UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS AGRÁRIAS E ENGENHARIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO

MIQUÉIAS FERNANDES

GuavaDB: O BANCO DE DADOS DA GENÔMICA DE *Psidium guajava* L.

ALEGRE – ES

2020

MIQUÉIAS FERNANDES

GuavaDB: O BANCO DE DADOS DA GENÔMICA DE *Psidium guajava* L.

Dissertação apresentada ao Programa de Pós-graduação em Genética e Melhoramento do Centro de Ciências Agrárias e Engenharias da Universidade Federal do Espírito Santo.

Orientador: Prof. Dr. Adésio Ferreira.

Coorientadora: Profª. Drª. Marcia Flores da Silva Ferreira.

ALEGRE – ES

2020

MIQUÉIAS FERNANDES

GuavaDB: O BANCO DE DADOS DA GENÔMICA DE *Psidium guajava* L.

Dissertação apresentada à Universidade Federal do Espírito Santo como requisito do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de Mestre em Genética e Melhoramento.

Aprovado em 10 de fevereiro de 2020.

COMISSÃO EXAMINADORA

Prof. Dr. Adésio Ferreira
Universidade Federal do Espírito Santo
Orientador

Profª. Drª. Marcia Flores da Silva Ferreira
Universidade Federal do Espírito Santo
Coorientadora

Prof. Dr. Tiago Antônio de Oliveira Mendes
Universidade Federal de Viçosa

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Fernandes, Miquéias, 1993-

F363g GuavaDB: o banco de dados da genômica de Psidium guajava L. / Miquéias Fernandes. - 2020.

62 f. : il.

Orientador: Adésio Ferreira.

Coorientadora: Marcia Flores da Silva Ferreira.

Dissertação (Mestrado em Genética e Melhoramentos) Universidade Federal do Espírito Santo, Centro de Ciências Agrárias e Engenharias.

1. goiaba. 2. banco de dados. 3. genética. 4. genômica. I. Ferreira, Adésio. II. Flores da Silva Ferreira, Marcia. III. Universidade Federal do Espírito Santo. Centro de Ciências Agrárias e Engenharias. IV. Título.

1. CDU: 631.523

---

as Melhores Conquistas sucedem os Maiores Desafios

AGRADECIMENTOS

Agradeço a todos que contribuíram para o êxito desse trabalho, entre eles:

À *Deus* o supremo regente do universo por orquestrar situações a todo momento para me manter firme;

À minha família que é responsável indiretamente por tudo que fiz neste trabalho;

Ao meu lar Alda-Isaac-Ruth por me acompanhar e compreender fornecendo tudo que precisei como tempo e companheirismo, vocês são TUDOOOOO para mim!

À FAPES especialmente, mas também a CAPES e CNPq por financiar todo esse trabalho colaborando para o crescimento intelectual, acadêmico e científico;

À UFES pelo PPGGMV que me proporcionou o ambiente para desenvolver a pesquisa;

Aos meus orientadores Adésio e Marcia por me nortear ao sucesso em ciências da vida;

Ao professor Tiago por compartilhar seus conhecimentos de bioinformática em disciplinas e helps;

Ao Galaxy-EU por fornecer infraestrutura para rodar a anotação funcional.

BIOGRAFIA

Miquéias Fernandes é filho de Luciene das Graças Fernandes e Antônio Fernandes, nascido em 01 de julho de 1993 em Cachoeiro de Itapemirim/ES. Interessou-se por tecnologia desde que abriu um rádio aos 5 anos, mas também por teologia.

Inspirado no seu principal interesse, em especial motores a vapor e de combustão interna desenvolveu seu próprio aos 15 anos. Com efeito estudou mecânica automotiva no SENAI em 2010 e por seu destaque em elétrica atuou na profissão de eletricista por 2 anos. Sob sugestão do professor Idálio realizou 3 cursos de informática na instituição, indo de eletricista para Técnico em informática - com o curso que realizou integrado ao 3º ano - que o habilitou desenvolver vários *websites* em trabalho autônomo. Em 2012, ingressou no curso de Sistemas de Informação na Universidade Federal do Espírito Santo determinado a trabalhar com desenvolvimento de sistemas *web*. Com o casamento com a bióloga Alda em 2013 decidiu entrar em uma área comum a ambos: bioinformática, de modo que em 2018 entrou para o mestrado na área de ciências da vida no Programa de Pós-graduação em Genética e Melhoramento Vegetal. Sob orientação do professor Adésio Ferreira e coorientação da Professora Marcia Flores realizou vários trabalhos de bioinformática com goiabeiras, entre eles o pioneiro genoma e transcriptoma da espécie que foi disponibilizado no sistema *web* GuavaDB em guava.ufes.br, objeto desta defesa em 2020.

FERNANDES, Miquéias. GuavaDB: O BANCO DE DADOS DA GENÔMICA DE *Psidium guajava* L. 2020. Dissertação (Mestre em Genética e Melhoramento). Universidade Federal do Espírito Santo. Orientador: Dr. Adésio Ferreira Coorientadora: Drª. Marcia Flores da Silva Ferreira.

RESUMO

A goiaba (*Psidium guajava* L.) está entre as frutas preferidas no Brasil para consumo *in natura*, sendo que ela ainda ocupa lugar de destaque nas indústrias de produtos alimentícios tendo em vista a gama de produtos oriundos de seu processamento. A fruta é apreciada não só pelo seu aroma e sabor, mas também pelo seu valor nutricional, uma vez que possui um dos mais altos teores de vitamina C. Neste trabalho o rascunho do genoma da espécie foi montado com uma abordagem hibrida que resultou em 2.097 contigs com N50 587.224. Depois 68% dos contigs foram ancorados nos 11 grupos de ligação do mapa genético, permitindo obter 686 scaffolds com N50 32.5Mb. A montagem de tamanho 394,5 Mpb corresponde a 85% do genoma de acordo com validação pelo BUSCO. Sequenciamentos do transcriptoma da espécie depositado no SRA foram utilizados para montagem *de novo* do transcriptoma. O transcriptoma montado foi utilizado pelo pipeline Seqping para identificação de 60.538 genes codificadores de proteínas. Além dos genes 3.557 elementos transponíveis e 14.188.576 segmentos duplicados forma identificados na anotação estrutural. A anotação funcional foi realizada em três abordagens de modo que 79% dos genes foram anotados sendo 39% deles com GOs e 42% com ortologos do EggNOG. Dados da montagem do genoma, da anotação estrutural e da anotação funcional foram persistidos no banco de dados GuavaDB disponível em guava.ufes.br. Várias funcionalidades como ferramentas de busca, página de downloads e gestão de usuários foram agregadas ao GuavaDB afim de facilitar a gerência e exploração dos dados armazenados. Nele a comunidade científica tem agora um poderoso recurso de referência para desenvolver pesquisas que carecem de informações genômicas da espécie.

Palavras-chave: goiaba; banco de dados; genética; genômica;

FERNANDES, Miquéias. GuavaDB: THE *Psidium guajava* L GENOMIC DATA BASE. 2020. Dissertação (Mestre em Genética e Melhoramento). Universidade Federal do Espírito Santo. Orientador: Dr. Adésio Ferreira Coorientadora: Drª. Marcia Flores da Silva Ferreira.

ABSTRACT

Guava (*Psidium guajava* L.) is one of the preferred fruits in Brazil for fresh consumption and occupies a prominent place in the food industries, considering the range of products from its processing. The fruit is appreciated not only for its aroma and flavor, but also for its nutritional value, since it has one of the highest levels of vitamin C. In this work the draft of the genome of the species was assembled with a hybrid approach that resulted in 2,097 contigs with N50 587,224, with a total length of 394.5 Mpb, that corresponds to 85% of the genome estimative according to BUSCO validation. We also anchored 68% of the contigs in the 11 linkage groups of the genetic map, allowing to obtain 686 scaffolds with N50 32.5Mb. Sequences of the RNA-Seq deposited in the SRA database were used to assembly the transcriptome of *P. guajava* using the *de novo* assembly strategy. The assembled transcriptome was used by the Seqping pipeline, allowing the identification of 60.538 protein coding genes. In the structural annotation, 3.557 transposable elements and 14.188.576 duplicated segments were identified. Functional annotation was performed in three approaches, where 79% of the genes were annotated, 39% of them with GOs and 42% with EggNOG orthologists. Genome assembly, structural annotation and functional annotation data were persisted in the GuavaDB database available at guava.ufes.br. Various features such as search tools, downloads page and user management have been added to GuavaDB in order to facilitate the management and exploitation of stored data. Now the scientific community has a powerful reference resource for developing research that lacks genomic information about the species.

Keywords: guava; banco de dados; genetics; genomics;

LISTA DE FIGURAS

LISTA DE TABELAS

ÍNDICE

2. INTRODUÇÃO

A goiaba (*Psidium guajava* L.) é uma fruta nativa dos trópicos americanos que pode ser encontrada em áreas tropicais e subtropicais do mundo (DECHEN; POMMER, 2006). A família *Myrtaceae* da qual a goiaba pertence possui mais de 500 espécies, contudo a fruta destaca-se entre outras das frutíferas da família por possuir alto valor nutritivo e industrial. De seu valor nutritivo pode- se dizer que possui alto conteúdo de vitaminas A e complexo B, sendo excepcionalmente rica em vitamina C (ácido ascórbico), com teor superior ao presente noutros sucos cítricos (NIMISHA et al., 2013).

Devido à variedade e alta concentração de óleos essenciais a goiaba tem aplicações medicinais e industriais; de modo que fármacos antioxidantes, cicatrizantes, antialérgicos, antidiabéticos, anticancerígenos e anti-inflamatórios podem ser produzidos a partir da fruta (JIMÉNEZ-ESCRIG et al., 2001; RAVI; DIVYASHREE, 2014). Quanto aos derivados industriais da fruta a goiabada, por exemplo, está entre os doces mais apreciados pelos brasileiros (SEBRAE, 2016). Ela ocupa lugar de destaque entre frutas mais exportadas do Brasil com 142 toneladas exportadas das 414.960 toneladas produzidas no ano de 2017, enquanto no mundo estima-se uma produção de cerca de 1,2 milhões de toneladas com o Brasil entre os principais produtores (KIST et al., 2018).

A espécie é diploide (2n = 22) e possui o conteúdo de DNA estimado de 465 Mb por citometria de fluxo (MARQUES et al., 2016). No *Genbank* (SAYERS et al., 2019) há um genoma com 386 Mb da espécie depositado (dados não publicados) e o cloroplasto de 158.841 bases caracterizado estruturalmente (JO et al., 2016). Quanto a transcriptoma existem sequenciamentos de um estudo (acesso pelo *BioProject*: PRJNA472130) não publicado do transcriptoma da espécie depositado no SRA (LEINONEN; SUGAWARA; SHUMWAY, 2011). Além destas informações CANAL (2019) proveu o primeiro estudo a respeito da genômica da espécie publicado. Mesmo assim a informação disponível é superficial, de modo que estudos genéticos com a espécie carecem de uma análise mais profunda, com validação de predições e realização de comparação estrutural e funcional com outras espécies próximas filogeneticamente com genoma disponível no banco de dados *NCBI* (AGARWALA et al., 2018).

Recentes avanços nas tecnologias de sequenciamento têm reduzido o custo de sequenciamento permitindo pesquisas genéticas em organismos não modelo, como é o caso deste trabalho com a goiabeira. Por outro lado, estas tecnologias tem alterado a maneira de como as pesquisas genéticas são realizadas (GOODWIN; MCPHERSON; MCCOMBIE, 2016) positivamente ao permitir pesquisas genéticas mais profundas e acuradas no genoma do organismo.

Para programas de melhoramento, conhecer a genômica da espécie é ganho substancial, desde que a genômica cria uma ponte entre o fenótipo e o genótipo de modo que programas modernos de melhoramento combinam várias abordagens de genômica para um melhoramento eficiente e redução do período necessário para realizar o melhoramento (KYRIAKIDOU et al., 2018). Assim, o sucesso deste trabalho abre oportunidades para conhecer a diversidade genética de acordo com padrões da agricultura moderna e consequentemente para o programa de melhoramento da cultura: i) analise de diversidade genética com reseqeunciamentos do genoma; ii) estudos de associação genômica (GWAS) para analises de QTL; iii) seleção genômica assistida por marcadores; e iv) modificações direcionadas no genoma com tecnologias de edição genica, resultando em introdução de variantes alélicas nas cultivares de goiabeiras (BARABASCHI et al., 2015; RASHEED et al., 2017).

Mas, o volume de informação relevante apresentado neste trabalho não tem valor disruptivo se disponibilizada desorganizada para a comunidade cientifica. Neste sentido a construção de bancos de dados biológicos tem se tornado comum entre os bioinformatas para persistência de dados e difusão do conhecimento (CHEN; HUANG; WU, 2017). Desse modo, com relação a espécie em estudo, além deste trabalho permitir estudos genéticos mais profundos e abrir novas possibilidades para programas de melhoramento, a implementação de um banco de dados público para persistência das informações geradas nas análises deste trabalho faz deste uma referência para a espécie na área da genômica sem precedentes.

3. REVISÃO BIBLIOGRÁFICA

O tamanho do genoma de plantas varia dramaticamente - de 63 Mbp a 148,8 Gbp. Como existem vários desafios com relação a, por exemplo, sequencias duplicadas, elementos repetitivos, ploidia e como estes desafios são proporcionais ao tamanho do genoma poucos genomas maiores que 10 Gbp foram montados (LI, F.-W.; HARKESS, 2018).

Mesmo entre os organismos que possuem tamanho genômico menor que 10 Gbp o número de genomas montados não é grande, devido ao fato de para ser considerado “montado” um genoma ter que estar em nível cromossômico sendo este outro desafio em projetos que não utilizem tecnologias de sequenciamento de terceira geração combinadas com mapas óticos por exemplo (BELSER et al., 2018). Com isso vários genomas de plantas são publicados com milhares de sequencias denominadas *scaffolds,* que são sequencias cromossômicas fragmentadas. Por outro lado, avanços nas tecnologias de sequenciamento aliado a ferramentas de bioinformática tem permitido obter sequencias de pseudomoléculas em escala cromossômica, estas por sua vez são os *scaffolds* ancorados ao seu devido cromossomo com orientação refinada por mapas óticos (MUKHERJEE et al., 2018), físicos e genéticos (YOU et al., 2018).

Neste sentido, o desenvolvimento de sequenciamentos de *high-throughput* revolucionou a genética e a genômica com baixos custos, permitindo incremento de projetos de sequenciamento de genoma completo (WGS) de várias espécies diferentes. Hoje diferentes tecnologias de sequenciamento de segunda e terceira geração e algoritmos para montagem de genoma que podem ser utilizadas ou combinadas para obter uma montagem de genoma de alta qualidade (PAAJANEN et al., 2019). Nesse trabalho, por exemplo, foi realizada uma abordagem hibrida de sequenciamento de segunda e terceira geração para montar o genoma da goiabeira.

Enquanto o sequenciamento de nova geração (NGS) estabeleceu-se como principal tecnologia de analise biológica, *workflows* e *pipelines* de análises destes dados exigem expertise dos bioinformatas para seleção de ferramentas apropriadas em consideração a: uso de paralelização, solução de armazenamento de dados e desenvolvimento de estratégias customizadas e automatizadas para

máxima exploração dos resultados que envolvem múltiplas condições experimentais (KULKARNI; FROMMOLT, 2017). Assim, referencias como LANTZ et al. (2018) que descreve as principais etapas para um bom projeto de montagem e anotação de genomas são referências indispensáveis para o desenvolvimento de um projeto de genômica como este. Para LANTZ et Al. (2018) os bioinformatas e outros envolvidos no projeto devem levar em consideração os dez tópicos a seguir:

1. Investigar propriedades do genoma em estudo, entre elas:
    1.1. tamanho do genoma;
    1.2. regiões repetitivas;
    1.3. heterozigozidade;
    1.4. ploidia;
    1.5. conteúdo GC.
2. Extrair amostras de DNA com qualidade, considerando:
    2.1. a modalidade *de novo;*
    2.2. pureza química;
    2.3. integridade estrutural do DNA.
3. Escolher a tecnologia de sequenciamento ideal, entre:
    3.1. sequenciamento de primeira geração (FGS);
    3.2. sequenciamento de segunda geração (SGS);
    3.3. sequenciamento de terceira geração (TGS);
    3.4. abordagens hibridas.
4. Estimar os recursos computacionais necessários.
5. Montar o genoma.
    5.1. montagem de *reads* curtas;
    5.2. montagem de *reads* longas;
    5.3. montagem de *scaffolds* e preenchimento de gaps;
    5.4. realizar controle de qualidade da montagem.
6. Identificar e anotar elementos repetitivos e transponíveis.
7. Anotar genes com evidência experimental.
8. Escolher formatos de arquivos de saída e persistir em bancos de dados.
9. Verificar se os métodos utilizados são repetíveis e reproduzíveis.
10. Investigar, analisar e anotar novamente.

Neste trabalho esta referência será muito utilizada por cobrir toda a parte de obtenção dos dados (genes, anotações, *scaffolds* etc.) que serão persistidos no banco de dados. Contudo, como aqui o genoma está sequenciado para o procedimento de análises de bioinformática apenas o tópico 4 (que será abordado na seção 4.2 Configuração do ambiente de trabalho) e os demais serão discutidos nas páginas seguintes.

2.1. Montagem de genomas

A montagem de um genoma pode ser comparada a um quebra-cabeça, com algumas peculiaridades como: número de peças extraordinariamente grande, peças de tamanho variável e repetidas. Dessa forma, montar um genoma é uma das etapas mais onerosas para os recursos computacionais. Nesse sentido, vários algoritmos, heurísticas e meta-heurísticas tem sido projetados e aperfeiçoados levando em consideração melhoria de desempenho e o avanço nas tecnologias de sequenciamento (SOHN; NAM, 2018).

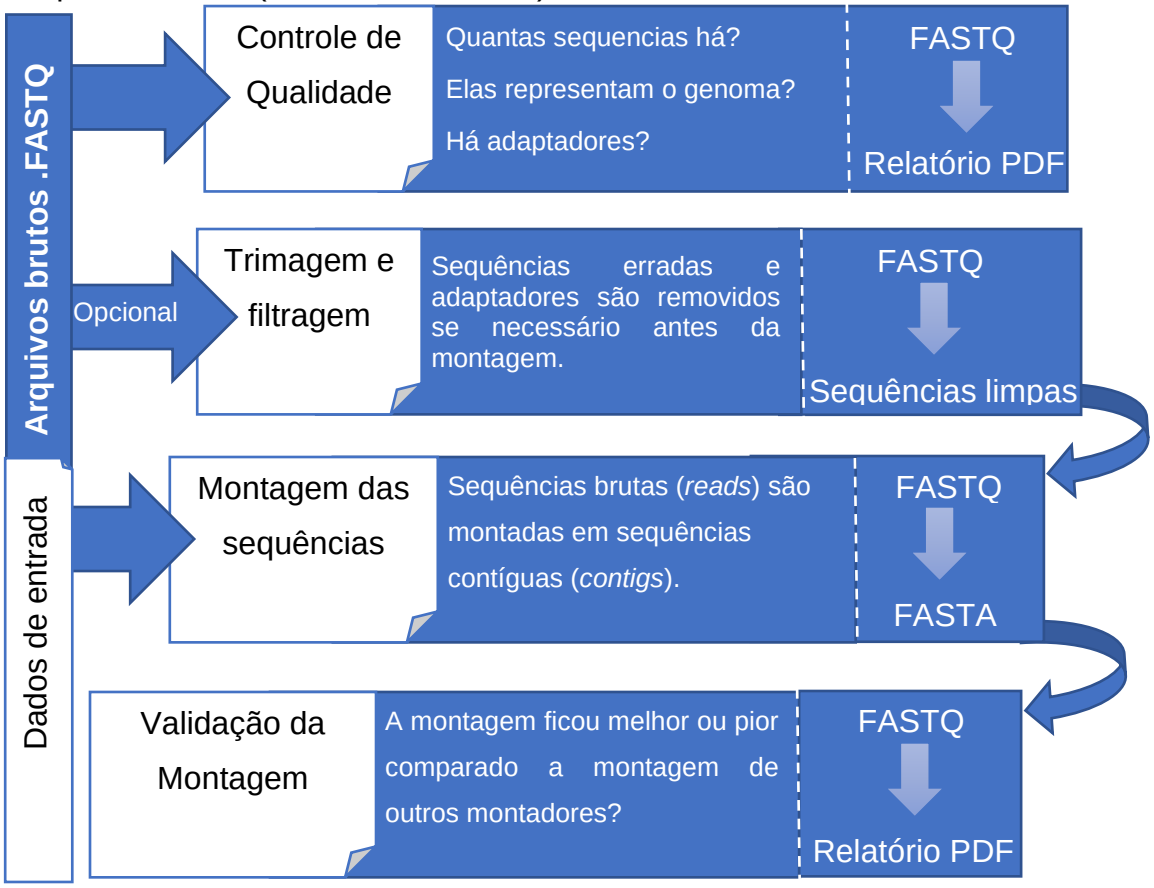

Figura 1 - Etapas da montagem de genomas adaptado de LANTZ et al (2018) que facilita a compreensão da ordem de softwares que devem ser executados para se obter o genoma montado a partir do sequenciamento e os respectivos arquivos gerados por cada um deles em cada etapa.

Apesar do aprimoramento dos algoritmos ter importância significativa na realização da montagem do genoma há outras tarefas indispensáveis antes e após a montagem que devem ser realizadas, conforme apresentado por LANTZ et al. (2018) na **Figura 1**. Nesta figura é possível conferir ainda os formatos de arquivos esperados em cada etapa envolvida na montagem de genomas. Pela **Figura 1** compreende-se que a etapa de montagem deve ser precedida por outras duas que são: controle de qualidade e filtragem, e sucedida por validação da montagem.

Nas etapas precedentes à montagem pode ser utilizado softwares como o *Trimmomatic* (BOLGER, A. M.; LOHSE; USADEL, 2014) ou *NGS QC Toolkit* (PATEL; JAIN, 2012) que são suítes para realização de várias etapas envolvidas no controle de qualidade do sequenciamento, que inclui mas não se limita a: remoção de adaptadores, remoção de reads sem qualidade de sequenciamento, corte de início ou fim de reads que possuem pares de base lidos com baixa qualidade, remoção de contaminações e remoção de reads não pareadas. Nesse momento o bioinformata deve levar em consideração a real necessidade de remoção de reads de baixa qualidade, pois de fato remover pode ser discutível (YANG et al., 2019). Consecutivamente, a qualidade pode ser atestada com auxílio de softwares como *FastQC* ou *FQC Dashboard* (BROWN; PIRRUNG; MCCUE, 2017) para visualização geral das reads que serão submetidas a montagem.

Com as *reads* limpas deve-se proceder a montagem. Nesse momento o bioinformata deve decidir entre vários montadores existentes não só o mais adequado, mas também o mais eficiente levando em consideração o recurso computacional disponível. Há montadores mais rápidos, mas que exigem mais memória RAM por exemplo. Isto ocorre devido as diferenças entre os algoritmos implementados por cada *software*, sendo que os algoritmos mais conhecidos utilizados por programas de montagem são: *Overlap-Layout-Consensus* (OLC), *grafos de Brujin* e caminhos eulerianos; entretanto devido à pouca sensibilidade por erros de sequenciamento este último não é muito utilizado (KYRIAKIDOU et al., 2018). O MASURCA (ZIMIN et al., 2013) é um exemplo de montador que utiliza o algoritmo OLC em combinação com *grafos de Brujin* para obter *scaffolds*.

Em geral, as sequências curtas precisam ser montadas em *contigs*, que então devem ser unidos em scaffolds (*scaffolding*) até alcançar o nível

cromossômico, com o fechamento de lacunas entre os *scaffolds* (*gap filling)* e ancoragem e orientação deles em cromossomos com mapas óticos/físicos/genéticos para obtenção de pseudo-moléculas.

Após a montagem, conforme a **Figura 1** deve-se realizar uma validação da montagem. Nesta etapa deve ser verificado métricas como N50, quantidade de *N*'s, tamanho acumulado por sequência, número e média de tamanho de sequencias. Para isso pode ser utilizado softwares como *Quast* (GUREVICH et al., 2013) que valida comparativamente montagens de genoma e permite além de visualizar, reportar as métricas da montagem em diversos formatos de arquivo e o BUSCO (WATERHOUSE et al., 2018) que valida a completude da montagem quanto a genes que espera-se encontrar em uma montagem. Como o BUSCO depende da predição de genes é interessante executá-lo após o processo apresentado a seguir.

2.2. Anotação de genomas

A anotação de genomas é o processo que identifica a estrutura e função de elementos e regiões ao longo da molécula de DNA. O momento ideal para realização deste processo é apos todas etapas do processo de montagem: *scaffolding*, *gap filling* e ancoragem em pseudo-moléculas. Isto porque na anotação estrutural algumas coordenadas de elementos identificados (como genes por exemplo) dependem de atributos da molécula onde foi identificado.

A primeira etapa da anotação de genomas é a anotação estrutural, onde são identificados os elementos repetitivos desde microssatélites a segmentos duplicados e elementos gênicos não funcionais (NC-RNA: tRNAs, rRNA, miRNAs entre outros) e funcionais (genes codificadores de proteínas). Após a anotação estrutural é possível visualizar os elementos identificados com navegadores de genoma como JBrowse (BUELS et al., 2016).

2.3. Anotação Estrutural

Há dois tipos de regiões no genoma que são: de baixa ou alta complexidade. A complexidade de uma região é dada pelo número de pares de base que compõe um motivo que se repete, se não há motivo a região é dita de alta complexidade. Assim: "AAAA" tem motivo A e complexidade 1 (baixa), já ATATATAT tem motivo AT e complexidade 2 (maior que a primeira). Para identificação desses elementos

repetitivos podem ser utilizados três softwares em paralelo: MISA (BEIER et al., 2017), REPET (FLUTRE et al., 2011) e SEDEF (NUMANAGIĆ et al., 2018) pois os processos são independentes. A identificação de microssatélites com motivos de mono a hexa podem ser realizadas com o software MISA que é de fácil utilização, apenas um script em *Perl* com possibilidade ainda de execução em serviço *web*. Já o pacote REPET é escrito em *Python2* e de difícil instalação e configuração. Por outro lado, o REPET reúne vários softwares especializados em predição de elementos transponíveis permitindo identificar com alta senilidade, agrupar e anotar várias classes de elementos transponíveis. Ele possui vários filtros para remoção de falsos positivos. Os segmentos duplicados são regiões repetitivas com tamanho superior a mil pares de base (NUMANAGIĆ et al., 2018), denotando teoricamente o maior tipo de elemento repetitivo no genoma. Para identificação destas regiões o SEDEF é uma boa opção por ser de fácil instalação, melhor usabilidade e possuir melhor desempenho comparado aos softwares legados.

A predição de genes é uma das mais importantes etapas do processo de anotação de genomas, de modo que vários softwares e pipelines foram desenvolvidos com diferentes abordagens para realização da predição de genes. Nesse contexto o pipeline *Seqping* (CHAN et al., 2017) foi apresentado como uma solução consenso ao unir em si os mais prestigiados preditores para eucariotos, que são *GlimmerHMM* (MAJOROS; PERTEA; SALZBERG, 2004), *SNAP* (KORF, 2004) e *AUGUSTUS* (STANKE; WAACK, 2003) para combinar com auxílio do *MAKER2* (HOLT; YANDELL, 2011) os genes preditos por cada um dos três softwares baseado na evidencia de transcritos de RNA-Seq montado. Apesar dessa estratégia, abordagens *ab initio* dificilmente identificam todas regiões transcritas e não traduzidas (UTR) de todos genes preditos em organismos não modelo. Isso acontece porque o Augustus tem dificuldade em identificar essas regiões para aqueles organismos cujo modelo de predição destas regiões não foi treinado (HOFF; STANKE, 2018), que é o caso da espécie estudada neste trabalho.

Para predizer os genes codificadores de *tRNA* podem ser utilizados os softwares *ARAGORN* (LASLETT; CANBACK, 2004), o *tRNAscan-SE* (LOWE; EDDY, 1997) ou o consenso de ambos. Já para predizer os genes codificadores de *rRNA* pode ser utilizado o software RNAMMER (LAGESEN et al., 2007). Outros

RNAs podem ser preditos com o software INFERNAL (NAWROCKI; EDDY, 2013) contra o sequencias que podem ser obtidas do banco de dados RFAM (SWEENEY et al., 2019) e RNAcentral (SWEENEY et al., 2019).

2.4. Anotação Funcional

Após a anotação estrutural pode-se proceder a anotação funcional para identificação da função biológica dos genes codificadores de proteínas (BOLGER, M. E.; ARSOVA; USADEL, 2018a). Nesta etapa as proteínas são analisadas alinhando-as contra vários bancos de dados para identificação domínios, famílias e relações evolutivas (ortólogos e grupos de ortólogos) entre elas conforme apresentado na **Figura 2**. Para isso podem ser utilizados os softwares Diamond (BUCHFINK; XIE; HUSON, 2014), InterproScan5 (JONES et al., 2014) e EggNOG-mapper (HUERTA-CEPAS et al., 2017) contra bancos de dados como NR RefSeq (O'LEARY et al., 2016), Uniprot SWISS/TREMBL (BATEMAN, 2019), EGNOGG

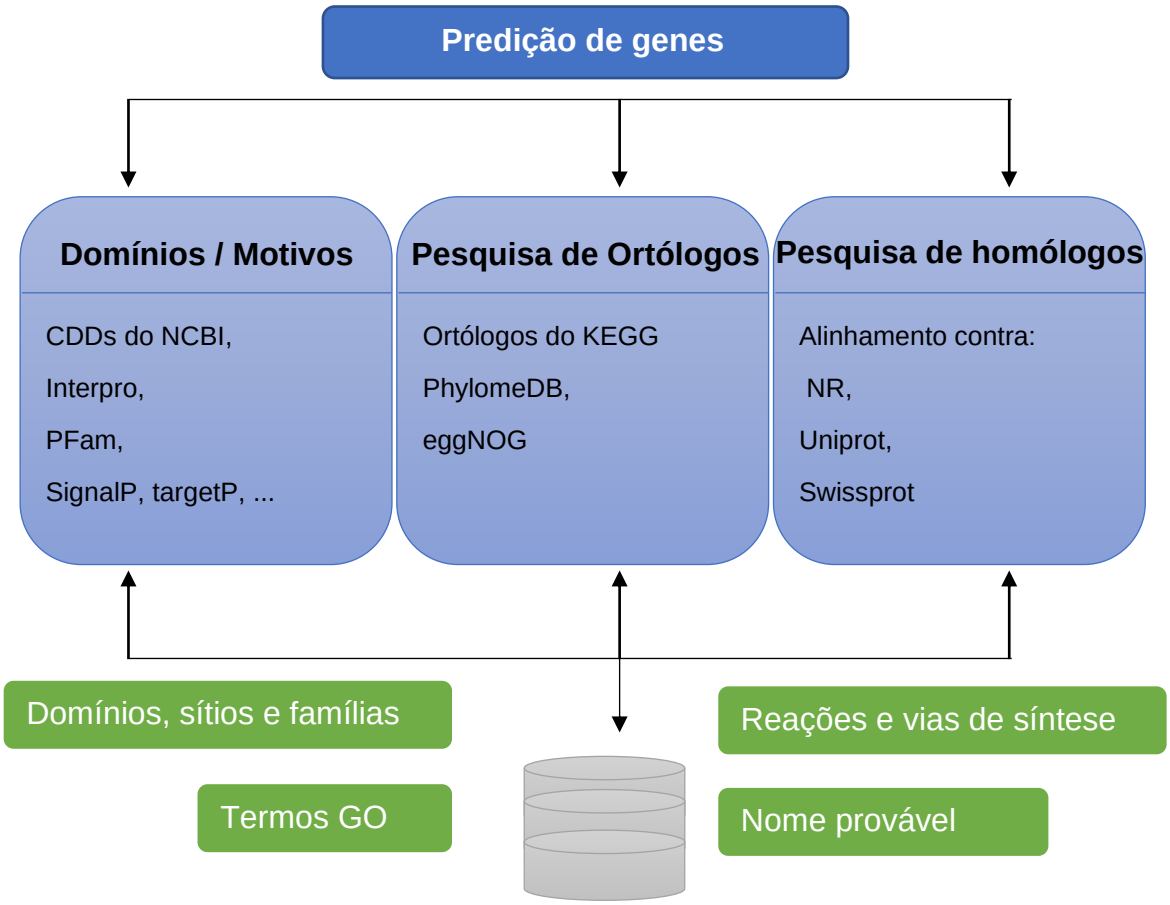

Figura 2 – Três abordagens da anotação funcional adaptado de LANTZ et al. (2018) que permite identificar com maior sensibilidade pelos softwares implícitos na figura: o Diamond, o EggNOG-mapper e InterproScan

(JENSEN et al., 2008), PANTHER (MI et al., 2019), PFam (EL-GEBALI et al., 2019) entre outros.

As análises podem ser executadas localmente ou em serviços web, sendo que existem ainda pipelines automatizados para anotação funcional como Blast2GO e KAAS (BOLGER, M. E.; ARSOVA; USADEL, 2018a). Com a anotação funcional realizada é possível atribuir funções e nomes aos genes cuja proteína foi anotada. Em geral um dos atributos mais importantes inferido a um gene é a ontologia. O consórcio Gene Ontology (GO) (ASHBURNER et al., 2000) é a maior convenção de pesquisadores para gerencia dessas ontologias. Eles definem um código para cada ontologia que pode ser atribuída a um gene. Uma vez que um gene tem seu nome como código alfanumérico pode ser validado facilmente e portado para diferentes softwares e análises sem qualquer preocupação que a informação possa não ser entendida pelo outro lado. Apesar de vantajoso o GO pode ser considerado complexo de modo que outras ontologias para anotação, como MapMan4 BIN (SCHWACKE et al., 2019), tem sido criadas.

A etapa de anotação funcional no processo de anotação do genoma é ilimitada, de modo que apesar de um gene ser anotado por homologia com alta identidade contra outro gene e/ou por similaridade de seu produto polipeptídico contra o de outro gene, ainda assim, análises mais profundas como de vias de síntese ou de redes de interação de proteínas podem apresentar funções mais importantes do gene para o organismo diferente da que foi indicado para ele por processos *in silico* que consideram apenas homologia, nesse sentido é necessário validação experimental e curadoria manual para determinar com certeza anotações dos genes.

2.5. Transcriptômica

O RNA-Seq é uma abordagem de alta resolução que pode revelar novas regiões de transcrição e processamentos alternativos de genes conhecidos, além de permitir gerar uma visão da extensão e complexidade do transcriptoma de eucariotos (WANG, Zhong; GERSTEIN; SNYDER, 2009). Uma vez que diferentes perfis do transcriptoma refletem as respostas do indivíduo sob fatores distintos (YU et al., 2017), estudar estes perfis pode fornecer *insights* sobre os mecanismos

moleculares que explicam como as plantas respondem ao fatores externos (LI, J. R. et al., 2018).

Para realização de um estudo de transcriptômica de sucesso é necessário primeiro definir um bom projeto experimental, com quantidade ideal de replicatas e definição de quais indivíduos sujeitos a quais fatores serão utilizados para criação de bibliotecas, precavendo-se ainda da possível contaminação das amostras na aquisição, pois pode trazer grandes prejuízos às análises (CONESA et al., 2016). Nesse projeto dados de perfis transcriptômicos depositados no NCBI serão utilizados para que os genes identificados possam ser verificados quanto a sua expressão ao consultá-los no navegador de genoma JBrowse (BUELS et al., 2016) citado no tópico 2.2.

2.6. Genômica comparativa e sintenia

A genômica comparativa é uma abordagem que permite elucidar a relação entre o genótipo e o fenótipo com auxílio de comparações do genoma de diferentes espécies (MOREIRA, 2015). A comparação genômica não é simples dado que há vários componentes e relações complexas na molécula de DNA, como o gene e suas partes. Como o paradoxo do valor-C (MOORE, 1984) prova que a complexidade de um organismo não é explicada pelo tamanho de seu genoma é intuitivo estudar em detalhes cada parte do genoma.

Quando o conteúdo gênico compartilhado entre espécies é apresentado - em geral - no diagrama de *Venn* é possível realizar um estudo detalhado que permite determinar várias classes de genes como, por exemplo, *housekeeping*, conservados e específicos (MOREIRA, 2015). No entanto, é sugestivo pela experiência com o paradoxo do valor-C que a quantidade de genes tem pouca explicação para o fenótipo, de modo que a comparação do genoma é extrapolada para a harmonia entre grupos de genes de um genoma com outro. A ordem direta ou reversa de um grupo de ortólogos entre duas espécies determina regiões em sintenia (MOREIRA, 2015). Já a ordem de conjuntos de grupos de ortólogos entre espécies determina a colinearidade entre elas.

Neste sentido, um genoma é comparado a outro em vários aspectos que podem ser apresentados: de forma circular como na **Figura 3-A,** em diagrama de

Venn como na **Figura 3-B** ou linear como na **Figura 3-C** e **Figura 3-D**. Estes gráficos permitem várias análises, que inclui, mas não se limita às comparações de: *i)* conteúdo nucleotídeo, *ii)* conteúdo não funcional, *iii)* conteúdo gênico, *iv)* conteúdo proteico, *v)* sintenia conservada, *vi)* eventos de rearranjo e *vii)* conteúdo de DNA repetitivo (WEI et al., 2002). A utilidade destas analises são indiscutíveis por inclusive colaborar no entendimento da história evolutiva do genoma ao permitir estudar genes conservados, seleção negativa e positiva, desvios evolutivos e eventos de especiação e duplicação (HARDISON, 2003).

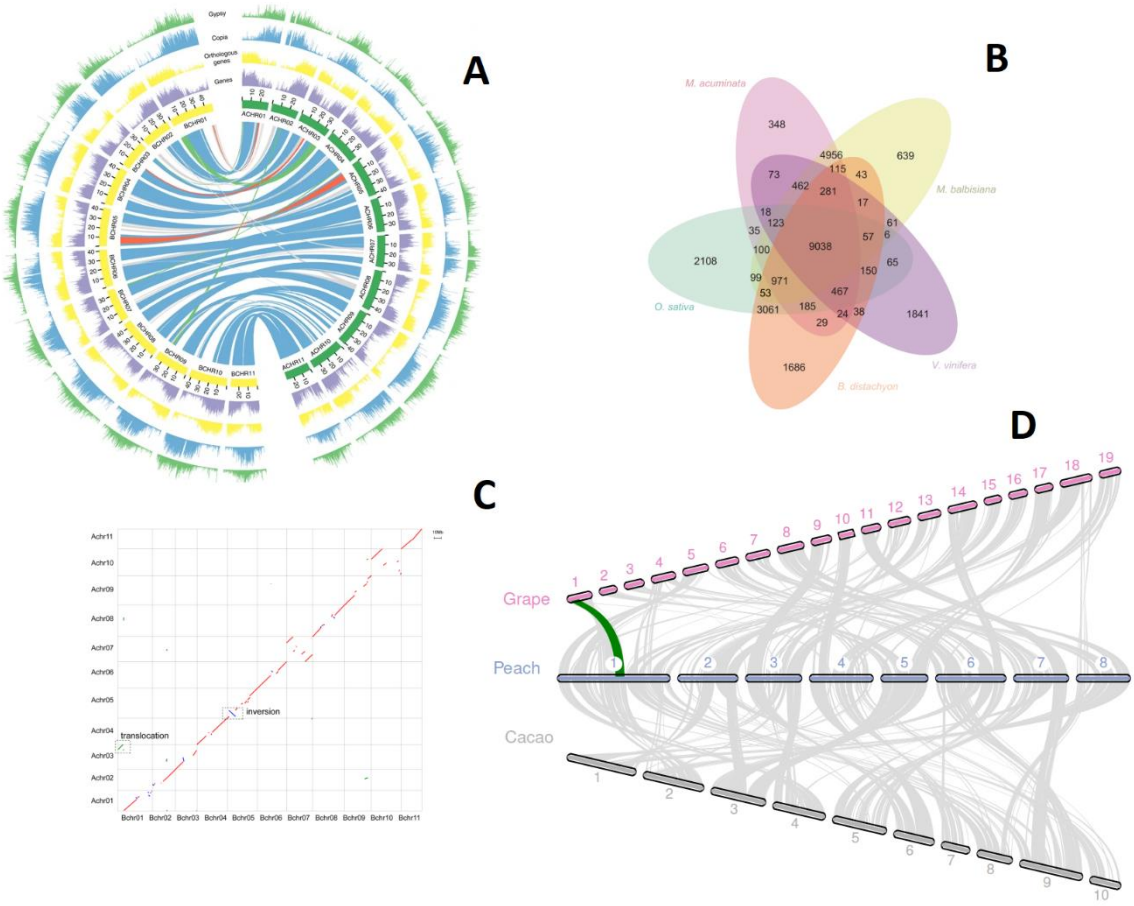

Figura 3 - Apresentações gráficas para genômica comparativa que são mais utilizadas na literatura, **A**) disposição circular dos genomas comparados, **B**) apresentação de conteúdo gênico compartilhado entre as espécies comparadas no diagrama de *Venn*, **C**) disposição linear de dois genomas com regiões sintênicas entre eles marcadas e **D**) Disposição linear de mais de um genoma com links entre regiões sintênicas entre eles. [1]

[1] **Figura 3** partes **A**, **B** e **C** foram extraídas da publicação do genoma da banana (WANG, Zhuo et al., 2019) já a parte **D** foi obtida de https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version).

4. OBJETIVOS

3.1. Objetivo Geral

Anotar o genoma da goiabeira e disponibilizar os resultados em banco de dados acessível para que a comunidade científica possa realizar estudos com a genômica da espécie.

3.2. Objetivos Específicos

- Obter uma montagem consenso do sequenciamento de diferentes cultivares da goiabeira;
- Identificar elementos repetitivos no genoma montado;
- Predizer regiões de RNA não codificador de proteínas (NC-RNA);
- Montar o transcriptoma da espécie;
- Identificar e quantificar genes codificadores de proteínas com validação do transcriptoma;
- Anotar função aos genes identificados com auxílio de diferentes estratégias e bancos de dados públicos;
- Dispor um banco de dados para consulta pública dos dados gerados.

5. MATERIAL E MÉTODOS

Com base em estatísticas de montagem (SOHN; NAM, 2018) e anotação funcional (BOLGER, M. E.; ARSOVA; USADEL, 2018b) - que são os processos mais onerosos - os recursos e prazos computacionais necessários para realização deste trabalho foram estimados conforme a **Tabela 1**. Com isso, este trabalho foi divido em cinco processos que aplicam conhecimentos das áreas de genômica, transcriptômica e informática. As etapas destes processos foram organizadas conforme o fluxo de trabalho exibido na **Figura 4**.

Tabela 1 - Estimativa preliminar do tempo necessário para execução das análises de bioinformática necessárias para obter os resultados deste trabalho.

Descrição	Tempo em semanas
Configuração do ambiente de trabalho	3
Controle de qualidade dos dados	2
Montagem do genoma	2
Ancoragem da montagem do genoma	1
Montagem do transcriptoma	2
Identificação dos elementos repetitivos	3
Predição de genes	2
Predição de NC-RNA	2
Alinhamento com blastp	5
Anotação com InterproScan 5	2
Anotação com eggNOG-mapper	2
Análises de transcriptômica	3
Construção do banco de dados	15
Total	44

4.1. Dados de Sequenciamento

Para realização da montagem do genoma foram obtidos dados de sequenciamento de goiabeira do grupo de pesquisa de melhoramento de goiabeira do laboratório de Biometria da UFES, e dados depositados no NCBI (AGARWALA et al., 2018). Estes dados e suas referências estão listados na **Tabela 2**.

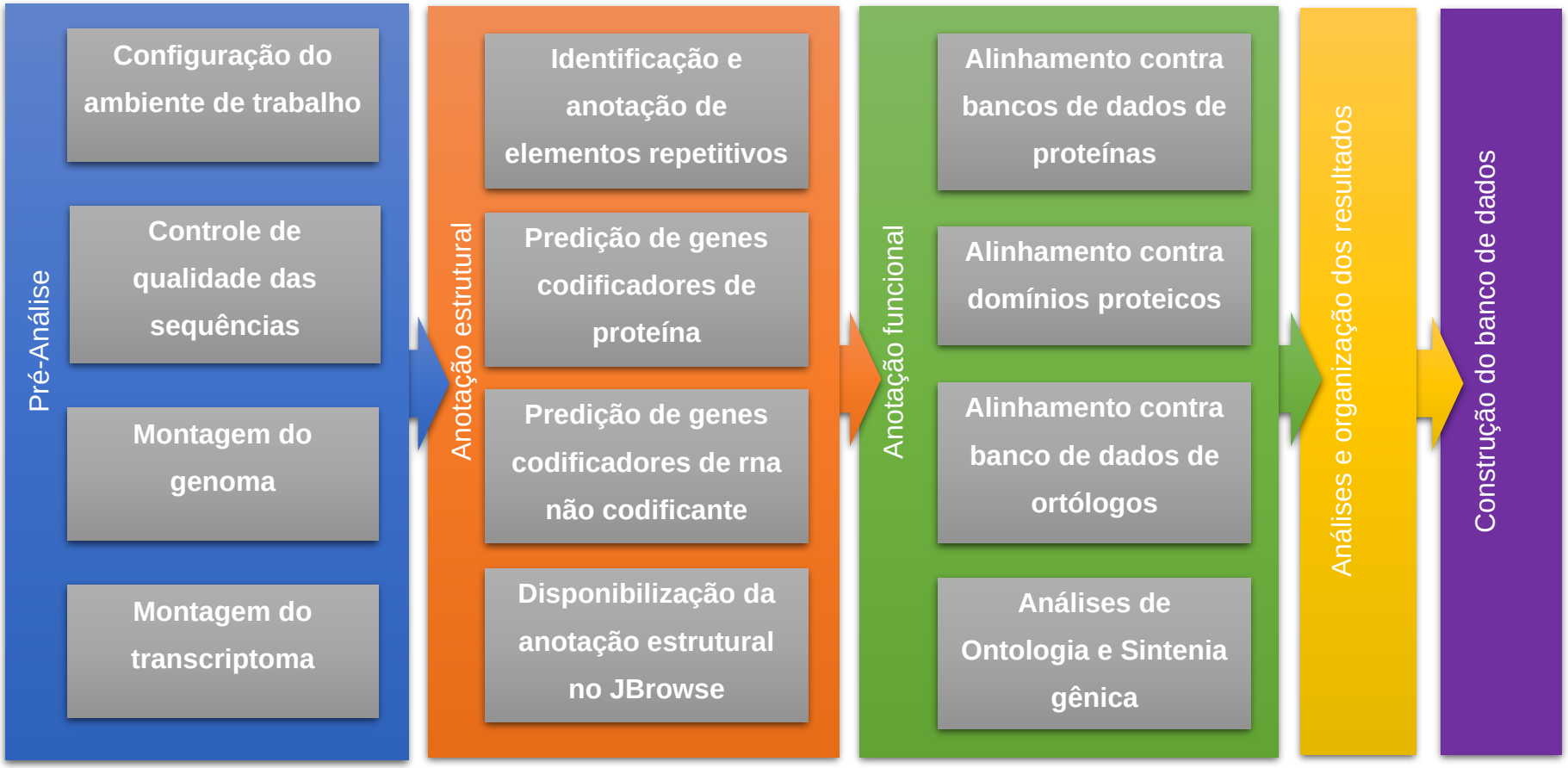

Figura 4 - Definição do fluxo de trabalho em processos que serão executados neste trabalho, iniciando com obtenção do rascunho do genoma na primeira análise, seguido de analises anotação estrutural e funcional que são sucedidas pela construção do banco de dados para persistência e disponibilização das informações.

Tabela 2 – Dados sequenciados e obtidos do NCBI que foram utilizados na montagem do rascunho do genoma e transcriptoma de *P. guajava* neste trabalho. TrimSampleALL.fq e subreads.fa foram sequenciados para este trabalho e os demais foram obtidos pelo *Genbank* e SRA do banco de dados NCBI. A cultivar Paluma foi amostrada no primeiro arquivo listado, enquanto a cultivar Cortibel foi amostrada nos dois primeiros arquivos listados. O terceiro arquivo é a montagem da cultivar Zhenzhu depositado no *Genbank*. Os arquivos da análise de transcriptoma são da cultivar *Allahabad safeda.* Mais detalhes das amostras biológicas podem ser obtidos de Canal (2019) para os dois primeiros arquivos e NCBI para os demais. Os dados listados nesta tabela não têm publicações associadas.

Arquivo	Código NCBI	Destino	Tecnologia**	Tamanho (GB)	# Pares de base***	Cobertura****
TrimSampleALL.fq	-	Genoma	Illumina-PE*	64	23.133.010.627	50x
subreads.fa	-	Genoma	Pacbio-SE	6,4	6.829.895.817	15x
NTGF01000000.fa	NTGF00000000.1	Genoma	Pacbio-SE	0,3	-	55x
folha_1.fq	SRR7186630	Transc.	Illumina-PE	3,4	808.699.730	-
folha_2.fq	SRR7186631	Transc.	Illumina-PE	2,8	691.364.076	-
folha_3.fq	SRR7186633	Transc.	Illumina-PE	2,8	665.875.481	-
flor_1.fq	SRR7186632	Transc.	Illumina-PE	2,2	495.799.337	-
flor_2.fq	SRR7186634	Transc.	Illumina-PE	2,4	542.927.955	-
flor_3.fq	SRR7186635	Transc.	Illumina-PE	2,0	466.313.634	-
fruto_1.fq	SRR7186629	Transc.	Illumina-PE	2,2	485.349.377	-
fruto_2.fq	SRR7186636	Transc.	Illumina-PE	2,6	600.885.314	-
fruto_3.fq	SRR7186637	Transc.	Illumina-PE	2,2	489.740.432	-

*Instrumento Illumina HiSeq 2000 os demais de tecnologia Illumina são HiSeq 2500; **PE designa *Paired-end*, e SE designa *Single-end*; ***Tamanho calculado após controle de qualidade; ****Baseado no tamanho real do genoma (465 Mpb)

4.2. Configuração do ambiente de trabalho

Neste trabalho foram utilizados dois computadores com dois processadores Xeon de 6 núcleos de 2Ghz cada denominados *Bioserver1* e *Bioserver2*. Apesar de vários softwares serem disponibilizados com serviços web alguns ainda exigem grande poder computacional inviabilizando a disponibilidade desses softwares em serviços web abertos ao público. Sendo assim estes computadores foram configurados com os seguintes softwares:

- Debian 9 como sistema operacional;
- OpenSSH Server: para acesso remoto e outras necessidades;
- SSHFS: para compartilhar diretórios para o programa REPET;
- Sun Grid Engine: para gerenciamento do cluster necessário para execução do programa REPET;
- MariaDB: para gerenciado de banco de dados para o programa REPET;
- Montador Masurca;
- Pacote REPET e dependências;
- Pipeline Seqping e dependências;
- Postgres: para gerenciador de bancos de dados
- Outros softwares adicionais

4.3. Controle de qualidade

Para realização do controle de qualidade das reads sequenciadas foram verificados a qualidade das sequências, o conteúdo GC, a presença de adaptadores, a super-representação de *K-mers* e a existência de sequências duplicadas com intuito de detectar erros de sequenciamento, artefatos do PCR ou contaminações (CONESA et al., 2016). Nesse sentido os dados brutos foram “trimados” com auxílio do software Trimmomatic 0.22 (BOLGER, A. M.; LOHSE; USADEL, 2014) a fim de obter reads de qualidade com o parâmetro -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:85. A qualidade das reads limpas foi atestada na visualização da qualidade das mesmas com auxílio do software FQC Dashboard 1.5.8 (BROWN; PIRRUNG; MCCUE, 2017).

4.4. Montagem do genoma

As reads subsequentes ao controle de qualidade foram submetidas ao processo de montagem do genoma. A montagem foi realizada em estratégia hibrida de modo a utilizar os dados de sequenciamento de DNA de diferentes tecnologias e cultivares, incluindo a cultivar Zhenzhu que não fora utilizada na montagem de CANAL (2019). Neste sentido, em primeiro momento foi realizado uma montagem dos dados do sequenciamento do genoma das cultivares Paluma e Cortibel (anotado com as iniciais P e C na **Figura 5**) com o software MaSuRCA 3.3.3 (ZIMIN et al., 2013). Os *contigs* da montagem da cultivar Zhenzhu (anotado com a inicial Z na **Figura 5**) são disponibilizados pelo NCBI e foram montados juntos com as sequencias montadas do Masurca e os *contigs* do sequenciamento de PacBio da cultivar Cortibel com o montador Flye 2.4.1 (KOLMOGOROV et al., 2019). Em seguida os *scaffolds* obtidos da montagem do Flye foram analisados e submetidos a ancoragem no mapa genético da goiaba (PADMAKAR et al., 2015) com auxílio do software ALLMAPS 0.8.12 (TANG et al., 2015) afim de obter scaffolds maiores, esta etapa é a última, representada no retângulo vermelho na **Figura 5**.

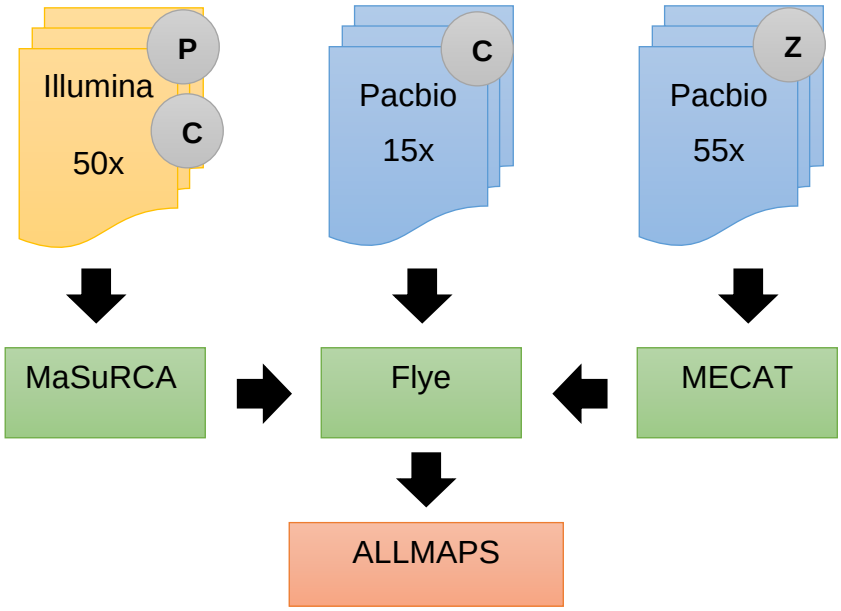

Figura 5 - Fluxograma da montagem do genoma onde os retângulos representam os softwares utilizados na montagem do rascunho do genoma, os arquivos representam as sequências após o controle de qualidade e as letras em cinza são as iniciais das cultivares amostradas em cada sequenciamento. Os retângulos em verde são os softwares de montagem do genoma e em vermelho a ancoragem dos scaffolds montados no mapa genético.

4.5. Montagem do transcriptoma

Dados de expressão dos genes nos tecidos folha, flor e fruto foram obtidos do SRA (AGARWALA et al., 2018) conforme listado na **Tabela 2**. Em primeiro momento os dados foram submetidos a controle de qualidade com o software TRIMMOMATIC com o parâmetro -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:85 (BOLGER, A. M.; LOHSE; USADEL, 2014) e sua qualidade foi atestada quando visualizada com o software FQC Dashboard 1.5.8 (BROWN; PIRRUNG; MCCUE, 2017).

Como não há transcriptoma da espécie montado disponível em repositórios públicos de dados biológicos foi realizada a montagem *de novo* do RNA-Seq com o montador TRINITY 2.8.4 (GRABHERR et al., 2011) para que o transcriptoma montado pudesse ser utilizado na identificação de janelas de leitura e validação dos genes preditos pelo pipeline Seqping (CHAN et al., 2017) na Anotação estrutural. Já para obtenção do arquivo de expressão dos genes para apresentação no JBrowse 1.16.6 (BUELS et al., 2016) as *reads* foram indexadas ao genoma montado com auxílio do programa Bowtie2 (LANGMEAD; SALZBERG, 2012) para então serem alinhadas no genoma com o software Tophat2 (KIM et al., 2013) afim de obter o arquivo BigWig que foi instalado nele.

4.6. Anotação estrutural

Em primeiro momento o genoma montado foi submetido a análise identificação e anotação de elementos repetitivos com auxílio dos pipelines TEdenovo.py e TEannot.py do software REPET (FLUTRE et al., 2011). O software MISA 1.0 (BEIER et al., 2017) versão local foi usado para identificação de microssatélites. Em seguida foi realizado análise de segmentos duplicados com o software SEDEF 1.1-24 (NUMANAGIĆ et al., 2018). Para predição de genes codificadores de proteína com evidência de transcritos o pipeline *Seqping* 1.45.1 (CHAN et al., 2017) foi utilizado com o conjunto de dados próprio para plantas obtido junto ao software. Para visualização gráfica dos elementos identificados na anotação estrutural uma figura circular foi gerada com software Circos 0.69-6 (KRZYWINSKI et al., 2009).

Em seguida foi realizado a identificação de tRNAs com o software Aragorn 1.2.38 (LASLETT; CANBACK, 2004) e tRNAscan-SE 1.4 (LOWE; EDDY, 1997), sendo que para o conjunto final as sobreposições de pelo menos um par de base foram resolvidas priorizando os tRNAs preditos pelo software Aragorn. Os rRNAs foram preditos com o software RNAmmer 1 (LAGESEN et al., 2007). A predição de snRNA e snoRNA foi realizada com o software INFERNAL 1.1.2 (NAWROCKI; EDDY, 2013) por pesquisa contra banco de dados Rfam14.1 (GRIFFITHS-JONES et al., 2003). Para identificação de HACA-box e CD-box foi utilizada a API do RNACentral v13.0 (SWEENEY et al., 2019). Para predição de miRNA *in silico* foram utilizados os pipelines *SUmirPredictor* e *SUmirLocator* (ALPTEKIN; AKPINAR; BUDAK, 2017) com os scripts *SUmirFind* e *SUmirFold* (LUCAS; BUDAK, 2012).

4.7. Anotação funcional

Com as sequencias de proteínas dos genes preditos foi realizada a anotação funcional em três abordagens diferentes a fim de obter uma abrangência maior da anotação funcional (BOLGER, M. E.; ARSOVA; USADEL, 2018). Na primeira abordagem foi utilizado o software Diamond 0.9.27 (BUCHFINK; XIE; HUSON, 2014) com os parâmetros evalue 1e-5 e tophit 5 para alinhar as proteínas contra os bancos de dados NR Refseq (O'LEARY et al., 2016) e SwissProt (BAIROCH, 2000).

Em seguida as proteínas foram anotadas utilizando os identificadores das sequencias alinhadas com a ferramenta Retrieve/ID mapping do banco de dados Uniprot (BATEMAN, 2019). Na segunda abordagem as proteínas foram processadas pelo software InterProScan 5 (JONES et al., 2014) para identificação de domínios conservados em bancos de dados de anotação como o PANTHER (MI et al., 2019) já integrados ao software.

Na terceira abordagem o software eggNOG-Mapper 1.0.3 (HUERTA-CEPAS et al., 2017) foi utilizado para identificação de grupos de genes ortólogos do banco de dados eggNOG 4.5 (HUERTA-CEPAS et al., 2016) e anotação funcional das proteínas. As três abordagens foram executas em paralelo. No fim das três abordagens de anotação funcional foi gerado uma tabela de anotação que contém a anotação mesclada, levando em consideração a maior significância estatística da informação obtida de cada abordagem.

As principais anotações atribuídas as proteínas são: famílias de domínios de proteínas do PFAM (EL-GEBALI et al., 2019), grupos de genes ortologos (JENSEN et al., 2008), vias metabólicas do KEGG (KANEHISA et al., 2016) e Reactome (FABREGAT et al., 2017) e ontologia genica do consorcio GO (ASHBURNER et al., 2000). As informações de ontologia dos genes foram utilizadas com a ferramenta GOSlim disponível na plataforma AgBASE (MCCARTHY et al., 2006) para obter uma visão geral das funções anotadas para os genes preditos.

4.8. Sintenia

Para análise de conteúdo gênico compartilhado entre espécies foi utilizado o OrthoVenn2 (XU et al., 2019) com parâmetros e-value 1e-5 e inflation value 1.5. Os organismos utilizados na análise foram *E. grandis* 2.0 cujos dados foram obtidos no Phytozome 12 (GOODSTEIN et al., 2012), *A. thaliana* e *V. vinifera* que possuem dados disponíveis no próprio serviço web.

Para obter o valor que determina um termo GO como enriquecido o OrthoVenn2 utilizou a distribuição hipergeométrica com significância estatística calculada pela fórmula:

$$P = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Onde N é o total de termos GO na anotação dos genes, M é a quantidade de termos em comum entre os genes, n a quantidade de termos GO não redundante e x a quantidade no conjunto distinto de termos GO em comum entre os genes. Como foi identificado enriquecimento de um termo GO relacionado a síntese de terpeno, analises adicionais de genômica comparativa foram realizadas com o software MCscan (TANG et al., 2008) para identificação de regiões sintênicas entre *P. guajava*, *V. vinifera* e *E. grandis*.

4.9. Banco de dados

O banco de dados foi construído com o gerador JHIPSTER (disponível em jhipster.tech), sendo que no *backend* foi utilizado Spring (disponível em spring.io) e no *frontend* Angular5 (disponível em angular.io) e Bootstrap (disponível em getbootstrap.com). Para gerenciador de banco de dados utilizou-se PostgreSQL (disponível em postgresql.org). A modelagem das entidades apresentada na **Figura 6** foi elaborada em linguagem própria do gerador JHIPSTER (JDL). O GuavaDB foi integrado ao JBrowse (BUELS et al., 2016) para que este possa suprir a funcionalidade de navegação no genoma. O JBrowse foi escolhido por ser de fácil instalação em ambiente *web* bastando apenas embutir a página do navegador.

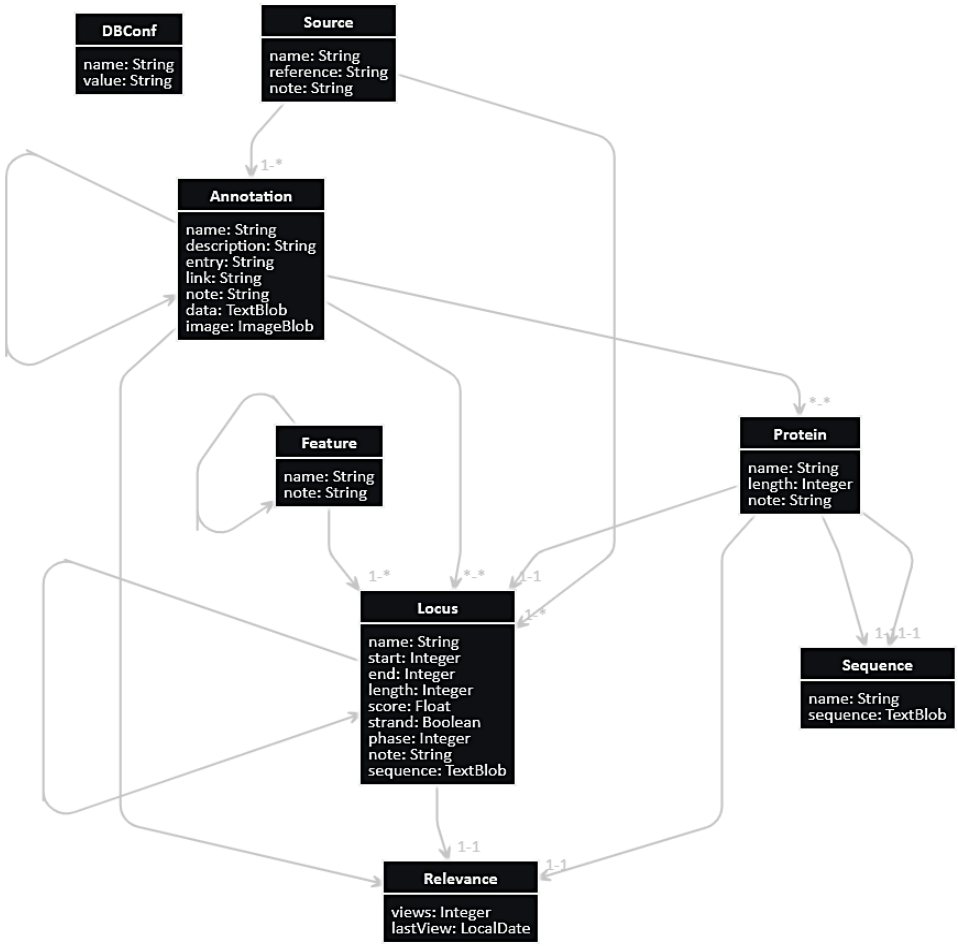

Figura 6 - Modelagem das entidades do banco de dados. As entidades abstraem informações e relações biológicas que serão persistidas no banco de dados, com exceção de *DBConf* que é responsável por registrar configurações variáveis do banco de dados e *Relevance* que registra a data e quantidade de visualizações de uma informação para apresentar no topo as informações mais relevantes para os usuários.

6. RESULTADOS

Para a montagem do genoma as sequencias oriundas de tecnologia Illumina foram submetidas a controle de qualidade com o Trimmomatic 0.22 (BOLGER, A. M.; LOHSE; USADEL, 2014) onde 10.5% das sequencias foram removidas, uma visão geral da qualidade final das reads Illumina, gerada pelo software FastQC 0.11.5 (BROWN; PIRRUNG; MCCUE, 2017) está na **Figura 7** a seguir.

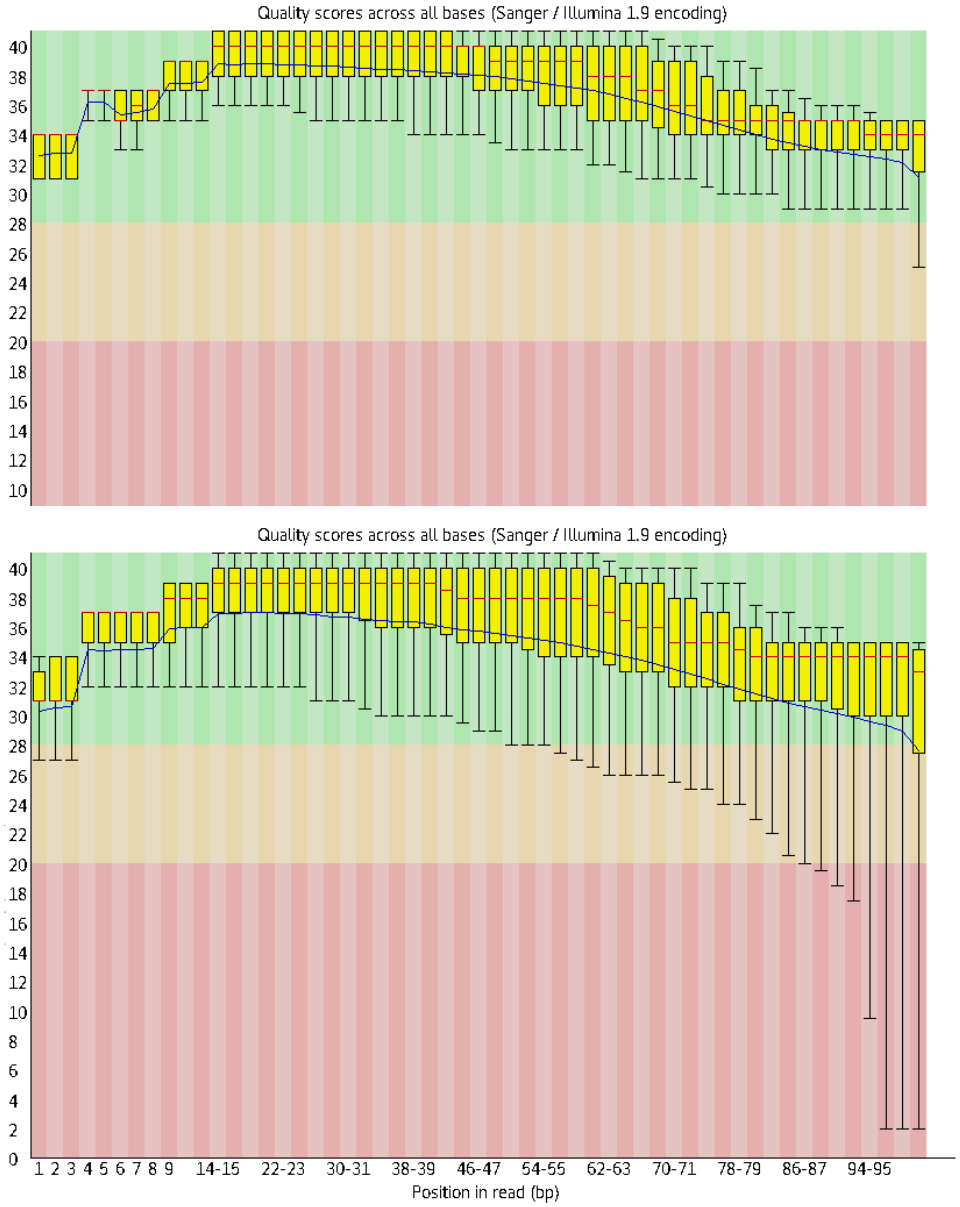

Figura 7 - Phred score das sequências do sequenciamento Illumina após controle de qualidade. No eixo Y o valor de Phred score e em X a posição do par de base na sequência lida pelo instrumento, a linha é a média. No topo as reads *foward* e abaixo as reads *reverse*.

As reads Illumina foram montadas em quatro horas no Bioserver2 (Xeon 2x6cores 2GHz e 128 GB RAM) com o montador MaSuRCA (ZIMIN et al., 2013) que resultou em uma montagem de 324,5 Mpb de 222.972 sequencias com N50 de 2.610 e media de tamanho de 1.455,67 pares de base. Essa montagem foi então submetida ao montador Flye (KOLMOGOROV et al., 2019) junto ao sequenciamento PacBio (*subreads.fasta*) e a montagem da espécie que há depositado NCBI (*GCA_002914565.1_Guava1.0_genomic.fna.fasta*) para obtenção do rascunho do genoma. A linha de comando necessária para execução do Flye é apresentada a seguir:

```
Flye \
    --pacbio-raw data/subreads.fasta \
    data/GCA_002914565.1_Guava1.0_genomic.fna.fasta \
    final.genome.scf.fasta  \
    --genome-size 465m \
    --out-dir out \
    --threads 24 1> log 2> err
```

O mapa genético da espécie *P. guajava* (PADMAKAR et al., 2015) disponível em figura foi digitalizado (apêndice 9.1) e as sequencias de primers (apêndice 9.2) foram obtidas no material suplementar do artigo do mapa genético para indicar ao ALLMAPS (TANG et al., 2015) as posições dos scaffolds. Assim, o mapa digitalizado determinou a posição genética do marcador e os *scaffolds* foram associados aos marcadores por alinhamento com o software *Blastn* (CAMACHO et al., 2009). Como os marcares estão associados aos grupos de ligação no mapa genético foi possível inferir qual grupo de ligação (LG) certo *scaffold* pertence, conforme apresentado na **Figura 8** a seguir.

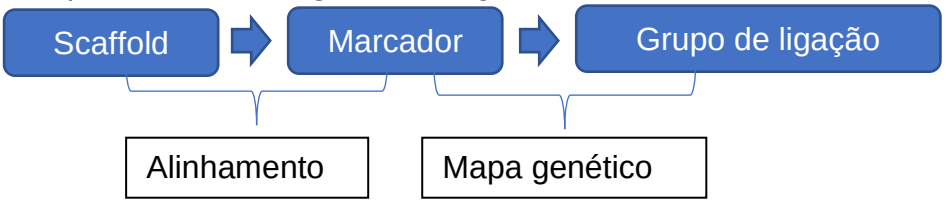

Figura 8 - Diagrama explicativo da associação dos *scaffolds* aos grupos de ligação intermediados por alinhamento e mapa genético. O grupo de ligação do scaffold é determinado pelo marcador em que foi alinhado, sendo que este está associado a um grupo de ligação pelo mapa genético. A direção e ordem dos scaffolds é resolvida pelo ALLMAPS considerando a posição genética.

O arquivo que associa os scaffolds aos grupos de ligação foi submetido ao software ALLMAPS que utilizou de artifícios de otimização linear para resolver vários conflitos de posição e orientação dos scaffolds e gerar o arquivo fasta com os scaffolds ancorados aos grupos de ligação. Assim, das 2.097 sequencias obtidas na montagem do rascunho do genoma, 1.422 (387,1 dos 391,5 Mb montados que corresponde a 98,9%) foram ancorados e 25,4% foram orientados em 11 pseudomoléculas, a **Figura 9** ilustra a ancoragem realizada pelo ALLMAPS para determinação da pseudomolécula LG8. No final o genoma contém, 2.097 contigs, 686 scaffolds e 11 pseudomoléculas que totalizam 394,5 Mb bases conforme apresenta a **Tabela 3**.

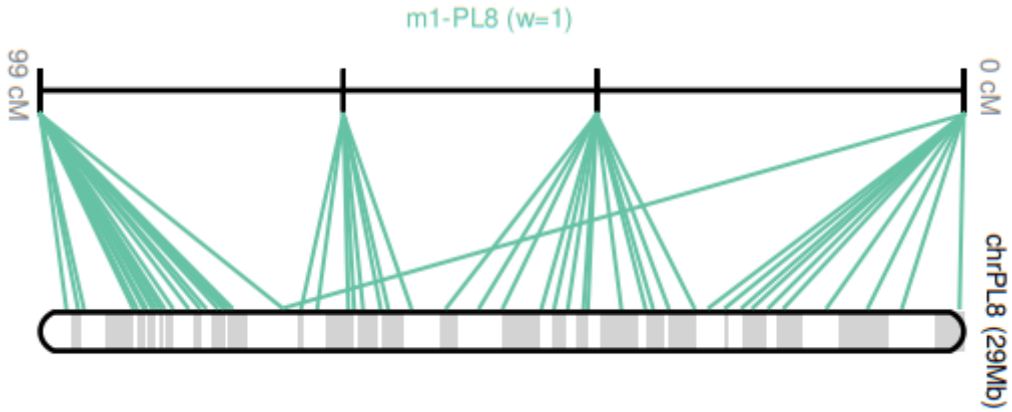

Figura 9 - Ancoragem de scaffolds na pseudomolécula LG8 baseado no mapa genético apresenta como o ALLMAPS posicionou os scaffolds na determinação desta pseudo-molécula.

A abordagem *de novo* foi utilizada na montagem do transcriptoma devido ser necessário ter o transcriptoma montado para predizer os genes, pois no caso de montagem por referência seria necessário informar ao *software* dados dos genes até então inexistentes. Para esta montagem foi executado o Trinity (GRABHERR et al., 2011) com as 95.755.870 *reads* pós controle de qualidade que totalizaram 11GB em disco. Estas sequencias de foram montadas em 167.376 sequencias que somam 249 Mpb apresentadas na **Tabela 3** a seguir.

Tabela 3 - Estatísticas gerais da montagem do genoma, ancoragem dos scaffolds nos 11 grupos de ligação do mapa genético e montagem do transcriptoma. As sequencias montadas são contiguas enquanto as ancoradas foram agrupadas em *scaffolds* que correspondem as pseudomoléculas.

	Genoma montado: contigs	**Contigs ancorados: scaffolds**	**Transcriptoma: contigs**
# sequencias	2.097	11 + 675	167.376
Tamanho Total	391.569.819	391.710.919	249.914.780
Maior sequência	3.012.728	49.678.869	27.923
N50	587.224	32.551.815	2.723
N75	286.884	30.474.487	1.532
L50	199	5	30.188
L75	435	8	60.137
GC (%)	39,53	39,53	44
N's por 100 kpb	0	36,02	0

5.1. Elementos repetitivos e gênicos identificados

A identificação das regiões de baixa complexidade no genoma iniciou-se com a execução do pacote REPET (FLUTRE et al., 2011) contra o rascunho do genoma montado, que identificou 3.557 elementos transponíveis cujos fragmentos totalizando 131,9 Mpb cobrem 33,7% do genoma detalhados na **Tabela 4**. O MISA (BEIER et al., 2017) foi executado para identificação de SSRs. Os 69.991 microssatélites que ele identificou é, em número, bem menor comparado aos 356.003 elementos identificados pelo pipeline REPET. Isso deve-se a abordagem de homologia para identificação dos motivos que os softwares executados pelo REPET aplicam ao invés de identificação de padrões aplicada pelo MISA. O SEDEF (NUMANAGIĆ et al., 2018) por sua vez identificou 14.188.576 segmentos duplicados no genoma o que denota a natureza repetitiva do genoma de plantas abordado na Revisão Bibliográfica.

Tabela 4 - Classificação dos elementos transponíveis identificados pelo pipeline REPET no genoma de *P. guajava*.

	Quantidade	Tamanho Total
Classe I	2.127	695.441
DIRS	131	17.333.838
LARD	131	15.138.870
LINE	243	7.734.103
LTR	1.437	55.277.797
SINE	82	701.416
TRIM	64	1.527.331
Classe II	757	3.103.075
Helitron	134	2.439.259
MITE	112	5.224.304
Maverick	2	70.998
TIR	470	19.117.602
não classificado	673	3.572.124
Total	3.557	131.936.158

Na predição dos genes codificadores de proteína o genoma foi divido em duas partes, a fim de aproveitar ao máximo o poder computacional dos servidores *Bioserver1* e *Bioserver2* do *BioCluster* em paralelo para ganhar tempo. O pipeline

Seqping (CHAN et al., 2017) é bastante oneroso por realizar vários alinhamentos - contra banco de sequências de elementos transponíveis (*TIGR Plant Repeat*, *RepBase* e *Gypsy Database*) - para remover falsos positivos e treinar e predizer comparativamente em mais de um software. A **Tabela 5** lista os mRNAS identificados pelo Seqping e outros tipos de RNAS preditos. O resultado dos mRNAs apresentados na tabela foram contabilizados pelo GAG (GEIB et al., 2018). A predição de tRNAs determinou o Aragorn (LASLETT; CANBACK, 2004) como mais sensitivo ao identificar 863 elementos contra 482 identificados pelo trnaScan (LOWE; EDDY, 1997). Outra vantagem do Aragorn é a imagem representativa de cada tRNA predito ser reportada em arquivo de texto conforme apresentado na **Figura 10** que permite estudos estruturais com esse tipo de RNA.

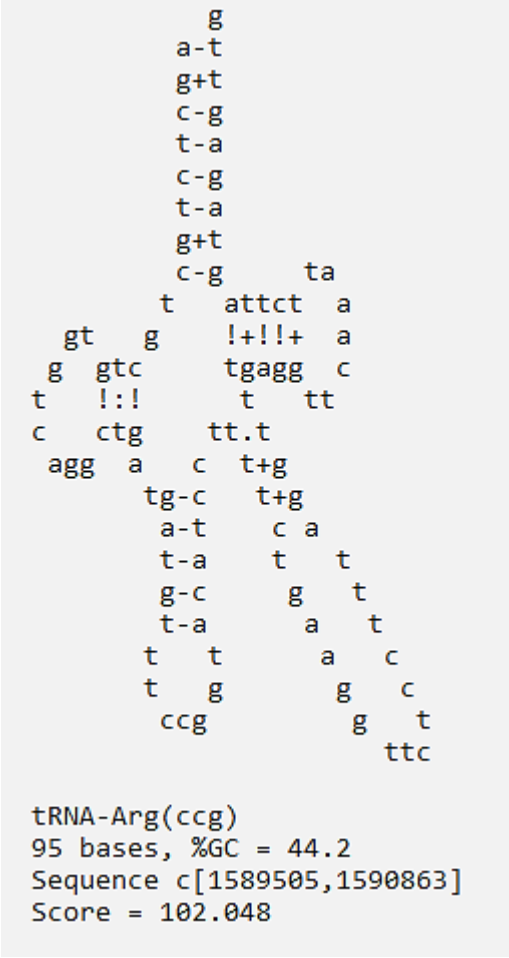

Figura 10 - Imagem ilustrativa de um dos tRNAs preditos pelo software Aragorn que apresenta a estrutura bidimensional do tRNA onde é possivel verficar partes do transcrito que anelarão sobre si mesma, o anti-codon e outras características desse tipo de RNA conhecidas na literatura.

Tabela 5 - Visão geral dos genes preditos no genoma de *P. guajava* na etapa anotação estrutural do genoma. Os snRNAs foram identificados com abordagem de homologia enquanto os demais em abordagem *ab initio*.

		Quantidade	Tamanho	
			Médio (pb)	Total (Mbp)
mRNA	Genes	60.538	3.169	191,8
	mRNA	70.346	3.702	260,4
	Éxons	216.681	282	82,2
	Íntrons	151.850	805	178,6
	5'UTR e 3'UTR	54.350	395	21,5
	CDS	252.822	746	52,5
miRNA		6.619	20	0,1
tRNA		1.345	256	0,3
rRNA	8s	13	113,2	1.472
	18s	15	1.940,3	29.104
	28s	14	5.436,3	76.108
snRNA	snoRNA	270	104,4	28.197
	HACA-box	15	115,6	1.734
	CD-box	188	96,9	18.226
	snRNA	91	142,7	12.983

A **Figura 11** a seguir apresenta os elementos gênicos e repetitivos contextualizados. Adicionalmente a expressão de cada gene em FPKM é apresentada no mapa de cores da track C para os tecidos que foram utilizados na montagem do transcriptoma para predição dos genes. É possível observar nela que os scaffolds menores, e que não ancoraram em nenhum grupo de ligação, tem menor densidade genica e elementos repetitivos à medida que decrescem em tamanho no LG*. Como a quantidade de segmentos duplicados identificados inviabilizam a plotagem da figura foi apresentado apenas os maiores que 50 mil pares de base.

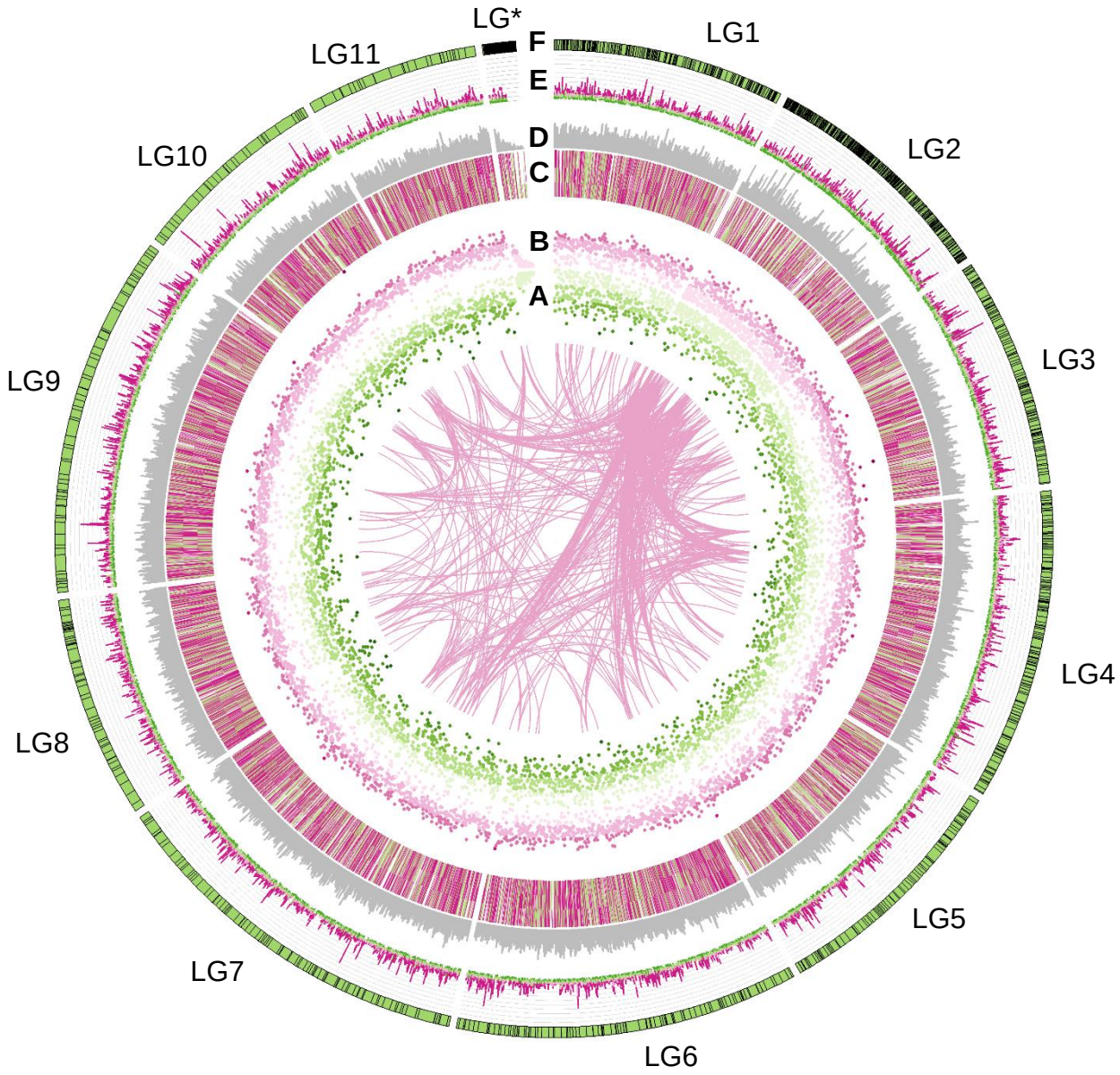

Figura 11 - Visão geral da anotação estrutural de *P. guajava*. No centro os segmentos duplicados identificados de tamanho maior que 50kpb. Em **A** densidade de microssatélites de verde claro (menor densidade) à verde escuro. Em **B** densidade de elementos transponiveis identificados. O heatmap em **C** é composto de tres partes de dentro para fora: FPKM em folha, flor e fruto. Em **D** densidade genica. Em **E** dendidade de SNP entre as cultivares paluma e cortibel. Em **F** os grupos de ligação com os scaffolds representados dentro em verde. O grupo LG* é composto dos scaffolds que não ancoraram em nenhum grupo de ligação. Todas densidades são calculadas em janelas deslizantes de 100kbp.

5.2. Elementos gênicos anotados

A anotação funcional foi realizada em três abordagens que foram executadas em paralelo. Na primeira abordagem o conjunto de proteínas dos genes identificados no genoma foi alinhado com o software Diamond (BUCHFINK; XIE; HUSON, 2014) no Galaxy (AFGAN et al., 2018) disponível no servidor usegalaxy.eu. Foi necessário solicitar cota acadêmica para obter espaço suficiente para utilizar os bancos SwissProt (BAIROCH, 2000), NR (O'LEARY et al., 2016) e TREMBL (BATEMAN, 2019) (~350GB). O resultado do alinhamento contra os bancos NR, SwissProt e TREMBL apresentados na **Tabela 6** foram obtidos com os parâmetros *evalue* 1e-5 e *top-hit* 5 e os demais como padrão.

Na segunda abordagem o conjunto de proteínas foi submetido a análise do InterproScan5 (JONES et al., 2014) que foi o software mais sensitivo ao anotar mais de 75% dos genes e proteínas - principalmente em bancos de dados de domínios conservados e família de proteínas - conforme listado na **Tabela 6**. Por último, o eggNOG-mapper (HUERTA-CEPAS et al., 2017) foi executado contra o proteoma com o parâmetro *--diamond* que permitiu a identificação dos 25.696 genes com ortólogos dos bancos de dados eggNOG e KEGG que estão contabilizados separadamente na **Tabela 6**.

Os ortólogos do eggNOG estão anotados em 12.859 grupos de ortologos (COG – *cluster of orthologs group*) que possuem anotação relacionado a sua função: uma excelente fonte para encontrar *loci* relacionados a funções de interesse. As sequências de ortólogos do eggNOG foram obtidas do banco de dados *online* para gerar árvore filogenética de cada proteína com ortólogos anotados. As árvores foram geradas com o software ETE3 (HUERTA-CEPAS; SERRA; BORK, 2016) que foi executado no pipeline do apêndice 9.3.

Com relação a anotação da ontologia gênica (GO) 9% dos termos GO estão anotados apenas na primeira abordagem, 12% estão anotados apenas da segunda abordagem (corroborando a maior sensibilidade do InterproScan5 mencionada) e 7% apenas última, que significa que 72% dos termos GO foram anotados nas três abordagens. Um resumo dos termos GO simplificados pela ferramenta *GOSlimm* do *AgBase* (MCCARTHY et al., 2006) está na **Figura 12**.

Tabela 6 - Quantificação de genes e proteínas anotados nas três abordagens da anotação funcional. Os bancos de dados NR, SwissProt e TREMBL foram utilizados na primeira abordagem, os domínios conservados, famílias de proteínas e vias de síntese de metabólitos (com exceção de BiGG Models) foram obtidos na segunda abordagem e os demais na terceira, com exceção de Gene Ontology que foi anotado nas três abordagens. A coluna % Genes refere-se a porcentagem com relação aos 60.538 genes codificadores de proteínas preditos no genoma.

Banco de dados	Proteínas	Genes	% Genes
Domínios conservados			
Interpro	32.056	24.277	40%
SMART	8.965	6.957	11%
CDD	9.735	7.738	13%
ProDom	359	316	1%
ProSitePatterns	5.142	4.085	7%
ProSiteProfiles	12.068	9.475	16%
Coils	7.205	5.983	10%
PRINTS	4.078	3.260	5%
Gene3D	23.403	17.988	30%
MobiDBLite	33.007	29.158	48%
Família de proteínas			
SFLD	206	174	0%
Hamap	560	434	1%
Pfam	27.073	20.684	34%
PIRSF	1.186	972	2%
TIGRFAM	2.771	2.265	4%
SUPERFAMILY	22.231	16.885	28%
PANTHER	35.698	26.739	44%
Vias de síntese de metabólitos			
MetaCyc	2.101	1.525	3%
KEGG Pathways	2.756	1.919	3%
Reactome	6.919	4.787	8%
BiGG Models	400	300	0%
Outros			
NR	40.583	31.072	51%
SwissProt	26.594	19.600	32%
TREMBL	38.286	28.890	48%
Gene Ontology	31.649	23.835	39%
Ortólogos do EggNOG	34.446	25.696	42%
Ortólogos do KEGG	16.342	11.913	20%
TOTAL não redundante	57.538	47.907	79%

Figura 12 - Abundância dos GOs anotados simplificado nas principais ontologias para plantas. O histograma de cima é a quantificação do aspecto processo biológico, no meio função molecular e em baixo componente celular.

As vias de síntese de metabolitos são importantes ferramentas para futuros estudos de metabolômica com a espécie, com as anotações de metabólitos, reações e vias de síntese é possível determinar qual gene tem relação com estes. Somados nas três abordagens foram anotados 6.225 genes (10%) e 8.896 proteínas (13%) não redundantes nos bancos de dados *MetaCyc* (CASPI et al., 2018), *KEGG Pathways* (KANEHISA et al., 2016), *Reactome* (JASSAL et al., 2019) e *BiGG Models* (KING et al., 2016). Para estender a anotação funcional foi realizado análise com o proteoma no iTAK (ZHENG et al., 2016) que identificou 3.151 proteínas que contém pelo menos um dos 66 fatores de transcrição (TFs) alinhados e 661 proteínas que contém algum dos 23 reguladores de transcrição (TRs) alinhados, totalizando 3.812 proteínas e 3.401 genes anotados.

5.3. Banco de dados GuavaDB em guava.ufes.br

O banco de dados foi gerado com o Jhipster (disponível em jhipster.tech) utilizando o *framework* Spring (disponível em spring.io) no *backend* e Angular (disponível em angular.io) no *frontend*. Foi implantado no *Bioserver2* junto de um serviço (apêndice 9.4) que foi habilitado no sistema para iniciar automaticamente o banco de dados após a inicialização do sistema operacional. Os dados foram persistidos no gerenciador de banco de dados orientado a objetos PostgreSQL (disponível em postgresql.org) e os arquivos FASTA para *download* foram armazenados em disco no *Bioserver2*. Para viabilizar a funcionalidade de alinhamento (*vide* guava.ufes.br/blast) foi necessário implementar o *script* em CGI do apêndice 9.5 para receber e processar as solicitações de execução de alinhamentos. Uma visão geral da interação entre os módulos é apresentada na **Figura 13**. O servidor Apache (disponível em apache.org) foi instalado para disponibilização dos arquivos de *download*, execução do *script* em CGI e execução do *JBrowse* (BUELS et al., 2016) que foi embutido nas páginas de detalhes dos *loci* (*vide* guava.ufes.br/locus) do banco de dados.

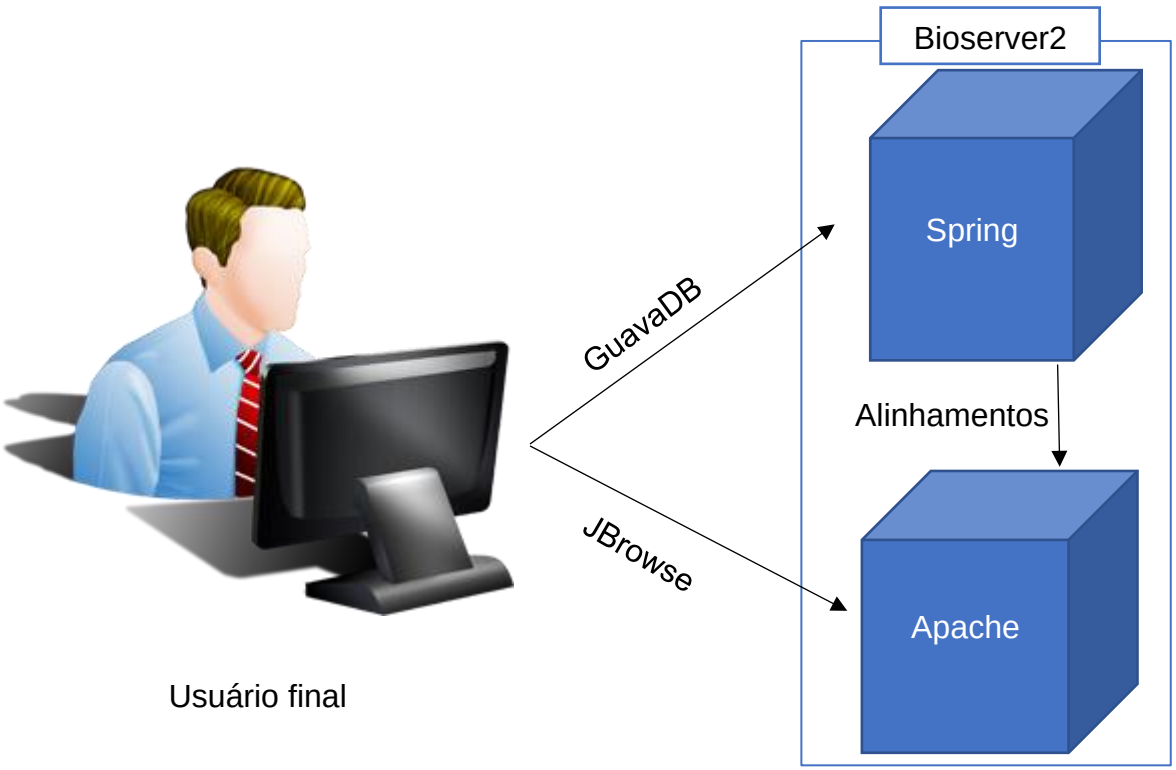

Figura 13 - Esquema de interação entre os módulos que compõem o banco de dados GuavaDB. O usuário acessa o banco de dados pela interface do Spring, que direciona requisições do usuário ao JBrowse no módulo Apache. O Spring delega ao Apache as requisições de alinhamentos do usuário pela interface do script CGI.

5.4. Estudo de caso 1: Buscar sequências de TPS no banco de dados

Este estudo de caso é a resolução de uma tarefa que surgiu ao longo das análises dos resultados genômicos de *P. guajava*. Verificou-se a necessidade de identificar quais genes da goiabeira tem relação com a síntese de óleos essenciais, em especial terpenóides (KARUNANITHI; ZERBE, 2019), por meio de homologia ou anotação das proteínas codificadas pelos genes. Com a identificação destes genes foi necessário curar o conjunto de possíveis genes levando em conta remover genes em que há ausência de domínios conservados (por exemplo pfam.xfam.org/family/Terpene_synth) em comum nesta família de genes (JIANG et al., 2019) e pseudogenes (TUTAR, 2012); que possuem as características estruturais contudo não são transcritos em mRNA tornando-os não funcionais, ou, desempenham papel regulatório que carece de outros estudos mais aprofundados para determinar sua relação com a síntese dos compostos de terpenos.

Apesar de já existir algoritmo para identificação e análise do "terpenoma" de plantas (PRIYA et al., 2018) a identificação manual na espécie *P. guajava* pelo banco de dados GuavaDB é bem simples. Primeiro deve ser obtido um arquivo multifasta contendo sequência de proteínas caracterizadas como composto de terpenos como os da lista disponibilizada no banco de dados de sequiterpenos (DURAIRAJ et al., 2019) utilizado aqui como exemplo. Esse arquivo multifasta deve ser alinhado ao proteoma da goiaba conforme apresentado na **Figura 14**. O alinhamento deste estudo de caso leva cerca de 5 minutos. Como uma proteína da goiabeira alinha em várias sequencias do multifasta é necessário baixar o arquivo do alinhamento e abrir no editor de planilhas (*Excel* no *Windows*, *Calc* no *Linux* ou *Numbers* no *MAC*) para remover as duplicatas. Com as duplicatas removidas identificou-se que foram alinhadas com significância estatística 120 proteínas da goiabeira. O nome dessas proteínas foi extraído da planilha, de modo que foi inserido na caixa de busca do banco de dados separado por vírgula conforme apresenta a **Figura 15**. Em seguida as proteínas foram verificadas quanto a transcrição detalhado na **Figura 16** e quanto a presença do domínio característico de terpenos conforme apresenta a **Figura 17**.

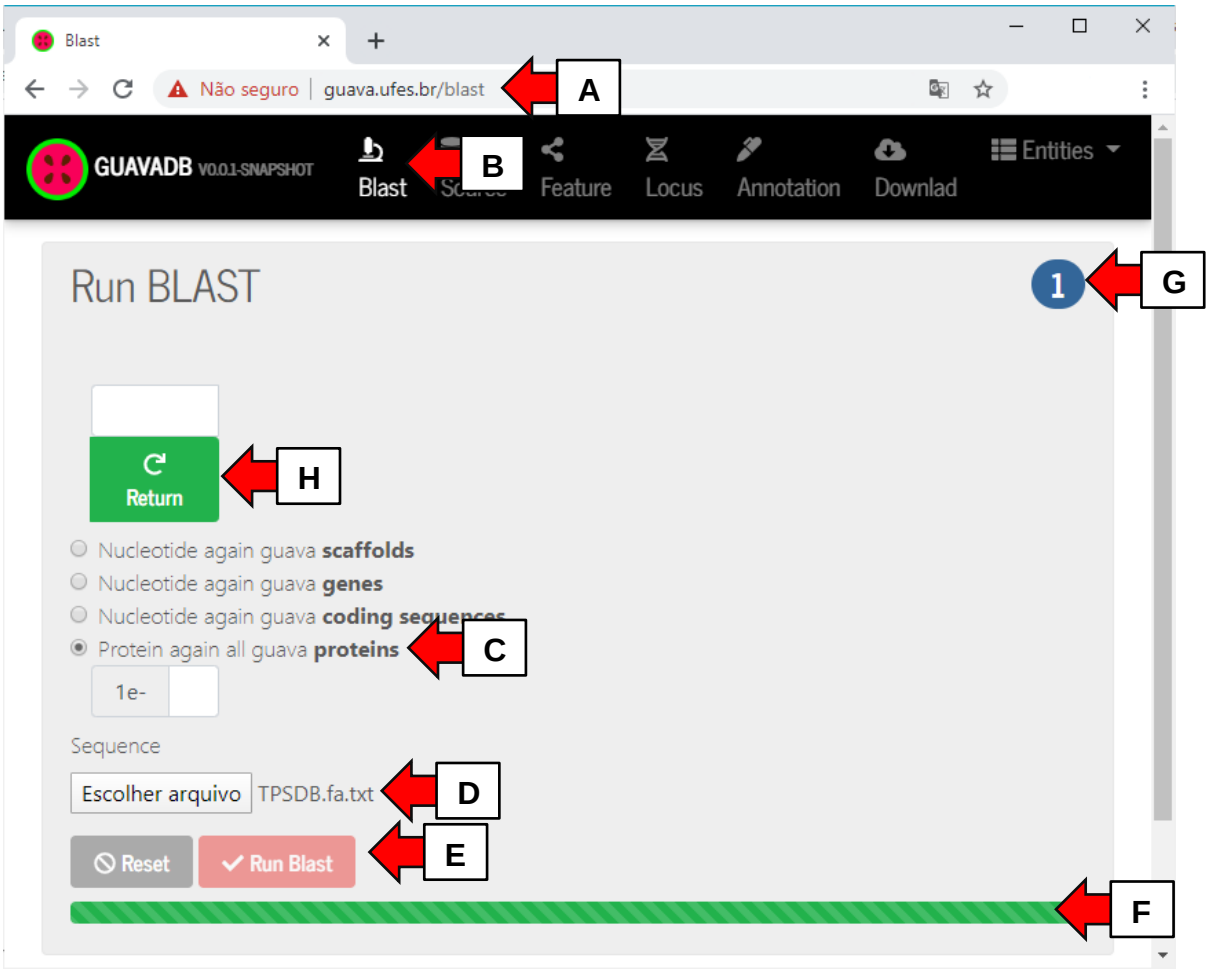

Figura 14 - Alinhamento do banco de dados de sequência de proteínas de terpenos contra as proteínas dos genes preditos em *P. guajava*. Primeiro o usuário deve acessar o banco de dados conforme a URL indicada pela seta **A**. Depois acessa a página de alinhamentos clicando no botão indicado pela seta **B**. Em seguida seleciona o conjunto de sequencias a ser alinhada (no caso deste estudo de caso será proteínas) indicado pela seta **C**. em **D** o usuário deve escolher o arquivo que contém as sequencias que deseja alinhar contra as sequencias do banco de dados. Em seguida clica no botão indicado pela seta **E** para executar o alinhamento. Enquanto o arquivo é enviado ao servidor a barra de progresso indicada pela seta **F** fica na cor rosa, em seguida sua cor fica em preto e verde intermitentemente enquanto o alinhamento é executado. Caso o usuário espere que o alinhamento seja muito demorado pode anotar o número indicado em **G** para poder sair do banco de dados e em outro momento retornar ao resultado das análises ao informar o número e clicar no botão “Return” indicado pela seta **H**.

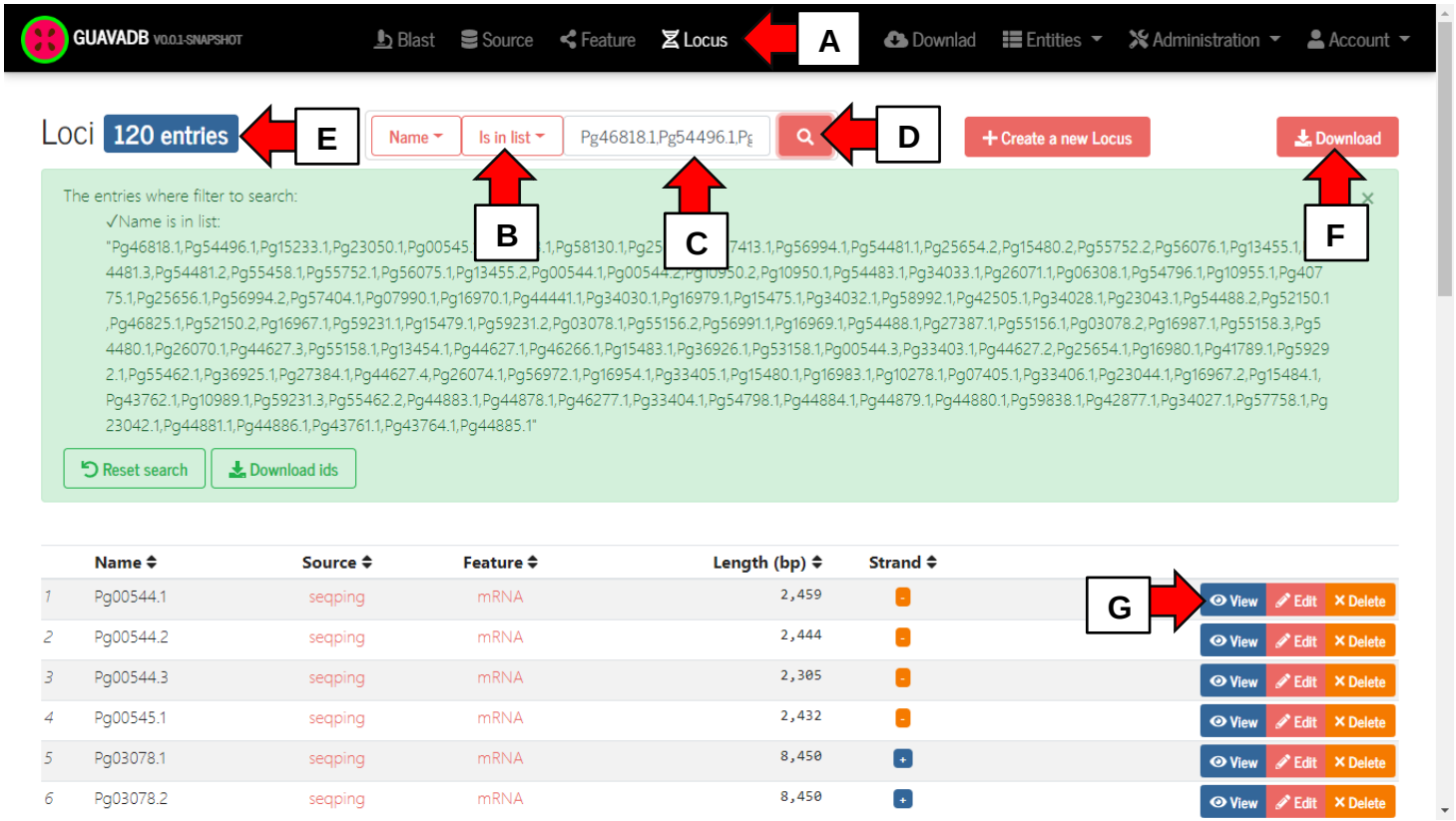

Figura 15 - Página Locus do GuavaDB utilizada no estudo de caso 1 para encontrar os 120 mRNAs que alinharam contra o arquivo multifasta. A página pode ser acessada pelo botão indicado pela seta **A**. Para pesquisar por uma lista de nomes de loci o usuário deve selecionar a opção “Está na lista” indicado pela seta **B**. Em seguida o usuário deve inserir a lista de nomes separados por virgulas no campo indicado pela seta **C** para em seguida clicar no botão pesquisar indicado pela seta **D**. com a pesquisa realizada o usuário deve verificar se todos foram encontrados, nesse estudo caso todos 120 mRNAs foram encontrados conforme apresentado pela seta **E**. O usuário pode então opcionalmente baixar um arquivo multifasta com todas as sequencias dos *loci* buscados indicado pela seta **F**. nesse estudo de caso cada loci será curado para obter um conjunto de genes de interesse, para isso cada mRNA deve ser analisado em detalhes ao clicar no botão indicado pela seta **G**.

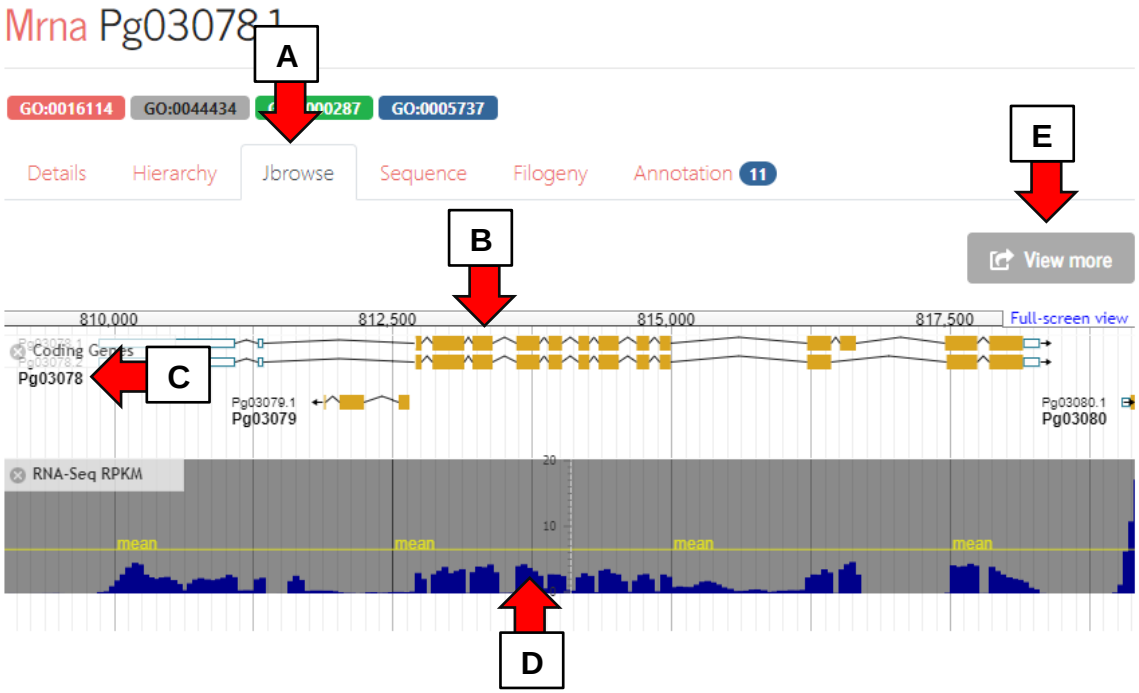

Figura 16 - Página de análise do RPKM dos mRNAs do GuavaDB. Para acessá-la basta dirigir-se a guia JBrowse indicado pela seta **A**. Após o carregamento do JBrowse o usuário pode verificar a estrutura do mRNA conforme apontado pela seta **B**. Observe-se que dois mRNAs do gene Pg03078 indicado pela seta **C** tem estruturas idênticas, com exceção de uma região CDS entre os pares de base 815.000 e 817.500 visualizável na imagem. Em **D** um histograma apresenta o RPKM que denota que este gene tem transcrição, observe-se que há transcrição apenas nas regiões codificadoras (blocos em âmbar na estrutura do gene) mas não nos íntrons e nas regiões UTR. Pelo botão indicado pela seta **E** o usuário pode contextualizar essa região do genoma com outras anotações estruturais, como: microssatélites, elementos transponíveis, mapeamento das *reads* utilizadas na montagem do genoma para verificação de cobertura do sequenciamento na região e análise de SNP entre duas cultivares para identificação de regiões polimórficas.

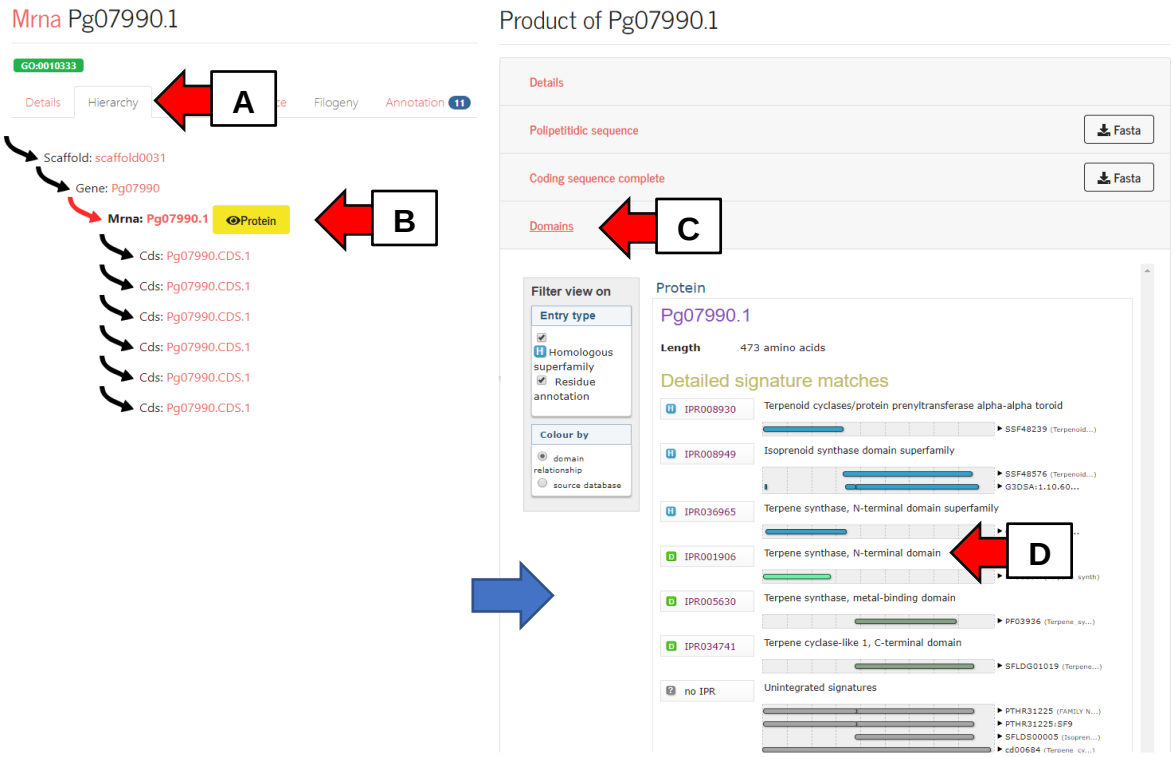

Figura 17 - Modo de acesso a página de domínios conservados de proteínas no GuavaDB. Neste estudo de caso o usuário deve acessar a aba de hierarquia do mRNA indicado pela seta **A** ainda na página de detalhes do lócus para em seguida clicar no botão que permite a visualização de detalhes da proteína indicado pela seta **B**. O usuário será então direcionado a outra página onde deve acessar as análises do InterproScan5 na seção indicada pela seta **C**. O resultado da análise é então apresentado e o usuário pode conferir os domínios conservados conforme o domínio indicado pela seta **D** que é o procurado neste estudo de caso.

5.5. Estudo de caso 2: Encontrar ortólogos no banco de dados

Este estudo de caso tenta explorar outras funcionalidades do GuavaDB ainda não discutidas, considerando ainda a necessidade de estudo mais aprofundados com genes relacionados a síntese de terpenos. Para isso é necessário realizar uma consulta por palavra chave no banco de dados de ortólogos do eggNOG (HUERTA-CEPAS et al., 2016) no GuavaDB para encontrar ortólogos ou grupos de ortólogos (OG) que tem relação com a síntese de terpenos conforme ilustra a **Figura 18**. Entre os resultados apresentados o usuário escolhe, para este estudo de caso, o OG "Terpene Synthase" a fim de encontrar ortólogos em *P. guajava* anotados com esse OG. Essa escolha leva-o do OG aos ortólogos que o eggNOG-mapper (HUERTA-CEPAS et al., 2017) identificou na Anotação funcional. Como esses ortólogos tem árvores filogenéticas reconstruídas com mRNAs de *P. guajava* é possível que o usuário navegue até o mRNA de *P. guajava* conforme orienta a **Figura 19**.

As árvores filogenéticas reconstruídas podem ser acessadas na guia "Filogeny" conforme apresentado na **Figura 20**. Outras informações podem ainda ser acessadas na guia "*Anotação*" como ortólogos do KEGG (KANEHISA et al., 2016), por exemplo. Nesse estudo de caso o mRNA Pg00544.1 tem a anotação K15797 que liga-o ao ortólogo do KEGG (KANEHISA et al., 2016) permitindo acessar outras informações relevantes no banco de dados KEGG (KANEHISA et al., 2016). Entre essas informações, a via de síntese do composto no KEGG Pathway (KANEHISA et al., 2016), por exemplo, denominada "*Sesquiterpenoid and triterpenoid biosynthesis"* pode ser acessada conforme apresentado na **Figura 21**. Outras informações ainda podem ser obtidas e contextualizadas com este ortólogo do KEGG (KANEHISA et al., 2016) que possui ortólogo identificado em *P. guajava*. Isto é interessante para contextualizar os genes relacionados a síntese de terpenos de *P. guajava* de modo a determinar seu papel, relação, influência e local de atuação na via de síntese do composto. Nesse sentido, com o GuavaDB a via de síntese de terpenos, e outras, podem ser reconstruídas para a espécie.

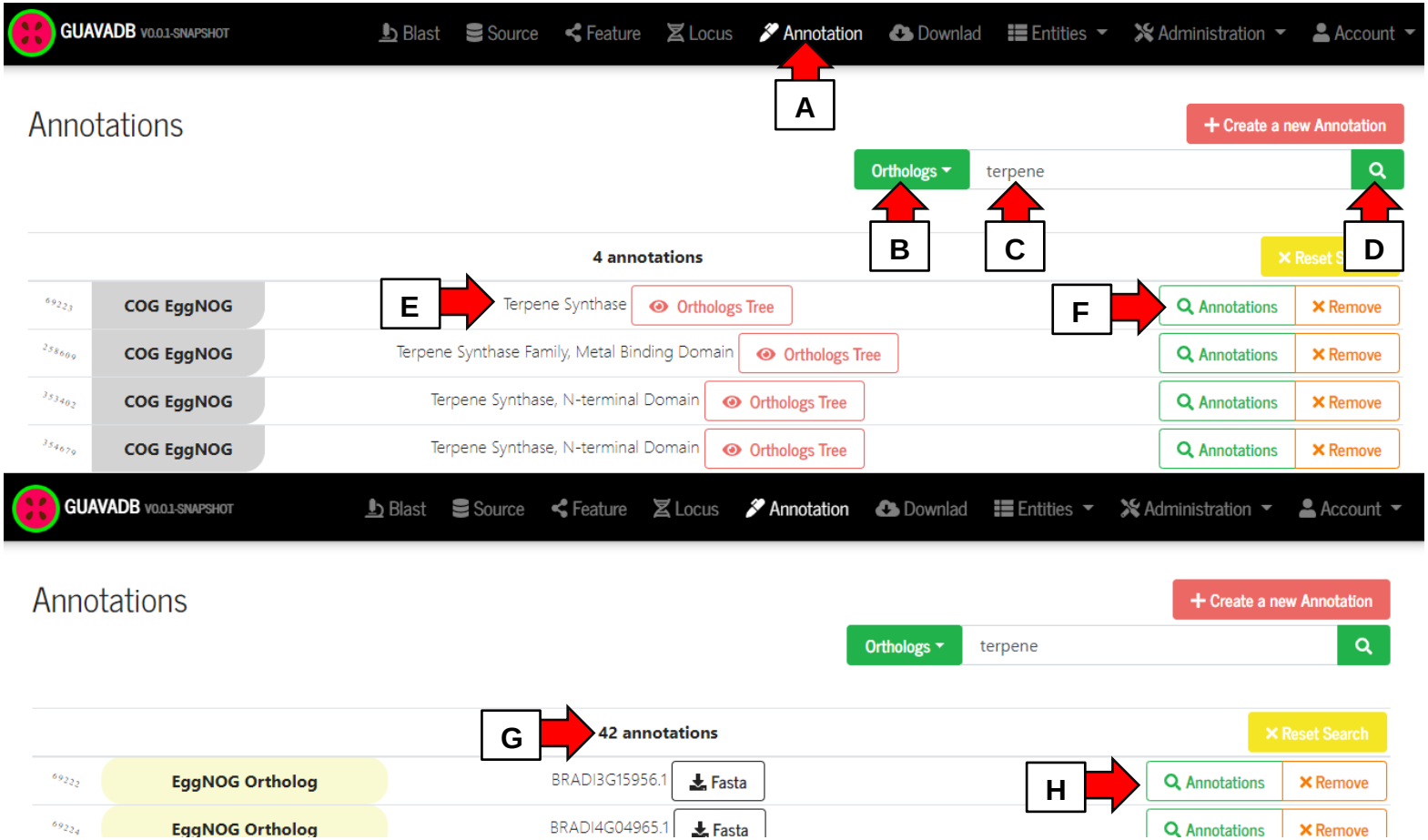

Figura 18 - Para Encontrar anotações de ortólogos ou OGs no GuavaDB o usuário deve acessar a página anotação indicada pela seta **A**. Nessa página o usuário deve selecionar no campo de pesquisa a opção "Orthologs" apontado pela seta **B** e em seguida digitar o texto "terpene", por exemplo neste estudo de caso, na caixa de texto indicada pela seta **C**. Ao pesquisar clicando no botão indicado pela seta **D** é apresentado ao usuário os resultados parcialmente listados na parte superior da figura, entre eles o de interesse nesse estudo de caso é o apontado pela seta **E**. Como são OGs do eggNOG o usuário tem a opção de ver a árvore filogenética dos ortólogos do eggNOG que compõem esse OG ou pode também acessar anotações deste OG clicando no botão indicado pela seta **F**, que resultará na tela apresentada na parte inferior da figura. Nessa tela aparecem 42 ortólogos que contém o OG pesquisado conforme apontado pela seta **G**. O usuário pode então pesquisar por anotações que contém esses ortólogos como, por exemplo, árvores filogenéticas clicando no botão indicado pela seta **H**.

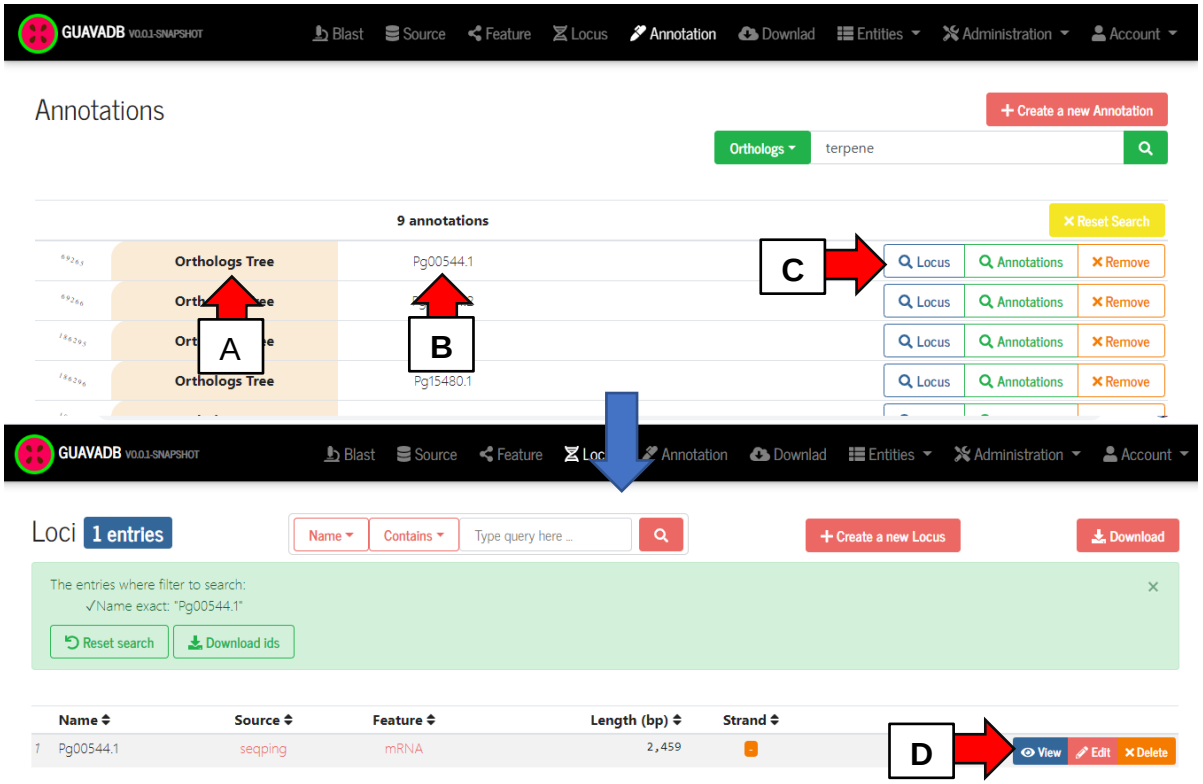

Figura 19 - Encontrar o *locus* com a ferramenta de pesquisa da página de anotações é possível ao navegar entre as anotações interligadas. Nesse estudo de caso a partir da anotação de OG de síntese de terpenos foram encontrados ortólogos, e, a partir dos ortólogos foram encontradas as árvores filogenéticas listadas na coluna indicada pela seta **A** na tela apresentada na parte superior desta figura. A coluna indicada pela seta **B** da tabela de resultados é o nome do Locus (mRNAs nesse caso) de *P. guajava* que está na arvore filogenética. Para visualizar a árvore o usuário deve clicar no botão Locus indicado pela seta **C** que redirecionará o usuário para a tela apresentada na parte inferior desta figura. Nessa tela o usuário poderá acessar detalhes do Locus clicando no botão “View” indicado pela seta **D**.

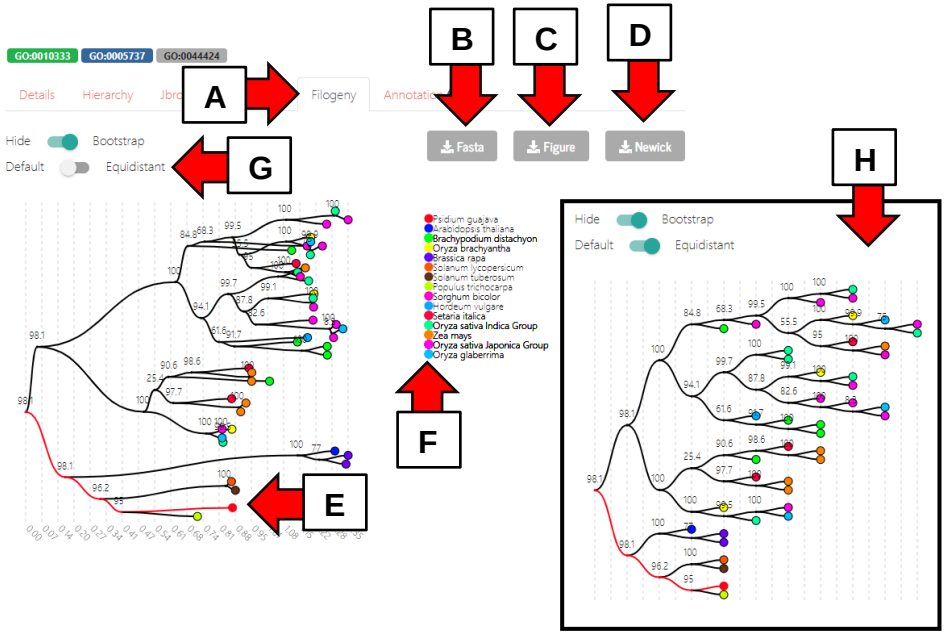

Figura 20 - Visualização da árvore filogenética na página de detalhes do mRNA de exemplo neste estudo de caso. Para acessá-la o usuário deve dirigir-se a guia “Filogeny” onde a arvore será exibida indicada pela seta **A**, caso a guia esteja desabilitada é devido ao eggNOG-mapper não ter identificado ortólogo para o mRNA em suas análises. Estas arvores foram geradas conforme citado nesta seção. O usuário pode opcionalmente baixar um arquivo fasta que contém sequências de proteínas de todos ortólogos apresentados na árvore filogenética indicado pela seta **B**, baixar o arquivo de imagem da árvore conforme indicado pela seta **C** ou ainda baixar o arquivo Newik clicando no botão indicado pela seta **D**. O arquivo Newik permite abrir a árvore filogenética em outros softwares de análises filogenéticas. O ortólogo de *P. guajava* é o indicado pela seta **E**. Ao mover o mouse sobre um nó folha o nome do ortólogo aparecerá e o círculo que indica a espécie do ortólogo irá aumentar seu diâmetro para identificação da espécie na legenda indicada pela seta **F.** A árvore exibe a distância dos nós de acordo com os passos evolutivos, o que pode ocasionar muita sobreposição dos nós folha dificultando entender as relações. Para resolver isso o usuário pode ativar a chave indicada pela seta **G** que faz com que as distancias entre os nós sejam proporcionais, desconsiderando a distância evolutiva, de modo que a árvore é apresentada conforme a figura indicada pela seta **H**.

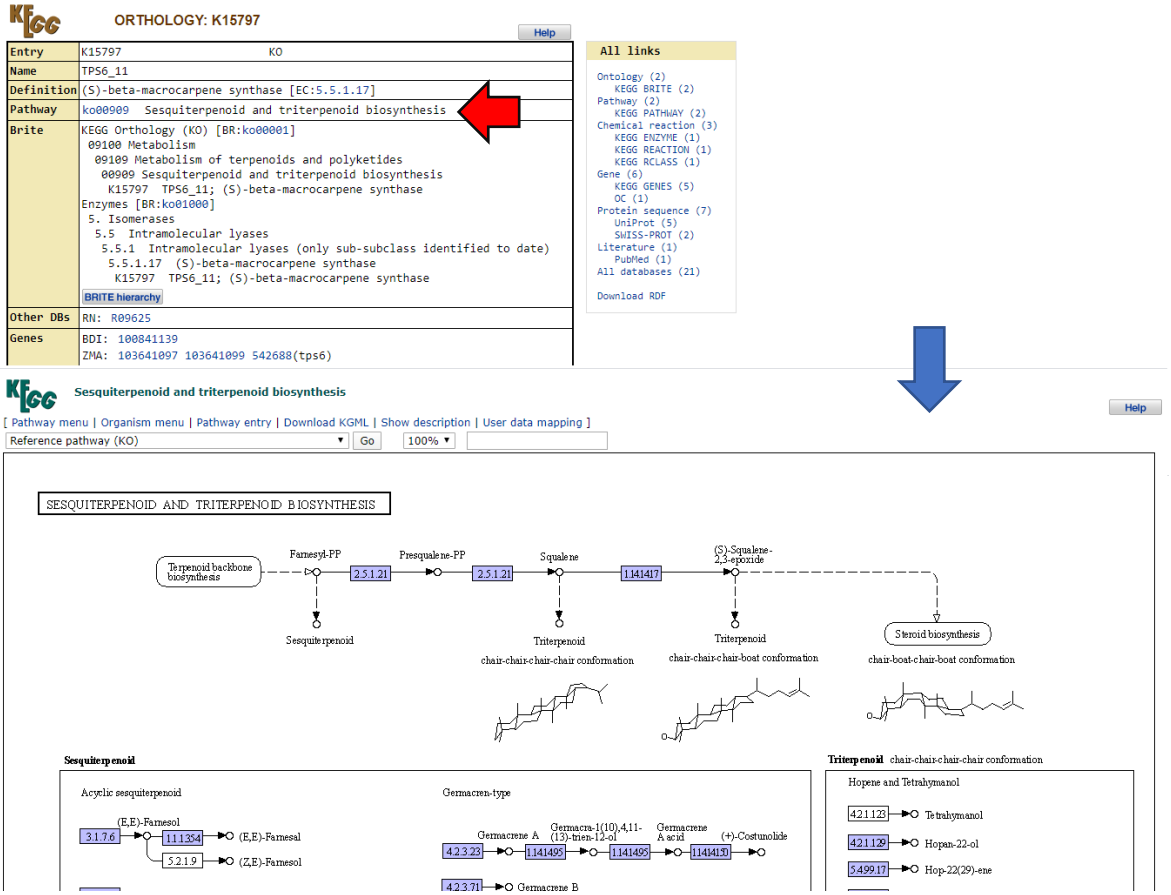

Figura 21 - Ortólogos do KEGG anotados aos mRNAs de *P. guajava*. Essa análise estendida a este estudo de caso permite por meio do ortólogo anotado ao mRNA de *P. guajava* encontrar outras informações no banco de dados KEGG, como por exemplo outros genes de outras espécies, um possível nome para o Locus ou ainda uma via de síntese associada ao ortólogo conforme, nesse caso, apontado pela seta vermelha. Ao clicar nesse link o usuário é direcionado para outra página dentro do banco de dados KEGG Pathways onde pode verificar em detalhes a via de síntese a que o ortólogo tem função relacionada.

6.1. Genômica Comparativa

As análises de conteúdo proteico compartilhado entre espécies no OrthoVenn2 (XU et al., 2019) permitiram identificar para o conjunto de genes preditos 2.994 genes cópia única, 24.868 genes com parálogos no genoma, 18.886 genes *housekeeping* e 15.493 genes ortólogos em *A. thaliana* ou *E. grandis* ou *V. vinifera*. As análises foram realizadas com os proteomas de *P. guajava* deste trabalho, de *Eucalyptus grandis* v2 obtido do Phytozome (GOODSTEIN et al., 2012)

e os demais disponibilizados no próprio OrthoVenn2, parâmetros *evalue* 1e-5 e os demais padrão. A **Figura 22** é gráfico de *Venn* que apresenta o conteúdo proteico compartilhado entre estas espécies.

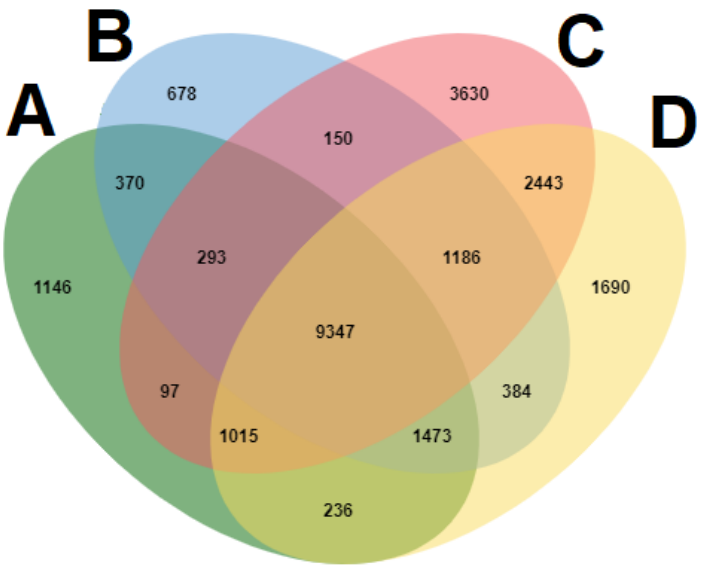

Figura 22 - Diagrama de Venn do conteúdo gênico compartilhado entre (**A**) *A. thaliana*, (**B**) *V. vinifera*, (**C**) *P. guajava* e (**D**) *E. grandis*. A grande quantidade de genes ortólogos com as demais espécies, em especial com *E. grandis* que é mais próximo evolutivamente da *P. guajava* denota a proximidade genética entre estes organismos.

Nas análises de enriquecimento de GO do conjunto de genes *housekepping* em *P. guajava* foram identificados 10 GOs enriquecidos. Entre os 10 termos GOs enriquecidos o termo *GO:0016114* "*Terpenoid biosynthetic process*" é o mais interessante por referir-se a um composto que desempenha papel importante em processos biológicos da planta como defesa contra patógenos, resposta a estreses abióticos e desenvolvimento (KARUNANITHI; ZERBE, 2019), o que implica na necessidade de estudar melhor essa família genes em *P. guajava*.

Como em *E. grandis* os genes da família de terpenos de mesma classe ocorrem agrupados (KÜLHEIM et al., 2015) é importante verificar a ocorrência desses genes identificados em *P. guajava*. Essa verificação pode ser realizada com uma análise de sintenia pelo o software MCscan (WANG, Y. et al., 2012), bastando marcar os genes da família de terpenos. O resultado desta análise é exibido na **Figura 23** onde é possível verificar 7 grupos de genes da família de terpenos que apresentam essa disposição na goiaba, sendo que para dois grupos o mesmo comportamento se repete ainda para *V. vinífera*.

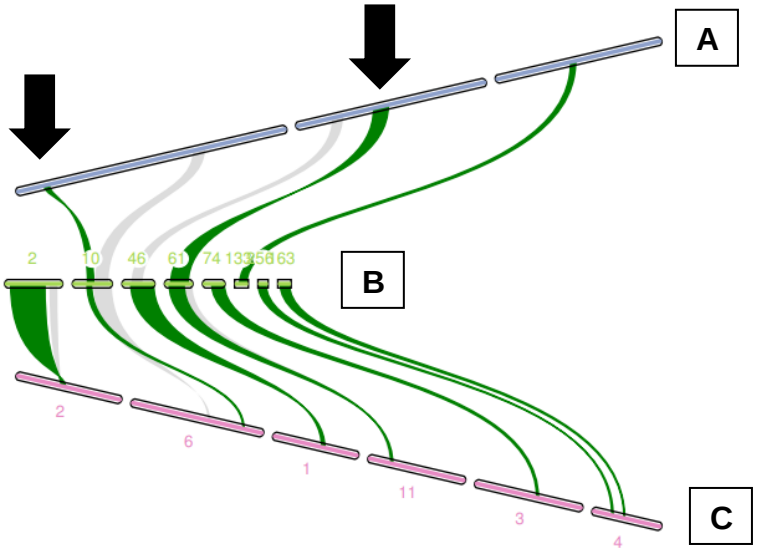

Figura 23 - Blocos sintênicos entre *V. vinifera* (**A**), *P. guajava* (**B**) e *E. grandis* (**C**) apresentam vários genes da família de terpenos dispostos agrupados. Cada traço desenhado para as espécies representa um fragmento de cromossomo ou scaffold onde ocorre regiões sintênicas. As regiões sintênicas são os links em cinza ou em verde entre os traços, sendo estes últimos as regiões sintênicas que possuem genes de síntese de terpenos. Cada link é um grupo de cerca de 30 genes ortólogos entre as espécies que determinam as regiões em sintenia entre elas. Duas regiões sintênicas que possuem mais de um gene da família de terpenos (links em verde) ocorrem em comum entre as três espécies conforme indicado pelas setas pretas.

7. DISCUSSÃO

A abordagem utilizada neste trabalho para montagem do genoma foi desenvolvida com experiencias entre os montadores aplicados. O montador Flye (KOLMOGOROV et al., 2019) por sua vez apresentou-se ser rápido e aplicou técnicas do estado da arte em montagem de genomas para tratar as regiões repetitivas da montagem. Apesar disso, o Flye não monta reads pareadas (em geral de tecnologia Illumina) de modo que o Masurca (ZIMIN et al., 2013) teve que ser combinado na abordagem apresentada neste trabalho.

A montagem ficou contígua (sem N's) contudo fragmentada em milhares de sequências, o que afeta o resultado da ancoragem dos contigs no mapa genético com o ALLMAPS (TANG et al., 2015) tornando-a experimental. Outro fator desafiador para a ancoragem foi alinhamento dos scaffolds contra os marcadores do mapa genético. Quando os marcadores são utilizados para ancoragem em montagens demasiadamente fragmentadas o tamanho de cerca de 30 pares de bases e a reduzida quantidade deles por grupo de ligação mitigam a resolução e a significância estatística. Por outro lado, o genoma montado é de excelente qualidade por utilizar sequenciamento de distintas cultivares. A começar pelo tamanho montado, quando comparado ao esperado por citometria de fluxo de 465 Mb (MARQUES et al., 2016) a montagem cobre 84%. Sendo que sua completude foi validada pelo BUSCO v3 (WATERHOUSE et al., 2018) usando o banco de dados para plantas, resultando em apenas 15% de genes ausentes. O mapeamento do sequenciamento Illumina contra o genoma resulta em 89% do das reads mapeadas.

Na anotação estrutural o REPET (FLUTRE et al., 2011), comparado aos demais, é o mais oneroso para instalar (levando em conta suas dependências) e utilizar. Isso deve-se principalmente a utilização de softwares legados pelo pipeline, para os quais não há mais suporte, afetando negativamente sua compatibilidade com sistemas em constante atualização. Como se não bastasse foi o mais lento para concluir as análises. Entretanto foi o mais sensitivo com relação a todas classes de elementos que identificou, isto é, elementos transponíveis e microssatélites. De forma que, mesmo sendo o mais oneroso os resultados foram melhores que a execução manual dos softwares que o pipeline executa (*Repeat*

Masker, *Repeat Modeler*, *Tandem Repeat Finder*, entre outros) e que noutro modo seriam executados manualmente.

O conjunto de genes preditos pelo pipeline Seqping (CHAN et al., 2017) em sua última etapa (7) foram em número condizente com a quantidade predita em outras espécies de plantas, contudo a verificação com o BUSCO (WATERHOUSE et al., 2018) acusava muitos genes faltantes. Como os resultados do BUSCO são importantes para definir a completude do genoma e o pipeline apresenta bons resultados com relação a esta validação no seu *benchmarking* (CHAN et al., 2017) foi realizado uma inspeção detalhada no resultado de cada etapa do pipeline. Nessa inspeção a avaria foi identificada na correção do GFF que ocorre no fim da etapa 7 já que na etapa 6 os resultados da análise do Busco eram coerentes. Com a análise do código verificou-se que um comando limitava em 10 estruturas (éxons, UTRs e CDS) por mRNA. No entanto muitos mRNAs têm mais de 10 estruturas preditas. O problema foi encontrado no fim da linha de código 888 do script *seqping.sh* que é apresentada parcialmente a seguir:

```
grep "${line}" $abspath_MTrainDIR/output/maker.gff3 | head >>
```

O comando **head** define que apenas as 10 primeiras linhas resultantes da instrução anterior devem ser emitidas à frente, isso fez com que várias estruturas de regiões codificadoras foram perdias, atrapalhando na verificação do BUSCO. Para resolver foi necessário remover o trecho `| head` desta linha de código. Com o problema resolvido o resultado da predição de genes com evidência de transcritos do pipeline ficou coerente.

6.1. Banco de dados para plantas

Em geral os bancos de dados que são desenvolvidos para um determinado genoma de planta, não são criados pensando na necessidade de um sistema, como por exemplo: alface (https://lgr.genomecenter.ucdavis.edu/), café (http://coffee-genome.org/) e arabidópsis (https://www.arabidopsis.org/). Isso deve-se ao fato de o foco ser disponibilizar os arquivos de dados para download desconsiderando futuras aplicações para o banco de dados como, por exemplo, o acesso programático por API. Nesse sentido, estes bancos de dados são portais que apesar de funcionais são bem restritos a extensões modulares de modo que

princípios e padrões de desenvolvimento de software (GAMMA et al., 1994; MARTIN, 2003) são parcialmente observados. No caso do sistema GuavaDB, desde o protótipo descartável construído para projeto do sistema atual vantagens têm se manifestado por ser arquitetado como sistema, entre elas: gestão de cadastro de usuários e acesso interativo aos dados, que são funcionalidades geralmente indisponíveis em portais.

Assim o GuavaDB: pelo ponto de vista tecnológico o sistema possui uma API *web,* de modo que novos módulos podem ser integrados facilmente permitindo que novas funcionalidades sejam incrementadas sem prejuízo de código fonte. Pelo lado administrativo o sistema permite monitoramento de seu estado em tempo real, controle de nível de logs, gestão de usuários entre outras funcionalidades importantes. Já no ponto de vista o usuário final a usabilidade e simplicidade para acessar traduz-se em acesso fácil e irrestrito aos dados persistidos nele.

8. CONCLUSÃO

O rascunho do genoma de *P. guajava* disponibilizado neste trabalho vem de uma montagem consenso de três cultivares: Paluma, Cortibel e Zhenzhu, e, híbrida de duas tecnologias que são Illumina e PacBio. A Estrutura do genoma foi anotada com a identificação de elementos repetitivos, RNA não codificador de proteínas e genes codificadores de proteínas com validação de transcritos. A disponibilização desta anotação estrutural na figura circular e no JBrowse permite comparar graficamente o genoma montado contra o genoma de outras espécies.

O rascunho do genoma montado torna-se aqui a principal fonte de informação para estudos com a genômica da espécie. Em programas de melhoramento com goiabeiras, por exemplo, genes com anotação funcional disponibilizados neste trabalho podem ser selecionados por suas relações com características de interesse. Esses genes podem ser identificados visando melhorar variedades para obter genótipos com maior produtividade e mais resistência a estresse.

As diferentes estratégias e bancos de dados utilizados para realizar a anotação funcional culminaram em 441.111 registros de anotação funcional no banco de dados. Eles estão interligados e classificados em taxonomia, proteínas, ortólogos, ontologia genica, famílias de proteínas e vias de síntese de metabolitos. Como essas classificações possuem vários níveis hierárquicos, o modelo relacional muitos-para-muitos utilizado para persistir esses dados poderia ser substituído por grafo. Desse modo, a ontologia de anotação funcional implícita neste trabalho poderia ser apresentada a comunidade cientifica.

O GuavaDB é maior fonte de informações ômicas para *P. guajava*. É ainda uma ferramenta para consulta das informações, permitindo acessá-las de maneira intuitiva, analisá-las de maneira fácil e obtê-las sem a necessidade linhas de comando no terminal. O código fonte do banco de dados foi escrito considerando futuro reuso em outros rascunhos de genoma, bastando algumas customizações opcionais. Dessa forma ele porta-se como um modelo que pode ser utilizado em outro software para gerar bancos de dados biológicos programaticamente.

9. REFERÊNCIAS

AFGAN, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. **Nucleic Acids Research**, v. 46, n. W1, p. W537–W544, 2 jul. 2018. Disponível em: <https://academic.oup.com/nar/article/46/W1/W537/5001157>. Acesso em: 28 jan. 2020.

AGARWALA, R. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 46, n. D1, p. D8–D13, 1 jan. 2018.

ALPTEKIN, B.; AKPINAR, B. A.; BUDAK, H. A comprehensive prescription for plant miRNA identification. **Frontiers in Plant Science**, v. 7, 24 jan. 2017.

ASHBURNER, M. et al. **Gene ontology: Tool for the unification of biology**. **Nature Genetics**. [S.l: s.n.]. , maio 2000

BAIROCH, A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Research**, v. 28, n. 1, p. 45–48, 1 jan. 2000.

BARABASCHI, D. et al. Next generation breeding. **Plant Science**, v. 242, p. 3–13, 17 abr. 2015.

BATEMAN, A. UniProt: A worldwide hub of protein knowledge. **Nucleic Acids Research**, v. 47, n. D1, p. D506–D515, 8 jan. 2019.

BEIER, S. et al. MISA-web: a web server for microsatellite prediction. **Bioinformatics (Oxford, England)**, v. 33, n. 16, p. 2583–2585, 15 ago. 2017.

BELSER, C. et al. **Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps**. **Nature Plants**. [S.l.]: Palgrave Macmillan Ltd. , 1 nov. 2018

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.

BOLGER, M. E.; ARSOVA, B.; USADEL, B. Plant genome and transcriptome

annotations: from misconceptions to simple solutions. **Briefings in bioinformatics**, v. 19, n. 3, p. 437–449, 2018a. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28062412>. Acesso em: 15 set. 2019.

BOLGER, M. E.; ARSOVA, B.; USADEL, B. Plant genome and transcriptome annotations: From misconceptions to simple solutions. **Briefings in Bioinformatics**, v. 19, n. 3, p. 437–449, 1 maio 2018b.

BROWN, J.; PIRRUNG, M.; MCCUE, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. **Bioinformatics (Oxford, England)**, 9 jun. 2017.

BUCHFINK, B.; XIE, C.; HUSON, D. H. **Fast and sensitive protein alignment using DIAMOND**. **Nature Methods**. [S.l.]: Nature Publishing Group. , 1 jan. 2014

BUELS, R. et al. JBrowse: A dynamic web platform for genome visualization and analysis. **Genome Biology**, v. 17, n. 1, 12 abr. 2016.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC bioinformatics**, v. 10, p. 421, 15 dez. 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20003500>. Acesso em: 15 set. 2019.

CASPI, R. et al. The MetaCyc database of metabolic pathways and enzymes. **Nucleic Acids Research**, v. 46, n. D1, p. D633–D639, 4 jan. 2018. Disponível em: <http://academic.oup.com/nar/article/46/D1/D633/4559117>. Acesso em: 28 jan. 2020.

CHAN, K. L. et al. Seqping: Gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data. **BMC Bioinformatics**, v. 18, 27 jan. 2017.

CHEN, C.; HUANG, H.; WU, C. H. Protein bioinformatics databases and resources. **Methods Mol. Biol.** [S.l.]: Humana Press Inc., 2017. v. 1558. p. 3–39.

CONESA, A. et al. **A survey of best practices for RNA-seq data analysis**. **Genome Biology**. [S.l.]: BioMed Central Ltd. , 26 jan. 2016

DECHEN, S. C. F.; POMMER, C. V. Goiaba no mundo. **O AGRONÔMICO**,

p. 75, 2006.

DURAIRAJ, J. et al. An analysis of characterized plant sesquiterpene synthases. **Phytochemistry**, v. 158, p. 157–165, 1 fev. 2019.

EL-GEBALI, S. et al. The Pfam protein families database in 2019. **Nucleic acids research**, v. 47, n. D1, p. D427–D432, 8 jan. 2019. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/30357350>. Acesso em: 15 set. 2019.

FLUTRE, T. et al. Considering transposable element diversification in de novo annotation approaches. **PloS one**, v. 6, n. 1, p. e16526, 31 jan. 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21304975>. Acesso em: 15 set. 2019.

GAMMA, E. et al. **Design Patterns: Elements of Reusable Object-Oriented Software**. [S.l.]: Addison-Wesley Professional, 1994.

GEIB, S. M. et al. Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. **GigaScience**, v. 7, n. 4, p. 1–5, 1 abr. 2018.

GOODSTEIN, D. M. et al. Phytozome: A comparative platform for green plant genomics. **Nucleic Acids Research**, v. 40, n. D1, jan. 2012.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. 2016 REVIEW 10 yrs NGS Nat Rev Genet Goodwin. **Nature reviews. Genetics**, v. 17, n. 6, p. 333–351, 2016. Disponível em: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=27184599&retmode=ref&cmd=prlinks%0Afile:///Users/sfrench/Dropbox/Papers/Papers2/Library.papers3/Files/F3/F3A4008C-A5EB-4061-87E9-E756E13A364D.pdf%0Apapers3://publication/doi/10.1038/>. Acesso em: 15 set. 2019.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature Biotechnology**, v. 29, n. 7, p. 644–652, jul. 2011.

GRIFFITHS-JONES, S. et al. **Rfam: An RNA family database**. **Nucleic**

Acids Research. [S.l: s.n.]. , 1 jan. 2003

GUREVICH, A. et al. QUAST: quality assessment tool for genome assemblies. **Bioinformatics (Oxford, England)**, v. 29, n. 8, p. 1072–5, 15 abr. 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23422339>. Acesso em: 27 set. 2019.

HARDISON, R. C. Comparative Genomics. **PLoS Biology**, v. 1, n. 2, p. e58, 17 nov. 2003. Disponível em: <https://dx.plos.org/10.1371/journal.pbio.0000058>. Acesso em: 17 nov. 2019.

HOFF, K. J.; STANKE, M. Predicting Genes in Single Genomes with AUGUSTUS. **Current Protocols in Bioinformatics**, p. e57, 22 nov. 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.57>. Acesso em: 27 jan. 2020.

HOLT, C.; YANDELL, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. **BMC Bioinformatics**, v. 12, n. 1, 22 dez. 2011.

HUERTA-CEPAS, J. et al. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. **Nucleic Acids Research**, v. 44, n. D1, p. D286–D293, 2016.

_____. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. **Molecular Biology and Evolution**, v. 34, n. 8, p. 2115–2122, 1 ago. 2017.

HUERTA-CEPAS, J.; SERRA, F.; BORK, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. **Molecular Biology and Evolution**, v. 33, n. 6, p. 1635–1638, jun. 2016. Disponível em: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw046>. Acesso em: 19 jan. 2020.

JASSAL, B. et al. The reactome pathway knowledgebase. **Nucleic acids research**, 6 nov. 2019. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/31691815>. Acesso em: 28 jan. 2020.

JENSEN, L. J. et al. eggNOG: automated construction and annotation of orthologous groups of genes. **Nucleic acids research**, v. 36, n. Database issue, p. D250-4, jan. 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17942413>. Acesso em: 15 set. 2019.

JIANG, S.-Y. et al. A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns. **Genome Biology and Evolution**, v. 11, n. 8, p. 2078–2098, 1 ago. 2019. Disponível em: <https://academic.oup.com/gbe/article/11/8/2078/5532225>. Acesso em: 29 jan. 2020.

JIMÉNEZ-ESCRIG, A. et al. Guava fruit (Psidium guajava L.) as a new source of antioxidant dietary fiber. **Journal of Agricultural and Food Chemistry**, v. 49, n. 11, p. 5489–5493, 2001.

JO, S. et al. Complete plastome sequence of Psidium guajava L. (Myrtaceae). **Mitochondrial DNA Part B: Resources**, v. 1, n. 1, p. 612–614, 2016.

JONES, P. et al. InterProScan 5: Genome-scale protein function classification. **Bioinformatics**, v. 30, n. 9, p. 1236–1240, 1 maio 2014.

KANEHISA, M. et al. KEGG as a reference resource for gene and protein annotation. **Nucleic acids research**, v. 44, n. D1, p. D457-62, 4 jan. 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/26476454>. Acesso em: 15 set. 2019.

KARUNANITHI, P. S.; ZERBE, P. **Terpene Synthases as Metabolic Gatekeepers in the Evolution of Plant Terpenoid Chemical Diversity**. **Frontiers in Plant Science**. [S.l.]: Frontiers Media S.A. , 1 out. 2019

KIM, D. et al. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. **Genome Biology**, v. 14, n. 4, 25 abr. 2013.

KING, Z. A. et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. **Nucleic Acids Research**, v. 44, n. D1, p. D515–D522, 4 jan. 2016. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1049>. Acesso em: 28 jan. 2020.

KIST, B. B. et al. **Anuário brasileiro da fruticultura 2018**. . [S.l: s.n.], 2018.

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

KORF, I. Gene finding in novel genomes. **BMC bioinformatics**, v. 5, p. 59, 14 maio 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15144565>. Acesso em: 27 set. 2019.

KRZYWINSKI, M. et al. Circos: An information aesthetic for comparative genomics. **Genome Research**, v. 19, n. 9, p. 1639–1645, set. 2009.

KÜLHEIM, C. et al. The Eucalyptus terpene synthase gene family. **BMC Genomics**, v. 16, n. 1, p. 450, 11 dez. 2015. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1598-x>. Acesso em: 29 jan. 2020.

KULKARNI, P.; FROMMOLT, P. Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. **Computational and structural biotechnology journal**, v. 15, p. 471–477, 2017. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29158876>. Acesso em: 26 set. 2019.

KYRIAKIDOU, M. et al. Current Strategies of Polyploid Plant Genome Sequence Assembly. **Frontiers in plant science**, v. 9, p. 1660, 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/30519250>. Acesso em: 26 set. 2019.

LAGESEN, K. et al. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Research**, v. 35, n. 9, p. 3100–3108, maio 2007.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, abr. 2012.

LANTZ, H. et al. Ten steps to get started in Genome Assembly and Annotation. **F1000Research**, v. 7, 2018.

LASLETT, D.; CANBACK, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. **Nucleic acids research**, v. 32, n. 1, p. 11–6, 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/14704338>. Acesso em: 27 set. 2019.

LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The sequence read archive. **Nucleic Acids Research**, v. 39, n. SUPPL. 1, jan. 2011.

LI, F.-W.; HARKESS, A. A guide to sequence your favorite plant genomes. **Applications in plant sciences**, v. 6, n. 3, p. e1030, mar. 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29732260>. Acesso em: 26 set. 2019.

LI, J. R. et al. Plant stress RNA-seq Nexus: A stress-specific transcriptome database in plant cells. **BMC Genomics**, v. 19, n. 1, 27 dez. 2018.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. **Nucleic Acids Research**, v. 25, n. 5, p. 955–964, 1 mar. 1997. Disponível em: <https://academic.oup.com/nar/article/25/5/955/5133591>. Acesso em: 27 set. 2019.

LUCAS, S. J.; BUDAK, H. Sorting the wheat from the Chaff: Identifying miRNAs in genomic survey sequences of Triticum aestivum chromosome 1AL. **PLoS ONE**, v. 7, n. 7, 17 jul. 2012.

MAJOROS, W. H.; PERTEA, M.; SALZBERG, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. **Bioinformatics (Oxford, England)**, v. 20, n. 16, p. 2878–9, 1 nov. 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15145805>. Acesso em: 27 set. 2019.

MARQUES, A. M. et al. Refinement of the karyological aspects of Psidium guineense (Swartz, 1788): a comparison with Psidium guajava (Linnaeus, 1753). **Comparative cytogenetics**, v. 10, n. 1, p. 117–28, 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/27186342>. Acesso em: 25 set. 2019.

MARTIN, R. C. **Agile Software Development: Principles, Patterns, and Practices**. USA: Prentice Hall PTR, 2003.

MCCARTHY, F. M. et al. AgBase: a functional genomics resource for agriculture. **BMC genomics**, v. 7, p. 229, 8 set. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16961921>. Acesso em: 15 set. 2019.

MI, H. et al. Protocol Update for large-scale genome and gene function

analysis with the PANTHER classification system (v.14.0). **Nature protocols**, v. 14, n. 3, p. 703–721, 2019. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/30804569>. Acesso em: 15 set. 2019.

MOORE, G. P. The C-Value Paradox. **BioScience**, v. 34, n. 7, p. 425–429, jul. 1984. Disponível em: <https://academic.oup.com/bioscience/article-lookup/doi/10.2307/1309631>. Acesso em: 17 nov. 2019.

MOREIRA, L. M. Ciências genômicas: fundamentos e aplicações. 1ªedição. **Sociedade Brasileira de Genética**, 2015.

MUKHERJEE, K. et al. Error correcting optical mapping data. **GigaScience**, v. 7, n. 6, 1 jun. 2018. Disponível em: <https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy061/5005021>. Acesso em: 27 jan. 2020.

NAWROCKI, E. P.; EDDY, S. R. Infernal 1.1: 100-fold faster RNA homology searches. **BIOINFORMATICS APPLICATIONS**, v. 29, n. 22, p. 2933–2935, 2013. Disponível em: <http://infernal.janelia.org.>. Acesso em: 29 out. 2019.

NIMISHA, S. et al. **Molecular breeding to improve guava (Psidium guajava L.): Current status and future prospective**. **Scientia Horticulturae**. [S.l.]: Elsevier. , 17 dez. 2013

NUMANAGIĆ, I. et al. Fast characterization of segmental duplications in genome assemblies. 1 set. 2018, [S.l.]: Oxford University Press, 1 set. 2018. p. i706–i714.

O'LEARY, N. A. et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. **Nucleic Acids Research**, v. 44, n. D1, p. D733–D745, 2016.

PAAJANEN, P. et al. A critical comparison of technologies for a plant genome sequencing project. **GigaScience**, v. 8, n. 3, 9 jan. 2019.

PADMAKAR, B. et al. Development of SRAP and SSR marker-based genetic linkage maps of guava (Psidium guajava L.). **Scientia Horticulturae**, v. 192, p. 158–165, 1 ago. 2015.

PATEL, R. K.; JAIN, M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. **PLoS ONE**, v. 7, n. 2, 1 fev. 2012.

PRIYA, P. et al. Terzyme: a tool for identification and analysis of the plant terpenome. **Plant Methods**, v. 14, n. 1, p. 4, 10 dez. 2018. Disponível em: <https://plantmethods.biomedcentral.com/articles/10.1186/s13007-017-0269-0>. Acesso em: 29 jan. 2020.

RASHEED, A. et al. **Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives**. **Molecular Plant**. [S.l.]: Cell Press. , 7 ago. 2017

RAVI, K.; DIVYASHREE, P. **Psidium guajava: A review on its potential as an adjunct in treating periodontal disease**. **Pharmacognosy Reviews**. [S.l.]: Medknow Publications. , 2014

SAYERS, E. W. et al. GenBank. **Nucleic Acids Research**, v. 47, n. D1, p. D94–D99, 8 jan. 2019.

SCHWACKE, R. et al. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. **Molecular Plant**, v. 12, n. 6, p. 879–892, 3 jun. 2019.

SEBRAE. **O cultivo e o mercado da goiaba | Sebrae**. Disponível em: <http://www.sebrae.com.br/sites/PortalSebrae/artigos/o-cultivo-e-o-mercado-da-goiaba,d3aa9e665b182410VgnVCM100000b272010aRCRD>. Acesso em: 14 set. 2019.

SOHN, J.-I.; NAM, J.-W. The present and future of de novo whole-genome assembly. **Briefings in bioinformatics**, v. 19, n. 1, p. 23–40, 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/27742661>. Acesso em: 15 set. 2019.

STANKE, M.; WAACK, S. Gene prediction with a hidden Markov model and a new intron submodel. 2003, [S.l: s.n.], 2003.

SWEENEY, B. A. et al. RNAcentral: A hub of information for non-coding RNA sequences. **Nucleic Acids Research**, v. 47, n. D1, p. D221–D229, 2019.

TANG, H. et al. ALLMAPS: Robust scaffold ordering based on multiple maps.

Genome Biology, v. 16, n. 1, 13 jan. 2015.

_____. **Synteny and collinearity in plant genomes**. **Science**. [S.l: s.n.]. , 25 abr. 2008

TUTAR, Y. **Pseudogenes**. **Comparative and Functional Genomics**. [S.l: s.n.]. , 2012

WANG, Y. et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. **Nucleic Acids Research**, v. 40, n. 7, abr. 2012.

WANG, Zhong; GERSTEIN, M.; SNYDER, M. **RNA-Seq: A revolutionary tool for transcriptomics**. **Nature Reviews Genetics**. [S.l: s.n.]. , jan. 2009

WANG, Zhuo et al. Musa balbisiana genome reveals subgenome evolution and functional divergence. **Nature Plants**, v. 5, n. 8, p. 810–821, ago. 2019.

WATERHOUSE, R. M. et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. **Molecular biology and evolution**, v. 35, n. 3, p. 543–548, 1 mar. 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29220515>. Acesso em: 26 out. 2019.

WEI, L. et al. Comparative genomics approaches to study organism similarities and differences. **Journal of biomedical informatics**, v. 35, n. 2, p. 142–50, abr. 2002. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/12474427>. Acesso em: 17 nov. 2019.

XU, L. et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. **Nucleic acids research**, v. 47, n. W1, p. W52–W58, 2 jul. 2019. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/31053848>. Acesso em: 29 out. 2019.

YANG, S.-F. et al. To Trim or Not to Trim: Effects of Read Trimming on the De Novo Genome Assembly of a Widespread East Asian Passerine, the Rufous-Capped Babbler (Cyanoderma ruficeps Blyth). **Genes**, v. 10, n. 10, p. 737, 23 set. 2019. Disponível em: <https://www.mdpi.com/2073-4425/10/10/737>. Acesso em: 27 set. 2019.

YOU, F. M. et al. Chromosome-scale pseudomolecules refined by optical,

physical and genetic maps in flax. **The Plant Journal**, v. 95, n. 2, p. 371–384, jul. 2018. Disponível em: <http://doi.wiley.com/10.1111/tpj.13944>. Acesso em: 27 jan. 2020.

YU, L. et al. Tissue-specific transcriptome analysis reveals multiple responses to salt stress in Populus euphratica seedlings. **Genes**, v. 8, n. 12, 8 dez. 2017.

ZHENG, Y. et al. **iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases**. **Molecular Plant**. [S.l.]: Cell Press. , 5 dez. 2016

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

10. APÊNDICES

9.1. Mapa genético de *P. guajava* digitalizado

#PL1		#PL2		#PL4		#PL7	
0.0	mPgCIR178	0.0	mPgCIR005	0.0	mPgCIR048	0.0	mPgCIR049
18.9	mPgCIR178	31.1	mPgCIR157	39.6	mPgCIR091	43.6	mPgCIR097
36.6	mPgCIR230	57.9	mPgCIR027	73.9	mPgCIR334	91.6	mPgCIR325
45.2	Me12Em9	71.0	mPgCIR027	96.6	mPgCIR257	110.5	mPgCIR139
45.2	Me10Em1	87.5	mPgCIR448	103.0	mPgCIR285	154.1	mPgCIR229
45.2	Me12Em9	104.0	mPgCIR205	103.0	mPgCIR285	204.6	mPgCIR041
45.2	Me4Em6	104.0	mPgCIR205	124.4	mPgCIR092		
45.2	Me10Em10	115.9	mPgCIR110	155.5	mPgCIR249		
45.2	Me10Em10	137.3	mPgCIR414	182.3	mPgCIR137		
45.2	Me13Em5	162.7	mPgCIR025	182.3	mPgCIR137		
45.2	Me13Em4	186.7	mPgCIR033	196.5	mPgCIR039		
45.2	Me12Em6	186.7	mPgCIR033	205.1	mPgCIR194		
45.2	Me10Em8	210.7	mPgCIR028	210.4	mPgCIR005		
45.2	Me12Em9	234.7	mPgCIR339	226.9	mPgCIR110	#PL8	
45.2	Me12Em6	262.9	mPgCIR090	250.9	mPgCIR165	0.0	mPgCIR191
45.2	Me3Em14	278.3	mPgCIR235	261.7	mPgCIR165	39.6	mPgCIR132
45.2	Me12Em6	278.3	mPgCIR235			53.8	mPgCIR180
45.2	Me13Em8	283.6	mPgCIR418			66.9	mPgCIR419
45.2	Me13Em4	283.6	mPgCIR418			99.6	mPgCIR404
45.2	Me11Em13	290.0	mPgCIR290				
45.2	Me12Em9	290.0	mPgCIR290				
45.2	Me11Em13	296.4	mPgCIR098				
45.2	Me13Em4	296.4	mPgCIR098				
45.2	Me13Em1	300.7	mPgCIR215				
45.2	Me12Em3	300.7	mPgCIR015	#PL5			
45.2	Me13Em8	300.7	mPgCIR009	0.0	mPgCIR097		
45.2	Me12Em9	300.7	mPgCIR215	24.0	mPgCIR177	#PL9	
45.2	Me4Em6	302.8	mPgCIR009	29.3	mPgCIR321	0.0	mPgCIR236
45.2	Me13Em4	317.0	mPgCIR237	34.6	mPgCIR177	37.8	mPgCIR102
45.2	Me12Em9	326.7	mPgCIR237	61.4	mPgCIR017	77.4	mPgCIR180
45.2	Me11Em13			91.0	mPgCIR047	121.0	mPgCIR426
45.2	Me12Em9			112.4	mPgCIR039	208.6	mPgCIR041
45.2	Me11Em13			139.2	mPgCIR153		
45.2	Me12Em9						
45.2	Me10Em10						
45.2	Me13Em4						
45.2	Me12Em9	#PL3				#PL10	
45.2	Me3Em14	0.0	mPgCIR022			0.0	mPgCIR016
45.2	mPgCIR419	37.8	mPgCIR046			41.5	mPgCIR022
45.2	Me12Em3	60.5	mPgCIR242			77.5	mPgCIR030
45.2	Me13Em4	80.7	mPgCIR228	#PL6		115.3	mPgCIR015
45.2	Me11Em12	96.1	mPgCIR030	0.0	mPgCIR414	117.4	mPgCIR257
45.2	Me12Em3	102.5	mPgCIR194	45.7	mPgCIR192	121.7	mPgCIR321
45.2	Me12Em3	106.8	mPgCIR230	54.3	mPgCIR192		
45.2	Me12Em9	121.0	mPgCIR025	93.9	mPgCIR046		
45.2	Me12Em9	132.9	mPgCIR352	129.9	mPgCIR228		
45.2	Me12Em3	141.5	mPgCIR253	173.5	mPgCIR277		
45.2	Me12Em9	146.8	mPgCIR018	181.0	mPgCIR277		
45.2	Me12Em3	146.8	mPgCIR018	229.0	mPgCIR094		
45.2	Me3Em14	151.1	mPgCIR011				
45.2	Me12Em3	151.1	mPgCIR011				
45.2	Me12Em3	164.2	mPgCIR031				
45.2	Me13Em4	191.0	mPgCIR099				
45.2	Me12Em9	191.0	mPgCIR099			#PL11	
45.2	Me3Em12	227.0	mPgCIR448			0.0	mPgCIR245
45.2	Me13Em5					63	mPgCIR034
45.2	Me4Em6					116.2	mPgCIR173
45.2	Me11Em12					172.3	mPgCIR175
45.2	Me13Em8						
47.3	mPgCIR253						
50.5	mPgCIR233						
55.8	mPgCIR233						
81.2	mPgCIR044						
81.2	mPgCIR044						
112.3	mPgCIR441						

9.2. Sequência dos primers utilizadas na ancoragem da montagem do genoma no mapa genético de *P. guajava*

Em10R	GACTGCGTACGAATTCAT
Em11R	GACTGCGTACGAATTCTA
Em12R	GACTGCGTACGAATTCTC
Em13R	GACTGCGTACGAATTCTG
Em14R	GACTGCGTACGAATTCTT
Em15R	GACTGCGTACGAATTGAT
Em16R	GACTGCGTACGAATTGTC
Em1R	GACTGCGTACGAATTAAT
Em2R	GACTGCGTACGAATTTGC
Em3R	GACTGCGTACGAATTGAC
Em4R	GACTGCGTACGAATTTGA
Em5R	GACTGCGTACGAATTAAC
Em6R	GACTGCGTACGAATTGCA
Em7R	GACTGCGTACGAATTCAA
Em8R	GACTGCGTACGAATTCAC
Em9R	GACTGCGTACGAATTCAG
Me10F	TGAGTCCAAACCGGAAA
Me11F	TGAGTCCAAACCGGAAC
Me12F	TGAGTCCAAACCGGAGA
Me13F	TGAGTCCAAACCGGAAG
Me1F	TGAGTCCAAACCGGATA
Me2F	TGAGTCCAAACCGGAGC
Me3F	TGAGTCCAAACCGGAAT
Me4F	TGAGTCCAAACCGGACC
Me5F	TGAGTCCAAACCGGAAG
Me6F	TGAGTCCAAACCGGACA
Me7F	TGAGTCCAAACCGGACG
Me8F	TGAGTCCAAACCGGACT
Me9F	TGAGTCCAAACCGGAGG

mPgCIR002	AGTGAACGACTGAAGACC	ATTACACATTCAGCCACTT
mPgCIR005	GCCTTTGAACCACATC	TCAATACGAGAGGCAATA
mPgCIR008	ACTTTCGGTCTCAACAAG	AGGCTTCCTACAAAAGTG
mPgCIR009	GCGTGTCGTATTGTTTC	ATTTTCTTCTGCCTTGTC
mPgCIR010	GTTGGCTCTTATTTTGGT	GCCCCATATCTAGGAAG
mPgCIR011	TGAAAGACAACAAACGAG	TTACACCCACCTAAATAAGA
mPgCIR015	TCTAATCCCCTGAGTTTC	CCGATCATCTCTTTCTTT
mPgCIR016	AATACCAGCAACACCAA	CATCCGTCTCTAAACCTC
mPgCIR017	CCTTTCGTCATATTCACTT	CATTGGATGGTTGACAT
mPgCIR018	TAAGCTGCATGTGTGC	ATGGCTTTGGATGAAA
mPgCIR019	AAAATCCTGAAGACGAAC	TATCAGAGGCTTGCATTA
mPgCIR022	CATAAGGACATTTGAGGAA	AATAAGAAAGCGAGCAGA
mPgCIR025	GACAATCCAATCTCACTTT	TGTGTCAAGCATACCTTC
mPgCIR027	AGCACTTAGGGACAAATTCA	CTCACTCTCCTCCATTCAAG
mPgCIR028	ACGTGAAAGCTGAAGTATGG	TGTCCAGCAAGTGAAGTAAA
mPgCIR029	CTCGCTTCAATCTCCATCTA	AGCGACACAGACTCTTCATT
mPgCIR030	CTCAAAGCACTATCATGTCG	CCTTGTGGGTTCTCTTTTG
mPgCIR031	TCTCACTGATGCAACTTTTC	CCCATTTTCATCTCAAAGTC
mPgCIR032	CGCCTTTCGTAAAAGAAGT	TCATATACTCGGACAAAACG
mPgCIR033	TGACATTCTCCTTCTTCGAT	GTGGGTTGTTAATGGATTTG
mPgCIR034	CTTTTAAGCACGTTTGACCT	GCGTATTAACAGCTTGGAGA
mPgCIR036	ATGACTTTGGAAGAGGATGA	AGCAGTTACAGTTGCTCGTT
mPgCIR037	GCGACGTTGTTGACTGAT	AGTGCGATAAAGACATTCAC
mPgCIR038	AGCCTGTTTTACGCCTTC	CGGCTGCTCTATTGTTATTT
mPgCIR039	GCTCACCTTACTCATTCAGC	CTGTTGCTAAGAGCTTTCGT
mPgCIR040	TGAATCTCCAGTGTCTTATCG	TGATTTCAACTGCGTATGTC
mPgCIR041	AAGTGGTGTCAGCAACTACC	CTTAGTTTGACCGCTCCAGT
mPgCIR042	CTCACCCAAAATCTACACAAG	AAGGGACTGGACGATGTT
mPgCIR043	GTCTTCCCGTCTGTCCTC	TCTCGACCTCAAGTCTTCC
mPgCIR044	TTCCAGGTCTATTGGATGTC	GGGGACACAAAACTTCATTA
mPgCIR045	TGGTGACCCTTTATGTCTTC	TATTCGGCGGTTTAGTTTTG
mPgCIR046	ATAGAACGCCATGTTACCAA	CAGGCTTATCTGTTACACCA
mPgCIR047	TGAACAGTAGCTTCATACCG	TCTGGATCACAAAACTTTCC
mPgCIR048	TGCAGCTTCTTCAATGTTC	AAAACTTGGCAACGTCAGT
mPgCIR049	TCCTTTCCCACTTTTCATC	ACAAGCTCCACACGTTACTC
mPgCIR089	TCGTCCAGAATCTTAAAAGC	AACTAGCATGTGACCAGGAG
mPgCIR090	ACGATGCAAGCTCTCCTAC	CTTTAGGGCTCTTCTTGAAA
mPgCIR091	GCGGTGGATTTGAATTTAG	CCAAGTAACCCACAACAATA
mPgCIR092	GTAAGTTTTCCGCATCCTG	ATATTGCACAGGGTGGTATC
mPgCIR093	GCATCATGTGTTTGAACGAT	AAGTGTGCGTTCTCCATCT
mPgCIR094	CAACCTTCCCGTGATTATT	CTAGCTTCTTCAGTGGGAAC
mPgCIR095	TGTGGAGATATAGGAAGGTGA	CAAGGATAAAGATGCACGTC
mPgCIR096	ACGCTGCAAACGATACTAAT	AACTCACACGAGCACAGAG
mPgCIR097	GACCTCAGTAGTTCAGCATGT	TAGAGTGGACGGGAGGAG
mPgCIR098	CATCAACTTTCCAGGCATA	CCATTCAGTCGGTTTGAC
mPgCIR099	TCAAAGTCCAAAACTCATGC	GGGATGGAGTAAAGATGAAA
mPgCIR100	CTAGAAGTCGAAGAATGGAA	TTTGTTAGTATCGGAGTCGAG
mPgCIR101	ATGGCTGTAAGAAGCAAAAG	GAAGAAATGTAGGTGCGTTC
mPgCIR102	AATTGGTGTAGCATCTGGA	GCCTACCATGAACAGAGAAA
mPgCIR104	ATTCCCGTGGATTATGTATC	ACAACCATTTTCTCCTCATC
mPgCIR105	CCTCCTTCGCTCTACATAAA	ATTACCCCACGAACATATCA
mPgCIR107	AAAGCCACTCATCTTCTTCA	ATCAGCTCGAAGGAAACAT
mPgCIR108	AGGACCTCACAGAAGTTCAC	CGCTGTTTACACTGTCGTT
mPgCIR109	AATTTCCACAGATCACAAGG	GGCATCTCCATCAAATACAT
mPgCIR110	GCCCCATTCTAAGAGACAAT	GAATGAAACCAGGTGTAGCA
mPgCIR111	CAACCTCGTTTGAGTCTTCT	AACATCATTGGGACCATTC
mPgCIR113	CATACACAATTCATGGCAGT	GCTTGTATTGACGGTGTCC
mPgCIR117	GCAGTATTGTAATGCCTTGA	CCTCTGCATAAGCCGTAAT
mPgCIR119	GCAGATATGGTCCTTTACAA	CTTCAGTGTGCATTCGTCTA
mPgCIR125	ATGTGGTGAGAGGAAGTGAG	TCCTATCCGTTCCAAACTAC
mPgCIR127	TGGAAACAATCTTGCTAGTG	ATAAACACAGCAGCATACAG
mPgCIR128	GCAAACCCTGCATCTCTT	AGGTGGGAGCAGGTTAGATA
mPgCIR130	CACTCATCACAAACACCATC	GTCTTGCCTCGTCAAAGTAT
mPgCIR132	TTTCTTGATCGCTCTTGC	CCCTTCACAGTTTTACCTCTT
mPgCIR135	CAGATAGCAAAACTGCCTCT	ATATCCCTCTCGCCTTCTT
mPgCIR137	GGGGAATGCAGAGATTGT	AGATGATGGTCTCGCTTTT
mPgCIR138	GGTGAGACCACTGAGTTCC	GAAAGACCAATGAAATCGAC
mPgCIR139	ATAATCCCCTCCCATAACTA	CCAACTCAACATGAGAAGC
mPgCIR140	GTGGTGAAGGAGTAAAGCTG	GCAGTATAAAGCAACAGATGG
mPgCIR143	TCTGGATTTTCTCCAATGC	GTATGCACCACCATCTGC
mPgCIR144	AGTCAAGCTGGTAATTGGAA	CTGATTTCCTAGAGTTGCAT
mPgCIR146	CGGATATTAGCAGGAGAAAA	AATGTTTGACGACTCGAAAG
mPgCIR148	CATACAGAGTCGGATGGTTT	GCTGCTGGTCTTAAAGCTAA
mPgCIR150	CCTAGTGACTCGAAGCAATC	TTGAGCCCTAGCATAGACAG
mPgCIR152	TTCGCAAGCATCTCAAGT	CCATAAGTTTGGGTGTCAA
mPgCIR153	GCCTCTGGTAAATCTGTTGA	ACATACGGATCAAGTCCAAA
mPgCIR155	GATATAAGATGGTTGCGAGT	CGAGTCATTGAAACGAAGA
mPgCIR157	AACCACCAAACCATACACC	CGACCAACCCTACATTCTG
mPgCIR158	ATCACCACTACTCCACTCGT	TAGAAGGTGCTCTAGGCTCA
mPgCIR160	TGGCTATAAGAATGGGAGAT	GACGAGCTTAGCCTCTGAAT
mPgCIR161	TCTCAAGGACCAACAAGAAG	AGGACTTAGCTTGGGTTTTC
mPgCIR162	CATAGAAAAGCGGGGTAAAC	AACTTGCGGAAACTCCAT
mPgCIR163	TCTTTGCACATCAAACTCG	CATGGTATCAATAGGTCAAGC
mPgCIR164	TCCCGTAAGTTGTTCTGTTT	CGTAAGAGATCGTGAAGGAG
mPgCIR165	TAAGGGATTCATTTCCGAGT	CTGGTGTGACGATGACTTTT
mPgCIR166	CTTTCCCATCAAAACGTAAG	CCAATTCATGCACTTAGACA
mPgCIR167	CAAGTTTGAAGGTGATGTAGAC	GTTATTTGCGGTTTGCAG
mPgCIR169	TTCAGGCAGATCGTGTTACT	GTGCCTAACCTACACCCTAA
mPgCIR171	TGACTTGCTCACCTAGATTGT	TCGATGGGAGATCAGAAGT
mPgCIR173	CGTGTTCTTTACATCCGTCT	TACCAGCAACACCAATGC
mPgCIR174	GCCACTGTGTAAGAGGATTG	ATTGTGGGAGATTGGAGAC
mPgCIR175	GCATTATGTGCCAAGCAA	TGCCAAGGTGTAATGTTGTA
mPgCIR176	TTCTCAATGATAGGTTACGG	ATGACTATTCTCCACCAGAT
mPgCIR177	TCTGAGAGTCTTGAAGGCATA	AGCACCCTTTTCTTCACAC
mPgCIR178	TACAGCGATTCTATCCCCTA	TCACGATGACAAGATAAAGC
mPgCIR179	GGGTCTCGACTAAAGAAGGA	CCTCCATTTGCATCAACTTT
mPgCIR180	CATGGATTCAACTCTTGTCG	CTACATTGGAAGCAGAATGG
mPgCIR182	GAGGAAGAAACCCGAAGTTA	GGTAGAAAGATCGGAAAGAC
mPgCIR183	GAAGAACAGAGACGCACCT	GCGCTAGGCTATAAATCATC
mPgCIR184	AAGCTACAATCGACGAAAAC	CACTATTAGCGAACCTGCAT
mPgCIR186	TCGAGCGGTTCTTGTCCT	ATCTTCTCCTCCTTCCATCC
mPgCIR187	AAGCCTAATAATGCGGAAGT	TCTTTCCCAGGATGGAGTA
mPgCIR188	TGGATGAATCAGGAGGATTA	TTGTGGGGAAGAAACTACTG
mPgCIR189	GAATTACCGTAGCGAACCAC	CCAGGGTATTACAATTTGCTTG
mPgCIR191	GACCCTCCCACTTATATTTTG	AAGCTGACATAACAGTCGAA
mPgCIR192	ACGCTAACTATCGAAATGCT	ACTACGCACTTGATGGAGAT
mPgCIR194	GCAGAGAATCGAAGCACTA	GCAAGCACAGGTTCTACTTT
mPgCIR196	TAACACCACGTCCATCAATA	GCGTTAAGCTACTTTGGCTA
mPgCIR198	CTCGATCAGAAGAACAACATC	ACTGTTCCTGATGGCTCTC
mPgCIR200	CCTTGCTTTGGTGAGGTC	GCTAATTCAGTCCTTCCAACT
mPgCIR201	TTTGCCTTCGAGCTTCTACT	ACAATTTCGTGGGCTCGT
mPgCIR202	CCATTAGAGCCGACAAAA	GACGAGAAACCCTAAACGTA
mPgCIR206	GGAAGTTTCAAAGTAACAGCAC	AGAATGAGTCCATGCTCAAA
mPgCIR208	AGCACGATGTAGGAAGAGAC	CGGCAAGCTCAACAATATAA
mPgCIR209	CTAAAGCCACATCCAGCA	CTAACATTTGCCTTCTACAGC
mPgCIR211	AGTGGGAGATTGTTAGGATG	GAATCCATCTTAGTTGTGAGC
mPgCIR212	CACTGTGCTTGAGTGAATGA	GGCCTCCCTTTGTAAACT
mPgCIR214	CTTGATGTTGCCTGGTTACT	TGCTGACGTATGAGTTCAAT
mPgCIR215	GGTCGAGACGATTACAAGAA	AGCGCAAGTAGAAGAGAATG

mPgCIR215	GGTCGAGACGATTACAAGAA	AGCGCAAGTAGAAGAGAATG
mPgCIR216	GATGGAACTGCAATGTATGA	AGACCTGCTCTGACTTGTGA
mPgCIR217	TTTCCTTTGTAGGACTTTCG	TTTTGCTTCGGGTTCATC
mPgCIR220	AGAGCAGTGGTTGCTATTTT	CCCATCTCTTACTTTTCTTGTG
mPgCIR221	CTAAGCCTGAAGTCCCAAAT	CCTCTTCTAAAGGCAACGAC
mPgCIR222	CCAGAATCAGACATAGTTAGAGC	CTGAAGACATCAACATGGAA
mPgCIR225	ATGGCTCAATGAAGATCCTA	GCTCATCCTTGTTAGGAAAG
mPgCIR226	AAACATATAGGGCGTCGAG	ACATCACCAACAAAAGTGGT
mPgCIR228	CAGAACAAAGAAGAGGATCTG	TGGATCAGTAGAATCGTTTG
mPgCIR229	GTCAAGCATACCTTCACCTC	CATCTCATCAACAGATTGGA
mPgCIR230	CACATTTGCTCCTGATTTTC	GCTCTTCAACGACCATCTT
mPgCIR231	CTCCAAGAAAATGGAAAGG	TGAAAACACCAAACAGCAC
mPgCIR233	GACTGAAGACCCAAATACCA	TTAGGCTGAAATGCTCCTTA
mPgCIR234	TTGGCTTGTCAACTACTGG	GCATCTCTTTCATTGGTGAT
mPgCIR235	TCTTCTGGACAACAAATTCC	GATTTAAGGCAAAGGATCG
mPgCIR237	AGATTCCATCTGCGATTGT	GCGGATCAAAACCTAATCT
mPgCIR238	GGTCCTCCTTCAACAGCTT	TCGACTTCACACTACTTCCA
mPgCIR242	TTAAGGTGGGACCAAGAAG	GACGTATCGGATCAAGTTTC
mPgCIR243	ACAGCAGGACACAAAGGA	GCTCTGAGGTGGTTTTCAT
mPgCIR244	CCCCACTTCAATCTTTCC	CAGCGTCATCATCCTCTC
mPgCIR245	CCAGACAAAATTCCAACG	AAATAGCCTCTCCAATCACA
mPgCIR247	CTTTGCACTCAATGTGTTTG	CCTTGTCACTCTTATAGCATCA
mPgCIR249	TTTGTCTGGTCGTCCTAGTT	CTTCAGTCCATCAGCAAAAT
mPgCIR250	AACTCGAATGGCTTCTGG	CTCGCGGTATTGAATGGT
mPgCIR251	TTTACAGCTTCTTCCTTTGC	GTCGTCTACTTCGATGGTTC
mPgCIR253	TGCCCTTAGCCTATAACTCA	CTTGTGGTCCAAGATGCTAT
mPgCIR254	GACTAGACAGAAATGGTAGGAT	TTTATGTTCTCCTCACCTG
mPgCIR255	TTGTTCTCTTTGTGGTTCCT	GTGACAACGTCAGTGGAGA
mPgCIR256	AGGTGCATGATTACGATTCT	CGAGGTTCTTGATGTTGTCT
mPgCIR271	TCAATGTGTGCATGAGTGA	CCGACTTTGATTTAGGTCAT
mPgCIR272	GTATGGGAGTATGGGTGCTA	AAACAACCCAAAGCAGAAC
mPgCIR274	TAACGAAAGGAAATGAGAGC	GCGCTTATTAGAGAAAGCAT
mPgCIR277	AGCCGATTATGATTACCTGA	CGATTCACTCCCTCATTACT
mPgCIR284	AACCCTTTCGGGGTCAAG	GATCCGATTGCGGAAGAG
mPgCIR285	CTTTGGGTGTCTCTGTGTCT	CAGCCAGGGCATAAGTATAG
mPgCIR287	GCTGGTGCAAAAGTAGTCA	GCAGTTCTTTTCCTTCTAACC
mPgCIR288	AAAGCAGAAAAGGGGTAAAC	GGAACCAGCACAGACATATT
mPgCIR290	AAATAGGAAGTAACCTGTGC	CCTGAACTTCTGTGTTTATATT
mPgCIR29	CTCGCTTCAATCTCCATCTA	AGCGACACAGACTCTTCATT
mPgCIR303	AAAACAAATTGGAAGAACCA	GAAGCGTTCAGGTTTCAG
mPgCIR316	GCTTCATATTACAAACCTTGG	GATCTAACTGACTTGCCAAAA
mPgCIR317	CAATAGCCACTAACTATGACATCT	AGCCAAAATCGTTCCATC
mPgCIR31	TCTCACTGATGCAACTTTTC	CCCATTTTCATCTCAAAGTC
mPgCIR326	AGAACAAGACACGAGAAGAG	AAAATCTACGCACAAACC
mPgCIR328	GTGACCGTGTTGGTTAATG	GATCGTGCCAATGTTCTAGT
mPgCIR32	CGCCTTTCGTAAAAGAAGT	TCATATACTCGGACAAAACG
mPgCIR332	GAAGCAATGGAGTCTTTTCTA	AGCTAGGGTTTTGAGATCAGT
mPgCIR334	GGATCTAACCGACCTTTCTT	AAACAGGACTGAGTTTCGAG
mPgCIR339	CCGAAGACGAGGAGATTA	TTAAGTGGAAAATCACAGTTG
mPgCIR343	CACTCACAGCATCGTAAA	GAAGTGGCAAGAATATCTGT
mPgCIR345	CTGGGAGACTTTTCAAGG	GAGTCCGATGTTGATGAAG
mPgCIR347	CTCTGAAAGGGAGAGGACTT	AGAATCTTCGCCTATTGCTT
mPgCIR34	CTTTTAAGCACGTTTGACCT	GCGTATTAACAGCTTGGAGA
mPgCIR350	CTACCCTGCTTCCTCCAT	GGGTTGTGGTGAGAGAAGT
mPgCIR352	TCGTAGTTAAACGAGTGATGC	GGACCAAACAGATATTCCAA
mPgCIR361	CTCACGTCCCTTATCAATGT	AATGTCTCGGTCAATTAGGA
mPgCIR361	CTCATGTCCCTTATCAATGT	AATGTCTCGGTCAATTAGGA
mPgCIR374	TTGGTTCCTGTCCAACTG	ATAGGCTGCTGCTGTTTT
mPgCIR377	CCAAACAAGGATTCACAAAT	TGTGGTTTAGTTCTTAGTTTTGC
mPgCIR378	GTCACTTGCCTGTTGATTTT	GCTTTCTCATTTCTTAACCAAG
mPgCIR383	CACATGGGGTGTCTTTGA	TTTTGTCTTCTTCCCATAATAACT
mPgCIR384	GGAGACTTTGTCATCGAAA	TTCCTTCATGGGTAAGTGAT
mPgCIR388	CTCGTCGTCATGCTCAGT	CTCAATTCTCCTCTCGTCTTT
mPgCIR389	TTTCTCCGTTGGTTTATGAC	ACGTGGATGTTTTACACCTG
mPgCIR392	AACTGTTTCTCTGGTTTGATG	ACGAAACGTCCTTGTTCTTA
mPgCIR399	AAAAGCTCATCCTTGTTGC	CCTCTCATTTTCTCCTCCA
mPgCIR39	GCTCACCTTACTCATTCAGC	CTGTTGCTAAGAGCTTTCGT
mPgCIR404	CTGCAACACATTGAACATTT	ATTCAGGCTTCTGGTTTGT
mPgCIR405	GAGATTTGGTTGAGGAAGGT	TTTCAAAATCAAGAGCCAAC
mPgCIR406	ACATTTGATGTCAGAAGGA	CCAATACAAGCTCAAAAACC
mPgCIR409	CCGGTGAAGTTGCTAATAAA	AACCTTACAATGATCCAACG
mPgCIR40	TGAATCTCCAGTGTCTTATCG	TGATTTCAACTGCGTATGTC
mPgCIR414	AACAACACGCTTTGAAGTTT	CCCAGAAAGATGAGACAAAG
mPgCIR416	TACTCTTGTCCTCTGGTTTT	AGGGAGAGCCAATATAATC
mPgCIR418	ACGTGACCTTCTCCAACAT	ACTGGATCGACTGAAGATGA
mPgCIR419	TCCTTGGTTTAATGTCTGATG	CTCACCTTCTGCTCTTCTTC
mPgCIR41	AAGTGGTGTCAGCAACTACC	CTTAGTTTGACCGCTCCAGT
mPgCIR420	CAACTTTGCTAGAGATGAAGC	ATGTAGTAATCGAAGAAATGGTT
mPgCIR422	CCGAAAACACAAGGGTTC	AGGTTCAAATGGTTTGTGG
mPgCIR426	TTCTGTCCTTCGTTGACTTT	GGTGTTACGGGTGAATAAGA
mPgCIR42	CTCACCCAAAATCTACACAAG	AAGGGACTGGACGATGTT
mPgCIR432	ATACAACGTGTCCTTGACG	CTATTTTGGTCACCCTTCAG
mPgCIR436	AGTCAGGTGTAGGTTTGGTG	CATCTGAGGTGAACCAGAGTA
mPgCIR437	ACAACAGTTCTGATCCCAAA	CTCGGAGACACAGAGGTCTA
mPgCIR439	GCATCTTGCTTCTGTCACTT	GGAGATGTGCAACGTATTTT
mPgCIR441	TAGGTATGGTTTGAAAGCTC	GTCTTTCTGCAAATATCCAT
mPgCIR446	GTGACCACTTTCTCAAAGCTA	ATCTCCATTTCTTGCTCGT
mPgCIR447	CTAACGAGTTTTCTCCGTGT	GTCCGGTCCTCCTCTATC
mPgCIR448	GCAGAGTAGGGTTTCAAAGG	CGTTTGTTGGACTTGGAG
mPgCIR448	GGAGAGTAGGGTTTCAAAGG	CGTTTGTTGGACTTGGAG
mPgCIR454	CCTTTTGCTTGTAGCTTTT	TCCTTCTTGTTCACATTTGTT
mPgCIR46	ATAGAACGCCATGTTACCAA	CAGGCTTATCTGTTACACCA
mPgCIR89	TCGTCCAGAATCTTAAAAGC	AACTAGCATGTGACCAGGAG
mPgCIR91	GCGGTGGATTTGAATTTAG	CCAAGTAACCCACAACAATA
mPgCIR92	GTAAGTTTTCCGCATCCTG	ATATTGCACAGGGTGGTATC
mPgCIR94	CAACCTTCCCGTGATTATT	CTAGCTTCTTCAGTGGGAAC
mPgCIR97	GACCTCAGTAGTTCAGCATGT	TAGAGTGGACGGGAGGAG
mPgCIR98	CATCAACTTTCCAGGCATA	CCATTCAGTCGGTTTGAC

9.3. Pipeline RunFilogeny.sh

```
#!/bin/bash
#author: Miquéias Fernandes 01/2020 bio@mikeias.net
#run filogeny and download Nwik tree from genome.jp tool

SERVER="https://www.genome.jp/tools-bin/ete"
TIMEOUT=5

FILE=$1

SEQ=${2:-"protein"}     #  nucleotide
workflow1=${3:-"mafft_default"}  #  mafft_einsi   mafft_linsi mafft_ginsi
clustalo_default  muscle_default
workflow2=${4:-"-none"} # -trimal001 -trimal01 -trimal02 -trimal05  -trimal_gappyout
workflow3=${5:-"-none"} # -prottest_default  -pmodeltest_full_ultrafast  -
pmodeltest_full_fast  -pmodeltest_full_slow  -pmodeltest_soft_ultrafast  -
pmodeltest_soft_fast -pmodeltest_soft_slow
workflow4=${6:-"-fasttree_default"} # -bionj_default  -fasttree_default  -
fasttree_full  -phyml_default  -phyml_default_bootstrap  -raxml_default  -
raxml_default_bootstrap

DATA="upload_file=@$FILE"  #DATA="sequence=`cat $FILE`"

JOB=$(curl -s \
    -F "seqtype=$SEQ" \
    -F "seqformat=unaligned" \
    -F "$DATA" \
    -F "workflow1=$workflow1" \
    -F "workflow2=$workflow2" \
    -F "workflow3=$workflow3" \
    -F "workflow4=$workflow4" \
    -F "workflow=$workflow1$workflow2$workflow3$workflow4" $SERVER | grep -m1
'ete?id=' | cut -d= -f2 | cut -d\" -f1)

if ((  `echo $JOB | awk '{print length}' ` > 10 )) ; then

    while (( `curl -s $SERVER'?id='$JOB | grep -c "Your job is still running"` > 0 ))
; do
        sleep $TIMEOUT
    done

    echo `curl -s $SERVER'?id='$JOB | grep -m1 midpoint_data | cut -d\" -f2`
fi
```

9.4. Serviço de inicialização do GuavaDB no servidor

```
#place on: /lib/systemd/system/guavadb.service
#enable: systemctl enable guavadb.service

[Unit]
Description=inicializar GuavaDB WEB SITE
After=network.target

[Service]
Type=simple
User=root
ExecStart=/usr/bin/java -jar
/home/cluster/Documentos/guavadb/target/guavadb-0.0.1-SNAPSHOT.jar
ExecStop=/usr/bin/killall java

[Install]
WantedBy=multi-user.target
```

9.5. Script CGI para execução de alinhamentos pelo GuavaDB

```
#!/bin/bash

## This script request run blast
##    mode[S{scaffold},G{gene},C{cds},P{protein},X{status}]
##    sequence[FASTA]
##    args[adicional args to blast] => return code

##  http://baleia.ufes.br:35700/align.sh?mode=P&sequence=/home/cluster/Guava
DB/query.faa&args=-evalue,1e-5&user=user
##  {"SEQUENCE":"/home/cluster/GuavaDB/query.faa","ARGS":"-evalue,1e-
5","CODE":"1"}
##  http://baleia.ufes.br:35700/align.sh?mode=X&sequence=1&user=user
##  {"STATUS":"ended"}
##  http://baleia.ufes.br:35700/results/user/1.csv


ROOT=/home/cluster/GuavaDB
RESULTS=$ROOT/results

echo "Content-type: application/json"
echo ""

####################################
####################################
saveIFS=$IFS
IFS='=&'
parm=($QUERY_STRING)
IFS=$saveIFS

for ((i=0; i<${#parm[@]}; i+=2))
do
    declare var_${parm[i]}=${parm[i+1]}
done
####################################
####################################

if [ ! -d "$RESULTS/$var_user" ]; then
   mkdir $RESULTS/$var_user
   echo 'Options -Indexes' > $RESULTS/$var_user/.htaccess
fi

DB=$ROOT/db/Pguajava

if [ "S" == "$var_mode" ] ; then
    DB=$DB.fasta
```

```
elif [ "G" == "$var_mode" ] ; then
    DB=$DB.genes.fasta
elif [ "C" == "$var_mode" ] ; then
    DB=$DB.cds.fasta
elif [ "P" == "$var_mode" ] ; then
    DB=$DB.proteins.faa
elif [ "X" == "$var_mode" ] ; then
    if [ -f "$RESULTS/$var_user/$var_sequence.fa" ]; then
        echo '{"STATUS":"'`[ -
f "$RESULTS/$var_user/$var_sequence.end" ] && echo 'ended' || echo 'running'
`'"}'
    else
        echo '{"STATUS":"ERROR"}'
    fi
    exit 0
else
    echo '{"ERROR":"DB_UNKNOWN","DB":"'$var_mode'"}'
    exit 0
fi

CODE=`ls -1a $RESULTS/$var_user | tail -n+3 | wc -l`
cp $var_sequence $RESULTS/$var_user/$CODE.fa

(blast`echo $var_mode | tr -s SGCP nnnp` \
    -db $DB \
    -query $RESULTS/$var_user/$CODE.fa \
    -outfmt 10 \
    -out $RESULTS/$var_user/$CODE.csv \
    `echo $var_args | tr , \  ` \
    && touch $RESULTS/$var_user/$CODE.end) \
    </dev/null >&/dev/null &

echo '{"SEQUENCE":"'$var_sequence'","ARGS":"'$var_args'","CODE":"'$CODE'"}'

exit 0
```