrados 852 SNPs compartilhados e presentes unicamente na cultivar IAC-100, que podem ser vistos na figura 15.

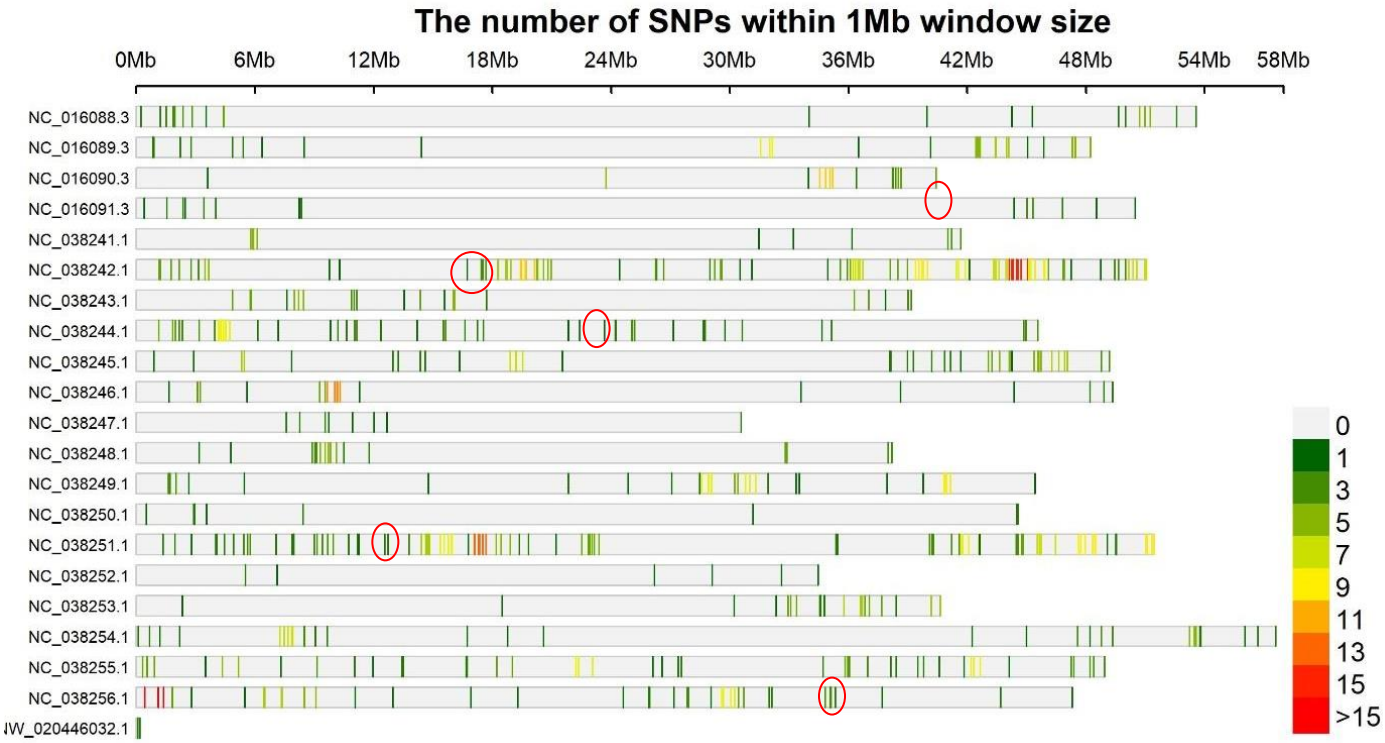


Figura 15: Representação das variantes compartilhadas encontradas nas linhagens resistentes, cromossomo 1 ao 20 de cima para baixo e mais um *scaffold*, círculos em vermelho representam regiões com QTLs relatados por Moller, 2010.

É interessante notar a região de 35 milhões de pares de base, do cromossomo I, relatado com um QTL, essa região está presente em todas as 25 linhagens resistentes, ao todo são 68 SNPs nesse cromossomo. A região dos dois QTLs relatados no cromossomo C2 (16029425 e 16985392 pares de base), apresenta apenas um SNPs exatamente entre as posições, entretanto considerando o haplótipo da região, existem vários SNPs. O cromossomo C2 apresenta no total 130 SNPs compartilhado, um número alto considerando o número total de SNPs, fazendo deste cromossomo um bom alvo para futuros estudos.

Outro QTL relatado é no cromossomo 15, próximo da posição 13,7 milhões de bases (66.5 cM), todo o haplótipo dessa região parece estar presente nas 25 linhagens mais resistentes, além disso o cromossomo apresenta 124 SNPs compartilhados, um número relativamente alto.

Os QTLs relatados nos cromossomos 7 (22.9 milhões de pares de base) e 4 (40.5 milhões de pares de base) estão em cromossomos com poucos SNPs compartilhados, 34 e 17 respectivamente, com o QTL do cromossomo 4 longe de qualquer SNP compartilhado, o que pode ocorrer devido a técnica utilizada, que foi uma biblioteca de representação reduzida do genoma ou um QTL falso positivo.

4 CONCLUSÕES

Os dois *assemblies* gerados neste trabalho, apenas com a tecnologia de *linked-reads*, assim como outros *assemblies* que usaram apenas essa abordagem, apresentam bons resultados a um custo relativamente baixo quando comparado com outras metodologias de sequenciamento, como *long-reads* ou a união de mais de um método de sequenciamento. Sendo assim, o uso dessa tecnologia aliada a esse *pipeline* pode ser uma ótima alternativa para a obtenção de genomas a um baixo custo, possibilitando um grande avanço para estudos de características específicas dos genótipos estudados.

Os haplótipos do genoma da IAC-100 presentes nas 25 linhagens mais resistentes são compatíveis com QTLs da literatura, desse modo essas regiões devem ser exploradas no melhoramento genético, via marcadores que podem ser desenvolvidos a partir das variantes únicas da IAC-100 e usados para mapeamento fino das regiões dos QTLs. Os haplótipos mais significativos podem ser usados em programas de melhoramento, além de diversos outros estudos genéticos para essa característica, como seleção genômica e associação genômica ampla.

Como gene GASA11 é ativo em regiões em desenvolvimento, como sementes e até mesmo em siliques (estruturas parecidas com vagens), e faz parte de uma família de genes responsiva a giberelina, fitohormônio conhecido por aumentar o número de vagens em soja, que é uma das características de cultivares resistentes ao complexo de percevejos, GASA11 assim como a giberilina, são bons candidatos a estudos futuros sobre resistência.

REFERÊNCIAS

ACE Darling, B Mau, FR Blattner, NT Perna ‘Mauve: multiple alignment of conserved genomic sequence with rearrangements’ **Genome research** 14 (7), 1394-1403, 2004

ADAMS, Keith L.; WENDEL, Jonathan F. Polyploidy and genome evolution in plants. **Current opinion in plant biology**, v. 8, n. 2, p. 135-141, 2005.

AHN, Yul-Kyun et al. Whole genome resequencing of Capsicum baccatum and Capsicum annuum to discover single nucleotide polymorphism related to powdery mildew resistance. **Scientific reports**, v. 8, n. 1, p. 1-11, 2018.

ALONGE, Michael et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. **Genome biology**, v. 20, n. 1, p. 1-17, 2019.

AMMIRAJU, Jetty SS et al. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. **The Plant Journal**, v. 52, n. 2, p. 342-351, 2007.

ARMSTRONG, E. E. et al. Cost-effective assembly of the African wild dog (Lycaon pictus) genome using linked reads**. Gigascience**. 2019; 8.

ARMSTRONG, Joel et al. Whole-genome alignment and comparative annotation. **Annual review of animal biosciences**, v. 7, p. 41-64, 2019.

BELKHADIR, Youssef; SUBRAMANIAM, Rajagopal; DANGL, Jeffery L. Plant disease resistance protein signaling: NBS–LRR proteins and their partners. **Current opinion in plant biology**, v. 7, n. 4, p. 391-399, 2004.

BENNETZEN, Jeffrey L.; WANG, Hao. The contributions of transposable elements to the structure, function, and evolution of plant genomes. **Annual review of plant biology**, v. 65, p. 505-530, 2014.

BERTOLIN, Danila Comelis et al. Increase of the productivity of the soybean crop with the application of biostimulants. **Bragantia**, v. 69, n. 2, p. 339-347, 2010.

BHARGAVA, Atul; FUENTES, F. F. Mutational dynamics of microsatellites. **Molecular biotechnology**, v. 44, n. 3, p. 250-266, 2010.

BISCHOFF, Volker et al. TRICHOME BIREFRINGENCE and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in Arabidopsis. **Plant physiology**, v. 153, n. 2, p. 590-602, 2010.

BOWERS, John E. et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. **Nature**, v. 422, n. 6930, p. 433, 2003.

BRUCE, Toby JA et al. The first crop plant genetically engineered to release an insect pheromone for defence. **Scientific reports**, v. 5, p. 11183, 2015.BUSCO: assessing genome

assembly and annotation completeness with single-copy orthologs. Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov **Bioinformatics**, published online June 9, 2015

CEPEA/ESALQ, ANDEF. Impacto econômico de pragas agrícolas no Brasil. In: 15° Enfisa - **Encontro de fiscalização e seminário sobre agrotóxicos**, Campos do Jordão, SP, Bras (2017)

CHAPMAN, Kaitlin M. et al. Abscisic and jasmonic acids contribute to soybean tolerance to the soybean aphid (Aphis glycines Matsumura). **Scientific reports,** v. 8, n. 1, p. 1-12, 2018.

CHEN, Risheng et al. “Whole genome sequencing and comparative transcriptome analysis of a novel seawater adapted, salt-resistant rice cultivar - sea rice 86” ***BMC genomics*** vol. 18,1 655. 23 Aug. 2017, doi:10.1186/s12864-017-4037-3

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

CINGOLANI, Pablo et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SNPEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80-92, 2012.

CM, BIRKETT MA, NAPIER JA, JONES HD, PICKETT JA. The first crop plant genetically engineered to release an insect pheromone for defence. ***Sci Rep.*** *25 Jun 2015;*

COMPEAU, Phillip EC; PEVZNER, Pavel A.; TESLER, Glenn. How to apply de Bruijn graphs to genome assembly. **Nature biotechnology**, v. 29, n. 11, p. 987-991, 2011.

CONAB, 2019 Disponível Em: https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/download/28423_f8d26395b1eb12f895718879e1aca901>. Acesso em: 1 de setembro de 2019.

CONTE, Matthew A. et al. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. **Gigascience**, v. 8, n. 4, p. giz030, 2019.

CORRÊA-FERREIRA, B. S.; KRZYZANOWSKI, F. C.; MINAMI, C. A. Percevejos e a qualidade da semente de soja **– série sementes**, 2009.

CORRÊA-FERREIRA, B. S.; PANIZZI, A. R. Percevejos da Soja e seu Manejo. **Circular técnica/Embrapa CNPSo, ISSN 0100-6703 n.24**, p. 45, 1999.

COSTA NETO, P. R. & ROSSI, L. F. S. Produção de biocombustível alternativo ao óleo diesel através da transesterificação de óleo de soja usado em fritura. **Química Nova**, v.23, p. 4, 2000

DALIO, R. J. D. et al. Efetores na interação planta-patógeno. Revisão **Anual de Patologia de Plantas**, Passo Fundo, v. 22, p. 25-68, 2014.

DASTMALCHI, Mehran; DHAUBHADEL, Sangeeta. Soybean chalcone isomerase: evolution of the fold, and the differential expression and localization of the gene family. **Planta**, v. 241, n. 2, p. 507-523, 2015.

DE BRUIJN, Nicolaas Govert. A combinatorial problem. In: Proc. **Koninklijke Nederlandse Academie van Wetenschappen**. 1946. p. 758-764.

DE BRUIJN; N. G. (1946). "A Combinatorial Problem". Koninklijke Nederlandse Akademie v**. Wetenschappen**. 49: 758–764.

DE LANNOY, Carlos; DE RIDDER, Dick; RISSE, Judith. The long reads ahead: de novo genome assembly using the MinION. **F1000Research**, v. 6, 2017.

DEL ANGEL, Victoria Dominguez et al. Ten steps to get started in Genome Assembly and Annotation. **F1000Research**, v. 7, 2018.

DODSWORTH, Steven; LEITCH, Andrew R.; LEITCH, Ilia J. Genome size diversity in angiosperms and its influence on gene space. **Current opinion in genetics & development**, v. 35, p. 73-78, 2015.

DRÖGE-LASER, Wolfgang et al. Rapid stimulation of a soybean protein-serine kinase that phosphorylates a novel bZIP DNA-binding protein, G/HBF-1, during the induction of early transcription-dependent defenses. **The EMBO Journal**, v. 16, n. 4, p. 726-738, 1997.

DU, Jianchang et al. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. **Bmc Genomics**, v. 11, n. 1, p. 113, 2010.

EMBRAPA 2018; Disponível Em: https://www.embrapa.br/busca-de-noticias/-/noticia/34944361/pesquisa-desenvolve-primeira-soja-tolerante-a-percevejo>. Acesso em: 25 de junho de 2018.

ERB M, VEYRAT N, ROBERT CA, XU H, FREY M, TON J, TURLINGS TC. Indole is an essential herbivore-induced volatile priming signal in maize**.** **Nat Commun** Feb 16, 2015

FEHR, Walter R.; CAVINESS, Charles E. Stages of soybean development. 1977.

FIERST, Janna L. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. **Frontiers in genetics**, v. 6, p. 220, 2015.

GAYOSO, Carmen et al. The Ve-mediated resistance response of the tomato to Verticillium dahliae involves H 2 O 2, peroxidase and lignins and drives PALgene expression. **BMC Plant Biology**, v. 10, n. 1, p. 232, 2010.

GHURYE, Jay et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. **bioRxiv**, p. 261149, 2019.

GILL, Navdeep et al. Molecular and chromosomal evidence for allopolyploidy in soybean. **Plant physiology**, v. 151, n. 3, p. 1167-1174, 2009.

GODOI, C. R. C. DE; PINHEIRO, J. B. Genetic parameters and selection strategies for soybean genotypes resistant to the stink bug-complex. **Genetics and Molecular Biology,** v. 32, n. 2, p. 328–336, 2009.

GOMEZ; ROGGIA 2012; Disponível Em: http://www.irac-online.org/content/uploads/folder-resistencia-percevejos.pdf >. Acesso em: 6 de fevereiro de 2018.

GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, 17.6: 333, 2016.

GREILHUBER, J. et al. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. **Plant biology**, v. 8, n. 06, p. 770-777, 2006.

GUO, Shuchun et al. Large-scale transcriptome comparison of sunflower genes responsive to Verticillium dahliae. **BMC genomics**, v. 18, n. 1, p. 42, 2017.

HANLEY, Mick E. et al. Plant structural traits and their role in anti-herbivore defence. **Perspectives in Plant Ecology, Evolution and Systematics**, v. 8, n. 4, p. 157-178, 2007.

HAPP, Mary M. et al. Generating High Density, Low Cost Genotype Data in Soybean [Glycine max (L.) Merr**.]. G3: Genes, Genomes, Genetics**, v. 9, n. 7, p. 2153-2160, 2019.

HARBORNE, J.B.; WILLIAMS, C.A. Advances in flavonoid research since. ***Phytochemistry*** 2000, *55*, 481–504. 1992.

HAWKINS, Jennifer S. et al. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. **Genome research**, v. 16, n. 10, p. 1252-1261, 2006.

HOFFMANN-CAMPO, C. B. et al. Pragas da soja no Brasil e seu manejo integrado. Londrina: **Embrapa Soja**, (Circular Técnica, 30).2000.

HOLT, Carson; YANDELL, Mark. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. **BMC bioinformatics**, v. 12, n. 1, p. 491, 2011.

HU, Haifei; SCHEBEN, Armin; EDWARDS, David. Advances in integrating genomics and bioinformatics in the plant breeding pipeline. , 8.6: 75, **Agriculture**, 2018

HUI, Li Xiao et al. Co-silencing of tomato S-adenosylhomocysteine hydrolase genes confers increased immunity against Pseudomonas syringae pv. tomato DC3000 and enhanced tolerance to drought stress. **Frontiers in plant science**, v. 6, p. 717, 2015.

HULSE-KEMP, Amanda M.,et al. Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. **Horticulture Research**, 2018, 5.1: 4.

JAKOBY, Marc et al. bZIP transcription factors in Arabidopsis. **Trends in plant science**, v. 7, n. 3, p. 106-111, 2002.

JIANG, Shukun et al. Resequencing and variation identification of whole genome of the japonica rice variety" Longdao24" with high yield. **PloS one**, v. 12, n. 7, p. e0181037, 2017.

JIAO, Wen-Biao; SCHNEEBERGER, Korbinian. The impact of third generation genomic technologies on plant genome assembly. **Current opinion in plant biology**, 36: 64-70, 2017.

JUNG, Hyungtaek et al. Tools and strategies for long-read sequencing and de novo assembly of plant genomes. **Trends in plant science**, v. 24, n. 8, p. 700-724, 2019.

JUNG, Hyungtaek, et al. Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes**. Trends in plant science**, 2019.

KANG, Z. et al. Ultrastructural and immunocytochemical investigation of pathogen development and host responses in resistant and susceptible wheat spikes infected by Fusarium culmorum. **Physiological and Molecular Plant Pathology**, v. 57, n. 6, p. 255-268, 2000.

KARLEY A. J.,MITCHELL C.,BROOKES C.,MCNICOL J.,O'NEILL T.,ROBERTS H.,ET AL.. Exploiting physical defence traits for crop protection: leaf trichomes of *Rubus idaeus* have deterrent effects on spider mites but not aphids. **Ann. Appl. Biol** 168 159–172, 2016

KERSEY, Paul Julian. Plant genome sequences: past, present, future. **Current opinion in plant biology**, v. 48, p. 1-8, 2019.

KIRST, MATIAS et al. "Coordinated Genetic Regulation of Growth and Lignin Revealed by Quantitative Trait Locus Analysis of cDNA Microarray Data in an Interspecific Backcross of Eucalyptus." ***Plant Physiology*** 135.4, 2004

KISHIKAWA, Toshihiro, et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. **Scientific reports**, 9.1: 1-10 , 2019

KOCH, Robert L. et al. Identification, biology, impacts, and management of stink bugs (Hemiptera: Heteroptera: Pentatomidae) of soybean and corn in the Midwestern United States. **Journal of Integrated Pest Management**, v. 8, n. 1, 2017.

KORTBEEK, Ruy WJ; VAN DER GRAGT, Michelle; BLEEKER, Petra M. Endogenous plant metabolites against insects. **European Journal of Plant Pathology**, v. 154, n. 1, p. 67-90, 2019

KREPLAK, Jonathan et al. A reference genome for pea provides insight into legume genome evolution. **Nature genetics**, v. 51, n. 9, p. 1411-1422, 2019.

KUROIWA, Tsuneyoshi et al. Genome size of the ultrasmall unicellular freshwater green alga, Medakamo hakoo 311, as determined by staining with 4′, 6-diamidino-2-phenylindole after microwave oven treatments: II. Comparison with Cyanidioschyzon merolae, Saccharomyces cerevisiae (n, 2n), and Chlorella variabilis. **Cytologia**, v. 81, n. 1, p. 69-76, 2016.

KWON, Daehong; LEE, Jongin; KIM, Jaebum. GMASS: a novel measure for genome assembly structural similarity. **BMC bioinformatics**, v. 20, n. 1, p. 147, 2019.

LI, Jianying et al. Transcriptome analysis reveals a comprehensive insect resistance response mechanism in cotton to infestation by the phloem feeding insect Bemisia tabaci (whitefly). **Plant biotechnology journal**, v. 14, n. 10, p. 1956-1975, 2016.

LIU, Yucheng et al. Pan-genome of wild and cultivated soybeans. **Cell**, 2020.

LYE, Zoe N.; PURUGGANAN, Michael D. Copy number variation in domestication. **Trends in plant science**, v. 24, n. 4, p. 352-365, 2019.

MARDIS, Elaine R. DNA sequencing technologies: 2006–2016. **Nature protocols**, 2017, 12.2: 213.

MAURI, Nuria et al. GEM, a member of the GRAM domain family of proteins, is part of the ABA signaling pathway. **Scientific reports**, v. 6, n. 1, p. 1-11, 2016.

MEDZHITOV, R. & JANEWAY, C. A.. Innate immunity: the virtues of a nonclonal system of recognition. **Cell** 91: 295–8. 1997

MENG, Fanli et al. QTL underlying the resistance to soybean aphid (Aphis glycines Matsumura) through isoflavone-mediated antibiosis in soybean cultivar 'Zhongdou 27'. **Theoretical and applied genetics**, v. 123, n. 8, p. 1459-1465, 2011.

METZKER, Michael L. Sequencing technologies—the next generation. **Nature reviews genetics**, v. 11, n. 1, p. 31, 2010.

MICHAEL, Todd P.; JACKSON, Scott. The first 50 plant genomes. **The plant genome**, v. 6, n. 2, 2013.

MITHÖFER, A. & BOLAND, W.. Recognition of herbivory-associated molecular patterns. **Plant Physiol**. 146: 825–31. 2008.

MOLLER M (2010) Mapeamento de locos de resistência quantitativa da soja ao complexo de percevejos. **Dissertação**, Universidade de São Paulo

NAHIRÑAK, Vanesa et al. Snakin/GASA proteins: involvement in hormone crosstalk and redox homeostasis. **Plant signaling & behavior**, v. 7, n. 8, p. 1004-1008, 2012.

OATES, Caryn N. et al. The transcriptome and terpene profile of Eucalyptus grandis reveals mechanisms of defense against the insect pest, Leptocybe invasa. **Plant and Cell Physiology**, v. 56, n. 7, p. 1418-1428, 2015.

OZEROV, Mikhail Yu, et al. Highly Continuous Genome Assembly of Eurasian Perch (Perca fluviatilis) Using Linked-Read Sequencing. ***G3: Genes, Genomes, Genetics***, 8.12: 3737-3743, 2018.

PANIZZI, A. R.; SLANSKY, F. J. Review of Phytophagous Pentatomids (Hemiptera: Pentatomidae) Associated with Soybean in the Americas **Florida Entomologist**, 1985.

PATIL, Gunvant B. et al. Whole-genome re-sequencing reveals the impact of the interaction of copy number variants of the rhg1 and Rhg4 genes on broad-based resistance to soybean cyst nematode. **Plant biotechnology journal**, v. 17, n. 8, p. 1595-1611, 2019.

PEI, Yakun et al. GhABP19, a novel germin-like protein from Gossypium hirsutum, plays an important role in the regulation of resistance to verticillium and fusarium wilt pathogens. **Frontiers in plant science**, v. 10, p. 583, 2019.

PELLICER, Jaume; FAY, Michael F.; LEITCH, Ilia J. The largest eukaryotic genome of them all?. **Botanical Journal of the Linnean Society**, v. 164, n. 1, p. 10-15, 2010.

PIMENTEL, D. Techniques for Reducing Pesticides: Environmental and Economic Benefits. **Chichester**, UK: John Wiley 1997

PINHEIRO, José Baldin; VENDRAMIM, José Djair; LOURENÇÃO, André Luiz. Programas geram cultivares de soja resistentes a insetos**, Visão Agricola** n° 5**,** 2016.

PIUBELLI, Giorla C. et al. Nymphal development, lipid content, growth and weight gain of Nezara viridula (L.)(Heteroptera: Pentatomidae) fed on soybean genotypes. **Neotropical Entomology**, v. 32, n. 1, p. 127-132, 2003 a.

PIUBELLI, Giorla Carla et al. Flavonoid increase in soybean as a response to Nezara viridula injury and its effect on insect-feeding preference. **Journal of chemical ecology**, v. 29, n. 5, p. 1223-1233, 2003b.

POP, Mihai; SALZBERG, Steven L.; SHUMWAY, Martin. Genome sequence assembly: Algorithms and issues. **Computer**, v. 35, n. 7, p. 47-54, 2002.

PROOST, Sebastian et al. Journey through the past: 150 million years of plant genome evolution. **The Plant Journal**, v. 66, n. 1, p. 58-65, 2011.

QI, X. *et al.* Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. ***Nat. Commun***. 5:4340 doi: 10.1038/ncomms5340 (2014).

REAPR: a universal tool for genome assembly evaluation, Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD, **Genome Biology** (2013, 14(5):R47).

RICE, Edward S.; GREEN, Richard E. New approaches for genome assembly and scaffolding. **Annual review of animal biosciences**, v. 7, p. 17-40, 2019.

ROBINSON, James T. et al. Integrative genomics viewer. **Nature biotechnology**, v. 29, n. 1, p. 24-26, 2011.

ROXRUD, Ingrid et al. GASA4, one of the 14-member Arabidopsis GASA family of small polypeptides, regulates flowering and seed development. **Plant and Cell Physiology,** v. 48, n. 3, p. 471-483, 2007.

ROXRUD, Ingrid et al. GASA4, one of the 14-member Arabidopsis GASA family of small polypeptides, regulates flowering and seed development. **Plant and Cell Physiology**, v. 48, n. 3, p. 471-483, 2007.

SANGER, F. et al. Nucleotide sequence of bacteriophage λ DNA. **Journal of molecular biology**, v. 162, n. 4, p. 729-773, 1982.

SANGER, F.,et al. Nucleotide sequence of bacteriophage λ DNA. **Journal of molecular biology**, 162.4: 729-773, 1982.

SANGER, Frederick, et al. The nucleotide sequence of bacteriophage φX174. ***Journal of molecular biology***, 125.2: 225-246, 1978.

SANGER, Frederick; NICKLEN, Steven; COULSON, Alan R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the national academy of sciences**, v. 74, n. 12, p. 5463-5467, 1977.

SANTOS, Michelle da Fonseca. **Mapeamento de QTL e expressão gênica associados à resistência da soja ao complexo de percevejos**. Tese de Doutorado. Universidade de São Paulo.212

SCHMUTZ, Jeremy et al. Genome sequence of the palaeopolyploid soybean. **nature**, v. 463, n. 7278, p. 178, 2010.

SHEN, Y.,Du, H.,Liu, Y. et al. Update soybean Zhonghuang 13 genome to a golden reference **Sci. China Life Sci**. 62: 1257, 2019.

SILVA, Ana Paula Mendes. **Expressão gênica associada à resistência da soja a Piezodorus guildinii**. Dissertação de mestrado. Escola Superior de Agricultura "Luiz de Queiroz.2014.

SILVA, J. P. G. F. et al. Characterization of antibiosis to the redbanded stink bug Piezodorus guildinii (Hemiptera: Pentatomidae) in soybean entries. **Journal of Pest Science**, v. 86, n. 4, p. 649–657, 2013.

SIMMONDS, M.S. Flavonoid-insect interactions: Recent advances in our knowledge. ***Phytochemistry*** , *64*, 21–30. 2003.

SIMMONDS, M.S.; STEVENSON, P.C. Effects of isoflavonoids from Cicer on larvae of Heliocoverpa armigera. *J. **Chem. Ecol**. , 27*, 965–977, 2001

SOSA-GÓMEZ, D.R.; MOSCARDI, F. **Retenção foliar diferencial em soja provocada por percevejos (Heteroptera: Pentatomidae**. Anais da Sociedade Entomológica do Brasil. **Anais**...Londrina: 1995

SOUZA, E. DE S. et al. Feeding preference of Nezara viridula (Hemiptera: Pentatomidae) and attractiveness of soybean genotypes. **Chilean journal of agricultural research**, v. 73, n. 4, p. 351–357, 2013.

SOYBASE. Disponível em: https://soybase.org/ >. Acesso em: 8 de fevereiro de 2018.

TANG, Haibao et al. ALLMAPS: robust scaffold ordering based on multiple maps. **Genome biology**, v. 16, n. 1, p. 3, 2015.

TAO, Yongfu et al. Exploring and exploiting pan-genomics for crop improvement. **Molecular plant**, v. 12, n. 2, p. 156-169, 2019.

TU, Xiongbing; LIU, Zhongkuan; ZHANG, Zehua. Comparative transcriptomic analysis of resistant and susceptible alfalfa cultivars (Medicago sativa L.) after thrips infestation. **BMC genomics**, v. 19, n. 1, p. 116, 2018.

ULLAH, Ihteram et al. Genome-wide identification and evolutionary analysis of TGA transcription factors in soybean. **Scientific reports**, v. 9, n. 1, p. 1-14, 2019.

VALLIYODAN, Babu et al. Landscape of genomic diversity and trait discovery in soybean. **Scientific reports**, v. 6, p. 23598, 2016.

VIVAN, L.M.; DEGRANDE, P.E. Pragas da soja. In: Fundação MT. **Boletim de Pesquisa de Soja**, v.11, p.239-297, 2011.

WEISENFELD, Neil I. et al. Direct determination of diploid genome sequences. **Genome research**, v. 27, n. 5, p. 757-767, 2017.

WENDEL, Jonathan F. et al. Evolution of plant genome architecture. **Genome biology**, v. 17, n. 1, p. 37, 2016.

WHITE C., EIGENBRODE S. D. Effects of surface wax variation in *Pisum sativum* on herbivorous and entomophagous insects in the field. **Environ. Entomol**. 29 773–780, 2000.

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

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

World Health Organization, United Nations Environment Programme. **Public Health Impact of Pesticides Used in Agricultu**re. Geneva: The World Health Organization; 1990.

WUYTS, N.; SWENNEN, R.; DE WAELE, D. Effects of plant phenylpropanoid pathway products and selected terpenoids and alkaloids on the behaviour of the plant-parasitic nematodes Radopholus similis, Pratylenchus penetrans and Meloidogyne incognita. ***Nematology*** *, 8,* 89–101,2006

XU, J. *et al.* Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC* ***Plant Biol****.* **14**, 83 (2014).

XU, Xiangyang; BAI, Guihua. Whole-genome resequencing: changing the paradigms of SNP detection, molecular mapping and gene discovery. **Molecular breeding**, v. 35, n. 1, p. 33, 2015.

YAMAMOTO, Naoki et al. Comparative whole genome re-sequencing analysis in upland New Rice for Africa: insights into the breeding history and respective genome compositions. **Rice**, v. 11, n. 1, p. 33, 2018.

YAN, Qiang et al. The soybean cinnamate 4-hydroxylase gene GmC4H1 contributes positively to plant defense via increasing lignin content. **Plant Growth Regulation**, v. 88, n. 2, p. 139-149, 2019.

YANDELL, Mark; ENCE, Daniel. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p. 329, 2012.

ZHANG, ShengChun; WANG, XiaoJing. Expression pattern of GASA, downstream genes of DELLA, in Arabidopsis. **Chinese Science Bulletin**, v. 53, n. 24, p. 3839-3846, 2008.

ANEXOS

Tabela suplementar1: Posição aproximada de variantes estruturais

Posição aproximada das variantes estruturais próximas de genes candidatos	Cromossomo
17.262 kb	06
17.787 kb	06
17.797 kb	06
17.803 kb	06
34.388 kb	20
34.481 kb	20
35.210 kb	20
35.442 kb	20
35.554 kb	20
35.596 kb	20
35.599 kb	20

Figura suplementar 1: Exemplo de variante estrutural única na IAC-100, posição 35.210 kb, cromossomo 20

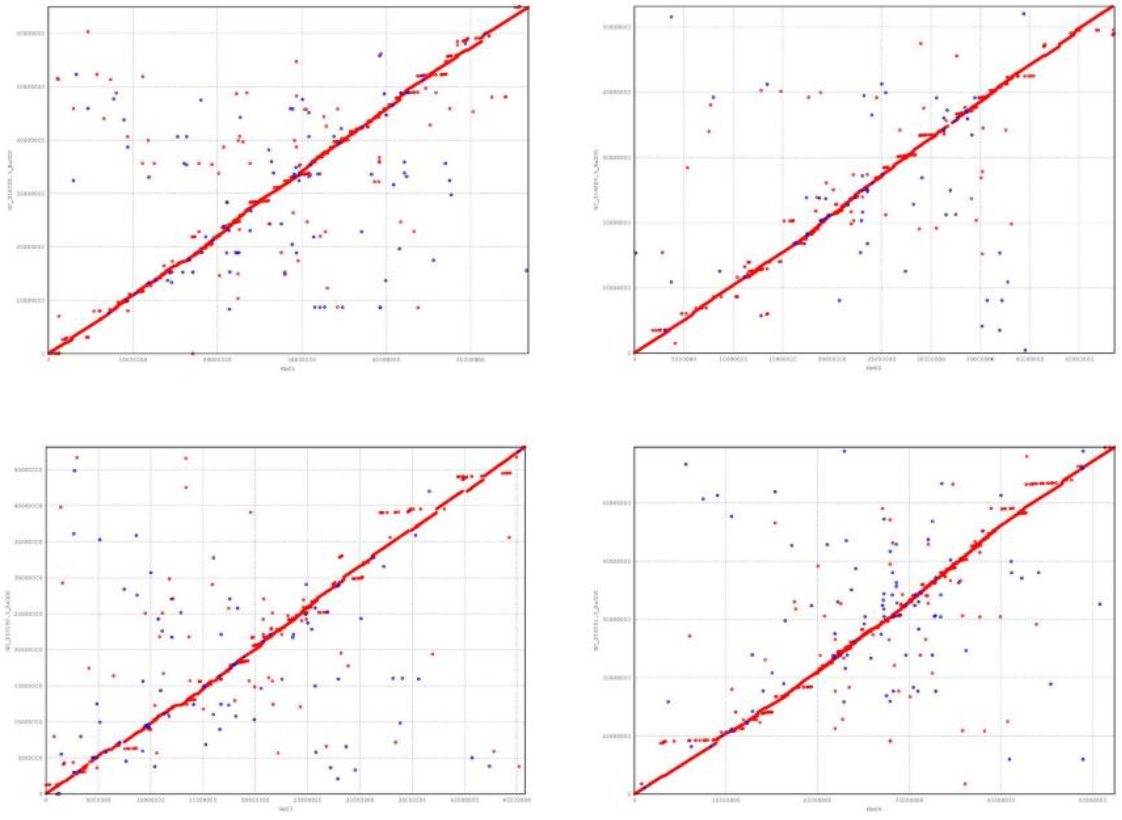


Figura Suplementar 2 : Plot do alinhamento entre os cromossomos 1 a 4 entre IAC-100 e Wiliams 82

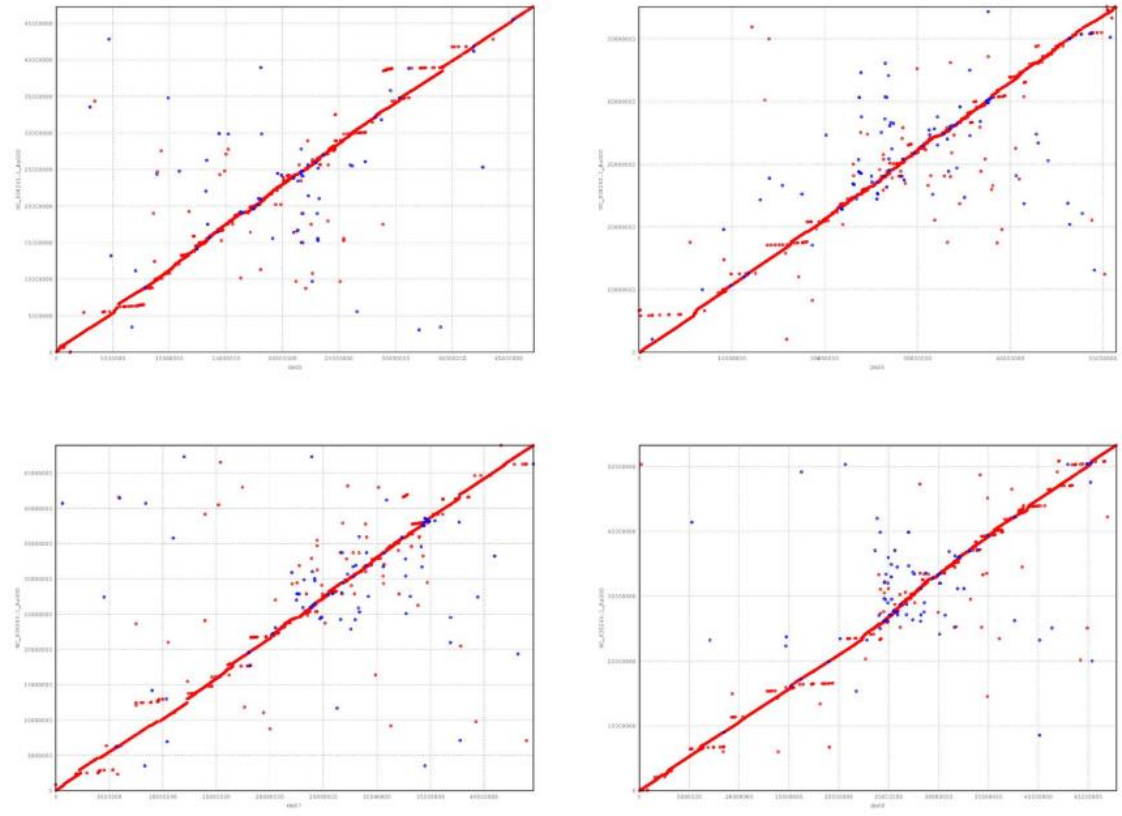


Figura Suplementar 3 : Plot do alinhamento entre os cromossomos 5 a 8 entre IAC-100 e Wiliams 82

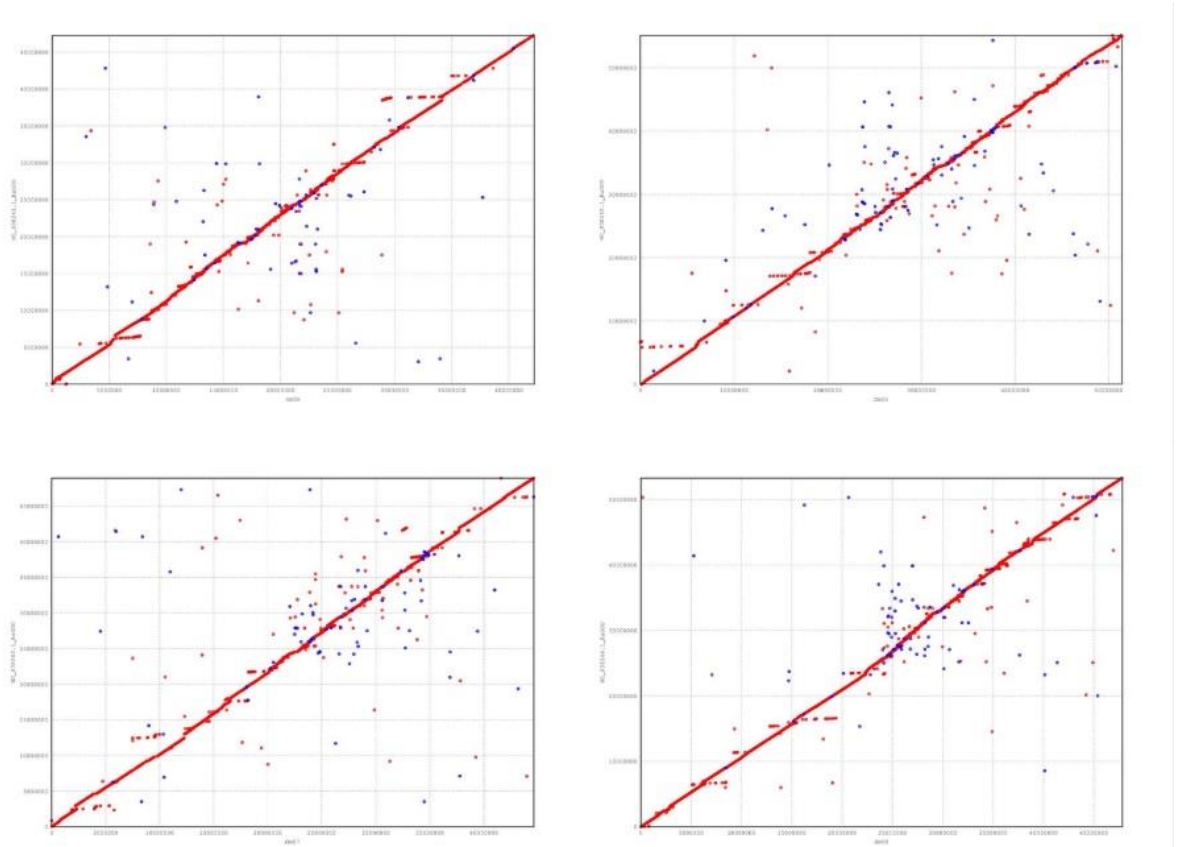
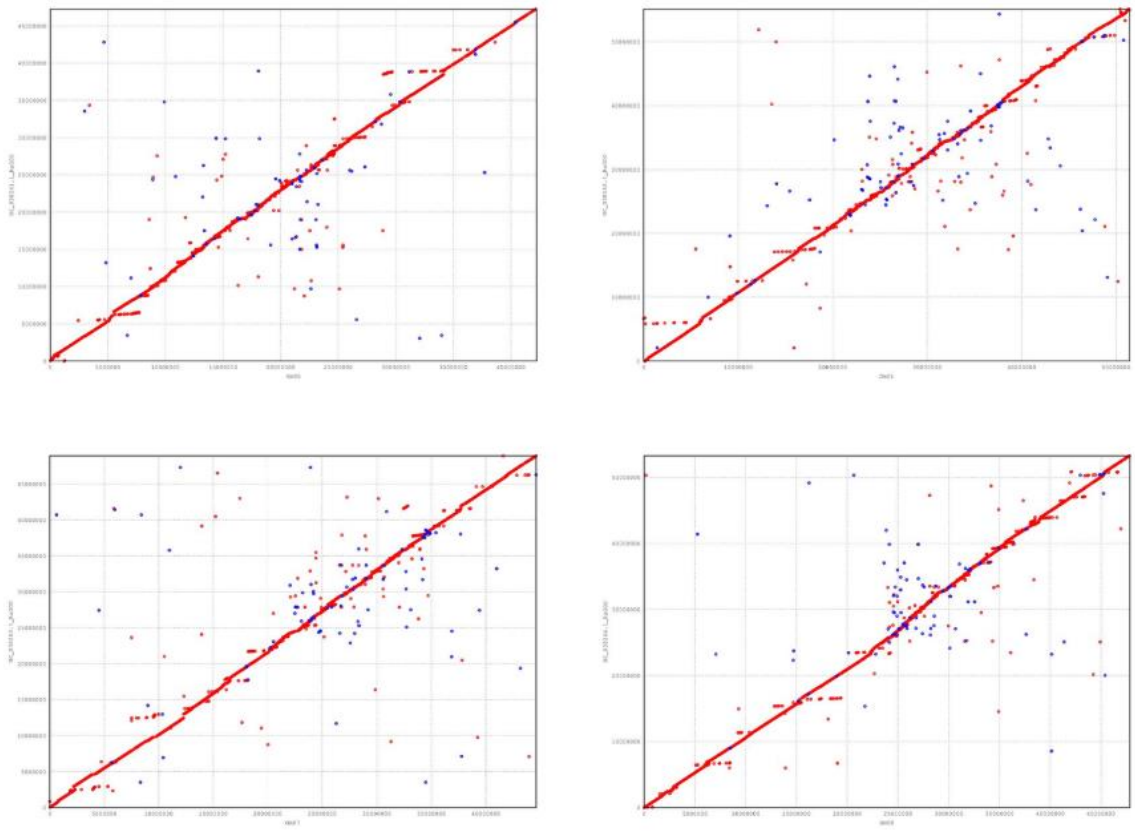


Figura Suplementar 4 : Plot do alinhamento entre os cromossomos 9 a 12 entre IAC-100 e Wiliams 82



,Figura Suplementar 5 : Plot do alinhamento entre os cromossomos 13 a 17 entre IAC-100 e Wiliams 82

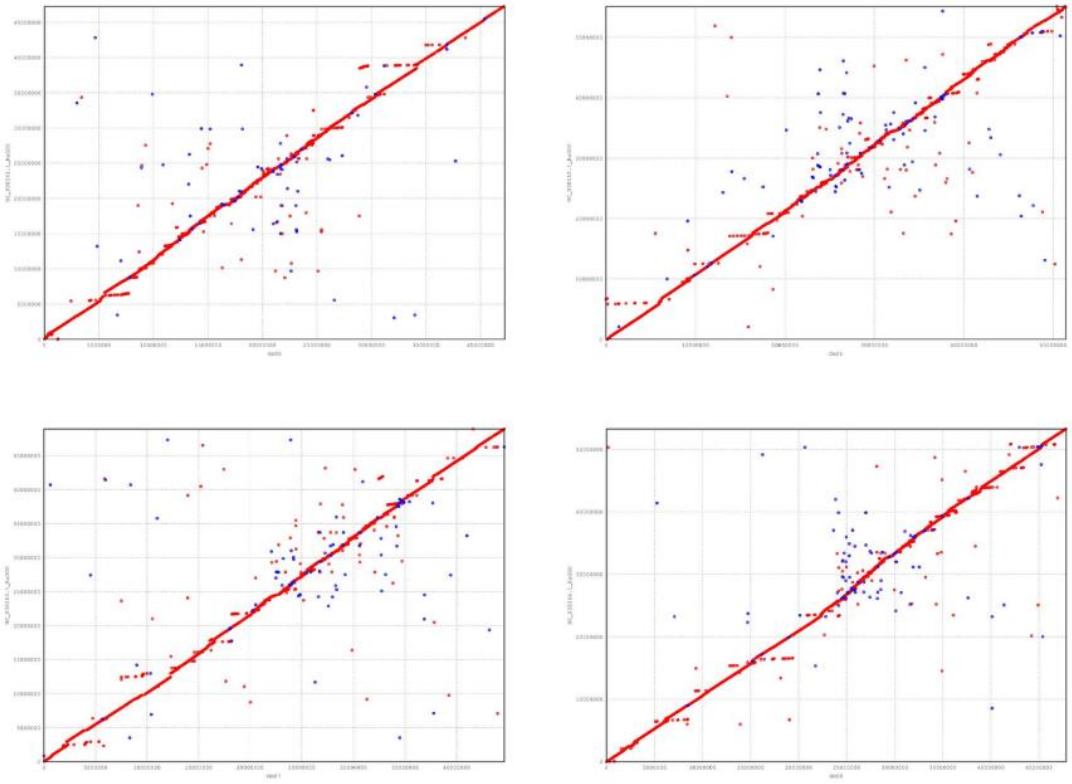


Figura Suplementar 6 : Plot do alinhamento entre os cromossomos 17 a 20 entre IAC-100 e Wiliams

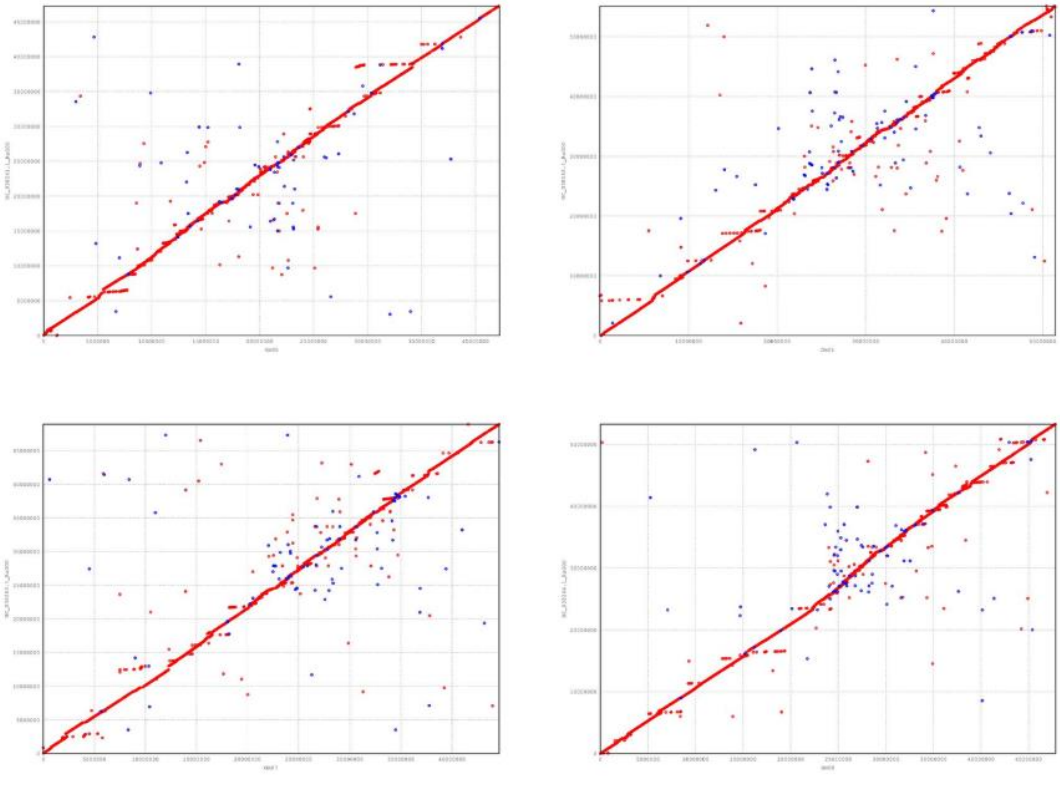


Figura Suplementar 7: Plot do alinhamento entre os cromossomos 1 a 4 entre CD-215 e Wiliams 82

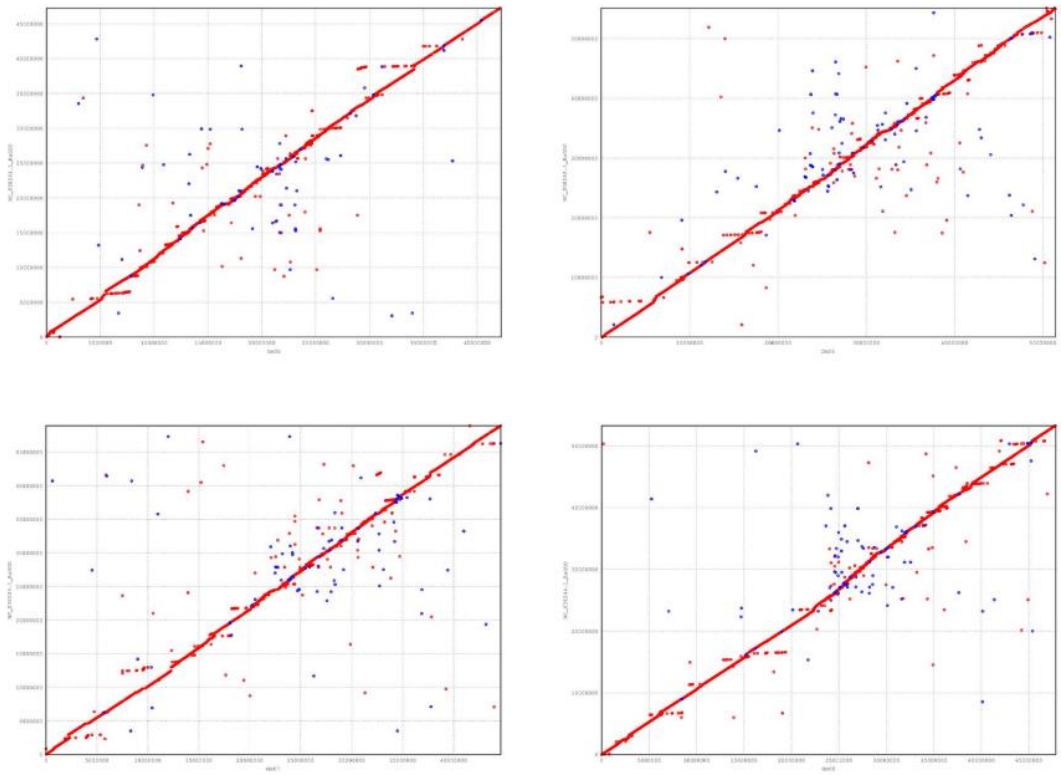


Figura Suplementar 8: Plot do alinhamento entre os cromossomos 5 a 8 entre CD-215 e Wiliamss82

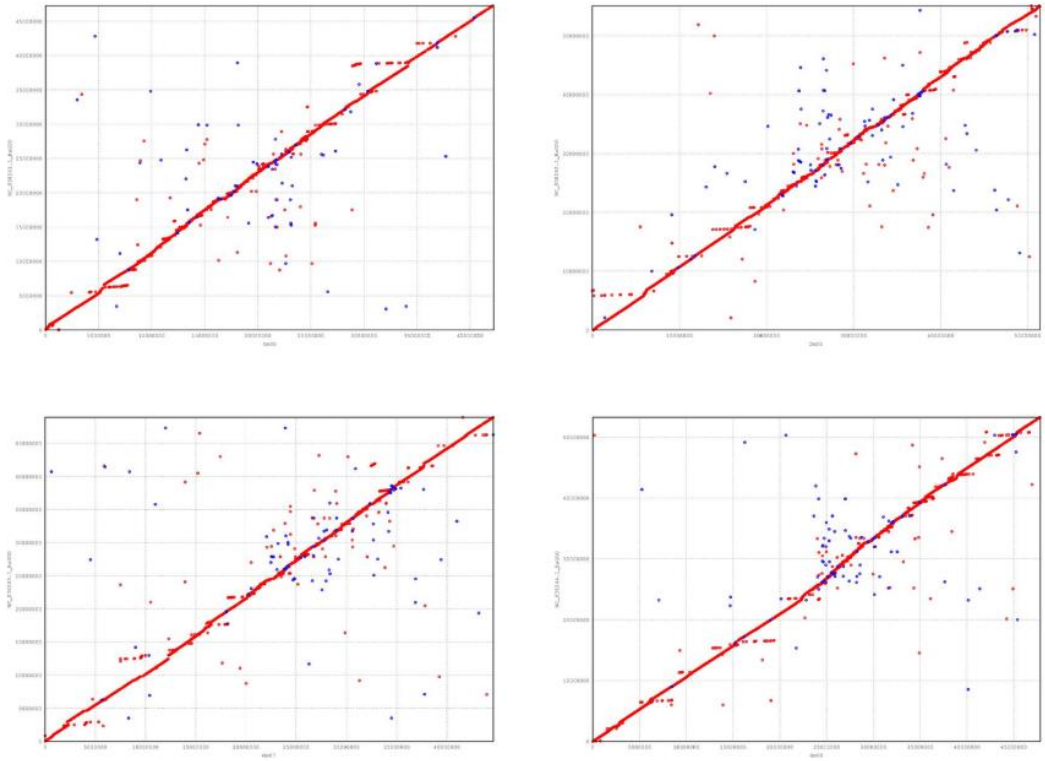


Figura Suplementar 9: Plot do alinhamento entre os cromossomos 9 a 12 entre CD-215 e Wiliamss 82

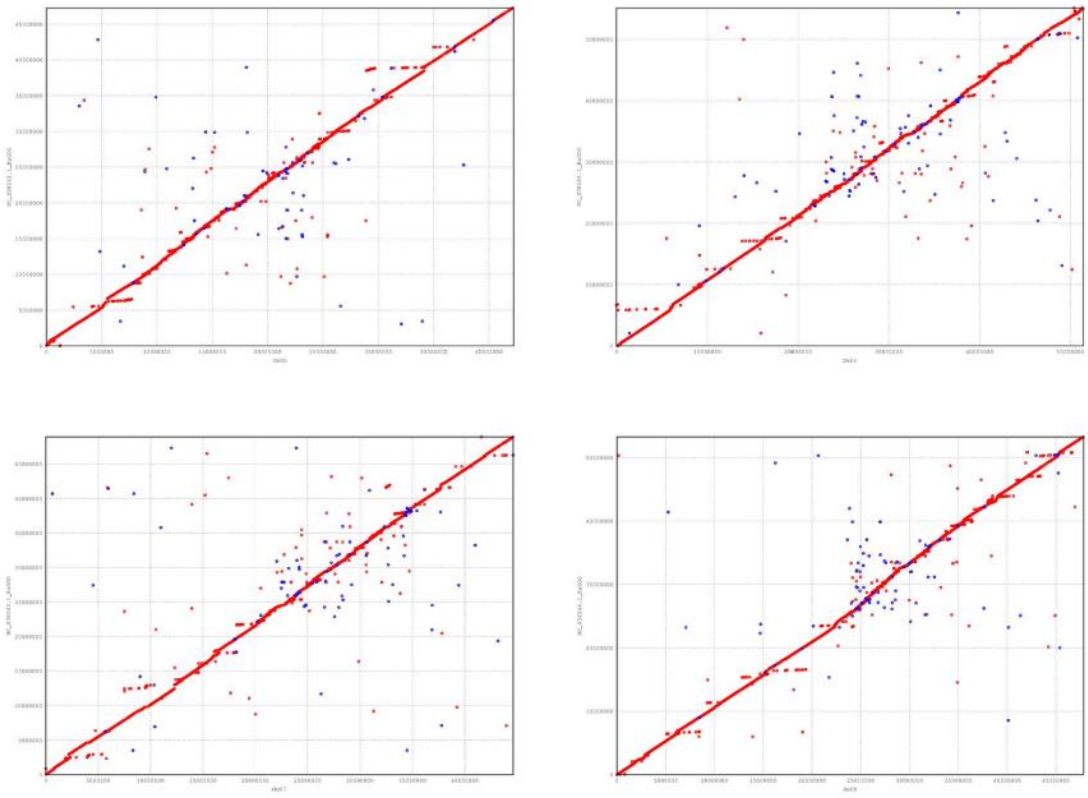


Figura Suplementar 10: Plot do alinhamento entre os cromossomos 13 a 16 entre CD-215 e Wiliams82

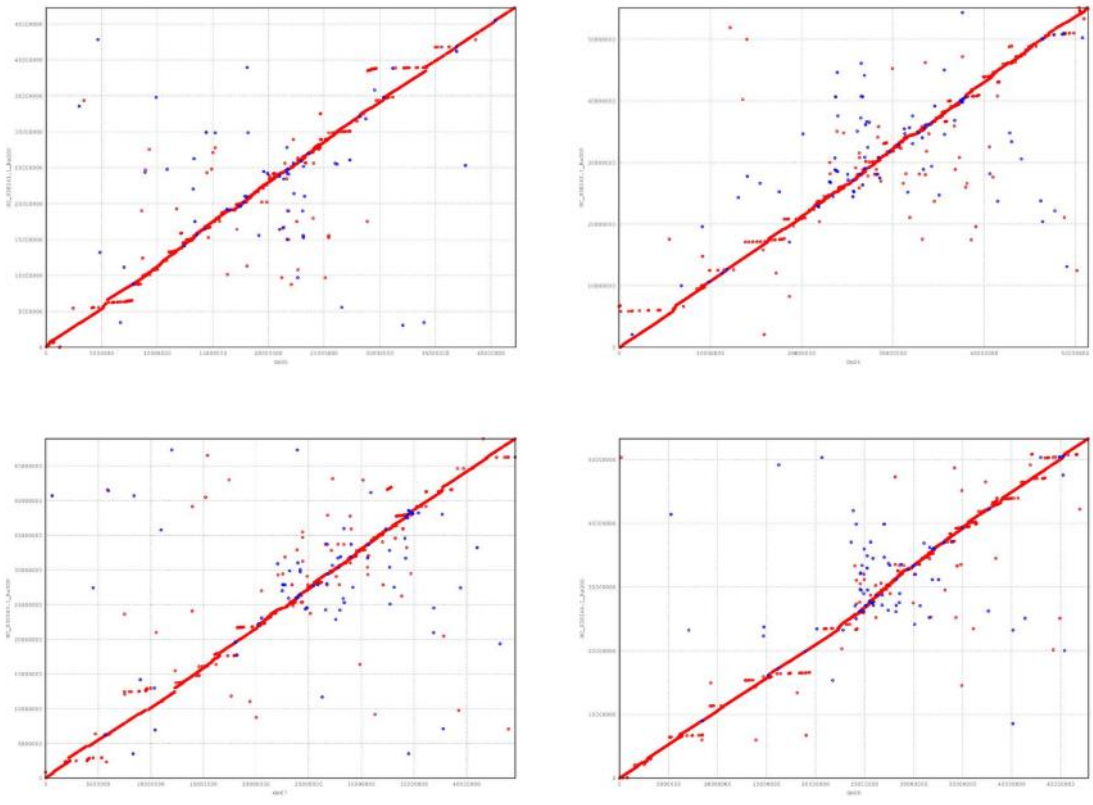


Figura Suplementar 11: Plot do alinhamento entre os cromossomos 17 a 20 entre CD-215 e Wiliams 82