**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Exploração de uma biblioteca genômica de *Passiflora edulis* f.
flavicarpa por sequenciamento de BAC-ends**

Anselmo Azevedo dos Santos

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

**Piracicaba
2013**

Anselmo Azevedo dos Santos
Bacharel e Licenciado em Ciências Biológicas

**Exploração de uma biblioteca genômica de *Passiflora edulis* f.
flavicarpa por sequenciamento de BAC-ends**

Orientadora:
Prof^a. Dra. **MARIA LUCIA CARNEIRO VIEIRA**

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

Piracicaba
2013

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

Santos, Anselmo Azevedo dos

Exploração de uma biblioteca genômica de *Passiflora edulis* f. *flavicarpa* por sequenciamento de BAC-ends / Anselmo Azevedo dos Santos.- - Piracicaba, 2013.

141 p: il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2013.

1. *Passiflora* 2. Biblioteca de BAC 3. BAC-end 4. Genômica 5. Bioinformática I. Título

CDD 634.425
S237e

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Aos meus pais e à minha irmã, por serem meu Alfa e Omega...

Dedico

À minha noiva Aline, por ser meu porto seguro...

Ofereço

AGRADECIMENTOS

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP) e ao Departamento de Genética, pela oportunidade e instalações oferecidos durante o desenvolvimento dos experimentos.

Ao Centro Francês de Recursos Genômicos Vegetais (CNRGV/INRA), pela hospitalidade e pela troca de conhecimentos, tão importantes para a realização deste trabalho.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro e pela bolsa concedida.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro.

À Professora Dra. Maria Lucia Carneiro Vieira, pela orientação e pela amizade.

À Professora Dra. Hélène Bergès, do CNRGV/INRA, pela orientação e pelo apoio nos experimentos realizados na França.

À Dra. Anete Pereira de Souza, do CBMEG do UNICAMP, pelo apoio na etapa sequenciamento de DNA.

À Dra. Maria Helena Pelegrinelli Fungaro, do CBMEG do UNICAMP, pelo apoio na etapa de sequenciamento de DNA.

À Dra. Claudia de Barros Monteiro-Vitorello, da ESALQ/USP, pelas dicas e sugestões que tanto contribuíram para a realização deste trabalho.

À técnica de laboratório Nadine Gautier, do CNRGV/INRA, pela amabilidade e pelo apoio na etapa de extração de DNA.

Às técnicas de laboratório Alessandra Regina Staffocker Oliveirae e Aline Moraes, do CBMEG do UNICAMP, pelo trabalho hercúleo realizado com o sequenciamento de DNA.

Ao doutorando Danilo Augusto Sforça, pela ajuda com a otimização das reações de sequenciamento.

Aos professores do Programa de Pós-Graduação em Genética e Melhoramento de Plantas da ESALQ/USP, pelos conhecimentos transmitidos.

Aos amigos do Laboratório de Genética Molecular de Plantas Cultivadas da ESALQ/USP, pela amizade, compreensão e companheirismo.

Aos meus pais e minha irmã, por me manterem forte.

À minha noiva Aline, por acreditar em mim nas horas em que eu mesmo já não mais acreditava.

"A vida sem ciência é uma espécie de morte."

Sócrates

SUMÁRIO

RESUMO	11
ABSTRACT	13
LISTA DE FIGURAS	15
LISTA DE TABELAS	19
ABREVIATURAS E SIGLAS.....	21
1 INTRODUÇÃO	23
2 REVISÃO BIBLIOGRÁFICA.....	27
2.1 A Família <i>Passifloraceae</i>	27
2.2 O Gênero <i>Passiflora</i>	29
2.3 Aspectos genéticos e citogenéticos	30
2.4 Bibliotecas genômicas inseridas em BACs e suas aplicações.....	36
2.5 Composição dos genomas vegetais	40
3 MATERIAL E MÉTODOS	45
3.1 Design do banco de dados SQL	45
3.2 Extração de DNA plasmidial e sequenciamento dos BAC-ends (BES).....	45
3.2.1 Cultura dos clones bacterianos	46
3.2.2 Miniprep	46
3.2.3 Confirmação da extração.....	47
3.2.4 Sequenciamento	47
3.3 Processamento pré-anotação	48
3.4 Anotação	50
3.4.1 Prospecção de candidatos a marcadores microssatélites	51
3.4.2 Identificação de elementos transponíveis.....	51
3.4.3 Identificação de regiões codificadoras de proteínas e anotação funcional.....	51
3.4.4 Mapeamento genômico comparativo	52
4 RESULTADOS E DISCUSSÃO.....	55
4.1 Banco de dados relacional.....	55
4.2 Extração dos BACs	57
4.3 Sequenciamento dos BAC-ends	62
4.4 Prospecção de candidatos a marcadores SSR.....	64
4.5 Elementos transponíveis.....	69
4.6 Regiões codificadoras de proteínas	72

4.7 Mapeamento genômico comparativo.....	84
4.8 Organização geral do genoma de <i>Passiflora edulis</i> f <i>flavicarpa</i>	92
5 CONCLUSÕES	95
REFERÊNCIAS	97
ANEXOS	111

RESUMO

Exploração de uma biblioteca genômica de *Passiflora edulis* f. *flavicarpa* por sequenciamento se BAC-ends

O maracujá-amarelo (*Passiflora edulis* f. *flavicarpa*) é uma frutífera de importância econômica no Brasil, sendo apreciado para a produção de suco concentrado e para o consumo *in natura*, além de ser usado pela indústria farmacêutica na extração da *passiflorina*. O presente trabalho visou à exploração da biblioteca genômica inserida em BACs (Ped-B-Flav) por meio da técnica de BAC-end *sequencing*, visando prover os primeiros *insights* sobre a composição e organização genômica da espécie, além de gerar novos candidatos a marcadores moleculares. Ao todo foram realizadas 9.979 reações de sequenciamento com eficiência média de 89 %, resultando em 8.821 BES de alta qualidade, com tamanho variando entre 100 pb e 1.255 pb, tendo em média 596 pb, totalizando cerca de 5,7 Mpb de informação genômica. Foram identificados, ao todo, 610 potenciais novos marcadores microssatélites. Os motivos de tetranucleotídeos foram os mais abundantes, ou seja, 28,9 % do total, sendo as repetições AATT aquelas observadas com maior frequência, com 131 ocorrências. Foram identificados e classificados 4.394 (19,69 %) elementos repetitivos. Dentre estes elementos, os grupos dos retrotransposons *gypsy* e *copia-like* foram os mais abundantes, correspondendo a 10,08 % e 7,93 % das ocorrências, respectivamente. Além disso, foram encontradas 767 (8,7 %) sequências com alta identidade a regiões codificadoras de proteínas. Estas sequências foram classificadas e anotadas de acordo com o vocabulário controlado GeneOntology. Análises de mapeamento genômico comparativo revelaram três regiões microssintênicas com o genoma de *Populus trichocarpa*, uma com o genoma de *Vitis vinifera* e uma com o genoma de *Arabidopsis thaliana*, além de evidenciarem uma série de regiões rearranjadas em relação aos genomas de referência. O presente estudo mostrou que os BES de *Passiflora edulis* são uma excelente fonte de informações sobre o genoma da espécie, principalmente no que tange à diversidade gênica, identificação de elementos transponíveis e ao potencial para o desenvolvimento de novos marcadores genéticos. Igualmente, foi possível empregar essas sequências na identificação de regiões microssintênicas entre o genoma do maracujá-amarelo e de outras espécies vegetais próximas.

Palavras-chave: *Passiflora*; Biblioteca de BAC; BAC-end; Genômica; Bioinformática

ABSTRACT

Exploitation of a genomic library of *Passiflora edulis* f. *flavicarpa* using BAC-end sequencing

Yellow passion fruit (*Passiflora edulis* f. *flavicarpa*) is of considerable economic importance to Brazil. It is used to produce juice concentrate and also marketed for consumption as a fresh fruit. In the pharmaceutical industry, it is used to produce passiflora extract. The aim of this study was to explore the BAC (Bacterial Artificial Chromosome) genomic library (Ped-B-Flav) using BAC-end sequencing (BES) to provide some initial insights into the composition and organization of the species genome, and to generate new candidates for molecular markers. Altogether, 9,979 sequencing reactions were performed, with an average efficiency of 89 %, resulting in 8,821 high-quality BES, of average length ranging from 100 bp to 1255 bp, and consisting of an average 596 bp, totaling some 5.7 Mb of genomic information. In all, we identified 610 potential new microsatellite markers. Tetranucleotide motifs (28.9%) were the most abundant and AATT was the most frequently observed motif, with 131 occurrences. We identified and classified 4,394 (19.69 %) repetitive elements. Retrotransposon gypsy (10.8%) and copia-like (7.93%) elements were the most abundant. Furthermore, we found 767 (8.7 %) sequences very similar to those of protein coding regions. These sequences were classified and annotated according to gene ontology controlled vocabulary. Comparative genomic mapping revealed three regions showing microsynteny with the genome of *Populus trichocarpa*, one with *Vitis vinifera* genome and one with the *Arabidopsis thaliana* genome. In addition it revealed a series of rearranged regions in comparison to the reference genomes. This study showed that *Passiflora edulis* BES form an excellent source of information on the genome of the species, especially in regard to genetic diversity, identification of transposable elements and potential for the development of new genetic markers. It was also possible, using these sequences, to identify regions showing microsynteny with other plant species.

Keywords: *Passiflora*; BAC library; BAC-end; Genomics; Bioinformatics

LISTA DE FIGURAS

- Figura 1 - Árvore filogenética representando a organização da ordem Mapighiales, de acordo com o sistema de classificação APG III. A seta destaca a posição da família Passifloraceae, e sua relação com as demais famílias da ordem. Os ramos marcados com “***” possuem suporte de bootstrap de 50 a 70 %, os marcados com “*” possuem suporte de bootstrap de 70 a 90% e os demais ramos possuem suporte de bootstrap maior que 90 %. Fonte: Missouri Botanical Garden (www.mobot.org).....28
- Figura 2 - Fluxograma das etapas de processamento pré-anotação e anotação das sequências de *P. edulis*.....51
- Figura 3 – Modelo ER (*Entity Relationship*) utilizado para a implementação do banco de dados para o armazenamento e gerenciamento das informações geradas pelo presente trabalho. Cada caixa representa uma das entidades que compõem o banco; dentro de cada uma delas encontram-se nomes representando os atributos que formam cada uma das entidades. Os símbolos sobre as barras que conectam as caixas indicam o tipo de relacionamento entre as entidades, podendo ser 1-∞ (um para muitos) ou ∞-∞ (muitos para muitos). Os asteriscos em frente aos nomes de alguns atributos indicam quais deles são “atributos chave” dentro da entidade....57
- Figura 4 – Gel de Eletroforese em Campo Pulsado mostrando os produtos de digestão com a enzima *NotI* para avaliação da qualidade dos eventos de extração. O primeiro grupo (esquerda) de amostras foi extraído utilizando o protocolo fornecido pelo fabricante, enquanto o segundo grupo (direita) foi extraído com o protocolo modificado, descrito na seção Material e Métodos. O tamanho dos fragmentos foi acompanhado utilizando o *Ladder MidRange 1* (NEB).....59
- Figura 5 – Géis de Eletroforese em Campo Pulsado mostrando os produtos de digestão com a enzima *NotI*. Os três géis mostram um grupo de 130 extrações aleatoriamente selecionadas a fim de avaliar, por amostragem, a qualidade das minipreparações.

O tamanho dos fragmentos foi acompanhado utilizando o *Ladder MidRange* 1 (NEB).....60

Figura 6 – Correspondência entre as placas de armazenamento dos clones da biblioteca Ped-B-Flav (384 poços) e as placas usadas na extração e sequenciamento dos Bac-ends (96 poços). A primeira linha (1-12) e primeira coluna (A-H) representam as marcações nas placas de 96 poços, enquanto a segunda linha (1-24) e segunda coluna (A-P) representam as marcações presentes nas placas de 384 poços. Cada placa de 384 poços foi subdividida em 4 quadrantes de 96 posições, representados na figura pelas cores e pelas iniciais Q1, Q2, Q3 e Q4.....61

Figura 7 – Perfis de eletroferogramas obtidos com o protocolo de sequenciamento conforme descrito na seção de Material e Métodos.....62

Figura 8 – Distribuição por tamanho dos BAC-ends de *P.edulis*. As 8.821 sequências analisadas, após editadas, variaram de 100 pb até 1.255 pb, com um tamanho médio de 596 pb.....63

Figura 9 – Distribuição das percentagens de microssatélites encontrados nos BES de *P. edulis*, categorizados quanto à classe de repetições, e comparação com as proporções reportadas em outras espécies de dicotiledôneas.....66

Figura 10 – Distribuição dos melhores hits (*top-hits*) dos BES de *P. edulis* contra sequências de todas as espécies vegetais presentes no banco “NR (*Non redundant*)”, do NCBI. As buscas foram realizadas por meio da ferramenta de BlastX.....72

Figura 11 – Análise da Ontologia Gênica dos BES de *P. edulis* para em nível 1.....74

Figura 12 – Análise da Ontologia Gênica dos BES de *P. edulis* para a categoria Molecular function.....75

Figura 13 – Análise da Ontologia Gênica dos BES de *P. edulis* para a categoria *Biological process*.....78

- Figura 14 – Análise da Ontologia Gênica dos BES de *P. edulis* para a categoria *Cellular Component*.....82
- Figura 15 – Representação esquemática do conteúdo da região de 79 kpb do cromossomo C de *Arabidopsis thaliana*, localizada entre as posições 50130 pb e 130159 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *A. thaliana*.....85
- Figura 16 – Representação esquemática do conteúdo da região de 177 Kpb do cromossomo 11 de *Vitis vinifera*, localizada entre as posições 2324226 pb e 2502417 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *V. vinifera*.....86
- Figura 17 – Representação esquemática do conteúdo da região de 217 Kpb do cromossomo 12 de *Populus trichocarpa*, localizada entre as posições 9673965 pb e 9890993 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *P. trichocarpa*.....88
- Figura 18 – Representação esquemática do conteúdo da região de 176 Kpb do cromossomo 14 de *Populus trichocarpa*, localizada entre as posições 9166211 pb e 9343042 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *P. trichocarpa*.....89
- Figura 19 – Representação esquemática do conteúdo da região de 306 Kpb do cromossomo 14 de *Populus trichocarpa*, localizada entre as posições 5539303 pb e 5845525 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *P. trichocarpa*.....89

Figura 20 – Composição estimada do genoma de <i>P. edulis</i> após a análise de 8.821 sequências oriundas do sequenciamento de BAC-ends.....	93
--	----

LISTA DE TABELAS

Tabela 1 – Comparação dos mapas de ligação de maracujá amarelo. Distâncias medidas em cM de acordo com a função de mapeamento de Kosambi.....	35
Tabela 2 – Características da biblioteca Ped-B-Flav, depositada no <i>Centre National de Ressources Génomiques Végétales</i> (CNRGV/INRA), em Toulouse, França.....	46
Tabela 3 – Estatística descritiva dos BES de <i>P. edulis</i>	64
Tabela 4 – Distribuição das percentagens de microssatélites encontrados nos BES de <i>P. edulis</i> , categorizados quanto à classe de repetições.....	65
Tabela 5 – Distribuição dos microssatélites presentes nos BES de <i>P. edulis</i>	68
Tabela 6 – Classificação e distribuição dos elementos genéticos repetitivos de <i>P. edulis</i>	70
Tabela 7 – Distribuição dos termos ontológicos de nível 3 dentro da categoria <i>Molecular function</i>	77
Tabela 8 – Distribuição dos termos ontológicos de nível 4 dentro da categoria <i>Molecular Function</i>	78
Tabela 9 – Distribuição dos termos ontológicos de nível 4 dentro da categoria <i>Biological process</i>	80
Tabela 10 – Distribuição dos termos ontológicos de nível 4 dentro da categoria <i>Biological Process</i>	81
Tabela 11 – Distribuição dos termos ontológicos de nível 3 dentro da categoria <i>Cellular Component</i>	83

Tabela 12 – Distribuição dos termos ontológicos de nível 4 dentro da categoria <i>Cellular Component</i>	83
Tabela 13 – Mapeamento comparativo de BAC-ends de <i>P. edulis</i> em outros genomas vegetais.	91
Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos <i>top-hits</i>	113

ABREVIATURAS E SIGLAS

BAC	Bacterial Artificial Chromosome
BES	BAC End Sequence
FISH	Fluorescent <i>In Situ</i> Hybridization
GO	Gene Ontology
QTL	Quantitative Trait Loci
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
TE	Transposable Element
YAC	Yeast Artificial Chromosome

1 INTRODUÇÃO

Em poucos anos, os cientistas vão dispor de sequências completas de um grande número de genomas de plantas superiores, incluindo os genomas de culturas agrícolas e plantas modelo. A análise dessas sequências permitirá o estabelecimento de associações entre a estrutura genômica e as variações morfológicas e fisiológicas, em níveis infraespecíficos ou filogenéticos, permitindo o uso destas informações por botânicos, geneticistas e evolucionistas. Segundo Paterson et al. (2010) o conhecimento sobre o *set* de genes e seu arranjo genômico, as variações nucleotídicas, além da identificação das famílias de genes e de genes duplicados no genoma das plantas superiores, trará avanços no conhecimento sobre a diversidade estrutural, funcional e regulatória desses genomas.

Atualmente, existem cerca de 25 genomas vegetais completamente sequenciados (HAMILTON; BUELL, 2012), e cerca de outros 15 genomas em processo de finalização (www.phytozome.net), e parte deste trabalho já vem sendo realizado para culturas agrícolas importantes como, por exemplo, em cereais (arroz, milho, sorgo e trigo), leguminosas (soja, feijão e alfafa), frutíferas (uva, maçã, melão e pêssego) e na cana-de-açúcar, uma cultura produtora de açúcar e etanol, de genoma complexo.

Recentemente, em colaboração com o Centre National de Ressources Génomiques Végétales (CNRGV) do INRA de Toulouse, França, o LGMPC (*Laboratório de Genética Molecular de Plantas Cultivadas* da ESALQ/USP) desenvolveu uma biblioteca genômica de maracujá amarelo inserida em BACs (*Bacterial Artificial Chromosomes*), denominada “Ped-B-Flav”. Esta biblioteca contém 82.944 clones, provendo uma cobertura de aproximadamente 6 vezes o tamanho total do genoma da espécie, que é de 3,16 pg (SOUZA et al., 2004). A fim de aproveitar o máximo das informações contidas na biblioteca, seria necessário sequenciar um alto número de BACs inteiros, randomicamente selecionados de forma a representar o genoma total. Além de muito oneroso, o sequenciamento total dos BACs demanda grande capacidade de armazenamento e processamento computacional para a montagem das sequências.

Uma vez que muitos projetos de sequenciamento têm como principal objetivo obter os primeiros *insights* sobre a composição do genoma da espécie em estudo, uma abordagem alternativa ao sequenciamento total é o sequenciamento das pontas dos insertos de BACs, conhecidas como BES (*BAC-End Sequences*) selecionadas randomicamente na biblioteca

(LAI et al., 2006). De posse de um grande número dessas sequências é possível investigar a composição e a complexidade genômica da espécie. A partir dos BES pode-se obter um grande volume de informações como, por exemplo, sobre a densidade gênica, conteúdo GC, presença de elementos repetitivos, e de possíveis regiões microssintênicas entre o genoma estudado e outros genomas vegetais de referência.

Além disso, esse tipo de estudo fornece material de excelente qualidade para outros trabalhos, como a seleção de sondas para hibridização fluorescente *in-situ* (FISH) em cromossomos metafásicos e paquitênicos, elucidação da estrutura de genes por meio de “chromosome walking” (CHEUNG et al., 2007), mapeamento de genes de cópia simples (PETERSON et al., 2000), desenvolvimento do mapa físico da espécie (WU et al., 2012), além de permitir a formação de um “scaffolding” que ajuda na montagem precisa do genoma, reduzindo em muito os custos gerados pelo fechamento de *gaps* e cobertura de regiões ricas em sequências repetitivas (RAMPANT et al., 2011).

É inegável a importância da obtenção de ao menos um “draft” do genoma de *P. edulis*. Porém, apesar dos avanços testemunhados com o advento da segunda geração de sequenciadores (SGS), o sequenciamento e principalmente a montagem de um genoma vegetal completo ainda é algo pouco trivial. Por isso, para espécies vegetais com genomas desconhecidos, como é o caso das passifloras, a utilização das novas plataformas de sequenciamento é ainda limitada. O tamanho das leituras produzidas por SGS é incompatível com a montagem de grandes genomas nucleares, que são também altamente repetitivos em plantas. Os poucos trabalhos realizados têm sido destinados ao sequenciamento de transcritos, ressequenciamento e sequenciamento “de novo” de genomas plastidiais, os quais são menores e contêm pouca quantidade de DNA repetitivo (CARVALHO; SILVA 2010). Wicker et al. (2006) demonstraram, trabalhando com BACs de cevada, que o sequenciamento com plataformas de segunda geração produz até nove vezes mais *gaps*, principalmente devido a erros em regiões de homopolímeros.

Os estudos em *Passiflora* carecem de qualquer tipo de referência genômica, tendo apenas algumas poucas centenas de sequências depositadas nos bancos de dados públicos. Dessa forma, a utilização de sequências BAC-ends se apresenta como uma alternativa rápida, mais barata que o sequenciamento total do genoma e capaz de subsidiar o desenvolvimento de uma miríade de novos estudos. Neste contexto, os objetivos principais deste trabalho foram a extração do DNA de, ao menos, 10 % dos clones da biblioteca Ped-B-Flav; o sequenciamento

do maior número possível de BAC-*ends*, pertencentes à estes clones extraídos; processamento e anotação dos dados gerados por sequenciamento; busca de sequências candidatas à novos marcadores moleculares SSR; busca e anotação de elementos transponíveis e repetitivos; anotação funcional das regiões codificadoras de proteínas; identificação de potenciais regiões de microssintenia entre o genoma de *P. edulis* e outros genomas vegetais.

Os dados aqui apresentados representam um grande salto em relação ao que se conhece sobre o genoma das passifloras, e também, disponibilizam, à comunidade científica, novas ferramentas que facilitarão o desenvolvimento de estudos teóricos, tais como a investigação da evolução do genoma da espécie, o esclarecimento das relações filogenéticas dentro e fora do gênero *Passiflora* e o enriquecimento de mapas de ligação e citogenéticos; e aplicados ao melhoramento da cultura, como a geração de novos marcadores moleculares e a produção de linhagens comerciais melhoradas.

2 REVISÃO BIBLIOGRÁFICA

2.1 A Família *Passifloraceae*

Passifloraceae é uma família de angiospermas dicotiledôneas, pertencente à ordem *Malvales*, com distribuição essencialmente neotropical, encontrada principalmente nas regiões mais quentes da América, com algumas ocorrências na Ásia, África do Sul, Nova Zelândia e Austrália e apenas uma espécie em Madagascar. Seus membros possuem como habitats mais comuns as florestas tropicais, savanas e várzeas, ocorrendo geralmente em regiões de baixas e médias altitudes (FEUILLET; MacDOUGAL, 2003). A Figura 1 apresenta uma árvore filogenética representando a organização da ordem *Malvales* e a posição da família *Passifloraceae* dentro do grupo.

Os membros da família *Passifloraceae* se unem sob uma série de características morfológicas, como a presença de gavinhas auxiliares, glândulas extraflorais ou nectários, folhas alternas, coroa de estaminódios, androginóforo e sementes ariladas (FEUILLET, 2004). No clado há uma grande variação no tamanho e na forma das folhas, assim como no tamanho, forma e coloração das flores. Os frutos, geralmente, são arredondados ou alongados, variando em coloração do verde ao alaranjado, com sementes pretas e achatadas, comumente envolvidas por um arilo gelatinoso e translúcido (JORGENSEN et al, 1984).

Os primeiros esforços para a caracterização e organização da família foram feitos pelo botânico Ellsworth Paine Killip, em seu trabalho “The American species of *Passifloraceae*”, publicado em 1938. Apesar de mais de 75 anos já terem se passado e cerca de 120 novas espécies terem sido descritas desde então, ainda não há consenso em relação ao número total de espécies que compõem a família. De acordo com o autor e a metodologia utilizada na classificação taxonômica do grupo, o número de membros pode variar de cerca de 530 espécies, como proposto por Watson e Dallwitz (1992) a 700 espécies, como proposto mais recentemente por Feuillet (2004). Também existe pouca conformidade em relação número de gêneros em que a família se divide, podendo variar de 18 gêneros, como proposto por Watson e Dallwitz (1992) e Feuillet (2004), até 23 gêneros, segundo a classificação proposta por Barroso et al. (1978).

Na América do Sul ocorrem duas espécies do gênero *Ancistrohyrsus*, seis espécies do gênero *Dilkea*, três espécies pertencentes ao gênero *Mitostemma* e mais de 400 espécies

pertencentes ao gênero *Passiflora* (CERVI, 1997). No Brasil ocorrem cerca de 140 espécies de passifloras, a maioria subordinada ao gênero *Passiflora* L.

Figura 1 - Árvore filogenética representando a organização da ordem *Mapighiales*, de acordo com o sistema de classificação APG III. A seta destaca a posição da família *Passifloraceae*, e sua relação com as demais famílias da ordem. Os ramos marcados com “**” possuem suporte de *bootstrap* de 50 a 70 %, os marcados com “*” possuem suporte de *bootstrap* de 70 a 90 % e os demais ramos possuem suporte de *bootstrap* maior que 90 %

Fonte: Missouri Botanical Garden website (www.mobot.org)

2.2 O Gênero *Passiflora*

Passiflora L é o gênero de maior importância econômica dentro da família *Passifloraceae*. Até o ano de 2010, cerca de 520 espécies haviam sido classificadas como pertencentes ao gênero (CERVI, 2010). Porém, desde então uma série de novas espécies vem sendo descobertas, aumentando ainda mais este número (CERVI; LINSINGEN, 2010).

As espécies pertencentes ao gênero *Passiflora* estão agrupadas em 4 subgêneros, *Astrophaea*, *Decaloba*, *Deidamiodes* e *Passiflora* (FEUILLET; MacDOUGAL, 2004), sendo *Decaloba* e *Passiflora* os subgêneros mais representativos, com cerca de 200 espécies cada e *Astrophaea* e *Deidamiodes*, com apenas algumas poucas dezenas de membros (ULMER; MacDOUGAL, 2004). Segundo Muschner et al. (2013), o gênero *Passiflora* teve sua origem a cerca de 40 milhões de anos atrás e a sua divergência em subgêneros é bastante antiga, tendo a primeira fissão ocorrido por volta de 38,3 milhões de anos atrás, quando *Astrophaea* se separou do que viriam a se tornar os demais subgêneros. A fissão mais recente data de cerca de 33,5 milhões de anos atrás, quando da separação de *Decaloba* e *Deidamiodes*.

Nativo da América do Sul, o gênero *Passiflora* possui na região norte do Brasil seu maior centro de diversidade, apresentando grande quantidade de espécies e uma ampla variabilidade genética e fenotípica. O número de espécies pertencentes ao gênero descritas no Brasil é de 135, sendo 85 destas espécies endêmicas da flora nacional, colocando o país em uma posição privilegiada em relação aos recursos genéticos do gênero (CERVI et al., 2010; FAJARDO et al.; 1998; VIANA et al.; 2003).

De modo geral, algumas das espécies do gênero são cultivadas por seu valor ornamental (ABREU et al., 2009) e para o consumo dos frutos (MELLETTI et al, 2005), que tem aroma e sabor bastante agradáveis. Seus frutos são ricos em sais minerais e vitaminas, sobretudo A e C, e suas folhas possuem propriedades farmacológicas associadas à produção de *passiflorina*, um calmante natural (BARBOSA et al, 2008).

Além das formas tradicionais de utilização, estudos recentes têm demonstrado propriedades medicinais dos maracujás até então desconhecidas. Segundo Ramos (2007), o extrato seco da casca do maracujá amarelo exerce influência benéfica no controle de diabetes tipo II, podendo ser utilizado como adjuvante em terapias tradicionais. Já Ichimura et al. (2006) demonstraram que o ácido Y-aminobutírico e a luteolina, presentes também na casca no maracujá amarelo, exerceram função anti-hipertensiva em cobaias naturalmente

hipertensas. Outra propriedade farmacêutica recentemente descoberta é a capacidade de reduzir o colesterol em pacientes que apresentam hipercolesterolemia, como observada por Ramos e colaboradores no ano de 2007.

O Brasil é o maior produtor mundial de maracujá (AGUIAR; SANTOS, 2001; LIMA et al., 2006), sendo o maracujá amarelo, *Passiflora edulis* f. *flavicarpa*, o maracujá roxo, *Passiflora edulis*, e o maracujá doce, *Passiflora alata*, as espécies mais cultivadas (revisado por VIEIRA; CARNEIRO, 2004). O maracujá amarelo, também conhecido como maracujá azedo, é a espécie mais expressiva comercialmente, dentre as passifloras cultivadas no Brasil

A cultura do maracujá amarelo ganhou importância após a década de 70, com as primeiras exportações de suco. Desde então, em razão da irregularidade na demanda, a cultura passou por vários ciclos de expansão e retração na área cultivada (RIZZI et al., 1998). Devido à qualidade dos seus frutos e à industrialização do suco, o maracujá amarelo está presente em 95% dos pomares comerciais, com uma produção anual, em 2009, de 731.515 toneladas, numa área cultivada de 50.795 ha, correspondendo à praticamente toda a área plantada e produção de maracujazeiros. Apesar do aumento nos preços, acumulado nos últimos 5 anos, é evidente que o consumo da fruta vem crescendo gradativamente no Brasil. O volume de frutos negociado pelo CEAGESP-SP, no mesmo período, teve aumento de 21%. (AGRIANUAL, 2012). Sua comercialização é feita na forma de fruta fresca para consumo “*in natura*” e na forma de suco (AGUIAR; SANTOS, 2001). A maior parte da produção mundial de maracujá é destinada à produção do suco, sendo que o Brasil, um dos principais exportadores, destina parte da sua produção à União Europeia. O país é considerado o maior produtor e consumidor de fruta fresca e processada de maracujá, necessitando eventualmente importar suco do Equador e da Colômbia para suprir a demanda (revisado por VIEIRA; CARNEIRO, 2005; SILVA; RUA, 2007).

2.3 Aspectos genéticos e citogenéticos

O gênero *Passiflora* foi muito pouco estudado do ponto de vista citogenético, uma vez que para cerca de 80 % das espécies, informações básicas, como número e tamanho dos cromossomos, ainda são desconhecidas (HANSEN et al., 2006). Ainda assim, os poucos

estudos cromossômicos têm sido de grande valia para determinar a evolução das relações em *Passiflora* (MELO et al 2001; CUCO et al. 2005; HANSEN et al. 2006; SOUZA et al. 2008).

Os primeiros trabalhos com citogenética de passifloras datam da década de 1940, com os estudos de Bowden (1945), estando estes focados apenas na contagem do número de cromossomos e breve descrição de sua morfologia. Desde então a caracterização cromossômica das espécies de passifloras foi, durante muito tempo, baseada unicamente em caracteres cariomorfológicos. No entanto, com o advento de técnicas de bandamento e de citogenética molecular, a caracterização cromossômica foi significativamente melhorada. Principalmente, estudos de bandamento CMA / DAPI (MELO et al., 2001; CUCO et al., 2005.) e sonda *FISH* (Hibridação Fluorescente *In situ*) (MELO; GUERRA 2003; CUCO et al 2005.; PRAÇA et al., 2008) foram empregados com o intuito de melhorar a identificação e a distribuição da heterocromatina nas espécies do gênero.

Dentro do gênero *Passiflora* são encontrados 4 números cromossômicos básicos: $x=6$, $x=9$, $x=10$ e $x=12$. A distribuição desses números básicos corrobora a divisão do gênero em quatro subgêneros, proposta por MacDougal e Feuillet (2004), uma vez que $x=6$ é observado apenas em *Decaloba*, $x=9$ em *Passiflora*, $x=10$ em *Astrophea* e $x=12$ em *Deidamioides* (MELO; GUERRA, 2003). Trabalhando com *FISH*, Melo et al., (2001) estabeleceram o número básico do gênero em $x=6$, sugerindo a ocorrência de eventos de poliploidia durante a evolução do grupo. Os números básicos $x=9$, $x=10$ e $x=12$ foram considerados secundários. Em 2006, Hansen et al. associaram dados citogenéticos e moleculares para contestar as conclusões de Melo e colaboradores, sugerindo $x=12$ como o número básico de cromossomos do gênero. Porém, segundo os próprios autores, qualquer tentativa de estipular o número básico de cromossomos do gênero seria deveras precipitada, dada a baixíssima quantidade de informações genéticas e citogenéticas disponíveis para o clado, principalmente para os subgêneros *Deidamioides* e *Astrophea*.

A grande maioria das espécies pertencentes ao gênero *Passiflora* é diplóide, com $2n=12$, $2n=18$ ou $2n=20$, havendo apenas algumas exceções tetraplóides, hexaplóides ou mesmo octaplóides (MELO; GUERRA, 2003). Dentre as espécies de interesse comercial, $2n=18$ é de ocorrência mais frequente, como em *Passiflora edulis* f. *flavicarpa*. Isto é interessante do ponto de vista do melhoramento genético via hibridização, uma vez que torna mais fácil a obtenção de progênies de cruzamentos interespecíficos. Algumas poucas espécies

de interesse comercial possuem $2n=12$ e 14 (*P. holosericea* L.), $2n=20$ e 22 (*P. foetida* L.), $2n=24$ e 36 (*P. suberosa* L.) ou $2n=84$ (*P. lutea* L.).

Utilizando técnicas de citogenética clássica e molecular, Cuco et al. (2005) estabeleceram o cariótipo *P. edulis* f. *flavicarpa*, observando constrições secundárias e satélites nos menores cromossomos: no braço longo do cromossomo 8 e no braço curto do cromossomo 9. O comprimento do conjunto haplóide foi de $22,66\ \mu\text{m}$ e o tamanho dos cromossomos variou de $3,16$ a $1,82\ \mu\text{m}$, sendo a maioria dos cromossomos metacêntricos. Mais recentemente, Praça et al. (2008) observaram valores médios maiores para o tamanho dos cromossomos, variando de $3,75\ \mu\text{m}$ a $2,22\ \mu\text{m}$, com o comprimento total do lote haploide de $26,65\ \mu\text{m}$. Essa discrepância deve-se, principalmente, ao fato de este trabalho ter utilizado metáfases mais distendidas do que as usadas por Cuco et al. (2005).

Variações nas metodologias empregadas e nas condições do material analisado também geraram discrepâncias nas descrições da posição e na localização das constrições secundárias presentes nos cromossomos de *P. edulis*. Oliveira (1996) descreveu apenas uma constrição secundária, localizada no cromossomo 8, enquanto Mayeda (1997) e Cuco et al. (2005) relataram constrições secundárias nos cromossomos 8 e 9; já Soares-Scott (1998) relatou constrições nos cromossomos 4 e 7 e Praça et al. (2008) nos cromossomos 1, 2, 7 e 8.

Pesquisas que aplicaram a metodologia de FISH com sondas de DNA repetitivo demonstraram também haver de dois a três sítios de DNAr 45S e um sítio de DNAr 5S em 20 espécies de *Passiflora* analisadas (MELO; GUERRA, 2003). Cuco et al. (2005) relataram sinais positivos correspondentes aos sítios de DNAr 45S nas constrições secundárias do cromossomo 8 e satélites do cromossomo 9 em *P. edulis* f. *flavicarpa*. Ainda, evidenciaram a existência de apenas um único sítio de DNAr 5S, detectado no braço longo do cromossomo 4 da espécie.

Mais recentemente, Penha (2012) estabeleceu um novo cariótipo para a espécie utilizando sondas derivadas de BACs, para a hibridização fluorescente *in situ*. Dessa forma foi possível estabelecer marcadores cromossomo-específicos para os cromossomos 1 (gene ERS), 3 (ACCO) e 4 (genes CYCD1 e G3PD), além de um sítio de DNAr 5S no cromossomo 5 e de 2 sítios de DNAr 45S, sendo um no cromossomo 7 e outro no cromossomo 8.

Estudos sobre o conteúdo de DNA também estão disponíveis para algumas espécies do gênero *Passiflora*. O C-valor e o tamanho do genoma são aspectos importantes da

biodiversidade das espécies, com significado biológico fundamental e muitos usos (BENNETT et al., 2011). Estudos clássicos definiram o C-valor de um organismo como a quantidade de DNA nuclear no seu núcleo gamético não replicado (SWIFT, 1950), independentemente do nível de ploidia do taxon.

Souza et al. (2004), trabalhando com citometria de fluxo, demonstraram que o conteúdo 2C de DNA, em picogramas, variou bastante entre as espécies estudadas. O valor de 1,83 foi encontrado para *P. suberosa*, 3,16 para *P. edulis* f. *edulis*, 3,40 para *P. mucronata*, 3,43 para *P. edmundoi*, 3,88 para *P. laurifolia*, 3,92 para *P. gilberti*, 5,36 para *P. quadrangularis* e 3,19 pg para *P. edulis* f. *flavicarpa*.

Mais recentemente Yotoko et al, (2011), observaram valores, de modo geral, menores do que os obtidos por Souza et al. (2004). Por exemplo, 2,516 para *Passiflora edulis* f. *edulis*, 3,024 para *P. mucronata*, 1,52 para *P. edmundoi* e 3,42 para *P. gilberti*. Este estudo não contemplou a análise do tamanho do genoma de *P. edulis* f. *flavicarpa*.

Em particular, o conteúdo do genoma (2C) observado por Souza et al. (2004) para *P. edulis* f. *flavicarpa* (3,19 pg) é bastante grande, em comparação com o conteúdo genômico de espécies filogeneticamente próximas, como *Vitis vinifera* (0,86 pg), *Populus trichocarpa* (0,98 pg), *Ricinus communis* (1,04 pg) e *Linum usitatissimum* (1,4 pg) (LODHI; REISCH, 1995; TUSKAN et al., 2006; HOUBEN et al., 2003; EVANS et al., 1972). Considerando-se a proporção de 1:1 de AT:GC e ignorando a presença de nucleotídeos modificados, pode-se empregar a relação estabelecida por Dolezel et al. (2003), onde 1 pg de DNA representaria $0,978 \times 10^9$ pares de bases. Nesses termos, levando-se em consideração apenas os valores observados por Souza e colaboradores, o genoma (2C) de *Passiflora edulis* f. *flavicarpa* teria, aproximadamente, 3.119 Mpb ou 3,1 Gpb.

Genomas de tamanhos elevados têm sido, geralmente, associados a eventos mutacionais, como crossing-over desigual (BENNETT et al., 2011), erros na replicação do DNA (KUNKEL et al., 1990), poliploidização (SOLTIS, 1999) e a propagação de elementos retrotransponíveis (LISCH, 2013). Porém, o “paradoxo do C-valor”, conhecido desde a década de 70, mostra que um maior volume de DNA não necessariamente reflete em maior complexidade ou aumento do número de genes (THOMAS, 1971). Por isso, espera-se que o genoma de *P. edulis* tenha quantidades de sequências não-codantes e elementos repetitivos maiores que o observado nas espécies supracitadas. Em contrapartida, espera-se que o número

de genes seja próximo ao encontrados nesses genomas, uma vez que a literatura recente tem mostrado que o número de genes não varia substancialmente entre as espécies vegetais.

O estabelecimento conteúdo de DNA, do cariótipo e a identificação de marcadores cromossômicos por FISH são passos importantes para o esclarecimento da estrutura genômica, da genética e diversidade das espécies de *Passiflora* podendo levar a avanços no campo teórico e a uma maior eficiência nos programas de melhoramento (CAPDEVILLE et al., 2009).

Do ponto de vista genético-molecular, alguns estudos foram realizados na última década, principalmente visando ao desenvolvimento de mapas de ligação para o maracujá amarelo. Estes mapas genéticos formam um arcabouço de marcas arranjadas em grupos de ligação, importantes para o melhoramento, uma vez que as populações de mapeamento podem ser levadas a campo para se proceder a avaliações fenotípicas. Isto feito, cada informação fenotípica dos indivíduos da população é analisada conjuntamente com a genotipagem molecular para os marcadores usados. Usando *softwares* apropriados, é possível identificar locos, de herança simples ou complexa (ditos QTL), associados ao controle dos caracteres avaliados que, então, são anotados nos grupos de ligação.

Os primeiros mapas genéticos de ligação para *Passiflora edulis* f. *flavicarpa* foram desenvolvidos por Carneiro et al. (2002), com base em marcadores dominantes do tipo RAPD (*Random Amplified Polimorfism DNA*), utilizando a estratégia conhecida como *duplo pseudo cruzamento teste*, onde 90 indivíduos de uma população F₁ segregante, derivada do cruzamento entre os acessos ‘IAPAR 06’ e ‘IAPAR 123’ foram genotipados, dando origem a dois mapas separados, um para cada genitor. O mapa do genitor feminino, ‘IAPAR 123’, teve um comprimento total de 727,7 cM, alocando 148 marcas RAPD distribuídas em 9 grupos de ligação, com distância média de 11,2 cM entre os locos. Já o mapa do genitor masculino, ‘IAPAR 06’, teve um comprimento total de 783,5 cM; foram alocadas 121 marcas RAPD em 9 grupos de ligação, com distância média de 12,2 cM entre os locos.

Em 2006, Lopes e colaboradores, utilizando os mesmos genitores de Carneiro et al. (2002), desenvolveram outros 2 mapas, baseados na mesma estratégia, porém com um número maior de indivíduos na população (117) e, também, uma maior quantidade de marcadores, esses do tipo AFLP (*Amplified Fragment Length Polimorfism*). Este trabalho rendeu mapas mais precisos, dado o maior tamanho da população de mapeamento, sendo que o mapa do genitor feminino possuía 140 marcas, distribuídas em 488,9 cM e o mapa do

genitor masculino 115 marcas, distribuídas em 790,2 cM. A distância média entre as marcas foi de 9,6 cM para o mapa do genitor ‘IAPAR 123’ e 12,6 cM para o genitor ‘IAPAR 06’.

Dois anos mais tarde, Oliveira et al. (2008) lançou mão de novas estratégias de mapeamento e de um grande número de marcadores dominantes (RAPD e AFLP) e alguns codominantes do tipo SSR (*Simple Sequence Repeat*) para gerar um mapa único integrado para o maracujá amarelo. Este mapa, além de mais que dobrar a cobertura do genoma, medindo 1.687,0 cM, também é muito mais informativo, uma vez que abriga certo número de marcadores bi-parentais. O mapa integrado possui 243 locos, entre marcadores dominantes e codominantes, distando em média 6,9 cM. A Tabela 1 resume as características dos mapas genéticos de ligação desenvolvidos para o maracujá amarelo.

Tabela 1 – Comparação dos mapas de ligação de maracujá amarelo. Distâncias medidas em cM de acordo com a função de mapeamento de Kosambi

Característica	Carneiro et al. (2002)		Lopes et al. (2006)		Oliveira et al. (2008)
	‘IAPAR 123’	‘IAPAR 06’	‘IAPAR 123’	‘IAPAR 06’	Mapa integrado
Número de Locos	148	121	140	115	243
Grupos de ligação	9	9	9	9	10
Tipos de marcadores usados	RAPD	RAPD	AFLP	AFLP	AFLP e SSR
Comprimento do mapa	727,7	783,5	488,9	790,2	1687
Comprimento médio dos Grupos	80,9	87,1	54,3	87,8	168,7
Distância média entre marcas	11,2	12,2	9,6	12,6	6,9

Fica claro, nos estudos subsequentes, que a tendência dos mapas genéticos é se tornarem cada vez mais informativos, melhorando em diversos aspectos, como na cobertura do genoma, em sua densidade e, principalmente, agregando marcadores moleculares mais informativos e funcionais. Assim, vale ressaltar que foram realizados esforços no intuito de gerar um mapa de ligação para outra espécie de interesse comercial (*Passiflora alata*), conhecida como maracujá doce. O primeiro mapa produzido para a espécie foi gerado por Pereira et al. (2013) e está grandemente enriquecido, com locos microssatélites e marcadores do tipo M-AFLP, AFLP, SNP (*Single Nucleotide Polymorphisms*) e RGA (*Resistance Gene Analogs*), além de conter sequências gênicas mapeadas. O mapa de ligação de *P. alata* conta com 549 locos, distribuídos em 2.073 cM, com distância média de 3,77 cM entre os locos.

Do ponto de vista filogenético, o primeiro trabalho a empregar dados moleculares para esclarecer as relações infragenéricas de *Passiflora* foi realizado por Muschner et al. (2003). Neste trabalho a filogenia do gênero *Passiflora* foi reconstruída com base em sequências de quatro regiões genômicas de 61 espécies. É digno de nota que os resultados deste trabalho já indicavam a divisão do gênero *Passiflora* em quatro grandes grupos, como proposto por FEUILLET E MacDOUGAL um ano mais tarde.

Pádua (2004), lançando mão de análises filogenéticas, morfométricas e marcadores microssatélites, também chegou à conclusão de que os dados moleculares e morfológicos discordavam da divisão infragenérica vigente na época, apontando, assim como os dados de Muschner et al. (2003), para a divisão do clado em quatro subgêneros.

Mais recentemente, Muschner et al. (2013), utilizando sequências de DNA derivadas de oito regiões diferentes do genoma de passifloras, não apenas reforçaram a ideia da divisão do gênero em 4 subgêneros, como também foram capazes de fazer inferências em relação ao tempo de divergência destes grupos e sua dispersão geográfica.

2.4 Bibliotecas genômicas inseridas em BACs e suas aplicações

O estudo dos genomas eucarióticos sofreu um grande avanço no início dos anos 80, com o advento das chamadas bibliotecas genômicas. Estas, basicamente, são uma coleção de clones que representam a totalidade do genoma de um organismo. Sua construção consiste, basicamente, em fragmentar o genoma em estudo, ligando os fragmentos resultantes a algum tipo de vetor. Esses fragmentos são, por sua vez, inseridos em organismos hospedeiros, geralmente bactérias ou leveduras. O conjunto desses fragmentos é denominado de “biblioteca genômica” (QUAIL et al., 2011).

Com o passar dos anos, uma série de vetores foi desenvolvida para a construção de bibliotecas de grandes insertos de DNA, cada um com seus usos, especificidades, vantagens e desvantagens. Dentre eles pode-se destacar os YACs (Burke et al., 1987), cosmídeos (COLLINS; HOHN, 1978), fosmídeos (KIM et al., 1992), bacteriófagos P1 (STEMBERG, 1990), BACs (SHIZUIA et al., 1992), PACs (IAONNOU et al., 1994) e PBCs (TAO; ZHANG, 1998). De acordo com o tipo de hospedeiros utilizados para a sua clonagem e

manutenção, esses vetores podem ser classificados em “Baseados em bactéria” ou “Baseados em leveduras”.

A utilização de vetores YACs (Cromossomos Artificiais de Levedura) e BACs (Cromossomas Artificiais de Bactérias) para a construção de bibliotecas genômicas de grandes insertos é muito mais frequente e apropriada em relação aos demais vetores disponíveis (CENSI et al., 2003). Também, bibliotecas em vetores BACs são geralmente preferíveis às bibliotecas genômicas em YACs, devido ao seu processo significativamente mais simples de produção, maior estabilidade e baixa frequência de quimerismo (WOO et al., 1994). Além disso, os clones armazenados em BACs podem ser facilmente isolados por meio de “miniprep” de lise alcalina, sendo diretamente reintroduzidos em células bacterianas. Mais importante, os sistemas de recombinação bacterianos são muito mais bem caracterizados que os sistemas de leveduras. (FRENGEN et al., 1999)

Bibliotecas genômicas em BACs são aplicadas em diversas áreas da genômica moderna, como para o sequenciamento de genomas completos (International Rice Genome Sequencing, 2005), clonagem posicional (PATOCCHI et al., 1999; JANDER et al., 2002, MONNA et al., 2002), isolamento de genes (LIANG et al., 2004; JANG et al. 2006.), integração de mapas de ligação e cromossômicos (FUCHS et al., 1998), prospecção de SNPs (OLLITRAULT et al., 2012), análises comparativas de estrutura gênica e sintenia (ILIC et al., 2003; WANG et al., 2005) e investigação da estrutura genômica por meio de hibridizações fluorescentes *in situ* (CHENG et al., 2002), tendo sido construídas para algumas das mais importantes espécies vegetais, como citrus (BAIG et al., 2009), tomate (BUDIMAN et al., 2000), café (NOIR et al., 2004), arroz (AMMIRAJU et al., 2006), girassol (BOUZIDI et al., 2006), quinoa (STEVENS et al., 2006) entre outras.

A utilização das bibliotecas genômicas em BAC tornou possível a condução de uma série de experimentos difíceis ou mesmo impossíveis de serem realizados usando bibliotecas convencionais. Isto inclui o desenvolvimento de mapas físicos para grandes genomas e o sequenciamento completo de genomas com um custo muito inferior ao realizado sem o auxílio de bibliotecas BAC (ZHANG et al., 2012).

Mapas físicos são representações da distância física entre genes, marcadores ou sequências de DNA, em pares de bases, e têm sido construídos para diversas espécies vegetais como *Arabidopsis thaliana* (MOZO et al., 1999), maçã (HAN et al., 2011), arroz (TAO et al., 2001), milho (COE et al., 2002), soja (WU et al., 2004) e trigo (PAUX et al., 2008). Estes são

fundamentais para o isolamento de genes de interesse, para assegurar a abrangência adequada de um sequenciamento genômico e para explorar a sintenia existente entre os genomas de espécies relacionadas.

Associadas as técnicas de sequenciamento de nova geração, as bibliotecas BAC permitem a redução dos custos de sequenciamento total de genomas, uma vez que é possível selecionar clones de interesse e favorecer o fechamento de *gaps*, além disso, permitem a descoberta, clonagem e caracterização de genes em larga escala (ZHANG et al., 2012).

Mesmo após o sequenciamento completo e a montagem dos genomas, a construção de bibliotecas BAC ainda é de grande valia para a realização de estudos genômicos, como é o caso do arroz, que mesmo após ter o primeiro *draft* de seu genoma publicado, ainda em 2002, foram contruídas e caracterizadas bibliotecas genômicas para 11 espécies selvagens e uma espécie cultivada do gênero *Oryza*, que depois de caracterizadas, foram usadas para construir mapas físicos das espécies, permitiram estudos sobre a evolução do gênero e organização do genoma, estudos de domesticação, redes de regulação gênica, e auxiliaram em esforços para a melhoria das culturas (WING et al., 2005; AMMIRAJU et al., 2006)

Bibliotecas BAC também podem ser utilizadas em conjunto com mapas de ligação a fim de identificar genes de interesse agrônômico. Uma vez que os mapas de ligação são frequentemente usados na identificação de marcadores associados a regiões que controlam a expressão de fenótipos específicos, as bibliotecas genômicas podem ser vasculhadas em busca de marcadores associados a estas regiões e, desta forma, favorecer o sequenciamento fino de regiões responsáveis pela expressão de um dado fenótipo. Um bom exemplo dessa aplicação foi o trabalho de Oblessuc (2013), que empregou esta estratégia na identificação genes relacionados à mancha angular em feijoeiro.

Uma estratégia bastante eficiente e viável para a obtenção dos primeiros *insights* sobre o conteúdo e a complexidade do genoma de uma espécie é oferecida pela construção de uma biblioteca genômica inserida em BACs, seguida do sequenciamento das regiões terminais de um grande número de clones, selecionados aleatoriamente. Tais sequências, conhecidas como BAC-ends (BES) podem ser usadas como suporte para a montagem de genomas sequenciados por meio da metodologia de *Shotgun* e para gerar mapas físicos comparativos (SHULTZ et al., 2007). A análise de dados provenientes dos BES também pode proporcionar uma visão geral da composição de um genoma, proporcionando informação sobre a sua densidade gênica, e a presença de potenciais elementos transponíveis (TEs) e microssatélites

(CAVAGNARO et al., 2009; HAN et al., 2008). Além de possibilitarem a identificação de diferentes regiões do genoma, os BAC-ends também são úteis para a busca de marcadores moleculares.

A primeira frutífera a ter seu genoma investigado por meio da abordagem do sequenciamento de BAC-*end* foi o mamão comum (*Carica papaya* L), quando Lai et al. (2006) analisaram cerca de 50 mil sequências derivadas de BAC-ends e obtiveram os primeiros *insights* sobre a composição e organização do genoma da espécie. Este trabalho serviu de base para uma série de estudos subsequentes que resultaram na construção de mapas genéticos mais densos (CHEN et al., 2007), no desenvolvimento de marcadores moleculares (EUSTICE et al., 2008), no desenvolvimento de um mapa físico e sua integração ao mapa de ligação (YU et al., 2009) e ao mapa citogenético da espécie (WAI et al., 2012), culminando no sequenciamento total de seu genoma (MING et al., 2012). Desde então, diversas outras espécies de frutíferas tem se beneficiado do uso de sequências de BAC-*ends*, como banana (CHEUNG et al., 2007), maçã (HAN et al., 2008), melão (GONZALEZ et al., 2010), citrus (Biwas et al., 2012), entre outras.

O sequenciamento de BAC-*ends* também ganhou popularidade nos últimos anos devido ao seu alto rendimento na prospecção *in silico* de marcadores moleculares, sejam estes gênicos, funcionais ou derivados de regiões heterocromáticas. Muitos trabalhos têm utilizado marcadores microssatélites derivados de BAC-*ends* para a construção, saturação e integração de mapas genéticos de ligação. Alguns exemplos da utilização desses marcadores derivados de BESs na construção de mapas de ligação podem ser encontrados em Frelichowski et al. (2006) e Shultz et al. (2007), que trabalharam com algodão e soja, respectivamente. Essa estratégia tem possibilitado o preenchimento de lacunas no mapeamento genético de espécies e facilitado o avanço de projetos de sequenciamento de genomas clonados em BACs (PETERSON et al., 2000; SCHATZ et al., 2010; VELASCO et al., 2010).

Como demonstrado recentemente por Ollitrault et al. (2012), BAC-*ends* também são uma excelente fonte para a mineração de marcadores SNPs. Neste trabalho 1.457 SNPs derivados de BAC-*ends* de *Citrus reticulata* foram genotipados na populações de outras espécies de citrus, a fim de testar a transferibilidade desses marcadores. Os resultados desse trabalho são bastante animadores, uma vez que marcadores SNPs, gerados a partir da biblioteca Ped-B-Flav, poderão ser aproveitados também em outras espécies do gênero *Passiflora*.

O sequenciamento de BES pela metodologia de Sanger permite a formação de um “scaffolding” que ajuda na montagem precisa do genoma, reduzindo em muito os custos gerados pelo fechamento de *gaps* e na cobertura de regiões ricas em sequências repetitivas (RAMPANT et al., 2011). Recentemente, estratégias híbridas têm sido utilizadas para o sequenciamento de genomas vegetais, aliando dados gerados pelo sequenciamento de BAC-ends com dados oriundos de técnicas de sequenciamento de segunda geração. Um bom exemplo dessa abordagem é genoma da maçã (VELASCO et al., 2010), em que se associou um grande número de BES às sequências geradas por SGS para a montagem do genoma. A mesma estratégia foi aplicada com sucesso nos genomas de álamo (KELLEHER et al., 2007), mamão (YU et al., 2009) e *Brachypodium* (GU et al., 2009).

Outra aplicação importante das sequências derivadas de BAC-ends é o chamado mapeamento genômico comparativo, que, basicamente, consiste na identificação de regiões de microssintenia entre o genoma da espécie à qual os BAC-ends pertencem e o genoma de uma espécie modelo, cujo genoma já tenha sido sequenciado. O conhecimento dessas regiões é de grande valia, pois serve de base para uma série de inferências filogenéticas e relacionadas à evolução dos genomas. Conhecendo a região correspondente a um BAC em um genoma de referência é possível se ter uma idéia de quantos e quais genes poderão ser encontrados com o sequenciamento daquele BAC. Além disso, BACs que possuem colinearidade com algum genoma de referência são excelentes candidatos para sequenciamento, uma vez que pode-se usar o genoma de referência para a sua montagem (HUO et al., 2009). Essa estratégia de mapeamento genômico comparativo vem sendo aplicada, com sucesso, em uma série de espécies vegetais, como por exemplo, na orquídea *Phalaenopsis* (HSU et al., 2012), cana-de-açúcar (Kim et al., 2013), melão (Gonzales et al., 2010) e carvalho (RAMPANT et al., 2011).

2.5 Composição dos genomas vegetais

O início da década de 1990 foi marcado pelo surgimento da tecnologia de sequenciamento automático de DNA, também conhecida como método de Sanger. Este avanço permitiu o surgimento de uma nova área das ciências biológicas, a genômica, que tem por principal objetivo estudar o genoma completo de um organismo, sua estrutura e dinâmica de evolução (HAMILTON; BUELL, 2012).

O bacilo de Pfeiffer, *Haemophilus influenza*, foi o primeiro organismo a ter seu genoma completamente sequenciado, ainda no ano de 1995 (FLEISCHMAN et al., 1995). Cinco anos mais tarde, no ano de 2000, foi publicada a primeira sequência genômica completa de um organismo vegetal, a planta modelo *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000). Desde então, cerca de 25 genomas vegetais já foram completamente sequenciados, possibilitando um melhor entendimento da estrutura e organização desses genomas (HAMILTON; BUELL, 2012).

Um dos principais questionamentos existentes antes da publicação dos primeiros genomas era “Quantos genes existem em um genoma vegetal?”. Embora o conteúdo genômico das angiospermas varie enormemente entre as espécies, há muito se sabe que a maior parte desta variação não está relacionada à quantidade ou ao tamanho dos genes presentes no genoma (BENNETZEN et al., 2005).

As primeiras estimativas do número de genes presentes do genoma de *A. thaliana* indicavam a existência de cerca de 25 000 genes. Com o avanço nos métodos de predição e com o aumento na disponibilidade de ESTs (*Expressed Sequence Tags*) este número subiu para os atuais 33.583 (NCBI, 2013). Já as primeiras predições feitas para o genoma do arroz, o segundo genoma vegetal a ser publicado, indicavam a existência de cerca de 50 000 genes, caindo para cerca de 40 mil (NCBI, 2013). Aparentemente angiospermas, de modo geral, possuem entre 25 e 45 mil genes, o que é maior do que a média dos demais eucariotos (STERK et al., 2007). A explicação mais provável para o número consideravelmente alto de genes nos genomas vegetais é o fato de que, a grande maioria dessas plantas, em algum momento de sua história evolutiva, passou por um evento de poliploidia (CIU et al., 2006).

Como citado anteriormente, a variação do conteúdo genômico em plantas tem pouca relação com a quantidade ou o tamanho dos genes presentes no genoma. Os grandes responsáveis pela dinâmica de variação do tamanho dos genomas vegetais são os chamados elementos transponíveis (LISCH, 2013).

Elementos transponíveis (TE's) são elementos móveis de DNA, que podem se mover e replicar dentro dos genomas hospedeiros, sendo amplamente prevalentes na maioria dos genomas eucarióticos. (FESCHOTTE et al., 2002).

Os elementos transponíveis foram observados pela primeira vez na década de 1940, pela botânica e geneticista Bárbara McClintock, ganhadora do prêmio Nobel de

Fisiologia/Medicina, em 1983, pela descoberta desses artefatos genômicos. Desde então, os elementos transponíveis passaram de uma simples curiosidade no genoma do milho para uma das principais forças motriz da evolução genômica em todos eucariotos (LISCH, 2013).

Longe de serem raros, os TE's chegam a compor mais de 50% do conteúdo total de alguns genomas. Esta relação pode ser ainda maior, passando dos 70% em alguns genomas de gramíneas (MEYERS, 2001). Elementos transponíveis já foram observados em virtualmente todos os organismos eucariotos observados, sendo *Plasmodium vivax* e algumas espécies relacionadas, as únicas exceções à regra (WICKER, 2007).

Embora os principais grupos de TE's sejam ancestrais e estejam presentes em basicamente todos os reinos, esses elementos apresentam uma diversidade extrema, chegando a possuir milhares de famílias diferentes apenas no reino vegetal (MORGANTE, 2005). De modo geral, em genomas vegetais, os elementos transponíveis são, sem sombra de dúvida, a parcela mais representativa e dinâmica do genoma. Sabe-se que ondas de expansão e retração no número de TE's podem resultar em diferenças drásticas entre genomas sabidamente próximos (BENNETZEN et al., 2005).

Além de sua grande influência na arquitetura dos genomas eucarióticos, a atividade desses elementos pode resultar em uma gama de efeitos fisiológicos, como mudanças no padrão de expressão de alguns genes, reestruturação de redes regulatórias e surgimento de novos genes (YAO, 2001).

De acordo com o tipo de transposição, mediada por RNA ou DNA, os TEs podem ser divididos em duas classes principais. Elementos de classe I ou retrotransposons, são o tipo mais comum de elemento genético móvel em genomas eucarióticos (FLAVELL et al., 1992) e se transpõe através de um intermediário de RNA por meio de um mecanismo de “Copiar e colar” (GRANDBASTIEN, 1992). Assim como nos demais eucariotos, Elementos de classe I são os mais abundantes em genomas vegetais (FESCHOTTE et al. 2002). Elementos de classe II, ou transposons de DNA, utilizam um mecanismo de “cortar e colar” envolvendo a reintegração do elemento no genoma através de um intermediário de DNA (BOWEN; JORDAN, 2002).

Ambas as classes se subdividem em famílias e superfamílias, dentro das quais membros compartilham semelhanças estruturais.

As grandes diferenças de tamanhos do genoma das espécies de plantas estão geralmente ligadas à presença de diferentes quantidades de retrotransposons. Geralmente, quanto maior o genoma vegetal, maior a chance de este conter uma grande quantidade de retroelementos (KUMAR; BENNETZEN, 1999). Por exemplo, genomas grandes, como o da cevada, podem chegar a ser compostos por até 70% destes elementos (VICIENT et al., 2001) enquanto em genomas pequenos, como o do arroz, estes representam apenas 17% da composição total (MCCARTHY et al., 2002).

Existem evidências de que alguns retrotransposons se acumulam preferencialmente em regiões específicas do genoma. Por exemplo, os Ty1-Copia são geralmente encontrados em regiões de eucromatina, em contraste com os Ty3-Gypsy que se acumulam preferencialmente em regiões de heterocromatina, como próximos a centrômeros e aos knobs de DNA em milho. Essas evidências nos levam a crer que cada grupo de TE's tenha desenvolvido seu próprio padrão de inserção dentro do genoma. Porém, ainda é incerto se essa acumulação em certas regiões se deve a algum mecanismo de inserção preferencial ou apenas a incapacidades dessas regiões do genoma de se “livrarem” de um grupo específico de elementos (KUMAR; BENNETZEN, 1999).

3 MATERIAL E MÉTODOS

3.1 Design do banco de dados SQL

Utilizando a linguagem SQL (*Structured Query Language*) foi desenvolvido um banco de dados relacional a fim de facilitar a armazenagem e a manipulação dos dados gerados durante o projeto. O design do modelo ER (*Entity-Relationship*) foi feito respeitando a nomenclatura proposta por Watson (2005).

O design do modelo esteve pautado, principalmente, em facilitar o armazenamento de uma grande quantidade de dados e também em interligar, da melhor forma possível, todas as informações relacionadas a uma mesma sequência. Uma das preocupações foi criar a possibilidade de, com *queries* simples, recuperar a maior quantidade de informação possível do banco de dados.

A implementação do banco de dados foi realizada por meio da plataforma MySQL (www.mysql.com). Essa plataforma foi escolhida devido a sua relativa facilidade de uso, pela grande documentação disponível e também por possuir licença GNU (www.gnu.org), sendo passível de modificação e de distribuição gratuita.

3.2 Extração de DNA plasmidial e sequenciamento dos BAC-ends (BES)

As sequências foram geradas a partir da biblioteca de clones em BACs Ped-BFlav, construída, por Penha (2012), a partir de folhas do acesso (IAPAR) 123 (*Passiflora edulis f. flavicarpa* 2n=2x=18), com tamanho médio de insertos de 110 kb. Para a confecção da biblioteca foi utilizada a enzima de restrição *HindIII* e os fragmentos clivados foram clonados e armazenados no vetor pIndigoBAC-5 *HindIII*, tendo como hospedeira *E. coli*, linhagem DH10B T1R. A Tabela 2 descreve as características básicas da biblioteca de BACs utilizada neste estudo.

Tabela 2 – Características da biblioteca Ped-B-Flav, depositada no Centre National de Ressources Génomiques Végétales (CNRGV/INRA), em Toulouse, França

Nome da Biblioteca	Ped-B-Flav
Vetor de Clonagem	pIndigoBAC-5 <i>Hind</i> III
Enzima de Restrição	<i>Hind</i> III
Hospedeiro	DH10B T1R
Antibiótico Seletivo	Cloramfenicol
Número de Clones	82944
Número de Placas (384 poços)	216
Tamanho Médio dos Insertos	108 kb
Tamanho Mínimo dos Insertos	50 kb
Tamanho Máximo dos Insertos	196 kb
Equivalentes Genômicos	6x

3.2.1 Cultura dos clones bacterianos

Os clones foram repicados a partir de microplacas de 384 poços e inoculados em storage plates (0.8 ml) de 96 poços, contendo 200 µl de meio 2x LB + 0.2 µl/ml de cloranfenicol. A inoculação foi realizada com o auxílio de replicadores plásticos de 96 pontas. Esses, por sua vez, foram mergulhados nas placas de 384 poços e, em seguida, deixados por instantes em contato com o meio de cultura. Em seguida, as placas foram incubadas a 37 °C sob constante agitação (325 rpm) por 7 h. Passadas 7 h da pré-cultura, foi feita uma nova repicagem, dessa vez para uma placa de 96 poços, contendo 1,5 ml de meio 2x LB + 1.25 µl/ml de cloranfenicol. Por sua vez, essas placas foram incubadas a 37 °C sob constante agitação (325 rpm), por um período mínimo de 18 h. Ao fim da incubação, as placas foram submetidas à centrifugação por 10min a 2.500 g. Por fim, o sobrenadante foi descartado e os *pellets* estocados -20 °C para posterior utilização no procedimento de miniprep.

3.2.2 Miniprep

Todas as minipreparações foram realizadas com kits Macherey-Nagel Nucleospin® 96 Flash. Os protocolos fornecidos pelo fabricante foram modificados, a fim de aumentar o rendimento e permitir a utilização dos kits em larga escala. Após o descongelamento de cada

pellet, os mesmos foram ressuspensos com a adição de 300 µl da solução F1+ RNase (solução de ressuspensão), seguida de agitação a 650 rpm por um período mínimo de 30 min. Em seguida, foi feita a adição de 300 µl da solução F2 (solução de Lyse) e a inversão das placas entre 10 a 15 vezes, deixando a solução descansar por 4 min. após a inversão. Em seguida, foi realizada a adição de 300 µl da solução F3 (solução de neutralização), seguida por mais 15 movimentos de inversão.

As soluções foram filtradas em NucleoVac 96, utilizando placas Flash filter (fornecidas junto com os kits), a 0.4 bar por 4 min. Ao produto filtrado foram adicionados 630 µl de isopropanol (0.7 volumes) e a solução foi centrifugada a 2500 g por 1 h. Após o descarte do sobrenadante, foram adicionados 500 µl de etanol 70% e a solução foi centrifugada a 2.500 g por 30min. Por fim, o sobrenadante foi descartado e os *pellets* foram deixados para secar a 37 °C por 1 h. Após este período, cada minipreparação foi ressuspensa em 35 µl de água ultrapura e deixada a 4 °C por 24 h, para posterior congelamento.

3.2.3 Confirmação da extração

A fim de avaliar a qualidade e o rendimento das minipreparações, foram selecionadas amostras aleatórias para serem submetidas a reações de restrição com a enzima *NotI*, seguidas da migração em eletroforese de campo pulsado (PFGE). Cada reação foi realizada contendo de 10 µl das minipreparações, 2 µl de tampão 10x NEB4 (New England Biolabs), 0.2 µl de BSA 100x, 0.11 µl de enzima *NotI* (NEB) sendo completadas com água ultrapura até atingir o volume total de 20 µl. As placas contendo as reações foram seladas e deixadas sob leve agitação (40 rpm) por 6 h a 37 °C.

A eletroforese em campo pulsado foi realizada em géis de 250 ml compostos por 0,25x e 0,8 % de agarose pulse field certified (Biorad). Cada poço foi preenchido com 15 µl do produto da digestão + 2 µl de tampão de carregamento. O tamanho dos fragmentos foi monitorado pela adição de Ladder MidRange 1 (NEB). As corridas foram realizadas a 6 V/cm, com ângulo de 120 °, *switch time* inicial de 5 s e *switch time* final de 15 s, por 16 h.

3.2.4 Sequenciamento

A totalidade das reações de sequenciamento foi realizada utilizando o kit ABI BigDye Terminator v3.1 (ABI, CA, USA). As condições de reação foram modificadas em relação

àquelas sugeridas pelo fabricante, a fim de melhorar a relação custo benefício e garantir um bom nível de qualidade e tamanho de sequência. Nas reações de sequenciamento, foram utilizados os primers universais T7-Forward (5'TAATACGACTCACTATAGGG 3') e M13r-Reverse (5' CAGGAAACAGCTATGAC 3').

Cada reação continha 2000 a 2500 ng de DNA plasmidial, 2 µl de tampão 5x, 1 µl de BigDye Terminator v3.1, 0.2 µl de primer e foi preenchida com água ultrapura até completar o volume de 12 µl. As amplificações foram realizadas em termociclador utilizando o seguinte programa: desnaturação inicial a 95 °C por 5 min, seguida de 99 ciclos de 95 °C por 30 s, 50 °C por 15 s e 60 °C por 4 min. As reações foram precipitadas em seguida pela adição de acetato de sódio (1,5 M) mais EDTA (250 mM), conforme as especificações do kit de sequenciamento, e etanol absoluto gelado, sendo cada *pellet* ressuspendido em 10 µl de formamida. A injeção e a eletroforese capilar nos sequenciadores automáticos ABI 3500 ou ABI 3730 XL (Applied Biosystems) foram realizadas na Universidade Estadual de Londrina (UEL, Londrina-PR), Universidade Estadual de Campinas (UNICAMP, Campinas-SP) ou no INRA-CNRGV (Toulouse, França).

3.3 Processamento pré-anotação

Todas as sequências obtidas no presente trabalho foram submetidas a um pipeline de processamento pré-anotação visando à avaliação de qualidade, mascaramento de vetores e primers, aparamento de pontas de baixa qualidade, filtragem por tamanho e por contaminantes.

Para o desenvolvimento do pipeline de processamento, foi utilizado o software Egene, desenvolvido por Durham et al. (2005). O software Egene é um sistema de geração modular de pipelines, que torna a sua implementação rápida e eficiente.

De posse dos arquivos contendo os eletroferogramas, foi dado início à primeira etapa do processamento das sequências, que consiste da execução dos procedimentos de base calling e avaliação da qualidade das bases. Por base calling, entende-se a interpretação de um eletroferograma, atribuindo a cada um dos picos a identidade de uma base nitrogenada, baseados na forma e intensidade do pico. Uma vez que a intensidade e a forma dos picos nem sempre é consistente e inequívoca, é atribuído um valor de qualidade a cada uma das bases. O índice de qualidade reflete o limite para considerar uma base como sendo verdadeira; por

exemplo, uma base com valor de qualidade “20” possui uma probabilidade de 1 em 100 de estar incorreta (EWING; GREEN, 1998; RICHTERICH, 1998).

O procedimento de base calling foi realizado por meio do software Phred (EWING, 1998), utilizando os parâmetros padrão do sistema. Os arquivos de saída do Phred foram então convertidos em arquivos FASTA formatados, por meio do software PHD2FASTA. Os arquivos FASTA, contendo as sequências foram encaminhados à etapa seguinte para a identificação e o mascaramento de trechos com alta identidade com os primers ou vetores usados na amplificação e clonagem dos fragmentos. O mascaramento desses trechos foi realizado pela ferramenta CrossMatch (<http://www.phrap.org/phredphrapconsed.html>), utilizando os parâmetros *standard* de penalidades e de alinhamento mínimo do software. Os bancos de referência de vetores e de primers foram modificados para conterem apenas a sequência do vetor utilizado na construção da biblioteca e os primer utilizados no sequenciamento dos BAC-ends. Esta medida foi tomada a fim de diminuir o tempo de busca por esses elementos.

A etapa seguinte consistiu na eliminação de bases que apresentaram valores baixos de qualidade. Para este propósito, os valores de Phred foram computados para todas as bases e comparados com um limite mínimo pré-estabelecido de 20. Uma “boa” sequência, neste caso, é aquela que possuiu uma porcentagem de bases válidas superior a 85%. Após o mascaramento das sequências, foi realizada a excisão dos trechos mascarados ou dos que apresentaram índices de qualidade inferiores aos parâmetros previamente estipulados. Após a excisão dos trechos indesejados, algumas sequências podem conter um número de bases insuficiente para serem utilizadas. Por esse motivo, foi necessário realizar uma filtragem das sequências por tamanho, a fim de eliminar aquelas que possam gerar viés nas análises. Para as sequências de BAC-ends, foi adotado o limite mínimo de 100 pb, visando maximizar o aproveitamento dos dados.

Após a filtragem por tamanho, as sequências foram mais uma vez submetidas a análise por meio do software CrossMatch (<http://www.phrap.org/phredphrapconsed.html>), para separar aquelas que apresentaram alta identidade com sequências de DNA mitocondrial, ribossomal e plastidial. Esse processo foi basicamente realizado através de uma busca por meio de alinhamento local contra um banco de sequências dessa natureza. Por fim, as sequências de DNA nuclear foram submetidas ao software CAP3 (HUANG; MADAN, 1999) a fim de avaliar a redundância dos dados.

3.4 Anotação

Por anotação, entende-se um processo múltiplo, pelo qual uma ou mais sequências brutas de DNA ou de aminoácidos são analisadas visando à atribuição de informações biológicas, ou seja, suas funções (STEIN, 2001). A anotação das sequências esteve focada principalmente na busca por elementos transponíveis, genes, e candidatos a marcadores microssatélites. A Figura 2 demonstra o fluxograma das etapas de processamento pré-anotação e anotação:

Figura 2 - Fluxograma das etapas de processamento pré-anotação e anotação das sequências de *P. edulis*

3.4.1 Prospecção de candidatos a marcadores microssatélites

Para a identificação das sequências contendo microssatélites, foi utilizado o aplicativo SciRoKo ver 3.4 (KOFER et al., 2007). Foram considerados insertos positivos aqueles que apresentaram repetições maiores ou iguais a 10 para mono-, 5 para di-, e 3 para tri-, tetra-, penta- ou hexanucleotídeos. A classificação, quanto ao número de repetições, foi baseada em cada motivo, incluindo todas as sequências complementares e as permutadas. Por exemplo, o motivo dinucleotídico AG representa as sequências AG, GA, e os complementos CT e TC e o motivo trinucleotídico GAA representa as sequências GAA, AAG, AGA e seus complementos TTC, TCT e CTT.

3.4.2 Identificação de elementos transponíveis

Os elementos repetitivos presentes nos BES foram identificados a partir de buscas por similaridade com sequências da seção Viridiplantae do banco de dados de repetições RepBase (release 04-18-2012), utilizando como ferramentas de busca os softwares RepeatMasker ver. 3.3.0 e Cross Match 0.990329, com um cut off de 1×10^{-20} .

A densidade de elementos repetitivos foi, então, calculada como a porcentagem de nucleotídeos na sequência com pelo menos um hit em uma repetição do banco de dados. As famílias de elementos repetitivos foram classificadas com base na anotação proposta pelo RepBase, respeitando as diretrizes para classificação e a identificação de elementos repetitivos fornecidas pelo “TIGR Codes for Plant Repetitive Sequences” (<http://www.tigr.org/tdb/e2k1/plant.repeats/repeat.code.shtml>). Uma vez identificadas, as regiões correspondentes aos elementos repetitivos foram mascaradas, a fim de evitar que essas pudessem atrapalhar as análises posteriores.

3.4.3 Identificação de regiões codificadoras de proteínas e anotação funcional

Após mascaradas para a ocorrência de qualquer sorte de elementos repetitivos, as sequências foram comparadas a todos os registros do TIGR Plant Gene Indices (QUACKENBUSH et al., 2000) e do TIGR Non-Identical Amino Acids Database que contém registros de proteínas não idênticas provenientes de diversos bancos de dados como GenBank,

RefSeq e Uniprot, utilizando a ferramenta de busca BLASTX (ALTSCHUL et al., 1997), com um cut off de 1×10^{-5} .

A anotação funcional dos dados foi realizada com base nas análises do software BLAST2Go (CONESA et al., 2005), atribuindo termos ontológicos de acordo com os resultados obtidos com o BLASTX, com um cut off de 1×10^{-6} . O Gene Ontology (GO) é um vocabulário controlado de termos funcionais que permite a anotação consistente de produtos gênicos, de acordo com três ontologias principais: *Molecular Function*, *Cellular Component* e *Biological Process* (ASHBURNER et al., 2000). Os vocabulários foram reduzidos com o auxílio da função GOSlim, utilizando o termo “*PLANT*”. A contagem das bases GC e a estimativa do conteúdo GC do genoma foi realizado por um script PERL (Practical Extraction and Report Language), desenvolvido em nosso laboratório. O conteúdo GC foi calculado de acordo com a fórmula:

$$\frac{100 \times (G + C)}{(G + C + A + T)}$$

A curadoria manual foi realizada com a ajuda da ferramenta Apollo (LEWIS et al., 2002), pois esta permite tanto a visualização, quanto a edição da anotação de forma eficiente e bastante simplificada.

3.4.4 Mapeamento genômico comparativo

Procurou-se identificar potenciais regiões de microsintenia entre os genomas do maracujá e algumas espécies filogeneticamente próximas. Para tanto, utilizou-se o algoritmo BWA-SW (Burrows-Wheeler Aligner's Smith-Waterman Alignment) desenvolvido por Durbin e Li (2010). As sequências genômicas completas das espécies *Arabidopsis thaliana*, *Populus trichocarpa* e *Vitis vinifera* foram todas obtidas nos bancos de dados do projeto Phytozome (GOODSTEIN et al., 2012). Apesar de não estar proximamente relacionada ao maracujá amarelo, as informações genômicas da espécie *A. thaliana* foram selecionadas para compor as análises, devido ao seu status de organismo modelo e a riqueza de informações genômicas disponíveis para a espécie.

Uma dada região foi considerada microssintênica, em relação ao alvo, se a ambas as extremidades estivessem mapeadas em um mesmo cromossomo, distando entre 15 kb e 350 kb uma da outra. Quando as duas extremidades foram corretamente orientadas em relação

uma a outra, as regiões foram consideradas colineares. Em caso contrário, as regiões foram consideradas rearranjadas entre as duas espécies

A curadoria manual dos alinhamentos foi realizada com a ajuda da ferramenta IGV – *Integrative Genomics Viewer 2.3* (THORVALDSDOTTIR et al., 2013), pois esta permite a visualização, no genoma de referência, das regiões de microssintenia e de suas imediações, além de permitir a avaliação da qualidade dos alinhamentos.

4 RESULTADOS E DISCUSSÃO

4.1 Banco de dados relacional

A fim de melhor organizar e gerenciar o grande volume de dados gerados pelo sequenciamento dos *Bac-ends* foi implementado um banco de dados relacional baseado na linguagem MySQL. O banco de dados foi projetado para uso interno, a fim de agregar os dados não apenas do presente trabalho, mas também das futuras investidas do nosso grupo de pesquisa. Sendo assim, o design do banco de dados foi feito para ser simples e capaz de armazenar dados genômicos de diferentes origens e tipos.

Como pode ser observado na Figura 3, o modelo ER é composto por oito entidades, inter-relacionadas, cada uma com um número específico de atributos. Como é comum neste tipo de modelo, os nomes das entidades e dos atributos são todos derivados da língua inglesa.

As duas entidades mais basais são *Study* e *Species*. A entidade *Study* armazena os dados referentes ao estudo que os dados pertencem. Seus atributos são *StdCod*, um código alfanumérico único que identifica um dado estudo; *StdNam*, o nome do estudo em questão; *Resp*, a identificação do responsável técnico pelo estudo e *Inst*, o nome do instituto, universidade ou empresa realizadora do estudo. A entidade *Species* armazena os dados referentes ao objeto de estudo. Seus atributos são *SpeCod* um código alfanumérico único que identifica a espécie em estudo; *SpeCiNam*, armazena o nome científico da espécie em estudo; *SpeCoNam*, armazena o nome popular da espécie em estudo; *NcbiTaxId*, armazena o código do táxon no banco de dados do NCBI; *GenCode*, contém o código genético da espécie; *MitGenCode*, contém o código genético mitocondrial da espécie.

À entidade *Library* fica delegada a função de armazenar as informações referentes à biblioteca de onde provêm as sequências. Seus atributos são *LibCod*, um código alfanumérico único que identifica a biblioteca em estudo; *LibNam*, nome da biblioteca em estudo; *LibType*, tipo de técnica usada para a construção da biblioteca; *LibCover*, cobertura genômica da biblioteca; *NumPlate*, número de placas que compõe a biblioteca; *NumClon*, número de clones que compõe a biblioteca; *Vector*, vetor utilizado para a clonagem dos fragmentos e *Enzyme*, enzima utilizada para a construção da biblioteca.

A entidade *Clone* armazena as informações referentes ao clone utilizado para gerar as sequências. Essa entidade é composta por quatro atributos, sendo esses *ClonCod*, um código alfanumérico único que identifica o clone em estudo; *Plaque*, que armazena a identificação da

placa que contém o referido clone e *Position*, que guarda a posição do clone dentro da placa onde está armazenado.

A entidade *Sequence* é a entidade central neste modelo, visto que a mesma é responsável por unir as informações contidas na maioria das demais entidades. Esta entidade é responsável por armazenar as informações a respeito da sequência em si, sendo composta por três principais atributos; *SeqCod*, um código alfanumérico único que identifica a sequência em estudo; *SeqTech*, que armazena o nome da técnica utilizada para o sequenciamento; *SeqSize*, que armazena o tamanho da sequência.

As três entidades restantes armazenam informações sobre os elementos contidos nas sequências, sejam eles genes, elementos transponíveis ou microssatélites.

A entidade *Gene*, como o próprio nome diz, armazena informações sobre os genes identificados em cada sequência. Seus atributos são *GeneBankAcc*, que armazena o número de acesso daquele gene no *GeneBank*, *UniGeneAcc*, que armazena o número de acesso daquele gene no banco *UniGene*; *GeneName*, que armazena o nome do gene e *GeneSymbol*, que armazena o símbolo do gene em estudo.

A entidade *TE* armazena os dados referentes aos elementos transponíveis identificados nas sequências. Esta entidade é composta pelos atributos *TECod*, que armazena um código alfanumérico único que identifica o elemento transponível; *TEType*, armazena o tipo de elemento encontrado e *TEfamily*, que armazena o nome da família a qual o elemento pertence. Por fim, a entidade *SSR* armazena as informações sobre microssatélites encontradas nas sequências analisadas. Esta entidade é composta pelos atributos *SatCod*, um código alfanumérico único que identifica o microssatélite; *SatNam*, que armazena um nome dado ao microssatélite encontrado; *SatType*, que armazena o tipo de microssatélite estudado; *Motive*, armazena o motivo repetitivo presente na sequência; *RepNum*, armazena o número de vezes que aquele motivo se repete; *SatSize*, guarda a informação sobre o tamanho do microssatélite. Por fim, os atributos *PrimF* e *PrimR* armazenam as sequências de primers *forward* e *reverse* utilizadas para amplificar o marcador SSR, caso estes existam.

Figura 3 – Modelo ER (Entity Relationship) utilizado para a implementação do banco de dados para o armazenamento e gerenciamento das informações geradas pelo presente trabalho. Cada caixa representa uma das entidades que compõem o banco; dentro de cada uma delas encontram-se nomes representando os atributos que formam cada uma das entidades. Os símbolos sobre as barras que conectam as caixas indicam o tipo de relacionamento entre as entidades, podendo ser 1-∞ (um para muitos) ou ∞-∞ (muitos para muitos). Os asteriscos em frente aos nomes de alguns atributos indicam quais deles são “atributos chave” dentro da entidade

4.2 Extração dos BACs

Com o intuito de melhorar o rendimento das extrações e a facilitar a execução dos procedimentos de extração para um grande número de amostras, foi feita uma série de modificações no protocolo de extração fornecido pelo fabricante dos kits Macherey-Nagel Nucleospin® 96 Flash. A primeira dessas alterações foi a adição de um período de pré-cultura dos clones em 200 µl de meio 2x LB + 0.2 µl/ml de cloranfenicol a 37 °C sob constante agitação (325 rpm) por 7 h. Esta medida teve o objetivo de garantir um bom rendimento final de clones para as minipreparações, uma vez que a etapa de cultura por 18 h se inicia com um número bem maior de clones. Os resultados foram bastante positivos, como será apresentado mais a frente.

O tempo de ressuspensão dos *pellets* foi padronizado em 30 min sob agitação a 650 rpm. Esta medida visou facilitar o processamento de um grande número de amostras ao

mesmo tempo. Esta etapa foi padronizada em 30 min, pois nossas observações mostraram que esse tempo era suficiente para se ter todos os *pellets* desfeitos; um tempo maior poderia favorecer a quebra do DNA, uma vez tratam-se de moléculas extremamente longas.

Outra modificação foi o aumento do tempo de centrifugação em álcool 70% de 10 min para 30 min. Experiências anteriores mostram que o maior tempo de centrifugação em álcool 70% resulta em maior rendimento final e em um DNA de maior pureza. Por fim, o número de inversões após as etapas de lise e neutralização foi aumentado, visto que, quando respeitadas as recomendações do fabricante, a lise das células era parcial e interferia no rendimento final das minipreparações.

Com o intuito de comparar os resultados obtidos usando o protocolo padrão e o protocolo modificado, foi selecionado um grupo aleatório de 22 clones, e os mesmos foram submetidos à extração com os dois protocolos. Usou-se *NotI* para a digestão, seguida da migração em eletroforese de campo pulsado (PFGE). A Figura 4 mostra a digestão do mesmo grupo de amostras, extraídas com o protocolo fornecido pelo fabricante (esquerda) e com o protocolo modificado (direita). Nota-se, claramente, que nas extrações feitas com o protocolo modificado (direita), a qualidade das minipreparações foi bastante semelhante, porém com um rendimento superior.

Uma vez otimizados os protocolos, foram selecionadas aleatoriamente 22 placas de 384 poços pertencentes à biblioteca Ped-B-Flav para que fossem realizados os procedimentos de extração. Para a execução das extrações e do sequenciamento foi necessário dividir cada uma das placas de 384 poços em 4 quadrantes de 96 poços cada. Portanto, a partir deste momento, a nomenclatura de cada sequência passa a ser composta da seguinte forma Pe (Iniciais de *Passiflora edulis*) + (número da placa) + (quadrante) + (posição na placa) + (orientação). Logo, uma sequência designada Pe122Q3F5R possui o identificador da espécie (Pe), o número de identificação da placa original (122), o quadrante ao qual pertence (Q3), sua posição na placa final (linha F, coluna 5) e sua orientação (se *forward* (F) ou *reverse* (R)). A Figura 6 mostra um diagrama de correspondência entre a placa original e os quatro quadrantes que compõem aquela placa.

Figura 4 – Gel de Eletroforese em Campo Pulsado mostrando os produtos de digestão com a enzima *NotI* para avaliação da qualidade dos eventos de extração. O primeiro grupo (esquerda) de amostras foi extraído utilizando o protocolo fornecido pelo fabricante, enquanto o segundo grupo (direita) foi extraído com o protocolo modificado, descrito na seção Material e Métodos. O tamanho dos fragmentos foi acompanhado utilizando o Ladder MidRange 1 (NEB)

Foram realizados 88 eventos de extração, sendo 22 placas compostas de 384 poços divididas em por 4 placas de 96 poços cada, totalizando 8.448 amostras. Como a checagem, por meio de digestão enzimática, de todas as 8.448 extrações seria inviável, visto o alto custo e grande volume de trabalho, foi selecionada uma amostra aleatória composta de 130 extrações para a digestão com *NotI*, seguida da corrida em eletroforese em campo pulsado. A verificação por meio da digestão enzimática mostrou que as extrações haviam sido bem sucedidas, tendo alto rendimento e ótima qualidade. Além disso, notou-se bastante uniformidade em relação à qualidade de todas as extrações, sinal de que todas as etapas foram bem executadas (Figura 5).

Figura 5 – Géis de Eletroforese em Campo Pulsado mostrando os produtos de digestão com a enzima *NotI*. Os três géis mostram um grupo de 130 extrações aleatoriamente selecionadas a fim de avaliar, por amostragem, a qualidade das minipreparações. O tamanho dos fragmentos foi acompanhado utilizando o Ladder MidRange 1 (NEB)

		1		2		3		4		5		6		7		8		9		10		11		12	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	A	A1Q1	A1Q2	A2Q1	A2Q2	A3Q1	A3Q2	A4Q1	A4Q2	A5Q1	A5Q2	A6Q1	A6Q2	A7Q1	A7Q2	A8Q1	A8Q2	A9Q1	A9Q2	A10Q1	A10Q2	A11Q1	A11Q2	A12Q1	A12Q2
	B	A1Q3	A1Q4	A2Q3	A2Q4	A3Q3	A3Q4	A4Q3	A4Q4	A5Q3	A5Q4	A6Q3	A6Q4	A7Q3	A7Q4	A8Q3	A8Q4	A9Q3	A9Q4	A10Q3	A10Q4	A11Q3	A11Q4	A12Q3	A12Q4
B	C	B1Q1	B1Q2	B2Q1	B2Q2	B3Q1	B3Q2	B4Q1	B4Q2	B5Q1	B5Q2	B6Q1	B6Q2	B7Q1	B7Q2	B8Q1	B8Q2	B9Q1	B9Q2	B10Q1	B10Q2	B11Q1	B11Q2	B12Q1	B12Q2
	D	B1Q3	B1Q4	B2Q3	B2Q4	B3Q3	B3Q4	B4Q3	B4Q4	B5Q3	B5Q4	B6Q3	B6Q4	B7Q3	B7Q4	B8Q3	B8Q4	B9Q3	B9Q4	B10Q3	B10Q4	B11Q3	B11Q4	B12Q3	B12Q4
C	E	C1Q1	C1Q2	C2Q1	C2Q2	C3Q1	C3Q2	C4Q1	C4Q2	C5Q1	C5Q2	C6Q1	C6Q2	C7Q1	C7Q2	C8Q1	C8Q2	C9Q1	C9Q2	C10Q1	C10Q2	C11Q1	C11Q2	C12Q1	C12Q2
	F	C1Q3	C1Q4	C2Q3	C2Q4	C3Q3	C3Q4	C4Q3	C4Q4	C5Q3	C5Q4	C6Q3	C6Q4	C7Q3	C7Q4	C8Q3	C8Q4	C9Q3	C9Q4	C10Q3	C10Q4	C11Q3	C11Q4	C12Q3	C12Q4
D	G	D1Q1	D1Q2	D2Q1	D2Q2	D3Q1	D3Q2	D4Q1	D4Q2	D5Q1	D5Q2	D6Q1	D6Q2	D7Q1	D7Q2	D8Q1	D8Q2	D9Q1	D9Q2	D10Q1	D10Q2	D11Q1	D11Q2	D12Q1	D12Q2
	H	D1Q3	D1Q4	D2Q3	D2Q4	D3Q3	D3Q4	D4Q3	D4Q4	D5Q3	D5Q4	D6Q3	D6Q4	D7Q3	D7Q4	D8Q3	D8Q4	D9Q3	D9Q4	D10Q3	D10Q4	D11Q3	D11Q4	D12Q3	D12Q4
E	I	E1Q1	E1Q2	E2Q1	E2Q2	E3Q1	E3Q2	E4Q1	E4Q2	E5Q1	E5Q2	E6Q1	E6Q2	E7Q1	E7Q2	E8Q1	E8Q2	E9Q1	E9Q2	E10Q1	E10Q2	E11Q1	E11Q2	E12Q1	E12Q2
	J	E1Q3	E1Q4	E2Q3	E2Q4	E3Q3	E3Q4	E4Q3	E4Q4	E5Q3	E5Q4	E6Q3	E6Q4	E7Q3	E7Q4	E8Q3	E8Q4	E9Q3	E9Q4	E10Q3	E10Q4	E11Q3	E11Q4	E12Q3	E12Q4
F	K	F1Q1	F1Q2	F2Q1	F2Q2	F3Q1	F3Q2	F4Q1	F4Q2	F5Q1	F5Q2	F6Q1	F6Q2	F7Q1	F7Q2	F8Q1	F8Q2	F9Q1	F9Q2	F10Q1	F10Q2	F11Q1	F11Q2	F12Q1	F12Q2
	L	F1Q3	F1Q4	F2Q3	F2Q4	F3Q3	F3Q4	F4Q3	F4Q4	F5Q3	F5Q4	F6Q3	F6Q4	F7Q3	F7Q4	F8Q3	F8Q4	F9Q3	F9Q4	F10Q3	F10Q4	F11Q3	F11Q4	F12Q3	F12Q4
G	M	G1Q1	G1Q2	G2Q1	G2Q2	G3Q1	G3Q2	G4Q1	G4Q2	G5Q1	G5Q2	G6Q1	G6Q2	G7Q1	G7Q2	G8Q1	G8Q2	G9Q1	G9Q2	G10Q1	G10Q2	G11Q1	G11Q2	G12Q1	G12Q2
	N	G1Q3	G1Q4	G2Q3	G2Q4	G3Q3	G3Q4	G4Q3	G4Q4	G5Q3	G5Q4	G6Q3	G6Q4	G7Q3	G7Q4	G8Q3	G8Q4	G9Q3	G9Q4	G10Q3	G10Q4	G11Q3	G11Q4	G12Q3	G12Q4
H	O	H1Q1	H1Q2	H2Q1	H2Q2	H3Q1	H3Q2	H4Q1	H4Q2	H5Q1	H5Q2	H6Q1	H6Q2	H7Q1	H7Q2	H8Q1	H8Q2	H9Q1	H9Q2	H10Q1	H10Q2	H11Q1	H11Q2	H12Q1	H12Q2
	P	H1Q3	H1Q4	H2Q3	H2Q4	H3Q3	H3Q4	H4Q3	H4Q4	H5Q3	H5Q4	H6Q3	H6Q4	H7Q3	H7Q4	H8Q3	H8Q4	H9Q3	H9Q4	H10Q3	H10Q4	H11Q3	H11Q4	H12Q3	H12Q4

Figura 6 – Correspondência entre as placas de armazenamento dos clones da biblioteca Ped-B-Flav (384 poços) e as placas usadas na extração e sequenciamento dos *Bac-ends* (96 poços). A primeira linha (1-12) e primeira coluna (A-H) representam as marcações nas placas de 96 poços, enquanto a segunda linha (1-24) e segunda coluna (A-P) representam as marcações presentes nas placas de 384 poços. Cada placa de 384 poços foi subdividida em 4 quadrantes de 96 posições, representados na figura pelas cores e pelas iniciais Q1, Q2, Q3 e Q4

4.3 Sequenciamento dos BAC-ends

Visto que o procedimento de extração de BACs tem rendimento limitado e que os custos de sequenciamento pela metodologia de Sanger são relativamente elevados, foram realizados testes de sequenciamento com diferentes quantidades de DNA e BigDye Terminator v3.1. Esses testes tiveram como principal objetivo otimizar as reações de sequenciamento, a fim de utilizar as menores quantidades possíveis de DNA e BigDye Terminator v3.1, mantendo a alta qualidade das sequências e um comprimento de sequência adequado para o processamento e anotação dos BAC-ends. Os melhores resultados foram obtidos utilizando 5 µl de DNA (aproximadamente entre 2.000 a 2.500 ng) e 0,5 µl de BigDye Terminator v3.1.

Para avaliar melhor essas condições, foram selecionadas, aleatoriamente, dentre o material extraído, 10 amostras para serem sequenciadas nos sentidos *forward* e *reverse*. As 20 reações de sequenciamento foram bem sucedidas, formando 10 *mated-pairs* de alta qualidade. Como pode ser observado na Figura 7, as sequências produzidas sob essas condições apresentaram picos de bom tamanho e muito bem definidos. O tamanho das sequências também foi muito satisfatório, tendo em média 700 pb, o que é considerado muito bom para este tipo de tecnologia.

Figura 7 – Perfis de eletroferogramas obtidos com o protocolo de sequenciamento conforme descrito na seção de Material e Métodos

Após a otimização dos procedimentos de sequenciamento, um total de 5.337 clones foi sequenciado, dos quais 4.660 foram sequenciados em ambas as direções, *forward* e *reverse*, e 677 clones foram sequenciados apenas em uma direção. Dentre as sequências não pareadas, 384 possuíam apenas a informação da ponta *forward* e 293 da ponta *reverse*. Dessa forma, 9.979 reações de sequenciamento foram realizadas, produzindo 8.821 BES de boa qualidade, indicando uma eficiência de 89%. O tamanho médio das sequências foi de 596 pb, variando de 100 pb, tamanho mínimo aceitável, até 1.255 pb (Figura 8). A redundância observada no set sequenciado foi de apenas 5,4%.

Figura 8 – Distribuição por tamanho dos BAC-ends de *P. edulis*. As 8.821 sequências analisadas, após editadas, variaram de 100 pb até 1.255 pb, com um tamanho médio de 596 pb

Levando-se em consideração o tamanho médio das sequências, pode-se inferir que foram gerados, aproximadamente, 5,7 Mpb (5.704.725 bp) de informação genômica, através do sequenciamento dos BES.

Depois de eliminadas as regiões de baixa qualidade e os trechos pertencentes ao vetor de clonagem, os dados foram comparados a bancos de dados contendo sequências pertencentes a mitocôndrias e cloroplastos de diversas espécies vegetais. Essa comparação revelou uma contaminação por genomas organelares de apenas 0,002 %, tendo sido encontrados 15 BES com alta similaridade a sequências de genomas cloroplastidiais e 2 BES com alta similaridade a sequências de genomas mitocondriais.

O conteúdo GC(%) total foi de 41,92%, enquanto o conteúdo GC(%) da porção gênica foi de 45,50 %. Esta relação corrobora a ideia postulada por Madigan e Martinko (2003) de que, de um modo geral, os genes são caracterizados por terem um elevado teor de GC, em contraste com o background do genoma.

O conteúdo GC(%) total observado no genoma do maracujá amarelo foi bastante alto em comparação com o de outras dicotiledôneas como papaia, carvalho e arabidopsis, que se

enquadram entre 33 % e 35 %. Essa estimativa é alta mesmo em comparação com espécies consideradas filogeneticamente próximas, como o álamo (33,72 %), mamona (32,5 %) e videira (34,55 %). Porém, como observado por Datema et al. (2008), a estimativa do conteúdo GC pode ser enviesada pela escolha da enzima de restrição utilizada na construção da biblioteca. Por outro lado, o conteúdo GC(%) observado no genoma do maracujá amarelo se aproxima da média normalmente encontrada em monocotiledôneas, como arroz, cana de açúcar e banana, que varia entre 43 % e 47 %. (MING et al., 2012). A Tabela 3 dá um apanhado geral dos resultados obtidos com o sequenciamento e processamento dos BAC-ends.

Tabela 3 – Estatística descritiva dos BES de *P. edulis*

Total de BES Sequenciados	9979
Sequências de alta qualidade	8821
Aproveitamento total	89,0 %
Redundância observada	5,4 %
Paired BES	9.320
Non-Paired BES	677
DNA cloroplastidial	11
DNA Mitocondrial	2
Tamanho médio dos BES	596 pb
Tamanho mínimo dos BES	100 pb
Tamanho máximo dos BES	1.255
Conteúdo GC médio (8821 BES)	41,92 %
Conteúdo GC gênico (767 BES)	45,50 %

4.4 Prospecção de candidatos a marcadores SSR

Antes do mascaramento das regiões repetitivas e dos elementos transponíveis, as sequências foram submetidas a uma busca por elementos repetitivos do tipo microssatélites SSR (*Simple Sequence Repeat*). Feita a busca automática pelos elementos repetitivos, cada cromatograma foi manualmente verificado, a fim de averiguar a qualidade dos picos e a confiabilidade da anotação automática.

Ao todo, foram encontrados 610 SSR nos 5.704.725 pb analisados. A densidade média estimada entre os microssatélites foi de 10,69 SSR a cada 100 Kpb (ou 1 SSR a cada 9,35 Kpb). Esta estimativa se aproxima das observadas em espécies como o nabo (10,46; HONG

et al., 2006), a batata (11,57; DATEMA et al., 2008) e a couve (11,13; <http://brassica.bbsrc.ac.uk/>), porém encontra-se bem abaixo das proporções observadas em espécies filogeneticamente mais próximas, como a videira (17,24; JAILLON et al., 2007) e o álamo (18,89; TUSKAN et al., 2006).

Em relação aos motivos encontrados, estes foram divididos em 6 classes, de acordo com o número de nucleotídeos. Do total, 15,2 % foram de mononucleotídeos, 18,7 % de dinucleotídeos, 12,7 % de trinucleotídeos, 28,9 % de repetições de tetranucleotídeos, 16,5 % de pentanucleotídeos e 8,0 % de hexanucleotídeos.

A título de comparação, foram feitas análises para a prospecção de candidatos a marcadores microssatélites em dados provenientes de projetos de sequenciamento de BAC-ends em cinco outras espécies de dicotiledôneas (*Vitis vinífera*, *Carica papaya*, *Cucumis melo*, *Malus x domestica* e *Manihot esculenta*), dados esses públicos obtidos junto ao site do NCBI (www.ncbi.nlm.nih.gov). As análises foram realizadas utilizando os mesmos softwares e parâmetros usados na prospecção de microssatélites nos BAC-ends de *P. edulis*. A distribuição do número de microssatélites categorizados quanto à classe, assim como a comparação com os dados observados nas análises realizadas nas demais espécies podem ser observadas na Figura 9 e Tabela 4.

Tabela 4 – Distribuição das percentagens de microssatélites encontrados nos BES de *P. edulis*, categorizados quanto à classe de repetições

	<i>Passiflora edulis</i>	<i>Vitis vinífera</i>	<i>Arabidopsis thaliana</i>	<i>Cucumis melo</i>	<i>Malus domestica</i>
Mono	15,2	6,0	40,3	39,7	17,2
Di	18,7	28,0	13,1	24,2	48,0
Tri	12,7	18,7	21,2	27,4	8,4
Tetra	28,9	14,9	3,9	4,9	6,9
Penta	16,5	21,4	15,3	2,9	19,4
Hexa	8,0	11,0	6,3	0,9	12,5
Nº de BES analisados	8.821	31.907	26.068	23.878	3.744

Um fato interessante a ser observado é a alta ocorrência de motivos tetranucleotídeos em relação às demais classes. Claramente, os motivos de tetranucleotídeos foram expressivamente mais abundantes que os demais, seguidos pelos dinucleotídeos e pentanucleotídeos. Esta disposição é pouco comum entre as dicotiledôneas, sendo normalmente os motivos de dinucleotídeos ou mononucleotídeos os mais abundantes.

Oliveira et al. (2008), trabalhando com *P. edulis* f *flavicarpa* observou uma proporção de 3,9 % de mononucleotídeos, 78,7 % de dinucleotídeos, 3,9 % de trinucleotídeos, 7,9 % de tetranucleotídeos e 3,2 % de repetições de hexanucleotídeos. Apesar das proporções serem bastante discrepantes em relação ao observado nas análises dos BES, as proporções observadas por Oliveira et al. (2008), segundo os próprios autores, são bastante enviesadas, uma vez que essas foram obtidas a partir de uma biblioteca genômica enriquecida com microssatélites, que foi confeccionada com enzimas e sondas que favorecem a obtenção de microssatélites com motivos de dinucleotídeos.

Figura 9 – Distribuição das percentagens de microssatélites encontrados nos BES de *P. edulis*, categorizados quanto à classe de repetições, e comparação com as proporções reportadas em outras espécies de dicotiledôneas

Em meio aos microssatélites pertencentes à classe dos mononucleotídeos, aqueles compostos por repetições A/T superaram em muito a proporção daqueles compostos por repetições G/C. Análises realizadas em diversas plantas, como *Arabidopsis thaliana*, *Brassica napus*, *Vitis vinifera* e *Populus trichocarpa*, tem mostrado que, de modo geral, dicotiledôneas apresentam alta frequência de microssatélites ricos em A/T, enquanto monocotiledôneas, como *Zea mays* e *Oryza sativa*, apresentam menores proporções desses motivos (HSU et al., 2011).

Como pode ser observado na Tabela 5, o motivo mais abundante em todas as categorias foi o tetranucleotídeo AATT, com um total de 131 ocorrências. Os mononucleotídeos mais abundantes foram os compostos por A/T, enquanto os dinucleotídeos mais frequentes foram aqueles compostos por AT, seguidos pelos compostos por AG. Em nenhuma das classes foi observada alta frequência de motivos ricos em GC. Essas observações estão de acordo com estudos feitos em maçã (HAN et al., 2008), *Arabidopsis*

(TAMANNA; KHAN, 2005), soja (SHULTZ et al., 2007), mamão (LAI et al., 2006), entre outras plantas, que mostram que motivos ricos em AT são muito mais prevalentes nos genomas das plantas superiores. A distribuição dos motivos dentro de cada classe em *P. edulis* segue a mesma linha do que normalmente é observado em estudos realizados com BAC-ends.

Tabela 5 – Distribuição dos microssatélites presentes nos BES de *P. edulis*

Motivo	Ocorrência	Motivo	Ocorrência
Monomeros	93	Pentameros	(continuação)
A/T	57	AAATC	1
C/G	36	ATACC	1
Dímeros	114	AAACC	1
AT	62	ATGAC	1
AG	45	AACAG	1
AC	7	AACCC	1
Trímeros	77	AATCT	1
AAG	47	AAGCT	1
AAT	13	AATAG	1
AAC	3	AATCC	1
AGG	6	ACACG	1
AGC	2	ACGGC	1
ACC	3	AGAGG	1
ACG	1	ATCTC	1
CCG	2	Hexameros	49
Tetrameros	176	AAAAAG	10
AATT	131	AAAGTC	4
AAAG	17	AAAAAT	3
AAAT	14	AAAGAG	2
AAAC	3	AAAAAC	4
ATAC	4	ACCACG	2
AACC	2	AAGATG	2
ACTC	1	AATTAC	1
AGCC	1	ACCTCG	1
ATCC	1	AGAGGG	1
AGGG	1	AACTGG	1
AATG	1	AACAGC	1
Pentameros	101	AAGGAG	2
AAAAG	29	AAGGTC	1
ATATC	17	AATGAG	1
AAAAT	14	AACAAG	1
AACCG	5	ACCCGT	2
AAGAG	4	AGGGGG	1
AAAAC	4	AAAATC	1
AAATT	4	AAAATT	1
AAATG	1	AAACCG	1
AATAT	3	AAAGAT	1
ACACG	1	AACACC	1
ACGGC	1	AATCCC	1
AGGGG	1	ACCAGC	1
AACTC	2	ATATCC	1
AACAT	1	AAAACC	1

4.5 Elementos transponíveis

Com o intuito de identificar elementos repetitivos presentes no genoma do maracujá amarelo, 8.821 sequências foram comparadas com o banco de dados RepBase, utilizando a fração do banco que corresponde apenas às sequências de organismos vegetais. Tal qual abordado no item de Revisão, elementos transponíveis (TEs) são elementos móveis de DNA, que podem se mover e replicar dentro dos genomas hospedeiros, sendo amplamente prevalentes na maioria dos genomas eucarióticos. De acordo com o tipo de transposição, mediada por RNA ou DNA, os TEs podem ser divididos em duas classes principais (WICKER et al., 2007). Elementos de classe I ou retrotransposons, são o tipo mais comum de elemento genético móvel em genomas eucarióticos e se transpõem através de um intermediário de RNA por meio de um mecanismo de “Copiar e colar”. Assim como nos demais eucariotos, Elementos de classe I são os mais abundantes em genomas vegetais (FESCHOTTE et al., 2002). Elementos de classe II, ou transposons de DNA, utilizam um mecanismo de “cortar e colar” envolvendo a reintegração do elemento no genoma através de um intermediário de DNA (BOWEN; JORDAN, 2002).

Como pode ser observado na Tabela 6, 19,69 % de todos os BAC-ends possuíam a presença de algum elemento repetitivo, somando 1.117.681 bases mascaradas. Apesar de a quantidade de DNA repetitivo encontrada no genoma do maracujá amarelo ser maior do que a observada em genomas pequenos como o de *Arabidopsis thaliana* (10%; The Arabidopsis Genome Initiative 2000), este valor ainda é muito inferior ao observado em outros genomas de dicotiledôneas já sequenciados, como álamo (42%; TUSKAN et al. 2006), mamão (51.9%; MING et al. 2012), e maçã (42.4%; VELASCO et al. 2010).

Dentre esses potenciais elementos transponíveis, os Transposons Classe I ou Retrotransposons foram expressivamente mais abundantes que os Transposons Classe II, sendo observados respectivamente em 18.44 % e 0.05 % de todos os BES analisados.

Tabela 6 – Classificação e distribuição dos elementos genéticos repetitivos de *P. edulis*

Classificação	Número de elementos	Bases mascaradas	Porcentagem de nucleotídeos
Retroelementos:	3.116	1.052.143 pb	18.44 %
LINES/SINES	69	19.683 pb	0.35 %
L1/CIN4	66	19.543 pb	0.34 %
Elementos LTR	3.047	1.032.460 pb	18.10 %
Ty1/Copia	1.486	452.581 pb	7.93 %
Gypsy/DIRS1	1.547	574.945 pb	10,08 %
DNA Transposons	36	2.767 pb	0.05 %
Não Classificados	7	1.192 pb	0.02 %
Pequenos RNAs	2	162 pb	0.00 %
Satélites	1	41 pb	0.00 %
Repetições simples	305	14.444 pb	0.25 %
Regiões de Baixa Complexidade	927	47.948 pb	0.84 %
Total	4.394	1.117.681 pb	19.59 %

As grandes diferenças no tamanho dos genomas das espécies de plantas estão, geralmente, associadas à presença de diferentes quantidades de retrotransposons. Quanto maior o genoma vegetal, maior a chance de este conter uma grande quantidade de retroelementos. Por exemplo, genomas grandes, como o da cevada, são compostos por até 70% desses elementos (VICIENT et al., 2001), enquanto em genomas pequenos, como o do arroz, esses representam apenas 17 % da composição total do genoma (McCARTHY et al., 2002).

Neste estudo, foram identificados retrotransposons em cerca de 18,54 % dos BES. Proporcionalmente, o valor não é alto, porém, deve-se levar em consideração que o genoma de *P. edulis* é maior que a média de tamanho do genoma das dicotiledôneas. Logo, 18,54 % de um genoma que pode chegar a 1,5 Gpb representa uma quantidade enorme de elementos transponíveis.

Dentre os elementos retrotransponíveis identificados, 69 pertencem à categoria dos LINE/SINES. SINES (*Short Interspersed Nuclear ElementS*) são sequências não-codificadoras, repetitivas, que possuem entre 70 e 500 pares de bases, enquanto os LINES (*Long Interspersed Nuclear ElementS*) podem chegar até 7.000 pares de bases. Estas

sequências têm grande importância na evolução dos genomas e podem ser usadas como marcadores aplicados a estudos de filogenia, como feito por Nikaido et al. (1999) que investigaram relações filogenéticas entre Cetáceos e outros mamíferos marinhos e Schimitz et al. (2001) que aplicaram os mesmos elementos na investigação de relações filogenéticas entre primatas.

SINES e LINES são inseridos na porção não-codificadora de um genoma. Assim, a presença de elementos similares em um mesmo loco, em espécies diferentes, implica que ambas espécies possuem um ancestral comum, no qual o evento de inserção ocorreu. Aparentemente, não existem mecanismos de perda desses elementos, salvo eventos que deletem grandes porções do genoma. Além disso, esses elementos não apenas indicam parentesco, mas também qual das espécies é mais primitiva, uma vez que o último ancestral comum que liga as duas espécies que compartilham um elemento deste tipo deve ter surgido depois do último ancestral comum que liga essas espécies à anterior, que não possui o elemento (LESK, 2008).

Um parâmetro comumente observado em estudos com *BAC-ends* é a proporção de retrotransposons Ty3-gypsy em relação a Ty1-copia. Estes grupos são baseados em similaridade de sequências e na ordem dos genes que codificam para a função da polimerase dentro da região dita poli-proteína (POL). Sabe-se que esses elementos são de grande importância para a evolução da arquitetura genômica de uma espécie (WILHELM;WILHELM, 2001).

Em maracujá, foi observado uma proporção aproximada de 1:1, o que indica que, aparentemente, as classes contribuíram igualmente para a evolução do genoma do maracujá amarelo. Esta proporção é a mesma encontrada em *Arabidopsis* (The Arabidopsis Genome Initiative, 2000), porém, diferentes proporções vêm sendo observadas em diversas espécies vegetais como 1:3 em maçã (HAN et al., 2008), 2:1 em arroz (International Rice Genome Sequencing Project 2005), 1:2 em carvalho (RAMPAMT et al., 2011), 2:3 em cana de açúcar (KIM et al., 2013) e 5:1 em orquídeas *Phalaenopsis* (HSU et al., 2011).

Outros dois fatos interessantes a serem observados é a baixa quantidade de elementos não classificados e de MITEs. O primeiro fato pode ser explicado, principalmente, pelo fato de os bancos de espécies relativamente próximas, como *Populus trichocarpa*, já estarem muito bem descritos. O segundo caso se deve, provavelmente, à enzima escolhida para a construção da biblioteca, *HindIII*. De acordo com Han et al. (2008), a detecção de MITEs pode ser significativamente afetada tanto pela enzima utilizada na produção da biblioteca

como por estruturas secundárias formadas pelos MITEs, que dificultam o seu sequenciamento.

4.6 Regiões codificadoras de proteínas

Com o objetivo de identificar regiões codificadoras de proteínas, as sequências dos BAC-ends foram comparadas às do banco de “NR” (Non redundant) pertencente ao NCBI, com o auxílio da ferramenta de busca Blast X. Foram identificadas 767 sequências, cerca de 8,69 % do total de BES, contendo alto índice de similaridade com genes de diversas espécies vegetais.

A grande maioria dos *top-hits* (ocorrência mais relevante para cada alinhamento) mostrou similaridade com as espécies *Populus trichocarpa*, *Ricinus communis*, *Prunus persica* e *Vitis vinifera*, com 28,8 %, 16,6 %, 13,1 % e 8,7 % do total de ocorrências, respectivamente. Esse cenário já era esperado, uma vez que essas são as espécies mais próximas de *P. edulis* para as quais existe uma abundância de dados genômicos. A distribuição das espécies cujas sequências apresentaram similaridade com as de *P. edulis* f. *flavicarpa* pode ser observada na Figura 9.

Figura 9 – Distribuição dos melhores hits (*top-hits*) dos BES de *P. edulis* contra sequências de todas as espécies vegetais presentes no banco “NR (Non redundant)”, do NCBI. As buscas foram realizadas por meio da ferramenta de BlastX

Na sessão de Anexos encontram-se todas as sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits* (Tabela A1).

Tal qual abordado no item de Revisão, o conteúdo genômico de *P. edulis* f. *flavicarpa*, relatado por Souza et al. (2004), é de 3,19 pg. Se considerarmos uma proporção de 1:1 de AT:GC e ignorarmos a presença de nucleotídeos modificados, pode-se utilizar a relação estabelecida por Dolezel et al. (2003) de $0,978 \times 10^9$ bases para cada um 1 pg de DNA. Nesses termos, o genoma (2C) de *Passiflora edulis* f. *flavicarpa* teria, aproximadamente, 3.119 Mpb ou 1.560 Mpb para o lote haplóide (1,56 Gpb).

O número BES homólogos a proteínas presentes do banco de dados “NR” (767) representa cerca de 8,7 % do tamanho total do *set* de BES sequenciados (5.704.725 pb), assumindo o tamanho médio dos BAC-ends de *P. edulis* como 596 pb. Dessa forma, pode-se presumir que aproximadamente 135,72 Mpb do genoma do maracujá amarelo é composto por sequências codantes. Assumindo que o tamanho médio de um gene de *P. edulis* é 3,4 Kpb, o mesmo do observado em *Vitis vinifera* (JAILLON et al., 2007), pode-se estimar o número total de genes da espécie em 40 mil.

Tanto a estimativa da proporção codante como a do número de genes, estão de acordo, por exemplo, com as propostas por Hsu et al. (2011) para *Phaleanopsis equestris*, espécie esta com genoma haplóide de tamanho bem próximo ao de *P. edulis*, cerca de 1.600 Mpb. No estudo de Hsu et al. (2011), a porção codante do genoma foi estimada em 9,99 % e o número total de genes em 47 mil. A estimativa do número de genes também se encontra dentro do intervalo de valores observados nos genomas próximos já sequenciados, de 30.434 genes em *Vitis vinifera* (JAILLON et al., 2007) e 45.555 genes em *Populus trichocarpa* (TUSKAN et al., 2006).

Uma vez identificada, por meio de BlastX, a similaridade dos BES com alguma sequência gênica presente no banco de dados “NR”, os 767 BES positivos foram submetidas à análise com o software Blast2Go, a fim de realizar a anotação funcional das sequências selecionadas. A anotação realizada pelo Blast2Go consiste, basicamente, em associar as sequências às ontologias criadas pelo *Gene Ontology Consortium*. Essas ontologias têm por objetivo padronizar a representação dos genes e de seus produtos para todos os sistemas biológicos, subdividindo-os em três principais categorias: *Biological process* – que se refere à atividade biológica com a qual o gene, ou seu produto, contribui; *Molecular function* – atividade bioquímica do gene ou de seu produto; *Cellular component* – local na célula onde o gene ou seu produto é ativo.

Ao todo, foram anotados 4.258 termos ontológicos associados a 686 das 767 sequências analisadas, o que resulta em uma média de 7,6 termos por sequência. A distribuição dos termos pelas sequências foi de 1 a 39, sendo os termos relacionados a *Molecular function* os mais abundantes e os termos relacionados a *Cellular component*, os menos abundantes. Do total de 683 sequências associadas a termos ontológicos, 558 sequências foram associadas a pelo menos um termo relacionado a *Molecular function*, 536 sequências foram associadas a pelo menos um termo relacionado a *Biological process*, apenas 387 sequências foram associadas a pelo menos um termo relacionado a *Cellular component* e 85 sequências não foram associadas a nenhum termo ontológico.

Para o nível hierárquico 1, o mais basal e, portanto, menos abrangente, foram anotados 596 ocorrências dos três termos mestres da ontologia gênica, sendo que 42% das ocorrências pertenciam ao termo *Molecular function*, 34% a *Biological process* e 24 % a *Cellular component*. A Figura 11 mostra as proporções e os valores absolutos das ocorrências no nível 1.

Figura 11 – Análise da Ontologia Gênica dos BES de *P. edulis* para em nível 1

Apesar de as proporções nem sempre serem idênticas, de modo geral, entre as dicotiledôneas, a categoria *Molecular function* é a mais frequente, seguida de perto pela categoria *Biological process*. Na grande maioria dos estudos a categoria *Cellular Component* é a de menor ocorrência dentre os termos de nível 1. Essas proporções podem ser claramente

observadas nos trabalhos realizados por Rampant et al. (2011) em carvalho, Han et al. (2008) em maçã e GAO et al. (2011), trabalhando com tremoço. Uma das poucas exceções à regra ocorreu na análise dos BAC-ends de noqueira, realizada por WU et al. (2011), onde *Biological process* obteve o maior número de ocorrências, seguido por *Cellular component* e *Molecular function* com o menor número de ocorrências.

Em relação ao nível hierárquico 2, dentro da categoria *Molecular function*, os termos de maior ocorrência foram *Binding* (GO:0005488), *Catalytic activity* (GO:0003824) e *Transporter activity* (GO:0005215), representando 49 %, 40 % e 4 % do total de termos dentro da categoria, respectivamente. A distribuição dos demais termos ontológicos pertencentes ao termo *Molecular function* pode ser observada na Figura 12.

Figura 12 – Análise da Ontologia Gênica dos BES de *P. edulis* para a categoria *Molecular function*

Segundo as definições fornecidas pelo próprio *Gene Ontology Consortium*, proteínas com a função *Binding* facilitam a interação seletiva, de forma não covalente, de uma molécula com uma região específica de outra molécula; já aquelas com a função *Catalytic activity*, como sugere o próprio nome, tem por função catalisar reações bioquímicas a temperaturas

fisiológicas e, por sua vez, as proteínas com a função *Transporter activity* são conhecidas por permitir o movimento dirigido de substâncias (por exemplo, macromoléculas, pequenas moléculas, ions, etc.) através das membranas da célula, ou entre células.

Ao descer na escala hierárquica do *Gene Ontology*, os termos ontológicos tornam-se menos abrangentes, referindo-se a eventos cada vez mais específicos em termos de funções metabólicas, processos biológicos e componentes celulares.

O termo *Binding* também foi o mais frequentemente observado em carvalho (RAMPANT et al., 2011), tremoço (GAO et al., 2011) e noqueira (WU et al., 2011), seguido por *Catalytic activity* nos três casos. O terceiro termo mais frequente em álamo e noqueira também foi *Transporter activity*, porém, em tremoço o terceiro termo mais frequente foi *Transferase activity*, sendo *Transporter activity* apenas o quinto termo de ocorrência mais frequente.

Apesar de muito importantes, a ocorrência de termos de ontologia gênica do nível hierárquico 3 em diante é muito pouco discutida. A grande maioria dos trabalhos disponíveis na literatura se atém em discutir as ocorrências dos dois primeiros níveis, realizando apenas a anotação dos demais termos. Apesar de pouco discutidos, quanto mais se desce na hierarquia do Gene Ontology mais específicos são os termos, possibilitando assim uma maior precisão ao relacionar um gene associado a um termo a um determinado evento fisiológico.

Dentre os termos de nível 3, dentro da categoria *Molecular function*, foram atribuídos 260 termos às sequências analisadas. Os termos observados com maior frequência foram *Small molecule binding* (GO:0036094), *Transferase activity* (GO:0016740) e *Nucleic acid binding* (GO:0003676), com 22,3%, 21,5% e 18,8% de todas as ocorrências, respectivamente. Via de regra, genes associados ao termo *Small molecule binding* estão relacionados à interação seletiva e não covalente com compostos de baixo peso molecular como, por exemplo, monossacarídeos. Genes associados ao termo *Transferase activity* codificam enzimas com a função de transferir grupos funcionais de um composto doador para um outro composto aceptor. Por sua vez, genes associados ao termo *Nucleic acid binding* estão relacionados à interação seletiva e não covalente com algum tipo de ácido nucléico. A Tabela 7 detalha todas as ocorrências de nível 3, dentro da categoria *Molecular function*.

Tabela 7 – Distribuição dos termos ontológicos de nível 3 dentro da categoria *Molecular function*

<i>Molecular Function Lvl-3</i>		
Termos	Ocorrências	Porcentagem (%)
<i>Small molecule binding</i>	58	22,3
<i>Transferase activity</i>	56	21,5
<i>Nucleic acid binding</i>	49	18,8
<i>Protein binding</i>	43	16,5
<i>Hydrolase activity</i>	42	16,2
<i>Signal transducer activity</i>	3	1,2
<i>Oxygen binding</i>	3	1,2
<i>Carbohydrate binding</i>	2	0,8
<i>Chromatin binding</i>	2	0,8
<i>Sequence-specific DNA binding transcription factor activity</i>	1	0,4
<i>Lipid binding</i>	1	0,4
Total	260	

Ao todo, foram anotados 118 termos de nível 4, dentro da categoria *Molecular Function*. Os termos ontológicos mais frequentes foram *Nucleotide Binding* (GO:0000166), *Transferase activity, transferring phosphorus-containing groups* (GO:0016772) e *DNA Binding* (GO:0003677), com 49,9 %, 22,9 % e 14,4 % das ocorrências, respectivamente. Como o próprio nome sugere, genes associados ao termo *Nucleotide binding* estão relacionados à interação seletiva e não covalente com nucleotídeos. Assim como no termo de nível 3, genes associados ao termo *Transferase activity* estão relacionados à produção de enzimas com a função de transferir grupos funcionais de um composto doador para um outro composto aceptor, porém, neste caso específico, as enzimas são especializadas na transferência de grupos fosfato. Por fim, o termo *DNA binding* relaciona-se a qualquer função molecular pela qual um produto gênico interage de forma seletiva e não covalente com uma molécula de DNA. A Tabela 8 detalha todas as ocorrências de nível 4, dentro da categoria *Molecular function*.

Tabela 8 – Distribuição dos termos ontológicos de nível 4 dentro da categoria *Molecular Function*

<i>Molecular Function Lvl-4</i>		
Termos	Ocorrências	Porcentagem (%)
<i>Nucleotide binding</i>	58	49,2
<i>Transferase activity</i>	27	22,9
<i>DNA binding</i>	17	14,4
<i>RNA binding</i>	8	6,8
<i>Hydrolase activity, acting on ester bonds</i>	5	4,2
<i>Hydrolase activity, acting on acid anhydrides</i>	2	1,7
<i>Receptor binding</i>	1	0,8
Total	118	

No que tange ao nível 2 da *Biological process*, os termos de maior ocorrência foram *Metabolic Process* (GO:0008152), com 25 % das ocorrências, *Cellular process* (GO:0009987) com 24 % e *Response to stimulus* (GO:0050896) com 10 %. A distribuição dos demais termos pertencentes a *Biological process* é apresentada na Figura 13.

Figura 13 – Análise da Ontologia Gênica dos BES de *P. edulis* para a categoria *Biological process*

Por definição, as proteínas que carregam o termo *Metabolic process* estão envolvidas na realização de processos metabólicos, sejam eles catabólicos ou anabólicos. Proteínas relacionadas ao termo *Cellular process* estão envolvidas em algum processo biológico em nível celular, porém não necessariamente restrito a uma célula apenas. Já as proteínas relacionadas ao termo *Response to stimulus* atuam em qualquer processo que resulta em uma mudança de estado ou atividade de uma célula ou de um organismo (em termos de movimento, secreção, produção de enzimas, expressão do gene, etc.), como consequência de um estímulo. O processo inicia-se com a detecção do estímulo e termina com uma mudança de estado ou de atividade na célula ou no organismo. Esta última categoria é muito interessante para os grupos de pesquisa que trabalham com maracujá amarelo, uma vez que grande parte dos genes de interesse agrônomo, como aqueles relacionados à resposta à seca, resistência a doenças, etc. estão agrupados dentro do termo *Response to stimulus*.

Assim como acontece com os termos de nível 2 pertencentes à categoria *Molecular function*, os dois termos mais frequentes da categoria *Biological process* seguem o padrão observado na literatura. *Metabolic process* e *Cellular process* representam a maior parte das ocorrências e alternam, de acordo com a espécie, entre a primeira e segunda posição, possuindo de modo geral frequências de ocorrência bastante próximas. Porém, em relação aos demais termos não existe tanta uniformidade. Em carvalho (RAMPANT et al. 2011) e nojeira (WU et al., 2011) o terceiro e quarto termos mais frequentes foram, respectivamente, *Localization* e *Biological regulation*. Já em tremço (GAO et al., 2011) as mesmas posições foram ocupadas respectivamente por *Multicellular organismal development* e *Cellular component organization and biogenesis*.

Ao todo, foram anotados 824 termos ontológicos de nível 3 para a categoria *Biological process*. Devido à grande quantidade de ocorrências, este nível se mostrou altamente diverso, sendo composto por 27 termos ontológicos distintos. Os termos mais frequentes foram *Primary metabolic process* (GO:0044238) com 12,5 % das ocorrências, *Cellular metabolic process* (GO:0044237) com 10,8 % das ocorrências e *Macromolecule metabolic process* (GO:0043170) com 9 %.

O termo *Primary Metabolic Process* relaciona-se às reações químicas e às vias metabólicas que envolvem compostos formados como parte dos processos anabólicos e catabólicos normais da célula. Já o termo *Cellular Metabolic Process* relaciona-se às reações químicas e às vias metabólicas pelas quais células individuais transformam substâncias químicas. O termo *Macromolecule Metabolic Process* relaciona-se às reações químicas e vias

metabólicas responsáveis pela criação, destruição e pelo processamento de moléculas de alto peso molecular, geralmente biopolímeros.

Cabe ressaltar que, dentro do nível hierárquico 3 da categoria *Biological Process* encontram-se diversos termos relacionados às características de interesse agrônomo, como resposta à estresse (*Response to stress*), resposta à estímulos abióticos e bióticos (*Response to abiotic stimulus*; *Response to biotic stimulus*; *Cellular response to stimulus*) e fixação de nitrogênio (*Nitrogen compound metabolic process*).

A Tabela 9 detalha todas as ocorrências de nível 3, dentro da categoria *Biological Process*.

Tabela 9 – Distribuição dos termos ontológicos de nível 4 dentro da categoria *Biological process*

<i>Biological Process Lvl-3</i>		
Termos	Ocorrências	Porcentagem (%)
<i>Primary metabolic process</i>	103	12,5
<i>Cellular metabolic process</i>	89	10,8
<i>Macromolecule metabolic process</i>	74	9,0
<i>Biosynthetic process</i>	61	7,4
<i>Cellular component organization</i>	48	5,8
<i>Response to stress</i>	45	5,5
<i>Establishment of localization</i>	41	5,0
<i>Nitrogen compound metabolic process</i>	40	4,9
<i>Multicellular organismal development</i>	39	4,7
<i>Response to abiotic stimulus</i>	31	3,8
<i>Regulation of biological process</i>	31	3,8
<i>Anatomical structure development</i>	30	3,6
<i>Cell communication</i>	28	3,4
<i>Cellular response to stimulus</i>	24	2,9
<i>Catabolic process</i>	23	2,8
<i>Cell cycle</i>	20	2,4
<i>Response to biotic stimulus</i>	19	2,3
<i>Reproductive process</i>	18	2,2
<i>Response to endogenous stimulus</i>	16	1,9
<i>Cellular developmental process</i>	13	1,6
<i>Secondary metabolic process</i>	12	1,5
<i>Cell growth</i>	8	1,0
<i>Response to external stimulus</i>	5	0,6
<i>Cell-cell signaling</i>	3	0,4
<i>Regulation of biological quality</i>	1	0,1
<i>Cell death</i>	1	0,1
<i>Cellular homeostasis</i>	1	0,1
Total	824	

No nível hierárquico 4 foi mapeada a ocorrência de 553 termos ontológicos pertencentes a categoria *Biological process*. O termos mais comuns foram *Cellular macromolecule metabolic process* (GO:0044260), que se refere às vias metabólicas de produção de biopolímeros, *Protein metabolic process* (GO:0019538), que refere a processos metabólicos envolvendo proteínas específicas, como, por exemplo, sua modificação, e *Macromolecule modification* (GO:0043412) que refere-se a processos que alteram de forma covalente um ou mais monômeros de um biopolímero. Estes termos representam 12,3 %, 11,2% e 8,7 % de todas as ocorrências, respectivamente. A Tabela 10 detalha todas as ocorrências de nível 4, dentro da categoria *Biological Process*.

Tabela 10 – Distribuição dos termos ontológicos de nível 4 dentro da categoria *Biological Process*

<i>Biological Process Lvl-4</i>		
Termos	Ocorrências	Porcentagem (%)
<i>Cellular macromolecule metabolic process</i>	68	12,3
<i>Protein metabolic process</i>	62	11,2
<i>Macromolecule modification</i>	48	8,7
<i>Transport</i>	41	7,4
<i>Nucleobase-containing compound metabolic process</i>	40	7,2
<i>Cellular nitrogen compound metabolic process</i>	40	7,2
<i>Post-embryonic development</i>	28	5,1
<i>Anatomical structure morphogenesis</i>	26	4,7
<i>Carbohydrate metabolic process</i>	26	4,7
<i>Signal transduction</i>	24	4,3
<i>Regulation of cellular process</i>	24	4,3
<i>Gene expression</i>	20	3,6
<i>Developmental process involved in reproduction</i>	17	3,1
<i>Lipid metabolic process</i>	14	2,5
<i>Cell differentiation</i>	13	2,4
<i>Cellular biosynthetic process</i>	11	2,0
<i>Macromolecule biosynthetic process</i>	11	2,0
<i>Regulation of metabolic process</i>	10	1,8
<i>Embryo development</i>	9	1,6
<i>Generation of precursor metabolites and energy</i>	9	1,6
<i>Photosynthesis</i>	6	1,1
<i>Response to extracellular stimulus</i>	4	0,7
<i>Pollination</i>	1	0,2
<i>Homeostatic process</i>	1	0,2
Total	553	

Dentre os termos pertencente a categoria *Cellular component* no nível hierárquico 2, os mais frequentes foram *Cell* (GO:0005623) com 43 % das ocorrências, *Organelle*

(GO:0043226) com 28 % e *Membrane* (GO:0016020) 22 %, como pode ser observado na Figura 14, em conjunto com a distribuição dos demais termos ontológicos pertencentes a *Cellular component*.

Figura 14 – Análise da Ontologia Gênica dos BES de *P. edulis* para a categoria *Cellular Component*

O termo *Cell* refere-se àquelas proteínas que atuam dentro do ambiente da célula, o termo *Organelle* se refere àquelas proteínas que atuam no núcleo, mitocôndrias, plastos, vacúolos, vesículas, ribossomos e no citoesqueleto, enquanto aquelas relacionadas ao termo *Membrane* atuam exclusivamente na membrana plasmática.

Ao contrário do que ocorre nas categorias *Biological process* e *Molecular function*, a ocorrência dos termos de nível 2, observada nos BES de *P. edulis*, variou muito pouco em relação ao observado na literatura. De modo geral, *Cell* é o termo mais frequente, seguido por *Organelle*, *Membrane*, *Macromolecular complex*, e *Extracellular region*, nesta ordem. Esta sequência se mantém mesmo entre as monocotiledôneas, como observado por KIM et al. (2013), trabalhando com cana-de-açúcar.

No nível hierárquico 3 foram mapeadas 230 ocorrências de apenas quatro termos ontológicos, sendo estes *Cell part* (GO:0044464), que refere-se às proteínas que atuam em qualquer parte da célula, com exceção de membranas e organelas, *Membrane-bounded organelle* (GO:0043227), que refere-se às proteínas que atuam ligadas às membranas plasmáticas da célula, *Non-membrane-bounded organelle* (GO:0043228), que refere-se às proteínas que estão nas membranas plasmáticas da célula, porém não ligadas a elas e *Organelle part* (GO:0044422) que refere-se às proteínas que atuam especificamente nas organelas celulares. A Tabela 11 detalha todas as ocorrências de nível 3, dentro da categoria *Cellular component*.

Tabela 11 – Distribuição dos termos ontológicos de nível 3 dentro da categoria *Cellular Component*

<i>Cellular Component Lvl-3</i>		
Termos	Ocorrências	Porcentagem (%)
<i>Cell part</i>	128	55,7
<i>Membrane-bounded organelle</i>	81	35,2
<i>Non-membrane-bounded organelle</i>	15	6,5
<i>Organelle part</i>	6	2,6
Total	230	

Por fim, foram mapeadas 296 ocorrências de sete termos ontológicos de nível 4, pertencentes ao termo *Cellular component*. Os termos de maior ocorrência foram *Intracellular* (GO:0005622) com 75 % das ocorrências, *Cell periphery* (GO:0071944) com 10,1 % das ocorrências e *Plasma membrane* (GO:0044237) com 10,1% das ocorrências. Esses termos referem-se, respectivamente aos produtos gênicos que compõem o ambiente intracelular, a região periférica da célula e a membrana plasmática.

Tabela 12 – Distribuição dos termos ontológicos de nível 4 dentro da categoria *Cellular Component*

<i>Cellular Component Lvl-4</i>		
Termos	Ocorrências	Porcentagem (%)
<i>Intracellular</i>	222	75,0
<i>Cell periphery</i>	30	10,1
<i>Plasma membrane</i>	27	9,1
<i>Organelle lumen</i>	6	2,0
<i>Intracellular organelle part</i>	6	2,0
<i>External encapsulating structure</i>	5	1,7
Total	296	

4.7 Mapeamento genômico comparativo

Um total de 8.821 BES foi comparado aos genomas das espécies *Vitis vinifera*, *Populus trichocarpa* e *Arabidopsis thaliana*, com o objetivo de identificar regiões microssintênicas entre a espécie em estudo e os genomas de referência. Os dois primeiros organismos foram escolhidos por possuírem genomas completos e serem filogeneticamente próximos do maracujá amarelo, além disso, as espécies *Populus trichocarpa* e *Vitis vinifera* estão entre as que mostraram os maiores números de *top-hits* nas buscas feitas no banco de dados do NCBI. Apesar de também figurarem nas primeiras colocações e serem filogeneticamente muito próximas de *P. edulis*, os genomas das espécies *Prunus persica* e *Ricinus communis* ainda se encontram em fases iniciais de montagem e anotação, o que prejudica enormemente as análises de mapeamento genômico comparativo. Já a espécie *Arabidopsis thaliana* foi escolhida por ser o genoma modelo padrão em plantas e, também, por ser o genoma vegetal mais bem anotado.

Os *matches* (ocorrência de similaridade entre um BES de *P. edulis* e uma região específica do genoma modelo) foram classificados em sete categorias, baseadas na classificação proposta por Rampant et al. (2011). As duas categorias mais basais são as das “Pontas únicas” ou “soltas”, que correspondem aos BES em que apenas uma das duas sequências, F (*Forward*) ou R (*Reverse*), possui correspondência no genoma de referência e “Pontas pareadas”, quando ambas as pontas possuem correspondência no genoma de referência, mas não necessariamente localizadas no mesmo cromossomo.

A categoria das “Pontas pareadas” foi subdividida em “Co-localizadas” e “Não Co-localizadas”, sendo que a primeira ocorre quando ambas as pontas encontram correspondência em um mesmo cromossomo e a segunda quando esta condição não é cumprida. Por sua vez, a classe das “Co-localizadas” foi subdividida em 3 categorias. “Interrompidas”, quando ambas as pontas se encontraram no mesmo cromossomo, porém com um espaçamento menor que 15 Kpb e maior que 350 Kpb, “Colineares” quando o espaçamento entre as pontas se encontra dentro do intervalo de 15 a 350 Kpb e ambas as pontas apresentam a mesma orientação observada no genoma do maracujá amarelo; e “Rearranjadas”, quando a orientação difere daquela observada no genoma sob estudo.

Em relação ao genoma de *A. thaliana*, foram obtidos 106 *matches* positivos, sendo 102 compostos por pontas soltas e apenas 2 compostos por duas pontas pareadas.

O par composto pelas sequências Pe69Q4G9F e Pe69Q4G9R foi ancorado no genoma cloroplastidial de *A. thaliana*, estando as sequências localizadas, respectivamente, nas posições 130159 pb e 50130 pb deste cromossomo. As sequências distam 80.029 pb, dentro do intervalo para serem consideradas colineares, e sua orientação confere com aquela observada em *P. edulis*. Dadas estas características, pode-se considerar as sequências como colineares e o intervalo, que abriga 55 genes, como provável região de microssintenia entre os genomas de *P. edulis* e *A. thaliana*. A Figura 15 fornece uma representação esquemática de região em questão.

Figura 15 – Representação esquemática do conteúdo da região de 79 kpb do cromossomo C de *Arabidopsis thaliana*, localizada entre as posições 50130 pb e 130159 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P. edulis* mapeadas no genoma de *A. thaliana*

O par composto pelas sequências Pe164Q1G7F e Pe164Q1G7R também foi ancorado no genoma cloroplastidial de *A. thaliana*, estando a ponta “F” localizada na região 72.358 pb e a ponta “R” localizada na região 106.018 pb. Por estarem localizadas no mesmo cromossomo, distarem cerca de 33 Kpb e estarem orientadas de forma diferente daquela apresentada no genoma do maracujá amarelo, esse par foi classificado como rearranjado. A pequena distância entre as sequências (muito inferior aos 100 kpb referente ao tamanho médio dos BACs de *P. edulis*) e a orientação das sequências (ambas orientadas para um mesmo lado) sugerem que a região passou por algum evento de rearranjo genômico durante a sua evolução, como uma quebra ou uma inversão cromossômica.

No mapeamento comparativo feito com o genoma de *V. vinifera* foram obtidos 189 *matches* positivos, sendo 181 compostos por pontas soltas e 8 compostos por pontas pareadas. Dos pares compostos por pontas pareadas, 1 foi classificado como colinear, 1 como rearranjado, 1 como interrompido e 1 como pareado não co-localizado.

O par composto pelas sequências Pe164Q2A6F e Pe164Q2A6F foi mapeado no cromossomo 11 do genoma de *V. vinifera*, estando a ponta “R” localizada na região 2.324.226 pb e a ponta “F” localizada na região 2.502.417 pb. Este par atende a todos os requisitos necessários para a identificação de uma região microssintênica, estando ambas as pontas localizadas no mesmo cromossomo, organizadas com a mesma orientação apresentada no genoma de *P. edulis* e distando cerca de 178 kb. Por esses motivos, essa região foi classificada como colinear. Uma vez que o genoma de *V. vinifera* já se encontra sequenciado, é possível investigar a região microssintênica e inferir sobre o conteúdo do genoma de *P. edulis*. Como pode ser observado na Figura 16, a região em questão é composta por 26 genes, com tamanhos que vão desde pequenos genes compostos por apenas um exon, até genes muito grandes, compostos por mais de 10 exons. O genoma da videira ainda encontra-se em processo de anotação, sendo assim, a identidade da maioria desses genes é incerta, havendo apenas inferências computacionais sobre sua função biológica.

Figura 16 – Representação esquemática do conteúdo da região de 177 Kpb do cromossomo 11 de *Vitis vinifera*, localizada entre as posições 2324226 pb e 2502417 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P. edulis* mapeadas no genoma de *V. vinifera*

Vale notar que o par Pe164Q1G7F/Pe164Q1G7R que em *A. thaliana* foi classificado como rearranjado e alocado no genoma cloroplastidial da espécie, aqui se encontra alocado no cromossomo 11. No genoma de *V. vinifera*, o par também apresentou orientação diferente da observada em *P. edulis*, porém nesse caso a distância entre as duas sequências foi ainda menor, distando apenas em 9 Kpb, menor que os 15 Kpb mínimos estipulados; dessa forma, classificado como interrompido.

O par composto pelas sequências Pe69Q4G9F e Pe69Q4G9R foi classificado como interrompido, pois apesar de ambas as sequências estarem localizadas em um mesmo

cromossomo, elas possuem uma orientação diferente da observada em *P. edulis* e estão separadas por mais de 435.860 pb. Esta conformação, muito provavelmente, é resultado de um evento de inversão cromossômica ocorrido no genoma de uma das espécies após sua separação em clados diferentes.

O par composto pelas sequências Pe214Q2A9F e Pe214Q2A9R foi classificado como pareado não co-localizado, uma vez que as sequências foram mapeadas em cromossomos diferentes. A ponta “F” do par foi localizada no cromossomo 16 do genoma da videira, enquanto a ponta “R” foi mapeada no cromossomo 1 da espécie. Uma vez que o o lote haplóide (n) do genoma de *Vitis* é composto por 19 cromossomos e o do genoma de *P. edulis* é composto por 9 cromossomos, uma explicação plausível para a ocorrência dessas sequências pareadas não co-localizadas seria a fissão de alguns cromossomos, em relação ao *set* do ancestral comum de ambas as espécies, dando origem a um número maior de cromossomos em *V. vinifera* e separando, assim, regiões cromossômicas.

Por fim, o mapeamento realizado com o genoma de *Populus trichocarpa* revelou um número de regiões sintênicas maior do que o observado nas comparações com os dois genomas anteriores. Ao todo, foram observados 326 *matches* positivos, sendo estes compostos por 294 pontas soltas e 16 pares de sequências.

Dentre as sequências classificadas como pontas pareadas, 22 foram classificadas como pareadas não co-localizadas. Isso se deve ao fato de, em todos os casos, as pontas pareadas terem sido localizadas em cromossomos diferentes. Assim como observado no genoma de *V. vinifera*, essas ocorrências se devem, muito provavelmente, a fissões nos cromossomos de *P. trichocarpa* durante o processo evolutivo. Esta hipótese é reforçada pelo fato de a espécie apresentar um lote haplóide $n = 19$, maior que o dobro do presente em *P. edulis*.

O par composto pelas sequências Pe173Q4A8F e Pe173Q4A8R foi classificado como rearranjado, uma vez que as sequências foram mapeadas em um intervalo de 195 Kpb, porém, com uma orientação diferente da observada no genoma de *P. edulis*.

Já o par composto pelas sequências Pe164Q1G7F e Pe164Q1G7R, além de não apresentar a orientação adequada, as sequências distavam mais de 6 Mpb, sendo assim classificado como interrompido. Vale notar que este par também foi mapeado nos genomas de *V. vinifera* e *A. thaliana*, nestes casos com intervalos bem menores entre as sequências.

Aparentemente, apesar de os trechos que flanqueiam a região serem bastante conservados, pode especular que esta região é altamente instável.

O par composto pelas sequências Pe216Q4A11F e Pe216Q4A11R atende a todos os pré-requisitos para ser classificado como colinear, estando ambas as sequências localizadas no cromossomo 12, separadas por cerca de 2179 Kpb e orientadas da forma correta. A região flanqueada pelas sequências é composta por 21 genes, como pode ser observado na Figura 17.

Figura 17 – Representação esquemática do conteúdo da região de 217 Kpb do cromossomo 12 de *Populus trichocarpa*, localizada entre as posições 9673965 pb e 9890993 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P. edulis* mapeadas no genoma de *P. trichocarpa*

Os dois últimos pares classificados como colineares foram anotados no cromossomo 14. O fato de estarem no mesmo cromossomo torna esses BACs excelentes candidatos para serem usados como sondas em estudos com BAC-FISH ou mesmo para serem sequenciados por completo. Além disso, essas sequências serão de grande valia para a montagem do cromossomo “14”, em um futuro projeto de sequenciamento do genoma de *P. edulis*.

O primeiro desses pares é composto pelas sequências Pe75Q4B6 (F/R), estando ambas separadas por aproximadamente 176 kpb e orientadas adequadamente. A região flanqueada pelas sequências abriga cerca de 29 genes, como pode ser observado na Figura 18.

Figura 18 – Representação esquemática do conteúdo da região de 176 Kpb do cromossomo 14 de *Populus trichocarpa*, localizada entre as posições 9166211 pb e 9343042 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *P. trichocarpa*

O segundo par é formado pelas sequências Pe164Q1F9F e Pe164Q1F9R. Estas encontram-se separadas por cerca de 306 Kpb, próximo do limite máximo estipulado para serem consideradas colineares. A região flanqueada pelas sequências abriga um total de 45 genes, relacionados com um grande número de processos fisiológicos diferentes. A Figura 19 fornece uma representação esquemática de região em questão.

Figura 19 – Representação esquemática do conteúdo da região de 306 Kpb do cromossomo 14 de *Populus trichocarpa*, localizada entre as posições 5539303 pb e 5845525 pb. As caixas azuis representam os genes presentes na região e acima das caixas são expostos os nomes de cada gene. As setas vermelhas representam as posições dos BAC-ends de *P.edulis* mapeadas no genoma de *P. trichocarpa*

Assim como acontece com o genoma de *V. vinifera*, o genoma de *P. trichocarpa* ainda se encontra em processo inicial de anotação. A grande maioria dos genes ainda está

identificada de forma putativa, pois foi baseada apenas em predições computacionais, sem qualquer embasamento experimental. Estas limitações prejudicam a riqueza de detalhes das inferências feitas com base no mapeamento comparativo. Porém, conforme avançam os esforços de anotação desses genomas, as informações geradas poderão beneficiar os estudos genômicos em *P. edulis*.

Já era esperado que fosse obtido um maior número de regiões potencialmente sintênicas entre o genoma de *P. edulis* e o de *P. trichocarpa*, em relação aos demais comparados, uma vez que, dentre os três, este se encontra muito mais próximo, evolutivamente falando, das passifloras. Estes resultados sugerem que as relações de colinearidade tendem a erodir conforme nos afastamos filogeneticamente do gênero *Passiflora*.

Tabela 13 – Mapeamento comparativo de BAC-ends de *P. edulis* em outros genomas vegetais

Genoma de Referência	Pontas Pareadas	Cromossomo de Referência	Coordenadas (pb)	Distância (pb)	Categoria	Conteúdo
<i>Arabidopsis thaliana</i>	Pe69Q4G9 (F/R)	Cromossomo C	50.130 - 130.159	80.029	Colinear	55 genes
	Pe164Q1G7 (F/R)	Cromossomo C	72.358 - 106.018	33.660	Rearranjado	
<i>Vitis vinifera</i>	Pe164Q2A6 (F/R)	Cromossomo 11	2.324.226 - 2.502.417	178.191	Colinear	26 genes
	Pe164Q1G7 (F/R)	Cromossomo 11	10.252.239 - 10.261.324	9.085	Rearranjado	
	Pe69Q4G9 (F/R)	Cromossomo 15	5.395.003 - 9.753.677	4.356.574	Interrompido	
	Pe214Q2A9 (F/R)	Cromossomos 16/1			Não Co-localizados	
<i>Populus trichocarpa</i>	Pe216Q4A11 (F/R)	Cromossomo 12	9.673.965 - 9.890.993	169.077	Colinear	21 genes
	Pe75Q4B6 (F/R)	Cromossomo 14	9.166.211 - 9.343.042	176.831	Colinear	29 genes
	Pe164Q1F9 (F/R)	Cromossomo 14	5.539.303 - 5.845.525	306.222	Colinear	45 genes
	Pe164Q1G7 (F/R)	Cromossomo 11	7.197.221 - 13.862.010	6.664.789	Interrompido	
	Pe173Q4A8 (F/R)	Cromossomo 09	10.462.257 - 10.657.680	195.423	Rearranjado	
	Pe164Q2A6 (F/R)	Cromossomos 18/06			Não Co-localizados	
	Pe1Q1G9 (F/R)	Cromossomos 05/02			Não Co-localizados	
	Pe214Q2A9 (F/R)	Cromossomos 12/15			Não Co-localizados	
	Pe93Q2D2 (F/R)	Cromossomos 08/10			Não Co-localizados	
	Pe93Q4D8 (F/R)	Cromossomo 04 / Scaffold 21			Não Co-localizados	
	Pe216Q1E3 (F/R)	Cromossomos 06/18			Não Co-localizados	
	Pe141Q2F4 (F/R)	Cromossomos 03/01			Não Co-localizados	
	Pe69Q1B4 (F/R)	Cromossomos 09/04			Não Co-localizados	
	Pe69Q1C6 (F/R)	Scaffold 104 / Cromossomos 10			Não Co-localizados	
	Pe69Q4G9 (F/R)	Cromossomos 11/13			Não Co-localizados	
	Pe75Q1F8 (F/R)	Cromossomos 04/17			Não Co-localizados	

4.8 Organização geral do genoma de *Passiflora edulis* f *flavicarpa*

Partindo do ponto que a biblioteca Ped-B-Flav é virtualmente aleatória e representativa do genoma do maracujá amarelo, os dados aqui apresentados fornecem uma visão geral sobre a composição do genoma da espécie. Foi possível extrapolar as proporções observadas na porcentagem sequenciada da biblioteca e projetá-las para o genoma total.

Assim, foi possível fazer as seguintes inferências sobre o genoma de *P.edulis*: é composto por aproximadamente 40 mil genes, ou seja, 8,70 % do genoma são genes.

A porção repetitiva é responsável por, aproximadamente, 19,59 % do genoma. Desta porção, tem-se que 18,44 % são formadas por elementos retrotransponíveis, divididos principalmente em LINES/SINES (0,35 %), LTR-Copia (7,93 %) e LTR-Gypsy (10,08 %). Apenas 0,05 % do genoma são formados por transposons, 0,01 % por pequenos RNAs e 0,001 % por repetições do tipo DNA satélite. Outros 1,09 % do genoma são preenchidos por repetições simples e regiões de baixa complexidade. Cerca de 71,7 % da composição de todo o genoma ainda permanecem obscuros, sendo necessários estudos mais aprofundados para reduzir esta proporção. A Figura 20 mostra um panorama da distribuição dos diferentes tipos de elementos genômicos identificados neste trabalho.

Em diversos aspectos, os dados obtidos para o genoma de *P.edulis* se alinham com aqueles pertencentes a outras espécies utilizadas como referência.

O número de genes inferido pela análise dos BAC-ends é bastante realista, levando-se em consideração os números observados nos genomas vegetais já sequenciados. A distribuição dos termos de ontologia gênica também não dissoa com os dados disponíveis na literatura, seja de projetos de geração de BAC-ends ou de sequenciamento completo do genoma.

Em alguns pontos específicos, o genoma do maracujá amarelo se mostra muito singular, como por exemplo, no conteúdo GC, relativamente alto para dicotiledôneas, estando mais próximo do observado na maioria das monocotiledôneas. A alta incidência de microsatélites do tipo tetranucleotídeo também é pouco relatada na literatura, aumentando a “*estranheza*” desse genoma que começa a ser esclarecido.

Figura 20 – Composição estimada do genoma de *P. edulis* após a análise de 8.821 sequências oriundas do sequenciamento de BAC-ends

A análise dos BAC-ends foi capaz de gerar um panorama inicial do genoma de *Passiflora edulis* e a cada estudo futuro, este quadro tende a se esclarecer, podendo culminar no sequenciamento completo do genoma e sua anotação.

De modo geral, o objetivo de elevar a um novo patamar o conhecimento genômico existente sobre a espécie foi alcançado com êxito. Este trabalho dá mais um passo na caminhada iniciada com a construção da biblioteca *Ped-b-Flav* e abre caminho para uma grande quantidade de novos estudos, sejam esses nos campos de ciência básica ou aplicada. Espera-se que os dados gerados sirvam de alicerce para uma miríade de conhecimentos ainda por serem desvendados.

5 CONCLUSÕES

O sequenciamento e a análise de BAC-ends são uma forma efetiva e com bom custo benefício para a obtenção dos primeiros *insights* sobre o genoma do maracujá amarelo.

A estratégia de BAC-end sequencing permite obter um grande número de informações sobre o genoma em estudo, principalmente no que tange à geração de novos marcadores SSR, identificação de elementos transponíveis e regiões codificadoras de proteínas, bem como avaliar a existência regiões microssintênicas entre o genoma de *P. edulis* e os genomas de outras espécies vegetais.

As inferências feitas com base nos diferentes tipos de elementos genômicos encontrados mostram que cerca de 9 % do genoma do maracujá amarelo é codante, com um total de cerca de 40 mil genes, 20 % do genoma é composto por DNA repetitivo, enquanto que cerca de 71 % do genoma ainda permanece obscuro.

A anotação funcional das regiões codificadoras de proteínas permite a aquisição de um conhecimento muito mais aprofundado sobre os genes identificados e cria possibilidades para a seleção genes candidatos para uma série de estudos nos mais variados campos.

621 BES de *P. edulis* possuem algum grau de similaridade com os genomas de *Populus trichocarpa*, *Vitis vinifera* e *Arabidopsis thaliana*, dentre estes, cinco pares atendem aos quesitos mínimos para serem considerados colineares.

REFERÊNCIAS

- ABREU, P.P.; SOUZA, M.M.; SANTOS, E.A.; PIRES, M.V.; PIRES, M.M.; ALMEIDA, A.A.F. Passion flower hybrids and their use in the ornamental plant market: perspectives for sustainable development with emphasis on Brazil. **Euphytica**, Wageningen, v. 166, p. 307-315, 2009.
- AGRIANUAL 2012. **Anuário da agricultura brasileira**. São Paulo: FNP-Consultoria e Comércio, 2011. p.482.
- AGUIAR, D.R.D.; SANTOS, C.C.F. Importância econômica e mercado. In: Bruckner, C.H.; Picanço, M.C. **Maracujá: tecnologia de produção, pós-colheita, agroindústria, mercado**. Porto Alegre: Cinco Continentes. 2001, p. 305-344.
- ALTSCHUL, S.F.; MADDEN, T.L.; SCHÄFFER, A.A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, London, v. 25, p. 3389-3402, 1997.
- AMMIRAJU, J.S.; LUO, M.; GOICOECHEA, J.L.; WANG, W.; KUDRNA, D.; MUELLER, C.; TALAG, J.; KIM, H.; SISNEROS, N.B.; BLACKMON, B.; FANG, E.; TOMKINS, J.B.; BRAR, D.; MACKILL, D.; MCCOUCH, S.; KURATA, N.; LAMBERT, G.; GALBRAITH, D.W.; ARUMUGANATHAN, K.; RAO, K.; WALLING, J.G.; GILL, N.; YU, Y.; SANMIGUEL, P.; SODERLUND, C.; JACKSON, S.; WING R.A. The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. **Genome Research**, New York, v.16, p.140-147, 2006.
- ASHBURNER, M.; BALL, C.A.; BLAKE, J.A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J.M.; DAVIS, A.P.; DOLINSKI, K.; DWIGHT, S.S.; EPPIG, J.T.; HARRIS, M.A.; HILL, D.P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J.C.; RICHARDSON, J.E.; RINGWALD, M.; RUBIN, G.M.; SHERLOCK, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature Genetics**, New York, v. 25, p. 25-29, 2000.
- BAIG, M.N.; YU, A.; GUO, W.; DENG, X. Construction and characterization of two Citrus BAC libraries and identification of clones containing the phytoene synthase gene. **Genome**, Ottawa, v. 52, p. 484-489, 2009.
- BISWAS, M.K.; CHAI, L.J.; MAYER, C.; XU, Q.; GUO, W.W.; DENG, X.X. Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in *Citrus*, **Molecular Biology Reports**, Basel, v. 39, p. 5373-5386, 2012.
- BARROSO, G.M. **Passifloraceae**. Rio de Janeiro: Livros Técnicos e Científicos; São Paulo: Editora da Universidade de São Paulo. 1978. **Sistemática de Angiospermas do Brasil v. 1**:p. 194-197.
- BARBOSA, P.R.; VALVASSORI, S.S.; BORDIGNON, C.L.; KAPPEL, V.D.; MARTINS, M.R.; GAVIOLI, E.C.; QUEVEDO, J.; REGINATTO, F.H. The aqueous extracts of *Passiflora alata* and *Passiflora edulis* reduce anxiety-related behaviours without affecting memory process in rats. **Journal of Medicinal Food**, Larchmont, v. 11, p. 282-288. 2008.
- BENNETT, M.D.; BHANDOL, P.; LEITCH, I.J. Nuclear DNA amounts in angiosperms and their modern uses, targest, trends and tomorrow. **Annals of Botany**, London, v. 86, p. 467-590, 2011.

- BENNETZEN, J.L.; MA, J.; DEVOS, K. M. Mechanisms of recent genome size variation in flowering plants. **Annals of Botany**, London, v. 95, p. 127-132, 2005.
- BOWDEN, M.W. A list of chromosome numbers in higher plants.II Menispermaceae to Verbenaceae. **American Journal of Botany**, Batimore, v. 32, p. 191-201, 1945.
- BOUZIDI, M.F.; FRANCHEL, J.; TAO, Q.; STORMO, K.; MRAZ, A.; NICOLAS, P.; MOUZEYAR, S. A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions. **Theoretical and Applied Genetics**, New York, v.113, p.81-89, 2006.
- BOWEN, N.J.; JORDAN, I.K. Transposable elements and the evolution of eukaryotic complexity. **Current Issues in Molecular Biology**, Wymondham, v. 4, p. 65-76, 2002.
- BUDIMAN, M.A.; MAO, L.; WOOD, T.C.; WING, R.A. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. **Genome Research**, New York, v.10, p.129-136, 2000.
- BURKE, D.T.; CARLE, G.F.; OLSON, M.V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. **Science**, Washington, v. 236, p. 806-811, 1987.
- CARNEIRO, M.S.; CAMARGO, L.E.A.; COELHO, A.S.G.; VENCovsky, R.; LEITE, R.P.; STENZEL, N.M.; VIEIRA, M.L.C. RAPD-based genetic linkage maps of yellow passion fruit (*Passiflora edulis Sims f. flavicarpa* Deg.). **Genome**, Ottawa, v. 45, n. 4, p. 670-678, 2002.
- CAPDEVILLE, G.D; SOUZA JÚNIOR, M.T.; SZINAY, D.; DINIZ, L.E.C.; WIJNKER, E.; SWENNEN, R.; KEMA, G.H.J.; JONG, H.D. The potential of high-resolution BAC-FISH in banana breeding. **Euphytica**, Wageningen, v.166, p. 431-443, 2009.
- CARVALHO, M.C.C.G.; SILVA, D C.G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, Santa Maria, v. 40, p. 735-744, 2010.
- CAVAGNARO, P.F.; CHUNG, S.M.; SZKLARCZYK, M.; GRZEBELUS, D.; SENALIK, D.; ATKINS, A.E.; SIMON, P.W. Characterization of a deep-coverage carrot (*Daucus carota* L.) BAC library and initial analysis of BAC-end sequences. **Molecular Genetics and Genomics**, Berlin, v. 281, p. 273-288, 2009.
- CENCIL, A.; CHANTRET, N.; KONG, X.; GU, Y.; ANDERSON, O.D.; FAHIMA, T.; DISTELFELD, A.; DUBCOVSKY, J. Construction and characterization of a half million clone BAC library of durum wheat (*Triticum turgidum ssp. durum*). **Theoretical and Applied Genetics**, New York, v. 107, p. 931-939, 2003.
- CERVI, A.C. Passifloraceae do Brasil: estudo do gênero *Passiflora* L. subgênero *Passiflora*. **Fontqueria**, Washington, v. 45, p. 1-92, 1997.
- CERVI, A.C.; Milward-de-Azevedo, M.A; Bernacci, L.C. 2010. Passifloraceae. In: **Lista de Espécies da Flora do Brasil**. Rio de Janeiro:Jardim Botânico do Rio de JaneiroDisponível em: <<http://floradobrasil.jbrj.gov.br/2012/FB012506>>Acesso em: 21 fev. 2013.
- CERVI, A.C.; LINSINGEN, L. *Passiflora kikiana*, uma nova espécie de *Passifloraceae* da Amazônia brasileira. **Acta Botanica Brasilica**, São Paulo, v. 4, p. 1062-1064, 2010.

- CHEN, C.; YU, C.; HOU, S.; LI, Y.; EUSTICE, M.; SKELTON, R.L.; VEATCH, O.; HERDES, R.E.; DIEBOLD, L.; SAW, J.; FENG, Y.; QIAN, W.; BYNUM, L.; WANG, L.; MOORE, P.H.; PAULL, R.E.; ALAM, M.; MING, R. Construction of a sequence-tagged high-density genetic map of papaya for comparative structural and evolutionary genomics in brassicales. **Genetics**, Austin, v. 177, p. 2481-2491, 2007.
- CHENG, Z.; PRESTING, G.G.; BUELL, C.R.; WING, R.A.; JIANG, J. High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice, **Genetics**, Austin, v. 157, p. 1749-1757, 2001.
- CHEUNG, F.; TOWN, C.D.; A BAC end view of the *Musa acuminata* genome. **BMC Plant Biology**, London, v. 7, p. 29-46, 2007.
- CIU, L.; WALL, P.K.; HEEBENS-MACK, J.H.; LINDSAY, B.G.; SOLTIS, D.E.; DOYLE, J.J.; SOLTIS, P.S.; CARLSON, J.E.; ARUMUGANATHAN, K.; BARAKAT, A. Widespread genome duplications throughout the history of flowering plants, **Genome Research**, New York, v. 16, p. 738–749, 2006.
- COE, E.; CONE, K.; MCMULLEN, M.; CHEN, S.S.; DAVIS, G.; GARDINER, A.; LISCUM, E.; POLACCO, M.; PATERSON, A.; SANCHEZ-VILLEDA, H.; SODERLUND, C.; WING, R.) Access to the Maize Genome: An integrated physical and genetic map. **Plant Physiology**, Lancaster, v. 128, p. 9-12, 2002.
- COLLINS, J.; HOHN, B. Cosmids: a type of plasmid gene-coding vector that is packageable in vitro in bacteriophage lambda heads. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 83, p. 4242-4246, 1978.
- CONESA, A.; GÖTZ, S.; GARCIA-GOMEZ, J.M.; TEROL, J.; TALON, M.; ROBLES, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, Oxford, v. 21, p. 3674-3676, 2005.
- CUCO, S.M.; VIEIRA, M.L.C.; MONDIN, M.; AGUIAR-PERECIN, M.L.R. Comparative karyotype analysis of three *Passiflora* L. species and cytogenetic characterization of somatic hybrids. **Caryologia**, Firenze, v.58, p. 220-228, 2005.
- DATEMA, E.; MUELLER, L.A.; BUELS, R.; GIOVANNONI, J.J.; VISSER, R.G.; STIEKEMA, W.J.; VAN HAM, R.C. Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. **BMC Plant Biology**, London, v. 8, 2008. . < > Acesso em: 8 mai. 2013.
- DOLEŽEL, J.; BARTOŠ, J.; VOGLMAYR, H.; GREILHUBER, J. Nuclear DNA content and genome size of trout and human. **Cytometry**, New York, v. 51, p. 127-128, 2003.
- DURBIN, R.; LI, H. Fast and accurate long-read alignment with Burrows-Wheeler transform. **Bioinformatics**, Oxford, v. 26, p. 589-595, 2010.
- DURHAM, A.M.; KASHIWABARA, A.Y.; MATSUNAGA, F.T.; AHAGON, P.H.; RAINONE, F.; VARUZZA, L. E GRUBER A. EGene: a configurable pipeline generation system for automated sequence analysis. **Bioinformatics**, Oxford, v.21, n. 12, p. 2812-2813, 2005.

EUSTICE, M.; YU, Q.Y.; LAI, C.W.; HOU, S.B.; THIMMAPURAM, J.; LIU, L.; ALAM, M.; MOORE, P.H.; PRESTING, G.G.; MING, R. Development and application of microsatellite markers for genomic analysis of papaya. **Tree Genetics and Genomes**, New York, v. 4, p. 333-341, 2008.

EVANS, G.M.; REES, H.; SNELL, C.L.; SUN, S. The relationship between nuclear DNA amount and the duration of the mitotic cycle. **Chromosomes Today**, New York, v. 3, p. 24-31, 1972.

EWING, B.; HILLIER, L.; WENDL, M.C.; GREEN, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research**, New York, v. 8, n. 3, p. 175-185, 1998.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome Research**, New York, v. 8, n. 3, p. 186-194, 1998.

FAJARDO, D.; ANGEL, F.; GRUM, M.; TOHME, J.; LOBO, M.; ROCA, W.M.; SANCHEZ, I. Genetic variation analysis of the genus *Passiflora* L. using RAPD markers. **Euphytica**, Wageningen, v. 101, p. 341-347, 1998.

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

FEUILLET, C. *Passifloraceae* (Passion Flower Family). In: SMITH, N.; MORI, S.A.; HENDERSON, A.; STEVENSON, D.W.; HELD, S.V. (Ed). **Flowering plants of the neotropics**. Oxford: Princeton University Press; New York Botanical Garden, 2004. p. 286-287.

FEUILLET, C.; MacDOUGAL, J.M. Checklist of recognized species names of passion flowers. **Passiflora**, Beltsville, v. 12, n. 2, p. 41-43, 2003.

FEUILLET, C.; MacDOUGAL, J.M. A new infrageneric classification of *Passiflora*. **Passiflora**, Beltsville, v. 14, p. 1-4, 2004.

FLAVELL, A.J.; DUNBAR, E.; ANDERSON, R.; PEARCE, S.R.; HARTLEY, R.; KUMAR, A. Ty1- copia group retrotransposons are ubiquitous and heterogeneous in higher plants. **Nucleic Acids Research**, London, v. 20, p. 3639-3644, 1992.

FRELICHOWSKI, J.E.; PALMER, M.B.; MAIN, D.; TOMKINS, J.P.; CANTRELL, R.G.; STELLY, D.M.; YU, J.; KOHEL, R.J.; ULLOA, M. Cotton genome mapping with new microsatellites from *Acala 'Maxxa'* BAC-ends. **Molecular Genetics and Genomics**, Berlin, v. 275, p. 479-491, 2006.

FRENGEN, E.; WEICHENHAN, D.; ZHAO, B.; OSOEGAWA, K.; VAN GEEL, M.; DE JONG, P.J. A modular, positive selection bacterial artificial multiple cloning sites. **Genomics**, San Diego, v. 23, p. 250-258, 1999.

FUCHS, J.; KUHNE M.; SCHUBERT, I. Assignment of linkage groups to pea chromosomes after karyotyping and gene mapping by fluorescent *in situ* hybridization. **Chromosoma**, New York, v. 107, p. 272-276, 1998.

- GAO, L.L.; HANE, J.K.; KAMPHUIS, L.G.; FOLEY, R.; SHI, B.J.; ATKINS, C.A.; SINGH, K.B. Development of genomic resources for the narrow leafed lupin (*Lupinus angustifolius*): construction of a bacterial artificial chromosome (BAC) library and BAC-endsequencing. **BMC Genomics**, London, v. 12, 2011. <<http://www.biomedcentral.com/1471-2164/12/521>> Acesso em: 8 mai. 2013.
- GENE ONTOLOGY CONSORTIUM. The Gene Ontology (GO) project in 2006. **Nucleic Acids Research**, London, v. 1, p. D322-D326, 2006.
- GOODSTEIN, D.M.; SHU, S.; NEUPANE, R.; HAYES, R.D.; FAZO, J.; MITROS, T.; DIRKS, W.; HELSTEN, U.; PUTMAN, N.; ROKHSAR, D.S. Phytozome: a comparative platform for green plant genomics. **Nucleic Acids Research**, London, v. 40, p. D1178-D1186, 2012.
- GONZÁLEZ, V.M.; RODRÍGUEZ-MORENO, L.; CENTENO, E.; BENJAK, A.; GARCIA-MAS, J.; PUIGDOMÈNECH, P.; ARANDA, M.A. Genome-wide BAC-end sequencing of *Cucumis melo* using two BAC libraries. **BMC Genomics**, London, v. 11, p. 618-629, 2010.
- GRANDBASTIEN, M.A.; Retroelement in higher plants. **Trends in Genetics**, Cambridge, v. 8, p. 103-108, 1992.
- GU, Y. Q.; MA, Y.; HUO, N.; VOGEL, J. P.; YOU, F. M.; LAZO, G.R.; NELSON, W. M.; SODERLUND, C.; DVORAK, J.; ANDERSON, O. D.; LUO, M. C. A BAC-based physical map of *Brachypodium distachyon* and its comparative analysis with rice and wheat. **BMC Genomics**, London, v. 10, p. 496, 2009.
- HAMILTON, J.P.; BUELL, C.R. Advances in plant genome sequencing. **Plant Journal**, Oxford, v. 70, p. 177-190, 2012.
- HAN, Y.; KORBAN, S.S. An overview of the apple genome through BAC-end sequence analysis. **Plant Molecular Biology**, Dordrecht, v. 67, p. 581-588, 2008.
- HANSEN, A.K.; GILBERT, L.E.; SIMPSON, B.B.; DOWNIE, S.R.; CERVI, A.C; JANSEN, R.K. Phylogenetic relationships and chromosome number evolution in *Passiflora*. **Systematic Botany**, East Melbourne, v. 31, p. 138-150, 2006.
- HONG, C.P.; PLAHA, P.; KOO, D.H.; YANG, T.J.; CHOI, S.R.; LEE, Y.K.; UHM, T.; BANG, J.W.; EDWARDS, D.; ROFT, B.I.; PARK, B.S.; LEE, J. A survey of the *Brassica rapa* Genome by BAC-end sequence; analysis and comparison with *Arabidopsis thaliana*. **Molecules and Cells**, Seoul, v. 22, p. 300-307, 2008.
- HOUBEN, A.; DEMIDOV, D.; GERNAND, D.; MEISTER, A.; LEACH, C.R.; SCHUBERT, I. Methylation of histone H3 in euchromatin of plant chromosomes depends on basic nuclear DNA content. **Plant Journal**, Oxford, v. 33, p. 967-973, 2003.
- HSU, C.C.; CHUNG, Y.L.; CHEN, T.C.; LEE, Y.L.; KUO, Y.T.; TSAI, W.C.; HSIAO, Y.Y.; CHEN, Y.W.; WU, W.L.; CHEN, H.H. An overview of the *Phalaenopsis* orchid genome through BAC end sequence analysis, **BMC Plant Biology**, London, v. 11, 2011, <<http://www.biomedcentral.com/1471-2229/11/3>> Acesso em: 12 abr. 2013.
- HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. **Genome Research**. New York, v. 9, p. 868-877, 1999.

HUO, N.; VOGEL, J.P.; LAZO, G.R.; YOU, F.M.; MA, Y.; MCMAHON, S.; DVORAK, J.; ANDERSON, O.D.; LOU, M.C.; GU, Y.Q. Structural characterization of *Brachypodium* genome and its syntenic relationship with rice and wheat, **Plant Molecular Biology**, Dordrecht, v. 70, p. 47-61, 2009.

IAONNOU, I.; AMEMIYA, C.T.; GARNES, J.; KROISEL, P.M.; SHIZUYA, H.; CHEN, C.; BATZER, M.A.; JONG, P.J. A new bacteriophage P1-derived vector for propagation of large human DNA fragments. **Nature Genetics**, New York, v. 6, p. 84-89, 1994.

ILIC, K.; SANMIGUEL, P.J.; BENNETZEN, J.L. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 100, p. 12265-12270, 2003.

ICHIMURA, T.; YAMANAKA, A.; ICHIBA.; TOYOKAWA, T.; KAMADA, Y.; TAMAMURA, T.; MARUYAMA, S. Antihypertensive effect of na extract of *Passiflora edulis* rind in spontaneously hypertensive rats. **Bioscience, Biotechnology and Biochemistry**, Tokyo, v. 70, p. 718-721, 2006.

International Rice Genome Sequencing Project. The map-based sequence of the rice genome. **Nature**, London, v. 436, p. 793-800, 2005.

JAILLON, O.; AURY, J.M.; NOEL, B.; POLICRITI, A.; CLEPET, C.; CASAGRANDE, A.; CHOISNE, N.; AUBOURG, S.; VITULO, N.; JUBIN, C.; VEZZI, A.; LEGEAI, F.; HUGUENEY, P.; DASILVA, C.; HORNER, D.; MICA, E.; JUBLOT, D.; POULAIN, J.; BRUYÈRE, C.; BILLAULT, A.; SEGURENS, B.; GOUYVENOUX, M.; UGARTE, E.; CATTONARO, F.; ANTHOUARD, V.; VICO, V.; DEL FABBRO, C.; ALAUX, M.; DI GASPERO, G.; DUMAS, V.; FELICE, N.; PAILLARD, S.; JUMAN, I.; MOROLDO, M.; SCALABRIN, S.; CANAGUIER, A.; LE CLAINCHE, I.; MALACRIDA, G.; DURAND, E.; PESOLE, G.; LAUCOU, V.; CHATELET, P.; MERDINOGLU, D.; DELLEDONNE, M.; PEZZOTTI, M.; LECHARNY, A.; SCARPELLI, C.; ARTIGUENAVE, F.; PÈ, M.E.; VALLE, G.; MORGANTE, M.; CABOCHE, M.; ADAM-BLONDON, A.F.; WEISSENBAACH, J.; QUÉTIER, F.; WINCKER, P.; The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. **Nature**, London, v. 27, p. 463-4677, 2007.

JANDER, G.; NORRIS, S.R.; ROUNSLEY, S.D.; BUSH, D.F.; LEVIN, I.M.; ROBERT, L.L. *Arabidopsis* map-based cloning in the post-genome era. **Plant Physiology**, Lancaster, v. 129, p. 440-450, 2002.

JANG, S.; ZHOU, F.; XIA, L.; ZHAO, W.; CHENG, H.; ZHOU, R. Construction of a BAC library and identification of *Dmrt1* gene of the rice field eel, *Monopterus albus*. **Biochemical and Biophysical Research Communications**, New York, v. 384, p. 775-780, 2006.

JORGENSEN, P.M.; LAWESSON, J.E.; HOLM-NIELSEN, L.B. A guide to collecting Passionflowers. **Annals of Missouri Botanical Garden**, Jefferson City, v. 71, p. 1172-1174, 1984.

KELLEHER, C. T.; CHIU, R.; SHIN, H.; BOSDET, I. E.; KRZYWINSKI, M. I.; FJELL, C. D.; WILKIN, J.; YIN, T.; DIFAZIO, S. P.; ALI, J.; ASANO, J. K.; CHAN, S.; CLOUTIER, A.; GIRN, N.; LEACH, S.; LEE, D.; MATHEWSON, C. A.; OLSON, T.; O'CONNOR, K.; PRABHU, A. L.; SMAILUS, D. E.; STOTT, J. M.; TSAI, M.; WYE, N. H.; YANG, G. S.; ZHUANG, J.; HOLT, R. A.; PUTNAM, N. H.; VREBALOV, J.; GIOVANNONI, J. J.; GRIMWOOD, J.; SCHMUTZ, J.; ROKHSAR, D.; JONES, S.J.; MARRA, M.A., TUSKAN, G. A., BOHLMANN, J.; ELLIS, B. E.; RITLAND, K.; DOUGLAS, C.J.; SCHEIN, J.E. A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. **Plant Journal**, Oxford v. 50, p. 1063–1078, 2007.

KILLIP, E.P. The American species of *Passifloraceae*. **Publications Field Museum of Natural History - Botanical Series**, London, v. 19, 1-613, 1938.

KIM, U.J.; SHIZUYA, H.; JONG, P.J.; BIRREN, B.; SIMON, M.I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. **Nucleic Acids Research**, London, v. 20, p. 1083-1085, 1992.

KIM, C.; LEE, T.H.; COMPTON, R.O.; ROBERTSON, J.S.; PIERCE, G.J.; PATERSON, A.H. A genome-wide BAC end-sequence survey of sugarcane elucidates genome composition, and identifies BACs covering much of the euchromatin. **Plant Molecular Biology**, Dordrecht, v. 81, p. 139-47, 2013.

KOFLER, R.; CHLÖTTERER, C.; LELLEY, T. SciRoKo: A new tool for whole genome microsatellite search and investigation. **Bioinformatics**, Oxford, v. 23, p. 1683-1685, 2007.

KUMAR, A.; BENNETZEN, J.L.; Plant retrotransposon. **Annual Review of Genomics and Human Genetics**, Palo Alto, v. 33, p. 479-532, 1999.

KUNKEL, T.A.; Misalignment-mediated DNA synthesis errors. **Biochemistry**, Washington, v. 29, p. 8003-8011, 1990.

LAI, C.W.; YU, Q.; HOU, S.; SKELTON, R.L.; JONES, M.R.; LEWIS, K.L.; MURRAY, J.; EUSTICE, M.; GUAN, P.; AGBAYANI, R.; MOORE, P.H.; MING, R.; PRESTING, G.G. Analysis of papaya BAC-end sequences reveals first insights into the organization of a fruit tree genome. **Molecular Genetics and Genomics**, Berlin, v. 276, p. 1-12, 2006.

LESK, A. **Introduction to Bioinformatics**. 3rd ed. Oxford:Oxford University Press, 2008. 432p.

LEWIS, S.E.; SEARLE, S.M.J.; HARRIS, N.; GIBSON, M.; IYER, V.; RICTER, J.; WIEL, C.; BAYRAKTAROGU, L.; BIRNEY, E.; CROSBY, M.A.; KAMINKER, J.S.; MATTHEWS B, PROCHNIK SE, SMITH CD, TUPY, J.L.; RUBIN, G.M.; MISRA, S.; MUNGALL, C.J.; CLAMP, M.E. Apollo: a sequence annotation editor. **Genome Biology**, London, v. 3, n. 12, p. research0082.1-research 0082.14, 2002.

LIANG, C.; XI, Y.; SHU, J.; LI, J.; YANG, J.; CHE, K.; JIN, D.; LIU, X.; WENG, M.; HE, Y.; WANG, B. Construction of a BAC library of *Physcomitrella patens* and isolation of LEA gene. **Plant Science**, Limerick, v. 167, p. 491-498, 2004.

LIMA, A.A.; CARDOSO, C.E.L.; SOUZA, J.S.; PIRES, M.M. **Comercialização do maracujazeiro**, 2006. Disponível em: <http://www.todafruta.com.br/>. Acesso em: 10 abr. 2013.

- LISCH, D. How important are transposons for plant evolution? **Nature Reviews-Genetics**, London, v. 14, p. 49-61, 2011.
- LODHI, M.A.; REISCH, B.I. Nuclear DNA content of *Vitis* species, cultivars and other genera of the *Vitaceae*. **Theoretical and Applied Genetics**, New York, v. 90, p. 11-16, 1995.
- LOPES, R.; LOPES, M.T.G.; CARNEIRO, M.S.; MATTA, F.P.; CAMARGO, L.E.A.; VIEIRA, M.L.C. Linkage and mapping of resistance genes to *Xanthomonas axonopodis* pv. *passiflorae* in yellow passion fruit. **Genome**, Ottawa, v. 49, p. 17-29, 2006.
- LOCKTON, S.; GAUT, B.S. Plant conserved non-coding sequences and paralogue evolution, **Trends in Genetics**, Cambridge, v. 21, p. 60-65, 2005.
- MCCARTHY, E.M.; LIU, J.; LIZHI, G.; MCDONALD, J.F. Long terminal repeat retrotransposons of *Oryza sativa*. **Genome Biology**, London, v. 3, p. Research0053.1-Research0053.11, 2002.
- MADIGAN, M.T.; MARTINKO, J.M. Brock biology of microorganisms. 10th ed. Harlow: Pearson-Prentice Hall, 2003. 1385p.
- MAYEDA, L.Y. **Estudo citogenético em dez táxons do gênero *Passiflora* L. (*Passifloraceae*)**. 1997. 89p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1997.
- MELETTI L.M.M.; SOARES-SCOTT M.D.; BERNACCI, L.C. Caracterização fenotípica de três seleções de maracujazeiro-roxo (*Passiflora edulis* Sims). **Revista Brasileira de Fruticultura**, Cruz das Almas, v. 27, p. 268-272, 2005.
- MELO, N.F.; CERVI, A.C.; GUERRA, M. Karyology and citotaxonomy of the genus *Passiflora* L. **Plant Systematics and Evolution**, Berlin, v. 226, p. 68-84, 2001.
- MELO, N.F.; GUERRA, M. Variability of the 5S and rDNA sites in *Passiflora* L. with species with distinct base chromosome numbers. **Annals of Botany**, London, v. 92, p. 309-316, 2003.
- MEYERS, B.C.; TINGEY, S.V.; MORGANTE, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. A detailed survey of maize retrotransposons that combines random genomic sample sequencing, hybridization to bacterial artificial chromosomes and data obtained from a large collection of ESTs. **Genome Research**, New York, v. 11, p. 1660-1676, 2001.
- MING, R.; HOU, S.; FENG, Y.; YU, Q.; DIONNE-LAPORTE, A.; SAW, J. H.; SENIN, P.; WANG, W.; LY, B. V.; LEWIS, K. L. T.; SALZBERG, S. L.; FENG, L.; JONES, M. R.; SKELTON, R. L.; MURRAY, J. E.; CHEN, C.; QIAN, W.; SHEN, J.; DU, P.; EUSTICE, M.; TONG, E.; TANG, H.; LYONS, E.; PAULL, R. E.; MICHAEL, T. P.; WALL, K.; RICE, D. W.; ALBERT, H.; WANG, M.-L.; ZHU, Y. J.; SCHATZ, M.; NAGARAJAN, N.; ACOB, R. A.; GUAN, P.; BLAS, A.; WAI, C. M.; ACKERMAN, C. M.; REN, Y.; LIU, C.; WANG, J.; WANG, J.; NA, J.-K.; SHAKIROV, E. V.; HAAS, B.; THIMMAPURAM, J.; NELSON, D.; WANG, X.; BOWERS, J. E.; GSCHWEND, A. R.; DELCHER, A. L.; SINGH, R.; SUZUKI, J. Y.; TRIPATHI, S.; NEUPANE, K.; WEI, H.; IRIKURA, B.; PAIDI, M.; JIANG, N.; ZHANG, W.; PRESTING, G.; WINDSOR, A.; NAVAJAS-PÉREZ, R.; TORRES, M. J.; FELTUS, F. A.; PORTER, B.; LI, Y.; BURROUGHS, A. M.; LUO, M.-C.; LIU, L.; CHRISTOPHER, D. A.; MOUNT, S. M.; MOORE, P. H.; SUGIMURA, T.; JIANG, J.; SCHULER, M. A.; FRIEDMAN, V.; MITCHELL-OLDS, T.; SHIPPEN, D. E.; DEPAMPHILIS, C. W.; PALMER, J. D.;

- FREELING, M.; PATERSON, A. H.; GONSALVES, D.; WANG, L.; AND ALAM, M. Genome of papaya, a fast growing tropical fruit tree. **Tree Genetics and Genomics**, New York, v. 8, p. 445-462, 2012.
- MONNA, L.; LIN, H.X.; KOJIMA, S.; SASAKI, T.; YANO, M. Genetic dissection of a genomic region for quantitative trait locus, Hd3, into two loci, Hd3a and Hd3b, controlling heading date in rice. **Theoretical and Applied Genetics**, New York, v. 104, p. 772-778, 2002.
- MORGANTE, M.; BRUNNER, S.; PEA, G.; FENGLER, K.; ZUCCOLO, A.; RAFALSKI, A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. **Nature Genetics**, New York, v. 37, p. 997-1002, 2005.
- MOZO, T.; DEWAR, K.; DUNN, P.; ECKER, J. R.; FISCHER, S.; KLOSKA, S.; LEHRACH, H.; MARRA, M.; MARTIENSSEN, R.; MEIER-EWERT, S. A complete BAC-based physical map of the *Arabidopsis thaliana* genome. **Nature Genetics**, New York, v. 22, p.271-275, 1999.
- MUSCHNER, V.C.; LORENZ, A.P.; CERVI, A.C.; BONATTO, S.L.; SOUZA-CHIES, T.T.; SALZANO, F.M; FREITAS, L.B. A first molecular phylogenetic analysis of *Passiflora* (*Passifloraceae*). **American Journal of Botany**, Baltimore, v. 90, p. 1229-1238, 2003.
- NCBI NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Disponível em: <<http://www.ncbi.nlm.nih.gov/genome/4>>. Acesso em: 24 mai. 2013.
- NOIR, S.; PATHEYRON, S.; COMBES, M.C.; LASHERMES, P.; CHALHOUB, B. Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). **Theoretical and Applied Genetics**, New York, v.109, p.225-230, 2004.
- NIKAIDO, M.; ROONEY, A.P.; OKADA, N. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 96, p. 10261-10266, 1999.
- OBLESSUC, P.R. **Mapeamento fino de locos associados à resistência à mancha angular em feijão (*Phaseolus vulgaris* L.)**. 2013. 171p. Tese (Doutorado em Genética e Biologia Molecular) – Instituto de Biologia, Universidade de Campinas, Campinas, 2013.
- OLIVEIRA, A.M.A. **Reprodução e citogenética de espécies de *Passiflora***. 1996. 148 p. Tese (Doutorado em Genética) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista “Júlio de Mesquita Filho”, São José do Rio Preto, 1996.
- OLIVEIRA, E.J.; VIEIRA, M.L.C.; GARCIA, A.A.F.; MUNHOZ, C.F.; MARGARIDO, G.R.A.; CONSOLI, L.; MATTA, F.P.; MORAES, M.C. An integrated molecular map of yellow passion fruit based on simultaneous maximum-likelihood estimation of linkage and linkage phases. **Journal of the American Society for Horticultural Science**, New York, v. 133, p. 35-41, 2008.
- OLLITRAULT, P.; TEROL, J.; GARCIA-LOR, A.; BERRARD, A.; CHAUVEAU, A.; FROELICHER, Y.; BELZILE, C.; MORILLON, R.; NAVARRO, L.; BRUNEL, D.; TALON, M. SNP mining in *C. clementina* BAC end sequences; transferability in the *Citrus* genus (Rutaceae), phylogenetic inferences and perspectives for genetic mapping. **BMC Genomics**, London, v. 13, 2012. <<http://www.biomedcentral.com/1471-2164/13/13>> Acesso em: 04 jan. 2013.
- PÁDUA, J.G. **Análises genéticas de espécies do gênero *Passiflora* L. com base em abordagens filogenéticas, morfométricas e em marcadores microssatélites**. 2004. 127p. Tese (Doutorado

em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2004.

PATERSON, A.H.; FREELING, M.; TANG, H.; WANG, X. Insights from the comparison of plant genome sequences. **Annual Review of Plant Biology**, Palo Alto, v. 61, p. 349-372, 2010.

PATOCCHI, A.; VINATZER, B.A.; GIANFRANCESCHI, L.; TARTARINI, S.; ZHANG, H.B.; SANSAVINI, S.; GESSLER, C. 1999. Construction of a 550 kb BAC contig spanning the genomic region containing the apple scab resistance gene Vf. **Molecular and General Genetics**, Berlin, v. 262, p. 884-891, 1999.

PAUX, E.; SOURDILLE, P.; SALSE, J.; SAINTENAC, C.; CHOULET, F.; LEROY, P.; KOROL, A.; MICHALAK, M.; KIANIAN, S.; SPIELMEYER, W.; LAGUDAH, E.; SOMERS, D.; KILIAN, A.; ALAUX, M.; VAUTRIN, S.; BERGES, H.; EVERSOLE, K.; APPELS, R.; SAFAR, J.; SIMKOVA, H.; DOLEZEL, J.; BERNARD, M.; FEUILLET, C. A physical map of the 1-gigabase bread wheat chromosome 3B. **Science**, Washington, v. 322, p. 101-104, 2008.

PENHA, H.A. **Construção de uma biblioteca genômica de *Passiflora edulis* f. *flavicarpa* inserida em BACs (Bacterial Artificial Chromosome) e mapeamento cromossômico usando hibridação in situ fluorescente**. 2012. 129p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.

PEREIRA, G.S.; NUNES, E.S.; LAPERUTA, M.F.; BRAGA, M.F.; PENHA, H.A.; DINIZ, A.L.; MUNHOZ, C.F.; GAZAFFI, R.; GARCIA, A.A.F.; VIEIRA, M.L.C. Molecular polymorphism and linkage analysis in sweet passion fruit, an outcrossing species. **Annals of Applied Biology**, London, v. 162, p. 347-361, 2013.

PETERSON, D. G.; TOMKINS, J. P.; FRISCH, D. A.; WING, R. A.; PATERSON, A. H. Construction of plant artificial chromosome (BAC) libraries: An illustrated guide. **Journal Agricultural Genomics**, Washington, v.5, p.1-3, 2000.

PRAÇA, M.M.; CARVALHO, R.C.; MARCELINO, F.C.; MENDONÇA, M.A.C. Morphological aspects of *Passiflora edulis* f. *flavicarpa* chromosomes using acridine orange banding and rDNA-FISH tools. **Caryologia**, Firenze, v. 61, p. 154-159, 2008.

QUACKENBUSH, J.; LIANG, F.; HOLT, I.; PERTEA, G.; UPTON, J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. **Nucleic Acids Research**, London, v. 28, p. 141-145, 2000.

QUAIL, M.A.; MATTHEWS, L.; SIMS, S.; LLOYD, C.; BEASLEY, H.; BAXTER, S.W. Genomic libraries: II. Subcloning, sequencing, and assembling large-insert genomic DNA clones. **Methods in Molecular Biology**, Totowa, v. 772, p.59-81, 2001.

RAMOS, A.T.; CUNHA, M.A.L.; SABAA-SRUR, A.U.O.; PIRES, V.C.F.; CARDOSO, M. A. A.; DINIZ, M. F. M.; MEDEIROS, C. C. M. Uso de *Passiflora edulis* f. *flavicarpa* na redução do colesterol. **Revista Brasileira de Farmacognosia**, São Paulo, v. 17, p. 592-597, 2007.

RAMPANT, P. F.; LESUR, I.; BOUSSARDON, C.; BITTON, F.; MARTIN-MAGNIETTE, M. L.; BODENES, C.; LE PROVOST, G.; BERGES, H.; FLUCH, S.; KREMER, A.; PLOMION, C. Analysis of BAC end sequences in oak, a keystone forest tree species, providing insight into the composition of its genome **BMC Genomics**, San Diego, v.12, p.292, 2011.

RICHTERICH, P. Estimation of errors in "raw" DNA sequences: a validation study. **Genome Research**, New York, v. 8, n. 3, p. 251-259, 1998.

RIZZI, L.C.; RABELLO, L.A.; MOROZINI FILHO, W.; SAVASAKI, E.T.; KAVATI, R. **Cultura do maracujá azedo**. Campinas: Coordenadoria de Assistência Técnica Integral, SAA, 1998. 235p.

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

SCHMITZ, J.; OHME, M.; ZISCHLER, H. SINE Insertions in Cladistic Analyses and the Phylogenetic Affiliations of *Tarsius bancanus* to Other Primates. **Genetics**, Austin, v. 157 p. 777-784, 2001.

SHIZUYA, H.; BIRREN, B.; KIM, U.J.; MANCINO, V.; SLEPAK, T.; TACHIIRI, Y.; SIMON M. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 89, p. 8794-8797, 1992.

SHULTZ, J.L.; KAZI, S.; BASHIR, R.; AFZAL, J.A.; LIGHTFOOT, D.A. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. **Theoretical and Applied Genetics**, New York, v. 114, p. 1081-1090, 2007.

SILVA, C.; RUA, P.S. Aspectos econômicos: Mercado Nacional e Internacional. In: LEONEL, S.; SAMPAIO, A.C. **Maracujá-doce: Aspectos técnicos e econômicos**. São Paulo: Fundação Editora da UNESP, 2007. p. 7-24.

SOARES-SCOTT, M.D. **Caracterização citogenética de algumas espécies e híbridos interespecíficos de *Passiflora***. 1998. 89p. Dissertação (Mestrado em Biologia Celular) – Instituto de Ciências Biológicas, Universidade Estadual de Campinas, Campinas, 1998.

SOLTIS, D.E.; SOLTIS, P.S. Polyploidy: recurrent formation and genome evolution. **Trends in Ecology & Evolution**, Cambridge, v. 14, p. 348-352, 1999.

SOUZA, M.M.; PALOMINO, G.; PEREIRA, M.G.; VIANA, A.P. Flow cytometric analysis of genome size variation in some *Passiflora* species. **Hereditas**, Landskrona, v.141, p.31-38, 2004.

SOUZA, M.M.; PEREIRA, T.N.S.; VIEIRA, M.L.C. Cytogenetics studies in some species of *Passiflora* L. (*Passifloraceae*). : a review emphasizing Brazilian species. **Brazilian Archives of Biology and Technology**, Curitiba, v. 51, p. 247-258, 2008.

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews-Genetics**, London, v. 2, p. 493-503, 2001.

STERCK, L.; ROMBAUTS, S.; VANDEPOELE, K.; ROUZE, P.; VAN DE PEER, Y. How many genes are there in plants (. . . and why are they there)? **Current Opinion In Plant Biology**. London, v. 10, p. 199-203, 2007.

STERNBERG, N. Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 87, p. 103-107, 1990.

STEVENS, M.R.; COLEMAN, C.E.; PARKINSON, S.E.; MAUGHAN, P.J.; ZHANG, H.B.; BALZOTTI, M.R.; KOOYMAN, D.L.; ARUMUGANATHAN, K.; BONIFACIO, A.;

FAIRBANKS, D.J.; JELLEN, E.N.; STEVENS, J.J. Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins. **Theoretical and Applied Genetics**, New York, v.112, p.1593-1600, 2006.

SWIFT, H. The constancy of deoxyribose nucleic acid in plant nuclei. **Proceedings of the National Academy of Sciences**, Washington, v. 36, p. 643-654, 1950.

TAMANNA, T.; KHAN, A.U. Mapping and analysis of Simple Sequence Repeats in the *Arabidopsis thaliana* Genome. **Bioinformation**, Singapore, v. 1, p.64–68, 2005.

TAO, Q.Z., ZHANG, H.B. Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. **Nucleic Acids Research**, London, v. 26, p. 4901-4909, 1998.

TAO, Q.; CHANG, Y.L.; WANG, J.; CHEN, H.; ISLAM-FARIDI, M.N.; SCHEURING, C.; WANG, B.; STELLY, D.; ZHANG, H.B. Bacterial Artificial Chromosome-Based Physical Map of the Rice Genome Constructed by Restriction Fingerprint Analysis. **Genetics**, Austin, v. 158, p. 1711-1724, 2001.

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, London, v. 408, p. 796-815, 2000.

THOMAS, C.A. The genetic organization of chromosomes. **Annual review of genetics**, Palo Alto, v. 5, p. 237-256, 1971.

THORVALDSDOTTIR, H.; ROBINSON, J.T.; MESIROV, J.P.; Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. **Briefings in Bioinformatics**, London, v. 14, p. 178-192, 2013.

TUSKAN, G.A.; DIFAZIO, S.; JANSSEN, S.; BOHLMANN, J.; GRIGORIEV, I.; ET AL. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). **Science**, Washington, v. 313, p. 1596-1604, 2006.

ULMER, T.; MacDOUGAL, J.M. **Passiflora**: passionflowers of the world. Cambridge: Timber Press, 2004. 430p.

VANDERPLANK, J. **Passion flowers**. Cambridge: MIT Press, Massachusetts Institute of Technology, 1991.460 p.

VELASCO, R.; ZHARKIKH, A.; AFFOURTIT, J.; DHINGRA, A.; CESTARO, A. The genome of the domesticated apple (*Malus × domestica* Borkh.) **Nature Genetics**, New York, v. 42, p.833–839, 2010.

VIANA, A.P.; PEREIRA, T.N.S.; PEREIRA, M.G.; SOUZA, M.M.; MALDONADO, J.F.M.; AMARAL JÚNIOR, A.T. Genetic diversity among yellow passion fruit commercial genotypes and among *Passiflora* species using RAPD. **Revista Brasileira de Fruticultura**, Cruz das Almas, v. 25, p. 489-493, 2003.

VIEIRA, M.L.C.; CARNEIRO, M.S. *Passiflora* spp. Passionfruit. In: LITZ, R.E. (Ed.). **Biotechnology of Fruit and Nut Crops**. Oxford: CABI Publishing, 2004. p. 435-453.

VICIENT, C.M.; JÄÄSKELÄINEN, M.; KALENDAR, R.; SCHULMAN, A.H. Active retrotransposons are a common feature of grass genomes. **Plant Physiology**, Lancaster, v. 125, p. 1283-1292, 2001.

WAI, C.M.; MOORE, P.H.; PAULL, R.E.; MING, R.; YU, Q.Y. An integrated cytogenetic and physical map reveals unevenly distributed recombination spots along the papaya sex chromosomes. **Chromosome research**, Oxford, v. 20, p. 753-767, 2012.

WANG, W.; TANURDZIC, M.; LUO, M.; SISNEROS, N.; KIM, H.R.; WENG, J.K.; KUDRNA, D.; MUELLER, C.; ARUMUGANATHAN, K.; CARLSON, J.; CHAPPLE, C.; PAMPHILIS, C.; MANDOLI, D.; TOMKINS, J.; WING, R.; BANKS, J.A. Construction of a bacterial artificial chromosome library from the spike moss *Selaginella moellendorffii*: a new resource for plant comparative genomics. **BMC Plant Biology**, London, v. 5, 2005.
<<http://www.biomedcentral.com/1471-2229/5/10> > Acesso em: 28 fev. 2012.

WATSON, L.; DALLWITZ, M.J. 1992 onwards. The families of flowering plants: descriptions, illustrations, identification, and information retrieval. Version: 19th December 2012

WATSON, R.T. **Data Management: Databases and Organizations** 5th ed. Wiley, 2005. 640p.

WICKER, T.; SABOT, F.; HUA-VAN, A.; BENNETZEN, J.L.; CAPY, P.; CHALHOUB, B.; FLAVELL, A.; LEROY, P.; MORGANTE, M.; PANAUD, O.; PAUX, E.; SANMIGUEL, P.; SCHULMAN, A.H. A unified classification system for eukaryotic transposable elements. **Nature Reviews-Genetics**, London, v. 8, p. 973-982, 2007.

WICKER, T.; SCHLANGENHAUF, E.; GRANER, A.; CLOSE, T. J.; KELLER, B.; STEIN, N. 454 sequencing put to the test using the complex genome of barley. **BMC Genomics**, London, v.7, p.275, 2006.

WILHELM, M.; WILHELM, F.X. 2001. Reverse transcription of retroviruses and LTR retrotransposons. **Cellular and molecular life sciences**, Basel, v. 58, p. 1246-1262, 2001.

WING, R.A.; AMMIRAJU, J.S.; LUO, M.; KIM, H.; YU, Y.; KUDRNA, D.; GOICOECHEA, J.L.; WANG, W.; NELSON, W.; RAO, K.; BRAR, D.; MACKILL, D.J.; HAN, B.; SODERLUND, C.; STEIN, L.; SANMIGUEL, P.; JACKSON, S. The **Oryza** map alignment project: the golden path to unlocking the genetic potential of wild rice species. **Plant Molecular Biology**, Dordrecht, v. 59, p. 53-62, 2005.

WOO, S.S.; JIANG, J.; GILL, B.S.; PATERSON, A.H.; WING, R.A. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. **Nucleic Acids Research**, London, v. 22, p. 4922-4931, 1994.

WU, C.; SUN, S.; NIMMAKAYALA, P.; SANTOS, F.A.; MEKSEM, K.; SPRINGMAN, R.; DING, K.; LIGHTFOOT, D.A.; ZHANGH, B. A BAC- and BIBAC-based physical map of the soybean genome. **Genome Research**, New York, v. 14, p. 319-326, 2004.

WU, J.J.; GU, Y. Q.; HU, Y.Q.; YOU, F.M.; DANDEKAR, A.M.; LESLIE, C.A.; ARADHYA, M.; DVORAK, J.; LUO, M.C. Characterizing the walnut genome through analyses of BAC end sequences **Plant Molecular Biology**, Dordrecht, v.78, p. 95-107, 2012.

YOTOKO, K.S.C.; DORNELAS, M.C.; TOGNI, P.D.; FONSECA, T.C.; SALZANO, F.M.; BONATTO, S.L.; FREITAS, L. Does variation in genome sizes reflect adaptive or neutral processes? New clues from *Passiflora*. **Plos One**, San Francisco, v. 6, 2011.
<<http://www.plosone.org/article/info:doi/10.1371/journal.pone.0018212> > Acesso em: 20 jan. 2012.

YAO, J.; DONG, Y.; MORRIS, B. A. Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. **Proceedings of the National Academy of Sciences**, Washington, v. 98, p. 1306-1311, 2001.

YU, Q.; TONG, E.; SKELTON, R. L.; BOWERS, J. E.; JONES, M. R.; MURRAY, J.E.; HOU, S.; GUAN, P.; ACOB, R. A.; LUO, M.C.; MOORE, P. H.; ALAM, M.; PATERSON, A.H.; MING, R.; A physical map of the papaya genome with integrated genetic map and genome sequence. **BMC Genomics**, London, v.10, p. 371, 2009.

ZHANG , H.B.; SCHEURING, F.C.; ZHANG, M.; ZHANG, Y.; WU, C.C.; DONG, J.J.; LI, Y. Construction of BIBAC and BAC libraries from a variety of organisms for advanced genomics research. **Nature Protocols**, London, v. 7, p. 479-499, 2012.

ANEXOS

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe216Q3C5F	trehalose-phosphate synthase	828	4,56E-168	87.95%
Pe216Q2H1R	sex determination protein tasselseed-2-like	965	1,76E-167	89.15%
Pe173Q4G7R	translation initiation factor eif-3 subunit 7	928	4,16E-166	94.2%
Pe185Q4G12F	anaphase-promoting complex subunit 8-like	806	3,02E-146	90.6%
Pe55Q1G10F	btb poz domain-containing protein at3g44820-like	965	1,93E-139	78.25%
Pe84Q2F10R	photosystem ii protein d2	645	6,70E-135	97.9%
Pe216Q4B2R	cytochrome p450	910	2,06E-127	72.15%
Pe164Q3H12R	pentatricopeptide repeat-containing protein chloroplastic-like	885	5,35E-126	83.65%
Pe164Q3B5R	pentatricopeptide repeat-containing protein at3g02330-like	935	2,08E-124	70.9%
Pe173Q4A8R	translocase of chloroplast chloroplastic-like	697	8,03E-123	92.55%
Pe84Q2G3R	galactoside 2-alpha-l-	1028	8,54E-122	76.9%
Pe216Q1G5F	polyol transporter 5-like	864	6,79E-117	78.5%
Pe214Q1F12F	udp rhamnose:anthocyanidin-3-glucoside rhamnosyltransferase	841	2,98E-115	72.8%
Pe101Q3A5F	pentatricopeptide repeat-containing protein	949	2,54E-114	61.2%
Pe75Q3A5R	octicosapeptide phox bem1p domain-containing protein	862	6,19E-113	73.3%
Pe212Q1E2R	e3 ubiquitin-protein ligase upl2-like	813	8,25E-113	75.2%
Pe214Q3D6R	pentatricopeptide repeat-containing protein mitochondrial-like	742	9,74E-112	77.3%
Pe84Q1G12R	lrr receptor-like serine threonine-protein kinase rch1-like	1015	5,38E-110	79.7%
Pe101Q3C4F	low quality protein: uncharacterized loc101215259	841	7,60E-107	84.85%
Pe214Q3E5R	probable apyrase 7-like	717	3,41E-104	75.6%
Pe69Q2H8R	pentatricopeptide repeat-containing protein at3g02330-like	786	8,81E-104	75.25%
Pe164Q3C5R	e3 ubiquitin-protein ligase pub23-like	822	5,31E-103	73.65%
Pe33Q2H1F	s-adenosyl-l-methionine-dependent methyltransferase domain-containing protein	1030	2,89E-98	76.4%
Pe84Q2E7R	ubiquitin carboxyl-terminal hydrolase-like protein	767	1,70E-95	69.6%
Pe93Q2H9F	protein iq-domain 14-like	714	3,80E-95	66.15%
Pe101Q3H7R	e3 ubiquitin-protein ligase upl5	864	1,31E-93	57.55%
Pe75Q1F8R	structural constituent of ribosome	1127	1,97E-93	74.05%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe212Q1E8R	serine threonine-protein kinase bri1-like 2-like	834	1,07E-92	74.05%
Pe173Q3A8R	translocase of chloroplast chloroplastic-like	614	1,91E-91	88.1%
Pe141Q2F2R	protein	923	1,07E-90	74.05%
Pe216Q2A8F	filament-like plant protein 4	705	1,44E-89	76.9%
Pe75Q3C7R	myb family transcription factor apl-like	868	1,45E-88	70.85%
Pe75Q4A9R	jasmonate o-	829	4,24E-87	62.9%
Pe216Q3C10F	cytochrome p450	827	9,57E-87	64.7%
Pe214Q3C2R	reticuline oxidase-like	556	3,02E-86	86.8%
Pe212Q3E12R	glutaredoxin family protein	619	1,25E-85	70.35%
Pe43Q4A5R	receptor-like protein kinase 2-like	648	4,89E-85	80.35%
Pe164Q2E4R	autophagy-related protein 9-like	664	1,34E-84	85.6%
Pe75Q1G12R	jasmonate o-	828	2,04E-84	62.85%
Pe75Q3B8R	detoxifying efflux carrier 35	899	1,75E-81	68.6%
Pe164Q2A6R	nac domain ipr003441	856	2,43E-81	76.35%
Pe84Q1C6F	transferring glycosyl	556	3,82E-81	88.2%
Pe101Q2G6R	katanin p60 atpase-containing subunit	1017	5,73E-81	79.05%
Pe164Q2G6F	cyclic nucleotide-gated ion channel 15-like	903	3,37E-80	79.3%
Pe75Q2C7F	probably inactive receptor-like protein kinase at2g46850-like	515	6,61E-80	71.85%
Pe141Q4H8F	e3 ubiquitin-protein ligase pub23-like	586	1,04E-79	73.45%
Pe164Q3H7F	wd40 and beach domain-containing protein	855	1,01E-78	62.25%
Pe141Q3F8R	dna repair and recombination protein	839	1,69E-78	63.95%
Pe69Q2D4R	pentatricopeptide repeat-containing protein chloroplastic-like	769	5,63E-78	88.9%
Pe214Q1E6R	wd repeat-containing protein 26-like	871	1,20E-77	81.6%
Pe101Q2B4R	cellulose synthase-like protein e1	825	1,51E-75	79.55%
Pe216Q4A1F	protein prd1-like	914	2,31E-75	63.15%
Pe216Q1G8F	uncharacterized wd repeat-containing	864	4,34E-75	72.55%
Pe1Q1F10F	leucine-rich repeat-containing	914	9,51E-73	75.9%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe214Q1E8F	pi-plc x domain-containing protein at5g67130-like	993	1,79E-72	72.7%
Pe214Q3A2R	probable methyltransferase pmt8-like	691	4,30E-72	67.65%
Pe212Q1E1F	type ii inositol 5-	915	8,19E-72	81.4%
Pe216Q2B5R	protein arabidillo 1-like	1062	1,15E-71	79.55%
Pe93Q3A3R	leucine-rich repeat disease resistance protein	582	6,00E-71	81.65%
Pe216Q1G1F	t7 bacteriophage-type single subunit rna polymerase	775	1,06E-70	77.95%
Pe185Q3A9F	mate efflux family protein dtx1-like	719	1,71E-70	64.9%
Pe216Q3B7F	benzyl alcohol o-benzoyltransferase-like	916	8,06E-70	69.85%
Pe75Q2G12F	protein ethylene insensitive 3-like	968	3,05E-69	90.75%
Pe101Q4D11R	fasciclin-like agp protein	695	3,70E-69	83.95%
Pe93Q1A4F	mip tip subfamily	744	1,21E-67	87.35%
Pe69Q2F9F	web family protein at5g55860-like	582	4,46E-67	77.55%
Pe141Q3E8R	pentatricopeptide repeat-containing protein	634	3,02E-65	66.1%
Pe75Q4C5R	protein	886	4,17E-65	74.4%
Pe93Q1B8R	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1-like	832	2,82E-64	73.05%
Pe84Q4D11R	at2g44900-like protein	752	6,31E-64	79.6%
Pe216Q2C7F	binding protein	912	9,47E-62	61.0%
Pe75Q1G11R	cc-nbs-lrr resistance protein	919	1,87E-61	63.9%
Pe84Q2G9F	pre-mrna-splicing factor atp-dependent rna helicase dhx16-like	531	3,63E-61	90.05%
Pe141Q1C2R	low quality protein: uncharacterized loc101220341	593	4,10E-61	72.55%
Pe75Q1A11F	probable inactive serine threonine-protein kinase lvsg-like	965	1,38E-60	76.2%
Pe1Q1E1R	enhancer of polycomb-like transcription factor protein	892	3,65E-60	67.4%
Pe101Q3C11R	serine threonine-protein kinase smg1	741	2,65E-58	64.75%
Pe212Q1E6R	cytochrome p450-like protein	794	1,96E-57	65.0%
Pe164Q4A9R	plastid division protein pdv2-like	697	3,43E-57	60.15%
Pe75Q1A5R	nitrate transporter -like	955	9,52E-57	78.8%
Pe84Q3H2R	protein	701	3,83E-56	61.7%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe93Q4E8R	at5g47580 mnj7_17	662	4,25E-56	70.0%
Pe84Q4D9R	probable s-acyltransferase at3g26935-like	609	9,52E-56	90.7%
Pe75Q4C10F	cyclin-dependent kinase d-1-like	788	2,46E-55	83.5%
Pe93Q2H4F	ap2 domain-containing transcription factor	563	9,20E-55	59.75%
Pe69Q2G8R	protein	836	3,92E-54	64.15%
Pe216Q2A3R	protein	863	5,11E-53	78.45%
Pe214Q2E7R	spx domain-containing membrane protein at4g22990-like	751	6,19E-53	93.35%
Pe93Q3A7F	atp-binding cassette	645	9,40E-53	82.45%
Pe214Q1B5F	protein kinase	1039	1,47E-52	84.75%
Pe93Q2H8F	protein iq-domain 14-like	633	3,19E-52	60.3%
Pe216Q2B10F	pwwp domain-containing protein	696	1,05E-51	57.15%
Pe173Q4A7R	polyprotein	881	3,55E-51	50.5%
Pe141Q2C1R	pheromone receptor-like protein	733	5,75E-51	70.25%
Pe216Q4C5F	magnesium transporter -like family protein	681	9,80E-51	86.85%
Pe216Q1E3F	uncharacterized aarf domain-containing protein kinase chloroplastic-like	850	1,39E-50	90.25%
Pe93Q3C6R	condensin-2 complex subunit d3-like	900	1,41E-50	72.0%
Pe216Q4G3F	cell wall invertase	804	1,87E-50	88.5%
Pe212Q4B4F	protein	820	2,09E-50	80.65%
Pe101Q2A3R	cytochrome p450	812	5,57E-50	78.75%
Pe216Q2B9F	kinesin heavy	747	6,16E-50	92.0%
Pe212Q4A6F	rcc1 and btb domain-containing protein	874	9,27E-50	67.15%
Pe93Q4E3F	metal tolerance protein	791	2,15E-49	59.65%
Pe1Q1G9R	protein	966	5,23E-49	85.0%
Pe75Q4B6R	vacuolar protein sorting-associated protein 51 homolog	947	6,98E-49	92.2%
Pe164Q2A6F	trna rrna	934	7,70E-49	69.45%
Pe216Q1B10F	amino-acid acetyltransferase-like	682	8,96E-49	75.75%
Pe164Q1F9R	neurobeachin-like protein	1092	1,69E-48	70.8%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe173Q2C8R	galactose-binding protein	651	1,83E-48	84.35%
Pe216Q4B11R	protein phosphatase 2c and cyclic nucleotide-binding kinase domain-containing protein	889	2,45E-48	81.55%
Pe93Q3E5R	glyceraldehyde-3-phosphate dehydrogenase	725	1,16E-47	57.7%
Pe75Q4F8R	probable pectinesterase pectinesterase inhibitor 51-like	873	3,10E-47	85.95%
Pe93Q1F10F	subtilase family protein	640	3,18E-47	74.85%
Pe214Q1A7R	probable calcium-binding protein cml25-like	791	4,35E-47	70.05%
Pe164Q4A8R	leucine-rich repeat-containing	961	8,83E-47	55.15%
Pe164Q1G7F	cell wall-associated hydrolase	767	1,57E-46	84.18%
Pe33Q2E11F	metal tolerance protein	673	3,02E-46	57.8%
Pe185Q2E5F	tetratricopeptide repeat-containing protein	790	5,91E-46	84.2%
Pe164Q4C11F	homeobox-leucine zipper protein revoluta	811	1,24E-45	59.95%
Pe33Q4B4R	probable inactive receptor kinase at1g27190-like	564	1,29E-45	65.25%
Pe164Q1G10F	ac082645_13 protein kinase	865	2,25E-45	72.8%
Pe216Q4E5R	photosystem i p700 apoprotein partial	936	3,08E-45	85.2%
Pe216Q1C2R	transmembrane protein 184c-like	767	9,04E-45	83.7%
Pe141Q1E2F	phosphomethylpyrimidine kinase thiamin-phosphate pyrophosphorylase	676	1,22E-44	81.85%
Pe216Q1F12F	protein	710	1,81E-43	84.1%
Pe214Q3B6R	transformation transcription domain-associated	783	2,84E-43	87.5%
Pe75Q3H9R	protein gamete expressed 3-like	855	3,76E-43	54.45%
Pe173Q3C4R	phytochelatin synthase	558	6,87E-43	77.4%
Pe216Q4B7F	pleiotropic drug resistance protein 1-like	884	7,02E-43	84.85%
Pe141Q2F4R	cytochrome p450	612	7,11E-43	88.85%
Pe141Q2H9R	aux1-like amino acid permease	823	1,97E-42	94.2%
Pe101Q4H11F	copalyl diphosphate synthase	891	2,78E-42	89.5%
Pe93Q2G2R	bed finger-nbs-irr resistance protein	732	1,83E-41	60.8%
Pe84Q3F11R	3-deoxy-d-arabino-heptulosonate 7-phosphate synthase	801	5,04E-41	97.3%
Pe212Q1H7F	leucine-rich repeat receptor-like protein kinase pxl2-like	459	7,27E-41	87.05%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe69Q3E9F	nac domain-containing	909	9,28E-41	80.75%
Pe43Q4E12F	pentatricopeptide repeat-containing protein	360	1,31E-40	76.1%
Pe141Q4E1R	leucine-rich repeat-containing	681	3,90E-40	57.95%
Pe214Q4C8R	serine threonine-protein kinase at5g01020-like	356	5,05E-40	77.3%
Pe69Q3C7F	nucleobase-ascorbate transporter 11-like	740	8,48E-40	47.86%
Pe84Q3E8R	f-box family protein	1001	1,27E-39	72.95%
Pe101Q4H5R	protein far1-related sequence 5-like	872	1,60E-39	50.9%
Pe212Q4E3R	protein far1-related sequence 5-like	472	1,66E-39	53.6%
Pe173Q4A8F	saur family protein	754	1,31E-38	78.2%
Pe212Q1H4R	cytochrome p450	925	1,69E-38	96.75%
Pe212Q4E6R	tetratricopeptide repeat domain-containing protein	464	1,81E-38	93.3%
Pe214Q2F10R	cytochrome p450	320	2,22E-38	76.45%
Pe214Q3G9R	conserved hypothetical protein	486	2,81E-38	64.1%
Pe173Q4E4F	cytochrome p450	826	3,19E-38	66.4%
Pe212Q1C7F	rna-dependent rna polymerase 1	965	4,70E-38	65.9%
Pe214Q4G10F	membrane-bound transcription factor site-2 protease-like	802	5,88E-38	64.8%
Pe75Q1D12R	protein sda1 homolog	697	8,04E-38	85.5%
Pe214Q1H5F	transducin beta-like protein 3-like	739	1,17E-37	77.95%
Pe69Q2H8F	protein	589	2,01E-37	70.35%
Pe33Q2D9R	leucine rich repeat receptor	526	3,32E-37	63.3%
Pe141Q3H2R	tpr repeat-containing thioredoxin ttl1-like	870	4,96E-37	84.1%
Pe55Q4H3R	protein	701	5,80E-37	69.45%
Pe93Q2D2R	mitochondrial substrate carrier family protein c-like	889	7,09E-37	92.5%
Pe173Q4A4R	squamosa promoter-binding-like protein 7-like	734	1,24E-36	66.28%
Pe141Q4A6R	uncharacterized oxidoreductase chloroplastic-like	883	4,11E-36	90.05%
Pe69Q1A3R	glycerol-3-phosphate dehydrogenase	723	6,78E-36	94.05%
Pe141Q4G12F	phospholipid glycerol acyltransferase family protein	887	9,77E-36	81.25%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe173Q4F11F	transmembrane 9 superfamily member 4-like	689	1,79E-35	69.3%
Pe216Q4A3F	pg1 homology to homo sapiens	829	1,97E-35	74.3%
Pe69Q2D9R	formin-like protein 1-like	786	3,40E-35	89.35%
Pe164Q3H3R	e3 ubiquitin-protein ligase hos1-like	879	3,53E-35	73.6%
Pe212Q2H9R	PREDICTED: uncharacterized protein LOC100248664	902	3,72E-35	65.95%
Pe55Q1H1F	casp-like protein poptrdraft_833824-like	738	6,23E-35	78.2%
Pe69Q1G6F	heme oxygenase- multi-helical protein	572	6,43E-35	66.4%
Pe214Q2H8R	bromodomain protein	946	8,18E-35	74.5%
Pe33Q2H8F	protein	745	8,73E-35	84.85%
Pe141Q2D9F	lrr receptor-like serine threonine-protein kinase rch1-like	317	1,47E-34	78.5%
Pe216Q2E9F	probable serine threonine-protein kinase at1g54610-like	789	1,51E-34	93.25%
Pe75Q4F5R	callose synthase 9	874	1,62E-34	76.55%
Pe23Q3E5Ft2	polyketide cyclase dehydrase and lipid transport protein	498	1,67E-34	79.45%
Pe164Q2G4F	kinesin heavy	743	6,82E-34	60.95%
Pe75Q2F12F	pr-10 type pathogenesis-related protein	902	9,68E-34	71.45%
Pe173Q3F5F	pentatricopeptide repeat-containing	589	1,61E-33	63.6%
Pe84Q1G6R	cytochrome p450 734a1-like	834	3,62E-33	85.55%
Pe216Q1E10R	zeaxanthin epoxidase	822	5,91E-33	92.95%
Pe141Q3F5R	uncharacterized loc101215791	783	6,47E-33	59.33%
Pe84Q3A7R	beta xylosidase	456	6,96E-33	90.4%
Pe164Q4B3F	athila retroelement orf1 protein	717	9,12E-33	60.6%
Pe33Q2G1R	cytochrome p450 734a1-like	941	9,20E-33	85.55%
Pe185Q1H10F	glutamate decarboxylase	947	1,44E-32	96.8%
Pe84Q2H11R	nbs-lrr resistance protein	634	1,83E-32	67.05%
Pe69Q1H3F	probable peptide transporter at1g52190-like	846	2,52E-32	83.55%
Pe141Q3E12R	uncharacterized loc101219594	539	5,25E-32	90.1%
Pe75Q3G8F	rossmann-fold nad -binding domain-containing protein	800	5,29E-32	62.6%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe43Q4G5F	calcium lipid binding	795	5,89E-32	87.3%
Pe69Q2E7F	protein	482	7,50E-32	71.55%
Pe214Q2E12F	gdp-mannose transporter gont2-like	598	8,16E-32	58.1%
Pe84Q1B2R	acyl-coenzyme a thioesterase 13-like	950	1,08E-31	76.75%
Pe75Q3F1F	mediator-associated protein 1-like	655	1,20E-31	60.8%
Pe69Q4G9R	nadh-plastoquinone oxidoreductase subunit 3	374	1,42E-31	87.4%
Pe164Q1F8F	respiratory burst oxidase homolog protein h-like	891	3,85E-31	75.3%
Pe101Q4E7R	protein	772	5,04E-31	73.6%
Pe69Q1C6R	lipase class 3 family protein	955	5,11E-31	68.75%
Pe214Q3A5F	mitochondrial phosphate transporter	749	9,99E-31	62.75%
Pe93Q3G4F	flavin-containing monooxygenase yucca6-like	631	2,09E-30	90.3%
Pe173Q3H4R	hypothetical protein VITISV_034229	662	2,27E-30	55.4%
Pe214Q2A9F	nac domain ipr003441	719	3,44E-30	69.65%
Pe214Q2A9R	protein nap1-like	972	5,08E-30	62.4%
Pe216Q4F9R	serine threonine-protein kinase tio-like	818	5,91E-30	88.5%
Pe101Q4A11R	amino acid permease	893	9,13E-30	76.15%
Pe75Q4G5R	protein	853	1,22E-29	49.25%
Pe23Q3D11Ft2	athila retroelement orf1 protein	919	1,47E-29	56.15%
Pe216Q2H7R	mediator of rna polymerase ii transcription subunit 23-like	775	2,37E-29	90.9%
Pe84Q4E8R	s-adenosylmethionine partial	430	6,92E-29	95.15%
Pe212Q4D10F	alpha-galactosidase alpha-n-	700	1,06E-28	85.9%
Pe214Q2H3F	reverse transcriptase like protein	698	1,08E-28	54.85%
Pe101Q3H6R	e3 ubiquitin-protein ligase upl5-like	828	1,60E-28	66.56%
Pe75Q2C4F	u-box domain-containing protein 50-like	754	1,77E-28	84.0%
Pe93Q3C3R	signal recognition particle receptor subunit alpha	617	5,23E-28	61.1%
Pe214Q2H7R	probable atp-dependent rna helicase dhx36-like	682	5,33E-28	88.25%
Pe75Q4G6R	homeobox protein knotted-	860	1,14E-27	83.65%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe69Q2B5F	protein far1-related sequence 5-like	529	1,49E-27	51.45%
Pe216Q3A12R	tubulin--tyrosine ligase-like protein 12-like	487	1,76E-27	80.15%
Pe216Q1B2R	ap2 erf domain-containing transcription factor	514	2,33E-27	72.8%
Pe69Q2F12R	cyanidin-3-o-glucoside 2-o-glucuronosyltransferase-like	750	2,62E-27	72.4%
Pe185Q2G6F	nadh	848	5,32E-27	89.63%
Pe101Q3H4R	protein argonaute 4	702	5,96E-27	77.75%
Pe185Q4E8F	probable inactive serine threonine-protein kinase lvsg-like	818	6,66E-27	76.6%
Pe55Q4F6R	alpha-glucan h isozyme-like	688	9,20E-27	96.05%
Pe164Q1F9F	dna repair protein rad51-like protein	734	1,07E-26	79.2%
Pe141Q2F8F	histone acetyltransferase	967	1,29E-26	74.3%
Pe164Q4C6F	beta-amyrin synthase	519	1,43E-26	84.45%
Pe69Q2E5F	methionine sulfoxide reductase type	756	4,49E-26	84.05%
Pe69Q4F2R	type-b response regulator	477	1,26E-25	56.36%
Pe93Q4C8F	exosome complex exonuclease rrp44-like	643	1,36E-25	90.2%
Pe75Q4G12F	probable galacturonosyltransferase 7-like	727	1,48E-25	79.4%
Pe33Q4A10R	rna polymerase ii subunit b1 ctd phosphatase rpap2 homolog	348	1,70E-25	68.0%
Pe26Q4E5F	wrky transcription factor	332	2,06E-25	70.9%
Pe43Q4A6F	chaperonin cpn60- mitochondrial-like	660	2,31E-25	84.05%
Pe101Q4E4F	f-box domain containing protein	734	3,29E-25	47.9%
Pe164Q4A12F	zinc cchc-type	857	3,32E-25	53.95%
Pe173Q4A3R	intron-binding protein aquarius	842	3,54E-25	48.15%
Pe173Q1H7R	polyneuridine-aldehyde esterase	718	3,55E-25	85.5%
Pe101Q3D8R	general transcription factor iie subunit	713	3,84E-25	74.95%
Pe69Q1B4F	calmodulin	518	6,04E-25	95.95%
Pe214Q2B7R	crs1 domain containing protein	712	1,75E-24	77.2%
Pe93Q4F11F	photosystem i assembly protein ycf3	609	1,82E-24	88.3%
Pe214Q3F6R	probable pectate lyase 4-like	757	1,89E-24	83.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe69Q4C11F	carboxyl-terminal proteinase	681	2,73E-24	70.85%
Pe212Q4G10F	receptor-like kinase	829	4,15E-24	45.75%
Pe93Q2H8R	crossover junction endonuclease mus81	836	5,31E-24	75.9%
Pe101Q4D3R	non-ltr retroelement reverse transcriptase	849	5,71E-24	45.35%
Pe33Q4H8F	sorting nexin 1	417	7,40E-24	58.5%
Pe93Q2F10R	flavine-containing monooxygenase	784	7,71E-24	57.75%
Pe141Q1D6F	protein	762	1,11E-23	85.45%
Pe141Q4H5R	cellular nucleic acid-binding partial	611	1,50E-23	50.85%
Pe84Q1G9F	carboxypeptidase d	707	1,57E-23	81.85%
Pe173Q2B11F	protein	792	2,60E-23	56.0%
Pe173Q4C4R	phytochelatin synthase	458	2,71E-23	67.25%
Pe75Q2H10R	protein far1-related sequence 5-like	863	2,92E-23	49.08%
Pe93Q2H11R	splicing factor 3a subunit 3-like	792	3,43E-23	97.2%
Pe141Q2E3R	fanconi anemia group i protein	919	3,70E-23	72.75%
Pe214Q1C5F	vamp-like protein ykt61-like	1030	4,91E-23	50.95%
Pe214Q1H10F	ring-h2 finger protein atl13-like	1037	7,46E-23	86.35%
Pe43Q4B1R	polyamine oxidase	450	7,96E-23	52.6%
Pe26Q4F7F	cytochrome p450 86b1-like	418	9,39E-23	58.6%
Pe214Q4E8R	at5g08400 f8l15_130	587	1,02E-22	63.75%
Pe69Q3F4R	ribonuclease h protein at1g65750-like	799	1,13E-22	47.3%
Pe84Q4H8R	protein	733	1,52E-22	91.2%
Pe101Q2H2R	eukaryotic translation initiation factor 3 subunit h-like	906	1,53E-22	67.5%
Pe26Q1B7R	heme-binding protein 2-like	596	1,54E-22	70.8%
Pe75Q1F3F	protein	586	1,68E-22	61.5%
Pe214Q2H4R	l-galactose dehydrogenase	743	2,24E-22	89.6%
Pe214Q4H9R	plant synaptotagmin	756	2,75E-22	53.8%
Pe212Q2H7R	glucan -beta-glucosidase	787	4,76E-22	82.05%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe141Q2A7F	PREDICTED: uncharacterized protein LOC101292944	441	5,30E-22	56.31%
Pe93Q2C5R	cbs domain-containing protein cbscbpb5-like	899	6,27E-22	52.9%
Pe101Q3F12R	retrotransposon gag protein	452	6,72E-22	57.24%
Pe212Q2F9R	reverse transcriptase	904	9,75E-22	57.4%
Pe75Q3E7R	retrotransposon protein	626	1,03E-21	53.65%
Pe26Q4H9F	retrotransposon gag protein	553	1,20E-21	54.45%
Pe214Q3D6F	rna polymerase ii-associated protein	791	2,51E-21	74.25%
Pe212Q1F5F	60s ribosomal protein l21-1-like	961	3,29E-21	91.05%
Pe101Q2F2R	probable methyltransferase pmt11-like	751	1,08E-20	84.1%
Pe173Q1D11R	pre-rrna-processing protein tsr2 homolog	734	1,12E-20	69.5%
Pe214Q3G10R	endoplasmic reticulum metalloproteinase 1-like	637	1,33E-20	70.0%
Pe185Q3E5F	tetratricopeptide repeat-containing protein	540	1,37E-20	70.35%
Pe164Q4C1R	ubiquitin carboxyl-terminal hydrolase 8-like	538	1,43E-20	65.4%
Pe185Q1F7R	beta-d-galactosidase	499	1,44E-20	74.5%
Pe173Q3G7R	translation initiation factor eif-3 subunit 7	699	1,47E-20	95.55%
Pe216Q1B6R	flavanone 3-dioxygenase-like	1049	3,56E-20	80.4%
Pe214Q2G4F	ubiquitin carboxyl-terminal hydrolase-like protein	703	4,05E-20	71.47%
Pe212Q2H6R	rop guanine nucleotide exchange factor 1-like	696	4,13E-20	55.56%
Pe141Q1E1R	bri1-kd interacting protein	632	4,26E-20	57.93%
Pe93Q2A4F	extra-large gtp-binding protein 3	759	5,78E-20	81.1%
Pe75Q4E7F	shaggy-related protein kinase alpha-like	849	1,05E-19	62.95%
Pe173Q2A5R	unnamed protein product	740	1,06E-19	78.5%
Pe75Q1H2R	ac099402_4 22 kda kafirin cluster ty3-gypsy type	686	1,45E-19	49.4%
Pe216Q1E2F	telomere repeat-binding protein 4-like	713	2,05E-19	88.7%
Pe93Q2F7R	probable serine incorporator-like	679	2,45E-19	92.15%
Pe214Q3G10F	uncharacterized loc101222640	557	2,53E-19	74.58%
Pe185Q2E1F	non-ltr retroelement reverse transcriptase	860	2,94E-19	45.2%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe216Q3D12F	uncharacterized loc101213119	686	2,96E-19	62.76%
Pe93Q2H9R	crossover junction endonuclease mus81	1069	3,58E-19	79.95%
Pe93Q4C12F	protein binding	752	3,94E-19	97.3%
Pe75Q1F8F	snf1-related protein kinase	622	3,98E-19	88.9%
Pe101Q4F12R	ac099402_4 22 kda kafirin cluster ty3-gypsy type	617	4,67E-19	51.05%
Pe33Q4E2F	athila transposon protein	383	4,86E-19	62.2%
Pe69Q2H2F	phospholipid-transporting atpase 4-like	545	6,20E-19	51.85%
Pe173Q2H6R	protein-tyrosine-phosphatase mkp1-like	418	7,17E-19	73.17%
Pe93Q1E9F	jag1	575	7,47E-19	88.0%
Pe164Q4G4F	atp-dependent dna helicase q-like 1-like	766	9,02E-19	86.95%
Pe164Q2H1F	cbs domain-containing protein cbscbpb5-like	651	9,27E-19	53.05%
Pe101Q2B2R	protein	872	9,52E-19	51.6%
Pe55Q1H11F	ramosa 1 enhancer locus 2	655	1,05E-18	74.0%
Pe33Q2E4R	12-oxophytodienoate reductase	560	1,05E-18	82.55%
Pe84Q3B9R	triose phosphate partial	456	1,15E-18	70.75%
Pe212Q2F7R	dna polymerase zeta subunit	613	1,29E-18	91.55%
Pe141Q2F4F	dna binding	622	1,47E-18	75.82%
Pe216Q3F2R	ribonuclease h protein at1g65750-like	700	1,63E-18	45.95%
Pe185Q1A5F	mlo-like protein 5-like	871	1,77E-18	51.45%
Pe214Q3A11R	exocyst complex component 5	758	2,05E-18	93.85%
Pe212Q4E2R	non-ltr retroelement reverse transcriptase	487	2,50E-18	50.35%
Pe185Q4A3F	probable s-acyltransferase at3g04970-like	867	3,77E-18	68.4%
Pe173Q1G1R	gag-pol protein	713	4,48E-18	50.55%
Pe185Q4D6F	retropepsin-like protein	441	4,63E-18	64.2%
Pe216Q2C3F	polypyrimidine tract-binding protein homolog 3-like	923	4,90E-18	54.25%
Pe173Q3A8F	saur family protein	527	4,94E-18	66.35%
Pe69Q2E2F	reverse transcriptase like protein	644	6,44E-18	64.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe93Q4A7F	cdk5rap3-like protein	750	8,08E-18	80.9%
Pe84Q2F4R	uncharacterized loc101216222	759	8,26E-18	81.9%
Pe216Q3C5R	adp-ribosylation factor gtpase-activating protein agd4-like	801	1,10E-17	70.65%
Pe214Q3B3F	e3 ubiquitin ligase-like protein	478	1,19E-17	91.4%
Pe93Q3E5F	predicted protein	540	1,38E-17	66.15%
Pe173Q1F4F	e3 ubiquitin-protein ligase at3g02290-like	612	1,74E-17	86.15%
Pe101Q4H2F	1510387aretrotransposon del1-46	854	2,23E-17	49.3%
Pe173Q4D12R	retrotransposon ty1-copia subclass	596	3,08E-17	62.9%
Pe212Q2A6R	retrotransposon unclassified	842	4,67E-17	48.65%
Pe173Q4D3F	retrotransposon ty3-gypsy subclass	522	6,22E-17	49.8%
Pe101Q3E4F	ras-related protein rabh1b	1009	6,37E-17	90.0%
Pe75Q2B9F	lipid binding	899	7,19E-17	79.15%
Pe216Q4E4R	alpha- -sialyltransferase 8b	856	8,34E-17	78.45%
Pe216Q3A9R	protein	573	8,70E-17	65.2%
Pe216Q4A10F	conserved hypothetical protein	785	8,76E-17	65.2%
Pe93Q4F12R	udp-n-acetylglucosamine transferase subunit alg14 homolog	615	9,06E-17	61.45%
Pe185Q2D10F	non-ltr retroelement reverse transcriptase	872	1,04E-16	45.35%
Pe173Q3F5R	2fe-2s iron-sulfur cluster binding domain-containing protein	546	1,13E-16	79.8%
Pe101Q4F10R	protein argonaute 16-like	796	1,38E-16	55.7%
Pe164Q3B5F	proline-rich receptor-like protein kinase perk8-like	688	1,49E-16	71.29%
Pe93Q2A10R	f-box fbd lrr-repeat protein	1029	1,51E-16	49.6%
Pe216Q4D8R	1211235ck orf 75	987	1,52E-16	84.42%
Pe185Q2B8R	retrotransposon ty3-gypsy subclass	610	1,53E-16	49.5%
Pe185Q2A10F	cc-nbs-lrr resistance protein	830	1,53E-16	73.35%
Pe93Q2E3R	retrotransposon ty3-gypsy subclass	724	1,77E-16	50.95%
Pe185Q2H2F	saur family protein	870	1,93E-16	78.45%
Pe185Q2A1F	cellular nucleic acid-binding protein	944	2,25E-16	51.45%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe173Q4C8R	stage iii sporulation protein aa	954	2,37E-16	86.2%
Pe164Q4A5R	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 2-like	822	2,70E-16	67.3%
Pe185Q1A5R	50s ribosomal protein chloroplastic-like	937	4,79E-16	88.45%
Pe173Q4A10F	glycine-rich protein	795	5,03E-16	80.1%
Pe69Q1E7F	phosphatidate cytidyltransferase	974	5,85E-16	96.1%
Pe216Q2H9F	protein	823	6,34E-16	74.25%
Pe164Q1F12R	reverse transcriptase	748	7,69E-16	49.75%
Pe216Q4B9R	u-box domain-containing protein 14-like	1068	8,10E-16	74.06%
Pe216Q3C2R	sorting nexin-16	808	8,24E-16	86.0%
Pe75Q2A4F	retrotransposon unclassified	702	8,33E-16	49.5%
Pe141Q4B7F	cop1-interacting protein 7	795	8,43E-16	70.24%
Pe216Q4B2F	retrotransposon unclassified	698	9,76E-16	49.35%
Pe101Q2B3R	ubiquitin-conjugating enzyme	451	1,29E-15	99.85%
Pe141Q2F6F	athila retroelement orf1 protein	733	1,63E-15	60.75%
Pe93Q3D4R	udp-glycosyltransferase 71c4	635	2,36E-15	72.6%
Pe212Q2E9R	protein	893	2,56E-15	71.13%
Pe216Q1E11R	class i chitinase	794	2,70E-15	58.7%
Pe173Q2E3R	ribosome biogenesis gtpase -like	460	3,09E-15	66.95%
Pe185Q1D4F	retrotransposon line subclass	854	3,71E-15	51.35%
Pe185Q2F5F	nucleoside-triphosphatase nucleotide binding protein	928	3,86E-15	60.3%
Pe212Q3H3F	nucleic acid binding	752	3,86E-15	75.5%
Pe23Q3G11R	profilin-like protein	310	3,91E-15	86.85%
Pe84Q2E7F	e3 ubiquitin-protein ligase rf298-like	653	4,12E-15	74.85%
Pe212Q3H11F	hypothetical protein PRUPE_ppa023432mg, partial	349	5,17E-15	51.75%
Pe33Q2F7R	retrotransposon ty1-copia subclass	663	5,20E-15	60.45%
Pe185Q1D4R	protein	721	5,65E-15	51.0%
Pe214Q2G10F	cc-nbs-lrr resistance protein	405	5,81E-15	65.85%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe212Q2H4R	pectinacylesterase family protein	835	6,24E-15	86.15%
Pe214Q3B8R	retrotransposon identical	670	6,31E-15	58.25%
Pe164Q4G5F	PREDICTED: uncharacterized protein LOC101292944	441	6,59E-15	58.8%
Pe84Q1G12F	mitogen activated protein kinase kinase kinase	621	9,95E-15	73.85%
Pe33Q4B8F	endoribonuclease dicer homolog 2-like	391	1,20E-14	75.35%
Pe75Q4H11F	multiple c2 and transmembrane domain-containing protein	715	1,27E-14	55.65%
Pe164Q2F2F	cell cycle checkpoint protein rad17-like	774	1,28E-14	83.38%
Pe173Q1H11R	pleckstrin homology domain-containing protein	712	1,31E-14	51.35%
Pe214Q1B4R	kinase family protein	928	1,43E-14	85.75%
Pe164Q3A6R	40s ribosomal protein s26-1-like	1022	1,58E-14	59.05%
Pe164Q2E8F	water chloroplastic-like	756	1,88E-14	91.0%
Pe141Q3E11R	tir-nbs-lrr resistance protein	762	1,93E-14	52.7%
Pe93Q4H2R	PREDICTED: uncharacterized protein LOC101267053	885	1,94E-14	60.8%
Pe33Q4H7R	conserved hypothetical protein	622	2,34E-14	96.35%
Pe212Q2C3R	auxin-induced protein 5ng4-like	1045	3,02E-14	66.7%
Pe214Q1D2R	at2g10940-like protein	890	3,64E-14	94.45%
Pe1Q1A8R	glutamate malate translocator	455	3,89E-14	83.35%
Pe93Q3H10R	probable xaa-pro aminopeptidase p-like	614	3,92E-14	96.65%
Pe101Q4F1R	retropepsin-like protein	682	3,99E-14	68.71%
Pe93Q3F1F	cellular nucleic acid-binding partial	655	4,53E-14	49.85%
Pe185Q1F4F	phenylalanine ammonia-lyase	323	4,61E-14	83.5%
Pe141Q1E7R	athila retroelement orf1 protein	393	4,73E-14	60.35%
Pe216Q3G7F	benzyl alcohol o-benzoyltransferase-like	567	5,18E-14	63.5%
Pe141Q2A10F	bifunctional nitrilase nitrile hydratase nit4b-like	734	5,46E-14	80.5%
Pe101Q4C6R	protein	813	5,47E-14	72.3%
Pe141Q1F12F	polyketide hydroxylase-like	644	5,71E-14	78.3%
Pe212Q1C8R	protein	739	9,24E-14	47.3%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe43Q4D11R	transcription factor lhw-like	571	1,17E-13	87.4%
Pe43Q4A1R	nephrin	645	1,19E-13	91.75%
Pe84Q3H7R	aldo keto	607	1,41E-13	80.7%
Pe185Q3B8F	nad h dehydrogenase s	889	1,79E-13	79.3%
Pe216Q2G1F	regulatory-associated protein of tor 1-like	888	1,90E-13	87.2%
Pe212Q3E10F	r2r3-myb transcription	570	2,07E-13	57.13%
Pe141Q4H11R	transferring glycosyl	872	2,07E-13	72.08%
Pe164Q4D5F	retrotransposon ty1-copia subclass	671	3,13E-13	64.55%
Pe141Q1F7F	protein	474	3,21E-13	49.65%
Pe23Q3G7R	protein	593	3,53E-13	67.2%
Pe75Q2E9R	atp binding	749	3,87E-13	69.8%
Pe164Q1C12F	ras-related protein rabelc-like	901	4,92E-13	61.85%
Pe101Q2G1R	multiple c2 and transmembrane domain-containing protein	478	6,28E-13	47.55%
Pe216Q1F9F	conserved hypothetical protein	609	6,80E-13	62.0%
Pe212Q3A12R	retrotransposon ty3-gypsy subclass	635	7,74E-13	49.75%
Pe185Q1F11R	cellular nucleic acid-binding partial	825	7,76E-13	51.1%
Pe75Q3D2R	PREDICTED: uncharacterized protein LOC101292944	458	8,43E-13	53.0%
Pe185Q2D9F	gag-pol identical	687	8,51E-13	58.05%
Pe216Q2G10F	protein	707	1,06E-12	61.3%
Pe75Q1E4F	hypothetical chloroplast rf1	1114	1,12E-12	65.71%
Pe173Q2E7F	1510387aretrotransposon del1-46	860	1,29E-12	56.7%
Pe214Q2C12R	retrotransposon ty3-gypsy subclass	633	1,57E-12	47.25%
Pe216Q4F5F	orf448	897	1,57E-12	63.8%
Pe75Q1C12F	protein	615	1,92E-12	73.6%
Pe216Q3D2F	gag-pol protein	524	2,02E-12	51.85%
Pe173Q4B12R	protein far1-related sequence 5-like	702	2,11E-12	47.5%
Pe164Q3A12F	gag-pol protein	574	2,42E-12	49.35%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe141Q1E8F	retrotransposon ty1-copia subclass	545	2,72E-12	57.7%
Pe43Q4H3F	trna-dihydrouridine synthase 2-like protein	681	2,83E-12	86.95%
Pe216Q4B3F	retrotransposon ty3-gypsy subclass	824	2,85E-12	50.15%
Pe75Q3C11R	rnase partial	860	2,87E-12	76.05%
Pe214Q3B4F	retrotransposon unclassified	374	2,93E-12	51.2%
Pe173Q2C1F	kinesin-like protein	565	3,06E-12	46.6%
Pe75Q1A2R	otu domain-containing protein at3g57810-like	925	4,00E-12	86.45%
Pe101Q3F2R	multiple c2 and transmembrane domain-containing protein	598	4,00E-12	56.59%
Pe101Q2F7R	cbs domain-containing protein	825	5,10E-12	79.92%
Pe93Q4A11F	retrotransposon ty3-gypsy subclass	779	6,16E-12	49.4%
Pe101Q4A12F	retrotransposon ty3-gypsy subclass	844	6,40E-12	49.2%
Pe1Q1G9F	atp binding	931	6,52E-12	70.0%
Pe93Q4E1F	seryl-trna synthetase	557	6,67E-12	56.5%
Pe69Q3F3R	multiple c2 and transmembrane domain-containing protein	510	6,73E-12	55.53%
Pe69Q4G9F	protein	435	6,86E-12	84.36%
Pe164Q3A7F	activating signal cointegrator 1 complex subunit 2-like	880	7,14E-12	64.0%
Pe216Q1H12F	1510387aretrotransposon del1-46	907	7,52E-12	50.07%
Pe141Q2G3F	ubiquitin-conjugating enzyme e2 variant 1a	985	8,12E-12	85.15%
Pe173Q3H7R	gdp-mannose transporter gonst2-like	782	9,17E-12	87.65%
Pe173Q2H9R	40s ribosomal protein sa-like	593	9,40E-12	87.6%
Pe214Q3A5R	carbonic anhydrase	789	9,44E-12	84.8%
Pe164Q2A3R	protein	941	1,11E-11	46.5%
Pe164Q3H6R	one helix protein 2	890	1,22E-11	94.5%
Pe101Q4B10F	atp-dependent dna helicase ddm1-like	863	1,44E-11	91.85%
Pe84Q2E5R	probable gmp synthase	710	1,65E-11	66.0%
Pe43Q4D4R	pol protein	370	1,66E-11	54.2%
Pe141Q3B10R	60s ribosomal protein l13a	782	1,70E-11	91.4%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe216Q4G6R	protein	1038	1,78E-11	78.85%
Pe141Q4E11F	glycosylphosphatidylinositol anchor attachment 1	828	1,89E-11	79.9%
Pe212Q1E1R	niemann-pick c1	721	1,97E-11	94.5%
Pe214Q2A8R	splicing factor u2af large	892	2,22E-11	51.17%
Pe214Q4F8F	ribonuclease 3-like protein 2-like	654	2,31E-11	67.05%
Pe185Q2H5F	gag-pol protein	694	2,86E-11	52.75%
Pe93Q3C5F	retrotransposon ty1-copia subclass	581	2,93E-11	61.7%
Pe173Q4E11R	retrotransposon ty3-gypsy subclass	714	3,02E-11	49.75%
Pe216Q3E2R	rna-directed dna polymerase (reverse transcriptase)-related family protein	470	3,10E-11	55.4%
Pe141Q4A6F	cleavage and polyadenylation specificity factor subunit 1-like	929	3,37E-11	88.25%
Pe185Q2E5R	non-ltr retroelement reverse transcriptase	647	3,69E-11	46.4%
Pe75Q4H3F	retrotransposon ty3-gypsy subclass	669	3,69E-11	47.2%
Pe173Q3F11R	protein	680	4,28E-11	90.8%
Pe55Q1D10F	retrotransposon gag protein	262	4,44E-11	65.75%
Pe212Q4F1R	probable wrky transcription factor 32-like	570	4,63E-11	61.15%
Pe101Q3B10F	retrotransposon ty1-copia subclass	813	4,86E-11	59.0%
Pe214Q1G4R	hypothetical protein	931	5,33E-11	43.85%
Pe93Q4D8F	protein	677	5,47E-11	84.67%
Pe43Q4H12R	at5g37260-like protein	683	5,99E-11	63.22%
Pe141Q4H1F	probable flavin-containing monooxygenase 1-like	842	6,72E-11	75.0%
Pe43Q4E11R	retrotransposon gag protein	548	1,05E-10	53.95%
Pe84Q2H2R	retrotransposon ty3-gypsy subclass	715	1,09E-10	51.05%
Pe212Q1C5F	snf1-related protein kinase	468	1,09E-10	75.15%
Pe26Q1D1R	mudra protein	742	1,13E-10	70.45%
Pe214Q2B1R	retrotransposon gag protein	724	1,15E-10	54.32%
Pe75Q3G8R	uncharacterized loc101209682	851	1,15E-10	88.3%
Pe75Q4C8F	sulfate bicarbonate oxalate exchanger and transporter sat-1	741	1,48E-10	90.8%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe212Q2B10R	predicted protein	891	1,71E-10	60.25%
Pe33Q2E3R	retrotransposon ty3-gypsy subclass	387	1,96E-10	51.3%
Pe185Q4G7F	retrotransposon ty3-gypsy subclass	735	2,16E-10	49.85%
Pe216Q3C2F	peptidyl-prolyl cis-trans isomerase cyp40-like	731	2,43E-10	91.5%
Pe164Q3H12F	solute carrier family 25 member 39-like	1003	2,55E-10	90.3%
Pe214Q4A7R	pfkb-type carbohydrate kinase family protein	836	2,90E-10	79.0%
Pe93Q2A12R	reverse transcriptase	523	3,24E-10	53.55%
Pe212Q4H7F	dna replication licensing factor mcm6-like	761	3,58E-10	80.11%
Pe141Q4G1R	single myb histone	800	3,71E-10	75.5%
Pe75Q2C5F	leucoanthocyanidin dioxygenase	1030	4,12E-10	86.2%
Pe214Q4E1F	non-ltr retroelement reverse transcriptase	509	4,20E-10	50.8%
Pe69Q3B8R	retrotransposon gag protein	585	4,59E-10	48.17%
Pe185Q1G2R	reverse transcriptase	743	4,63E-10	50.9%
Pe173Q3E10R	ferredoxin-dependent glutamate synthase	524	5,24E-10	51.15%
Pe33Q2G1F	protein club	332	5,49E-10	69.79%
Pe101Q3G12F	fyve finger-containing phosphoinositide	863	6,24E-10	64.33%
Pe75Q4G9R	endonuclease exonuclease phosphatase domain-containing protein	866	6,52E-10	87.93%
Pe55Q1H3F	hypothetical protein VITISV_000323	390	6,69E-10	50.2%
Pe173Q2E2R	retrotransposon ty1-copia subclass	787	6,72E-10	60.2%
Pe55Q2F12R	transcription initiation factor tfiid subunit 6-like	667	6,75E-10	90.65%
Pe75Q2E1R	retrotransposon ty3-gypsy subclass	784	8,18E-10	48.75%
Pe93Q1F8F	at4g33690 t16l1_180	964	8,28E-10	83.2%
Pe84Q1A6R	squalene expressed	595	1,03E-09	88.6%
Pe101Q4H5F	AC009317_2T30E16.4	889	1,13E-09	68.6%
Pe33Q4F7F	calcium ion binding	359	1,15E-09	56.33%
Pe93Q4C10F	hypothetical protein PRUPE_ppa014973mg, partial	703	1,19E-09	43.63%
Pe141Q2B2F	transmembrane protein g5p	719	1,32E-09	86.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe216Q2F1F	regulatory-associated protein of tor 1-like	695	1,38E-09	86.3%
Pe216Q1B5F	reverse transcriptase like protein	817	1,45E-09	57.4%
Pe101Q3F10R	protein	658	1,67E-09	62.0%
Pe185Q2H5R	low quality protein: uncharacterized loc101222143	685	1,68E-09	70.25%
Pe101Q2B8R	oxysterol-binding protein 2a-like	771	1,77E-09	77.71%
Pe33Q2F1F	protein club	741	1,94E-09	80.58%
Pe173Q4C5R	copia-like polypeptide	773	2,01E-09	49.6%
Pe26Q1H6R	actin-related protein 7-like isoform 2	1057	2,04E-09	97.5%
Pe43Q4F9F	retrotransposon ty3-gypsy subclass	361	2,53E-09	51.4%
Pe173Q4C12R	at1g65720 f1e22_13	427	2,55E-09	60.5%
Pe69Q2E7R	pentatricopeptide repeat-containing protein mitochondrial-like	806	2,80E-09	59.0%
Pe55Q4D4F	hypothetical protein NEUTE2DRAFT_76221	422	2,85E-09	74.82%
Pe55Q4C1F	indole-3-acetic acid inducible 26	633	3,09E-09	73.15%
Pe101Q3C9F	syntaxin, putative	892	3,56E-09	60.0%
Pe216Q4B7R	protein	916	3,58E-09	79.65%
Pe101Q3E11F	retropepsin-like protein	473	4,71E-09	50.0%
Pe214Q3E2R	retrotransposon ty3-gypsy subclass	669	5,48E-09	46.85%
Pe101Q3B2R	gag-pol protein	642	5,69E-09	52.55%
Pe141Q4E12F	retrotransposon ty1-copia subclass	751	7,06E-09	57.5%
Pe1Q1D8F	predicted protein	249	7,13E-09	56.0%
Pe101Q3D8F	two-component response regulator-like apr2-like	683	7,19E-09	85.2%
Pe69Q4G11R	hypothetical protein PRUPE_ppa023432mg, partial	321	8,12E-09	53.67%
Pe164Q2B7R	ac099402_4 22 kda kafirin cluster ty3-gypsy type	739	8,69E-09	49.25%
Pe93Q3D9R	nuclear control of atpase protein	794	8,71E-09	72.55%
Pe33Q2E9R	protein	628	8,77E-09	76.25%
Pe216Q4C12F	reverse transcriptase like protein	862	9,59E-09	64.57%
Pe212Q3E2F	retrotransposon gag protein	372	9,81E-09	52.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe93Q1D4R	reverse transcriptase	716	9,97E-09	50.75%
Pe69Q2A6F	retrotransposon ty1-copia subclass	507	1,05E-08	64.65%
Pe33Q2H2R	athila retroelement orf1 protein	657	1,05E-08	51.55%
Pe93Q2F3R	retrotransposon ty3-gypsy subclass	785	1,06E-08	43.75%
Pe216Q2D6F	retrotransposon ty1-copia subclass	694	1,08E-08	66.59%
Pe69Q1D7F	PREDICTED: uncharacterized protein LOC101218085	602	1,12E-08	57.56%
Pe69Q4D12R	adomet-dependent rRNA methyltransferase spb1-like	642	1,24E-08	77.56%
Pe75Q1A11R	nuclear export mediator factor nemf-like	682	1,28E-08	91.5%
Pe214Q4B3R	transparent testa 12 protein	514	1,28E-08	73.95%
Pe214Q3E1F	retrotransposon ty3-gypsy subclass	606	1,29E-08	52.0%
Pe93Q2C11F	protein	484	1,34E-08	45.33%
Pe93Q4F7F	ribosomal protein l10	764	1,34E-08	87.4%
Pe69Q3E6R	PREDICTED: uncharacterized protein LOC101298139	760	1,47E-08	45.0%
Pe141Q1D2R	60s ribosomal protein l36-2	623	1,48E-08	80.89%
Pe69Q2H9R	1510387aretrotransposon del1-46	795	1,55E-08	54.88%
Pe214Q4B3F	40s ribosomal protein s8-like	616	1,64E-08	80.9%
Pe216Q3C12R	hypothetical protein MTR_2g085130	747	1,68E-08	61.5%
Pe141Q1H3F	ap2 erf domain-containing transcription factor	536	1,70E-08	57.55%
Pe216Q4F8F	retropepsin-like protein	903	1,79E-08	61.0%
Pe69Q4C11R	ndr family protein	602	2,03E-08	39.5%
Pe141Q1D12F	protein trigalactosyldiacylglycerol chloroplastic-like	803	2,07E-08	79.71%
Pe43Q4H11F	predicted protein	812	2,18E-08	79.0%
Pe216Q3E11R	retrotransposon ty1-copia subclass	841	2,21E-08	61.83%
Pe214Q2D11R	calcium calmodulin dependent protein kinase	612	2,22E-08	83.53%
Pe173Q1H9R	40s ribosomal protein sa-like	863	2,26E-08	83.2%
Pe84Q4F9R	virion binding	761	2,27E-08	70.56%
Pe93Q3B11F	epstein-barr nuclear	540	2,54E-08	80.75%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe141Q4C2R	protein	871	2,56E-08	52.27%
Pe216Q1B8R	retrotransposon ty3-gypsy subclass	554	2,77E-08	54.19%
Pe55Q4E3F	predicted protein	520	2,95E-08	61.0%
Pe185Q3F9F	retrotransposon ty3-gypsy subclass	778	3,08E-08	48.4%
Pe93Q4C2R	dna binding	701	3,08E-08	54.5%
Pe23Q3E6Ft1	retrovirus-related polyproteins	247	3,15E-08	74.0%
Pe84Q4A4R	hypothetical protein VITISV_006250	513	3,22E-08	58.0%
Pe43Q4D2R	retrotransposon ty3-gypsy subclass	428	3,28E-08	47.65%
Pe69Q4A3F	protein	551	3,30E-08	49.35%
Pe185Q2E6F	retrotransposon gag protein	783	3,32E-08	56.6%
Pe185Q2H7R	retrotransposon ty3-gypsy subclass	861	3,40E-08	48.15%
Pe33Q4D6R	hypothetical protein CHLNCDRAFT_23080	661	3,45E-08	83.0%
Pe173Q2D4R	replication protein a 70 kda dna-binding subunit-like	331	3,48E-08	85.17%
Pe101Q4E12R	hypothetical protein PRUPE_ppb020897mg	754	3,51E-08	48.5%
Pe185Q2E4F	predicted protein	843	3,62E-08	55.0%
Pe93Q1G2R	b3 domain-containing transcription factor vrn1-like	665	3,70E-08	60.73%
Pe75Q2E8F	copia-like retrotransposable element	839	3,77E-08	46.33%
Pe141Q1F3R	retrotransposon ty1-copia subclass	490	3,98E-08	73.75%
Pe173Q2E6F	hypothetical protein VITISV_028342	844	4,02E-08	54.65%
Pe216Q3H6F	hypothetical protein PRUPE_ppb020037mg	923	4,43E-08	62.8%
Pe69Q2H4F	flavanone 3-dioxygenase-like	522	4,52E-08	75.0%
Pe216Q4E2F	pol protein	1007	5,22E-08	62.8%
Pe185Q2G10R	nadh dehydrogenase subunit 2	802	5,23E-08	75.15%
Pe216Q3B8R	serine threonine-protein kinase	890	5,26E-08	48.6%
Pe173Q1E6F	hypothetical protein VITISV_022906	711	5,34E-08	54.35%
Pe1Q1G8R	hypothetical protein PRUPE_ppa022673mg	900	5,38E-08	54.5%
Pe185Q1F9R	hypothetical protein MTR_001s0023	681	5,39E-08	64.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe75Q1H5R	hypothetical protein PRUPE_ppa015000mg	870	5,49E-08	52.67%
Pe26Q4B7F	solute carrier family 25 member 39-like	470	5,52E-08	50.0%
Pe69Q4D12F	bel1 homeotic	480	5,80E-08	88.0%
Pe141Q4C6F	gag-pol identical	786	5,85E-08	46.65%
Pe93Q4A3R	mitochondrial translocator assembly and maintenance protein 41 homolog	610	5,93E-08	63.06%
Pe55Q1A3F	non-ltr reverse transcriptase	728	6,32E-08	46.0%
Pe164Q1G6F	hypothetical protein PRUPE_ppa022673mg	690	6,32E-08	54.33%
Pe185Q1F5R	cell wall-associated partial	823	6,52E-08	61.4%
Pe43Q4E10R	athila retroelement orf1 protein	565	7,30E-08	50.9%
Pe164Q2F5F	pentatricopeptide repeat-containing protein mitochondrial-like	579	7,54E-08	78.08%
Pe214Q3D11R	rna binding	665	7,81E-08	76.15%
Pe26Q1B12R	a-kinase anchor protein 17a-like	1018	7,99E-08	72.57%
Pe101Q3H7F	glutamate synthase	802	9,19E-08	88.7%
Pe164Q1G1F	nadh-plastoquinone oxidoreductase subunit 7	901	9,42E-08	95.25%
Pe101Q3F12F	gag-pol protein	686	9,57E-08	50.57%
Pe212Q4E5F	pleckstrin homology and lipid-binding start domain-containing protein	621	9,61E-08	92.05%
Pe84Q3F6R	hypothetical protein PRUPE_ppa008581mg	598	9,97E-08	80.0%
Pe214Q4E11F	retrotransposon ty3-gypsy subclass	492	1,01E-07	48.85%
Pe212Q3A5R	rnase h domain-containing protein	715	1,01E-07	46.15%
Pe164Q3A12R	retrotransposon gag protein	745	1,13E-07	44.8%
Pe26Q4C1F	rna-directed dna polymerase ribonuclease	559	1,15E-07	53.0%
Pe141Q1B4F	9-cis-epoxycarotenoid dioxygenase	577	1,24E-07	89.2%
Pe141Q3B11R	retrotransposon ty1-copia subclass	710	1,24E-07	58.3%
Pe75Q3E9F	retrotransposon protein	758	1,31E-07	50.0%
Pe69Q4A10R	rna polymerase sigma-70 factor	544	1,43E-07	78.0%
Pe75Q4H10F	hypothetical protein TRIUR3_21348	807	1,45E-07	74.0%
Pe33Q2E6R	multiple c2 and transmembrane domain-containing protein	456	1,53E-07	65.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe141Q4C5R	predicted protein	900	1,54E-07	85.0%
Pe185Q3C9F	retrotransposon unclassified	583	1,55E-07	48.45%
Pe216Q4C11R	resistance protein mg13	1011	1,66E-07	46.1%
Pe185Q1D7F	multiple c2 and transmembrane domain-containing protein	780	1,75E-07	55.8%
Pe212Q4A4F	retrotransposon gag protein	890	1,75E-07	53.29%
Pe23Q3E10Ft2	athila retroelement orf1 protein	629	1,77E-07	61.55%
Pe33Q2E1R	multiple c2 and transmembrane domain-containing protein	528	1,80E-07	55.0%
Pe69Q3H8F	uncharacterized loc101222732	637	1,90E-07	92.7%
Pe216Q1C10R	sister chromatid cohesion 1 protein 1-like	997	2,10E-07	79.8%
Pe23Q3F2R	probable xyloglucan endotransglucosylase hydrolase protein 30-like	209	2,12E-07	69.67%
Pe23Q3D7Ft2	unnamed protein product	825	2,12E-07	60.0%
Pe185Q1H6F	hypothetical protein PRUPE_ppa020976mg	836	2,16E-07	57.0%
Pe212Q3D4F	athila retroelement orf1 protein	564	2,21E-07	50.9%
Pe216Q2E6F	retrotransposon ty3-gypsy subclass	715	2,23E-07	44.55%
Pe164Q1D12R	endoplasmic reticulum	797	2,53E-07	85.68%
Pe75Q2D9F	retrotransposon gag protein	650	2,57E-07	54.0%
Pe173Q4F11R	hypothetical protein	720	2,59E-07	53.0%
Pe69Q2D8R	chloroplast biotin carboxylase	803	2,74E-07	78.25%
Pe84Q2F4F	myb domain protein 4r1	583	2,74E-07	71.75%
Pe164Q4D6F	rna-directed dna polymerase (reverse transcriptase)-related family protein	642	2,78E-07	41.95%
Pe93Q1A12R	retrotransposon ty1-copia subclass	751	3,31E-07	56.2%
Pe141Q3B9R	retrotransposon protein	626	3,58E-07	70.0%
Pe185Q3C5F	retrotransposon gag protein	639	3,65E-07	63.67%
Pe101Q4B11R	hypothetical protein PRUPE_ppb020037mg	716	4,47E-07	52.17%
Pe75Q2A2F	ac099402_4 22 kda kafirin cluster ty3-gypsy type	472	4,49E-07	48.25%
Pe101Q4E11F	hypothetical protein F775_52415	751	4,50E-07	62.0%
Pe33Q4G7R	retrotransposon ty3-gypsy subclass	581	4,59E-07	48.6%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe93Q2C2R	conserved hypothetical protein	612	5,24E-07	52.0%
Pe93Q3A5R	retrotransposon ty1-copia subclass	736	5,39E-07	59.0%
Pe185Q2H2R	predicted protein	715	5,39E-07	86.0%
Pe75Q1D7F	retrotransposon ty3-gypsy subclass	604	5,61E-07	55.8%
Pe75Q2G1F	integrase core domain containing protein	764	5,69E-07	64.0%
Pe214Q3G2R	Urease, putative	771	5,70E-07	66.0%
Pe23Q3F10Ft2	hypothetical protein ARALYDRAFT_918466	807	5,81E-07	90.0%
Pe185Q2D7F	hypothetical protein PRUPE_ppa014973mg, partial	691	5,92E-07	55.0%
Pe216Q1C9F	glyoxylate hydroxypyruvate reductase hpr3-like	726	6,03E-07	75.86%
Pe173Q3H4F	26s protease regulatory subunit 8	628	6,16E-07	75.9%
Pe93Q1C7R	retrotransposon protein	841	6,17E-07	57.5%
Pe69Q4D9R	benzyl alcohol o-benzoyltransferase-like	426	6,31E-07	75.35%
Pe101Q3G10F	dna-directed rna polymerase ii subunit rpb3-a-like	680	6,33E-07	70.68%
Pe93Q1F10R	chaperone protein dnaj	688	6,35E-07	76.29%
Pe141Q1H5F	retroelement pol poly	761	6,74E-07	65.1%
Pe69Q1D10R	retrotransposon gag protein	441	6,83E-07	75.5%
Pe173Q1A12F	protein with unknown function	825	6,86E-07	68.0%
Pe214Q1D11F	retrotransposon protein	582	6,86E-07	47.5%
Pe69Q1F7F	retrotransposon ty1-copia subclass	691	6,87E-07	60.43%
Pe43Q4E4F	integrase core	399	6,98E-07	69.33%
Pe173Q2F5F	wd repeat-containing protein 48-like	947	7,08E-07	47.55%
Pe101Q4C5R	hypothetical protein MTR_6g026690	735	7,10E-07	44.75%
Pe43Q4F1F	virion binding	549	7,15E-07	69.6%
Pe216Q4A1R	udp-glycosyltransferase 75d1-like	892	7,25E-07	72.05%
Pe75Q1C5F	pol protein	675	7,32E-07	64.88%
Pe55Q4E8R	miraculin-like protein	826	7,39E-07	55.0%
Pe185Q1D10F	ribonuclease h protein at1g65750-like	759	7,60E-07	48.33%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe75Q3G11R	gag-pol protein	866	9,48E-07	49.5%
Pe69Q3D9F	hypothetical protein PRUPE_ppa015000mg	872	1,01E-06	55.0%
Pe164Q2E10F	conserved hypothetical protein	491	1,01E-06	57.0%
Pe164Q1F5R	v-type proton atpase 16 kda proteolipid subunit-like	1089	1,03E-06	86.0%
Pe43Q4E10F	retrotransposon gag protein	446	1,03E-06	70.0%
Pe216Q2G2R	conserved oligomeric golgi complex subunit 3-like	851	1,05E-06	83.85%
Pe214Q4F6R	athila transposon protein	714	1,06E-06	54.75%
Pe164Q4H6R	retrotransposon ty1-copia subclass	640	1,08E-06	64.7%
Pe214Q2A1R	conserved hypothetical protein	834	1,09E-06	53.0%
Pe26Q4F2F	ankyrin repeat-containing protein at5g02620-like	460	1,12E-06	80.0%
Pe214Q1B10R	conserved hypothetical protein	718	1,15E-06	58.0%
Pe212Q1D1R	protein	780	1,16E-06	49.25%
Pe141Q2B8R	non-ltr retroelement reverse transcriptase	661	1,17E-06	45.71%
Pe164Q4A9F	serine threonine-protein kinase	823	1,19E-06	83.6%
Pe84Q3B6R	retrovirus-related polyproteins	507	1,20E-06	73.0%
Pe55Q4B9R	atp synthase cf0 subunit iii	734	1,23E-06	96.15%
Pe214Q2D1F	probable lrr receptor-like serine threonine-protein kinase at1g56130-like	779	1,25E-06	45.0%
Pe216Q2A9R	protein	869	1,29E-06	46.0%
Pe185Q1H1R	btb poz domain-containing protein fbl11-like	842	1,30E-06	69.33%
Pe33Q4E1R	retrotransposon ty1-copia subclass	494	1,32E-06	64.9%
Pe173Q2D8F	predicted protein	739	1,33E-06	54.0%
Pe173Q3C9F	hypothetical protein PRUPE_ppa019381mg	551	1,40E-06	50.0%
Pe75Q3G9R	retrotransposon ty1-copia subclass	717	1,43E-06	58.82%
Pe185Q1C5F	hypothetical protein PRUPE_ppa004403mg	950	1,46E-06	80.0%
Pe185Q2D8F	protein	701	1,47E-06	51.4%
Pe185Q1F5F	hypothetical protein CARUB_v10025819mg	719	1,48E-06	68.0%
Pe164Q2C9F	rna-directed dna polymerase (reverse transcriptase) polynucleotidyl ribonuclease h fold	709	1,52E-06	50.88%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe69Q2B3R	hypothetical protein PRUPE_ppa016093mg, partial	680	1,52E-06	75.0%
Pe93Q2G1F	hypothetical protein PRUPE_ppa011514mg	788	1,55E-06	76.0%
Pe69Q1A12R	retrotransposon ty3-gypsy subclass	380	1,59E-06	49.1%
Pe185Q2C12F	type-b response regulator	685	1,61E-06	49.5%
Pe69Q2C5R	predicted protein	818	1,63E-06	57.0%
Pe216Q1A10R	receptor protein	1112	1,68E-06	78.5%
Pe69Q2F1R	hypothetical protein VITISV_010121	751	1,70E-06	68.5%
Pe141Q2H3F	hypothetical protein PRUPE_ppa006410mg	629	1,71E-06	57.0%
Pe75Q4F9R	sodium calcium exchanger family protein	811	1,74E-06	91.2%
Pe141Q1B3R	hypothetical protein PRUPE_ppa016658mg, partial	373	1,80E-06	61.0%
Pe185Q1G12R	gag-pol protein	850	1,80E-06	49.15%
Pe75Q2C8R	rnix1_arath ame: full= ribonuclease h protein at1g65750	831	1,88E-06	58.2%
Pe141Q4F7R	polyprotein	480	1,88E-06	56.2%
Pe212Q2C12R	retrotransposon gag protein	777	1,91E-06	54.0%
Pe185Q4H9F	retroelement pol poly	821	1,91E-06	56.67%
Pe101Q3D1R	predicted protein	806	1,94E-06	67.0%
Pe75Q1C10F	PREDICTED: uncharacterized protein LOC101294544	737	2,04E-06	71.0%
Pe69Q2C5F	e3 ubiquitin-protein ligase pub23-like	646	2,08E-06	82.4%
Pe141Q4E10R	rrna methyltransferase ylbh	580	2,13E-06	83.0%
Pe173Q1F5F	wd repeat-containing protein 48-like	902	2,23E-06	75.5%
Pe214Q2C11F	hypothetical protein VITISV_003173	519	2,28E-06	57.95%
Pe173Q1F12R	hypothetical protein PRUPE_ppa023432mg, partial	728	2,29E-06	55.67%
Pe93Q2A5F	hypothetical protein MTR_5g043070	683	2,42E-06	58.0%
Pe216Q1D10R	ac007203_3 reverse-transcriptases	884	2,54E-06	57.0%
Pe55Q4A1R	PREDICTED: uncharacterized protein LOC101267053	733	2,74E-06	60.0%
Pe164Q3B4R	ethylene-insensitive 3f	978	2,78E-06	61.67%
Pe75Q4H1F	receptor for activated protein kinase	838	2,80E-06	97.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Continua)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe185Q3E4F	f-box and wd40 domain	702	2,80E-06	77.0%
Pe212Q4A12R	hypothetical protein PRUPE_ppa022673mg	576	2,87E-06	49.0%
Pe101Q3G3R	predicted protein	922	2,87E-06	60.0%
Pe69Q1F12R	retrotransposon ty1-copia subclass	350	3,24E-06	65.25%
Pe26Q1A3R	retrotransposon ty3-gypsy subclass	745	3,26E-06	48.5%
Pe185Q4F7F	polyprotein	779	3,38E-06	68.0%
Pe84Q2H10R	hypothetical protein PRUPE_ppb020528mg, partial	452	3,38E-06	67.0%
Pe23Q3A6R	vacuolar protein sorting-associated	170	3,56E-06	78.67%
Pe164Q3C4F	hypothetical protein PRUPE_ppb016975mg	764	3,58E-06	61.0%
Pe75Q1B1F	hypothetical protein VITISV_040272	658	3,58E-06	55.0%
Pe33Q4G4R	hypothetical protein PRUPE_ppa015000mg	435	3,67E-06	49.5%
Pe185Q3B6F	quinone-oxidoreductase-like protein	830	3,73E-06	91.0%
Pe216Q1E3R	peptidyl-trna hydrolase ptrhd1-like	744	3,77E-06	86.81%
Pe185Q1E1F	hypothetical protein PRUPE_ppb013312mg, partial	654	3,92E-06	56.5%
Pe214Q2E2R	1510387aretrotransposon del1-46	398	4,00E-06	50.56%
Pe214Q1C11F	1510387aretrotransposon del1-46	757	4,00E-06	56.0%
Pe55Q1B11F	atp-dependent rna helicase dhx29-like	647	4,03E-06	75.15%
Pe93Q2A11F	at1g67700 f12a21_30	763	4,09E-06	94.0%
Pe43Q4B1F	nucleobase-ascorbate transporter 6-like	660	4,32E-06	93.8%
Pe101Q3G1R	retrotransposon gag protein	605	4,36E-06	45.38%
Pe164Q3B3R	ac099402_4 22 kda kafirin cluster ty3-gypsy type	809	4,49E-06	54.0%
Pe214Q2G8F	glucose-6-phosphate 1-dehydrogenase chloroplastic-like	867	4,71E-06	74.38%
Pe93Q2F4F	hypothetical protein PRUPE_ppa021229mg	707	5,21E-06	69.67%
Pe185Q2F10R	nadh dehydrogenase subunit 2	580	5,54E-06	76.1%
Pe84Q1H11F	hypothetical protein PRUPE_ppa022673mg	591	6,00E-06	42.0%
Pe93Q3B4F	predicted protein	626	6,15E-06	83.0%
Pe23Q3G12Ft2	1510387aretrotransposon del1-46	429	6,19E-06	48.0%

Tabela A1 – Sequências que obtiveram similaridade com algum membro do banco de dados “NR” e seus respectivos *top-hits*. (Conclusão)

Nome da sequência	Descrição da sequência	Tamanho da sequência	E-Value	Similaridade
Pe75Q1C5R	retrotransposon ty1-copia subclass	901	6,55E-06	68.15%
Pe141Q4B6F	transferring glycosyl	788	7,28E-06	79.0%
Pe164Q1C12R	PREDICTED: uncharacterized protein LOC101218085	1046	7,85E-06	58.0%
Pe214Q4A10R	predicted protein	365	7,89E-06	56.0%
Pe216Q3D4R	protein binding	884	8,20E-06	82.88%
Pe75Q4F7F	ergosterol biosynthetic protein 28-like isoform 1	653	9,31E-06	78.0%
Pe216Q2B9R	retrotransposon ty1-copia subclass	824	9,32E-06	66.6%
Pe75Q4B6F	probable pre-mrna-splicing factor atp-dependent rna helicase-like	978	9,41E-06	96.3%
Pe185Q1A9F	disease resistance protein	735	9,42E-06	50.85%
Pe185Q2C10R	retrotransposon ty3-gypsy subclass	541	9,65E-06	53.85%
Pe93Q4E11R	predicted protein	570	9,82E-06	54.0%