University of São Paulo
"Luiz de Queiroz" College of Agriculture

Genomic studies in *Passiflora edulis* (Passifloraceae)

Zirlane Portugal da Costa

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2018

Zirlane Portugal da Costa
Bsc. in Agronomic Engineering

Genomic studies in *Passiflora edulis* (Passifloraceae)

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **MARIA LUCIA CARNEIRO VIEIRA**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2018

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

Costa, Zirlane Portugal da

Genomic studies in *Passiflora edulis* (Passifloraceae) / Zirlane Portugal da Costa. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2018.

92 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. *Passiflora edulis* 2. Maracujá 3. Genoma do maracujá 4. Malpighiales 5. Transposição 6. Biblioteca genômica 7. BAC I. Título.

DEDICATORY

A Deus,
pela vida, saúde e força
A minha família e amigos,
por todo o apoio e amor

ACKNOWLEDGEMENTS

É com muita satisfação e orgulho que concluo esta etapa da minha vida. Sinto muita gratidão por todo o aprendizado profissional e pessoal adquiridos nesse período e por todas as pessoas que fizeram parte dessa realização.

Primeiramente, agradeço a Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP) e ao Departamento de Genética, pela oportunidade e pelo suporte acadêmico.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelas bolsas e auxílio financeiro concedidos para a realização dos projetos.

A Prof. Dra. Maria Lucia Carneiro Vieira pela orientação em todo o período de pós-graduação. Muito obrigada por todo o apoio, confiança e oportunidades concedidas durante o desenvolvimento deste trabalho.

A todos os professores do Departamento de Genética e do Programa de Pós-Graduação em Genética e Melhoramento de Plantas por todos os conhecimentos compartilhados.

Aos funcionários do Departamento de Genética pelo convívio amigável e pela colaboração durante a execução do trabalho.

Ao Alessandro Varani pelo auxílio nas análises de bioinformática e pela disponibilidade para discussões.

A Prof. Dra. Marie-Anne Van Sluys e ao Geovani Tolfo pela disponibilidade e auxílio nas análises dos elementos transponíveis.

Ao Carlos Alberto de Oliveira pelo auxílio técnico durante a realização do trabalho e pela amizade.

A todos os amigos do Laboratório de Genética Molecular de Plantas Cultivadas, pelo auxílio nas análises e pela amizade e companheirismo.

E a minha família pelo apoio e amor.

SUMMARY

RESUMO

Estudos genômicos em *Passiflora edulis* (Passifloraceae)

Passiflora edulis, popularmente conhecido como maracujá azedo, é a espécie do gênero *Passiflora* mais cultivada, tendo importância econômica no Brasil tanto para a produção de suco industrializado quanto para o consumo da fruta fresca. Apesar da relevância da espécie, faltam pesquisas, principalmente nas áreas básicas. Para superar isso, nosso grupo tem realizado diversos estudos genéticos que incluem a estimativa dos níveis de polimorfismos moleculares, o estudo de locos quantitativos envolvidos no controle da produção e qualidade dos frutos e o mapeamento de genes de resistência à mancha bacteriana causada por *Xanthomonas axonopodis*. Além disso, para conhecer o genoma de *P. edulis*, foi construída uma biblioteca genômica inserida em BACs (82.944 clones, com cobertura de 6× do genoma haploide da espécie) que é mantida no CNRGV/INRA em Toulouse, França. Inicialmente, cerca de 10.000 BES (BAC-end sequences) foram sequenciadas, gerando aproximadamente 6,2 Mb de informação genômica, fornecendo a primeira visão do genoma de *P. edulis* sobre o conteúdo de genes e da porção repetitiva. Com o objetivo de obter informações mais completas, decidiu-se selecionar cerca de 100 insertos de BACs para o sequenciamento completo. A análise genômica realizada, especialmente a anotação estrutural e funcional dos genes e a identificação e caracterização dos elementos de transposição, constituíram os objetivos deste estudo. Os dados gerados pelo sequenciamento completo de 10.4 Mb de *P. edulis* representam uma fonte rica de informações que foram exploradas nesta tese de doutorado.

Palavras-chave: *Passiflora edulis*; Maracujá; Genoma do maracujá; Malpighiales; Transposição; Biblioteca genômica; BAC

ABSTRACT

Genomic studies in *Passiflora edulis* (Passifloraceae)

Passiflora edulis, popularly known in Brazil as sour passion fruit is the most widely cultivated species of the genus *Passiflora*, and is of economic importance in Brazil for industrial production of juice and fresh fruit for consumption. Despite its economic importance, little research has been conducted on this species, even on the most basic aspects. To fill in this gap, our group conducted various genetic studies, including estimating levels of molecular polymorphism, studying quantitative loci that control fruit yield and quality, and mapping genes conferring resistance to bacterial spot disease caused by *Xanthomonas axonopodis*. In addition, to enhance our knowledge of the *P. edulis* genome, a genomic library inserted into bacterial artificial chromosomes (BAC) has been constructed (82,944 clones, with coverage of 6× the species' haploid genome). The library is kept at the French Plant Genomic Resources Center (CNRGV/INRA) in Toulouse. Initially, some 10,000 BES (BAC-end sequences) were sequenced, generating approximately 6.2 Mb of genomic information and providing an initial overview of *P. edulis* genome in terms of gene content and repeat sequences. With the aim of obtaining more comprehensive information, it was decided to select around 100 BAC inserts for complete sequencing. The aim of this study was to carry out genomic analysis in order to elucidate the structural and functional annotation of the genes, and identify and characterize transposable elements. The data generated by fully sequencing 10.4 Mb of *P. edulis* provide a rich source of information which has been taken full advantage of in this doctoral thesis.

Keywords: *Passiflora edulis*; Passion fruit; Passion fruit genome; Malpighiales; Transposition; Genomic library; BAC

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The genus *Passiflora* is the largest in the *Passifloraceae* family, including some 600 species that are widely distributed in tropical and subtropical regions of the Neotropics, especially in South and Central America; about 22 species occur in Southeast Asia, Australia and the Pacific Islands (Ulmer and Macdougal, 2004). In Brazil, 142 species are found, of which 83 are endemic (see Bernacci et al., 2015).

Passiflora edulis, popularly known as passion fruit, is the most cultivated species, representing ~95% of the production in the last decades (Meletti et al., 2010). In Brazil, it has gained importance for both juice production and *in natura* consumption. The country is the world's largest producer and consumer, with a cultivation area of 56,000 ha producing about 823,000 tons per year (Agrianual, 2017). Furthermore, an additional value lies in the production of industrialized passion fruit juice to be exported and used as an essential exotic ingredient in blends. The main destinations are the European countries (see Meletti et al., 2010).

Despite *P. edulis* economic importance, little research has been conducted on this species, even on the most basic aspects. Genetic (Carneiro et al., 2002; Moraes et al., 2005; Oliveira et al., 2008) and molecular-based studies have been carried out in our laboratory (Munhoz et al., 2015; Santos et al., 2014) in order to satisfy the needs of a wide range of breeders to boost passion fruit crop production and fruit quality.

Our group has also established a linkage map for mapping genes involved in the response to bacterial spot disease caused by *Xanthomonas axonopodis* (*Xap*), one of the limiting diseases affecting the orchards (Lopes et al., 2006). More recently, Munhoz et al. (2015) characterized a set of transcripts differentially expressed in response to *Xap* inoculation. This analysis constituted the first information on the transcriptional reprogramming during passion fruit-*Xap* interaction, and generated sequence data that became available for subsequent studies. For instance, Costa et al. (2017) developed putative functional SSR (Simple Sequence Repeats) and SNP (Single Nucleotide Polymorphism) markers from the transcript sequences identified in Munhoz et al. (2015), including the SNP found in the lipoxygenase-2 gene, which encodes the most differentially expressed enzyme in response to the bacterium.

With the aim of enhancing our knowledge of the *P. edulis* genome, a genomic library inserted into bacterial artificial chromosomes (BAC) has been constructed (82,944 clones,

with coverage of 6× the species' haploid genome). A BAC genomic library consists of thousands of DNA fragments that represent the entire genome of a genotype. The *P. edulis* genomic library, denoted Ped-B-Flav, is kept at the French Plant Genomic Resources Center (CNRVG/INRA) in Toulouse. It contains 82,944 clones with coverage of 6× the species' haploid genome, according to the estimative of genome size (DNA 1C value= 1,230 Mb) reported by Yotoko et al. (2011)

Later on, Santos et al. (2014) sequenced 9,698 BES (BAC-end sequences), generating approximately 6.2 Mb of genomic information about the species. The size of the sequences ranged from 100 to 1,255 bp, with an average size of 587 bp. Analysis of these sequences provided the first overview of the genome of *P. edulis*. It was found that 19.6% of the sequences were composed of repetitive elements, of which 94.4% are transposable elements, and 9.6% had similarity to plant genes. It was possible to attribute 940 ontological terms to half of the sequences that had similarity to plant genes. In addition, 610 SSRs were identified, been potential sequences for the development of microsatellite markers. The GC content of the genome was estimated to be 41.9%. In addition, some BES were compared with the genome of *Arabidopsis thaliana*, *Vitis vinifera* and *Populus trichocarpa*, identifying microsyntenic regions.

With the aim of obtaining more comprehensive information, it was decided to select around 100 BAC inserts for complete sequencing. Then, in this study we carried out genomic analysis in order to elucidate the structural and functional annotation of the genes, and identify and characterize transposable elements. The data generated by fully sequencing 10.4 Mb of *P. edulis* provide a rich source of information which has been taken full advantage of in this doctoral thesis.

References

Agrianual (2017). *Anuário da agricultura brasileira*. 22 ed. , ed. F. C. & Agroinformativos São Paulo.

Bernacci, L. C. ., Cervi, A. C. ., Milward-de-Azevedo, M. A. ., Nunes, T. S. ., Imig, D. C. ., and Mezzonato, A. . (2015). Passifloraceae. *List. espécies da flora do Bras.* Available at: http://reflora.jbrj.gov.br/jabot/floradobrasil/FB182 [Accessed November 15, 2017].

Carneiro, M. S., Camargo, L. E. A., Coelho, A. S. G., Vencovsky, R., Rui, P. L. J., Stenzel, N. M. C., et al. (2002). RAPD-based genetic linkage maps of yellow passion fruit (*Passiflora edulis* Sims. f. flavicarpa Deg.). *Genome* 45, 670–678. doi:10.1139/g02-035.

Costa, Z. P., Munhoz, C. de F., and Vieira, M. L. C. (2017). Report on the development of putative functional SSR and SNP markers in passion fruits. *BMC Res. Notes* 10, 445. doi:10.1186/s13104-017-2771-x.

Lopes, R., Lopes, M. T. G., Carneiro, M. S., Matta, F. D. P., Camargo, L. E. A., and Vieira, M. L. C. (2006). Linkage and mapping of resistance genes to *Xanthomonas axonopodis* pv. passiflorae in yellow passion fruit. *Genome* 49, 17–29. doi:10.1139/G05-081.

Meletti, L. M. M., Oliveira, J. C., and Ruggiero, C. (2010). Maracujá. *Série Frutas Nativ.* 6.

Moraes, M. C., Gerald, I. O., Matta, F. P., and Vieira, M. L. C. (2005). Genetic and phenotypic parameter estimates for yield and fruit quality traits from a single wide cross in yellow passion fruit. *HortScience* 40, 1978–1981.

Munhoz, C. F., Santos, A. A., Arenhart, R. A., Santini, L., Monteiro-Vitorello, C. B., and Vieira, M. L. C. (2015). Analysis of plant gene expression during passion fruit-*Xanthomonas axonopodis* interaction implicates lipoxygenase 2 in host defence. *Ann. Appl. Biol.* 167, 135–155. doi:10.1111/aab.12215.

Oliveira, E. J., Vieira, M. L. C., Garcia, A. A. F., Munhoz, C. F., Margarido, G. R. A., Consoli, L., et al. (2008). An integrated molecular map of yellow passion fruit based on simultaneous maximum-likelihood estimation of linkage and linkage phases. *J. Am. Soc. Hortic. Sci.* 133, 35–41. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-37849007543&partnerID=40&md5=92f8347ed25bb95a14157a5b4912095d.

Santos, A., Penha, H., Bellec, A., Munhoz, C. De, Pedrosa-Harand, A., Bergès, H., et al. (2014). Begin at the beginning: A BAC-end view of the passion fruit (*Passiflora*) genome. *BMC Genomics* 15, 816. doi:10.1186/1471-2164-15-816.

Ulmer, T., and MacDougal, J. M. (2004). *Passiflora*: passionflowers of the world. *Timber Press*, 430p.

Yotoko, K. S. C., Dornelas, M. C., Togni, P. D., Fonseca, T. C., Salzano, F. M., Bonatto, S. L., et al. (2011). Does variation in genome sizes reflect adaptive or neutral processes? New clues from *Passiflora*. *PLoS One* 6, e18212. doi:10.1371/journal.pone.0018212.

2. A GENE-RICH FRACTION ANALYSIS OF THE *PASSIFLORA EDULIS* GENOME

ABSTRACT

Passiflora edulis is the most widely cultivated species of passionflowers. It is cropped mainly for industrialized juice production and fresh fruit consumption. Climatic conditions for natural occurrence of the species range from cool subtropical (purple variety) to warm tropical (yellow variety). Despite its commercial importance, little is known about the genome structure of *P. edulis*. To fill in this gap in our knowledge, we have built a large-insert genomic library in collaboration with the French Plant Genomic Resources Center (https://cnrgv.toulouse.inra.fr/Library/Passiflora). The library has approximately 83,000 clones and cover 6 × the haploid genome of *P. edulis*. After providing initial insights into the *P. edulis* genome using BAC-end sequence data as a major resource, we are now continuing to study this genome using the SMRT sequencing platform, and have completely sequenced over 100 BAC-inserts. A total of 10.4 Mb of sequencing data were assembled from long sequence reads, and structural sequence annotation resulted in the prediction of 1,883 genes, with 1,502 showing significant levels of similarity with plant proteins. Gene sequences represented 44% of all data. The GC content of this data was 41.09%, in agreement with previously reported. Functional annotation resulted in 3,178 ontology terms assigned to 1,191 genes, related to several processes. The search for microsatellites resulted in the identification of 11,020 and 1,762 SSRs in all sequence data and CDS respectively. All classes of microsatellite repeats were found, with trinucleotides been the most frequent in both data. Our study not only provides a first accurate gene set for *P. edulis*, but also opens the way for new studies on the evolutionary issues in *Passiflora* genomes.

Keywords: Plant genome; *Passiflora*; SMRT sequencing; Gene-rich fraction; Malpighiales

2.1. Introduction

The Passifloraceae family belongs to the Malpighiales order and is a member of the Rosids clade, according to classical and molecular phylogenetic analysis. The family consists of some 700 species, classified in 16 genera. The majority of species belong to the genus *Passiflora* (~530 species), popularly known as passion fruits (Ulmer and MacDougal, 2004). This genus is widely distributed in tropical and subtropical regions of the Neotropics. Approximately 150 species are native to Brazil, which is acknowledged to be an important centre of diversity (Bernacci et al., 2015).

Among the American tropical species of *Passiflora*, some 60 fruit-bearing species are marketed for human consumption. Moreover, several species and hybrids have been produced for ornamental purposes (see www.passiflora.it; Abreu et al., 2009), and pharmacologists have found that passion fruit vines contain bioactive compounds that are used in traditional folk medicines as anxiolytics and antispasmodics (Deng et al., 2010). *Passiflora edulis* is the major species of passionflowers grown for fresh fruit consumption and

juice production in climates ranging from cool subtropical (purple variety) to warm tropical (yellow variety). Species grown particularly in Brazil include *P. edulis* (sour passion fruit) and *P. alata* (sweet passion fruit). Because of the quality of its fruit and yield for processing into commercial juices, *P. edulis* is grown in 90% of the commercial orchards. The most recent agricultural production survey showed that 58,089 hectares were planted with passion fruits, yielding 838,444 tons per year (IBGE, 2015).

P. edulis is a diploid (*2n*= 18) (Cuco et al., 2005), self-incompatible species (Madureira et al., 2012; Suassuna et al., 2003), with perfect, insect-pollinated flowers. Over the last two decades, our research group has carried out studies for estimating the genetic parameters of experimental populations (Moraes et al., 2005), as well as constructing genetic maps (Carneiro et al., 2002; Oliveira et al., 2008) and mapping quantitative loci associated with the response to *Xanthomonas axonopodis* infection (Lopes et al., 2006). Munhoz et al. (2015) were able to determine which gene expression patterns were significantly modulated during the *P. edulis-X. axonopodis* interaction.

Despite its commercial success, little is known about the genome structure of *P. edulis*. The genome size has been estimated at ~1,230 Mb (1C DNA content= 1.258 pg by flow cytometric analysis) (Yotoko et al., 2011). To fill in this gap in our knowledge, we have built a large-insert genomic BAC (Bacterial Artificial Chromosomes) library denoted Ped-B-Flav (https://cnrgv.toulouse.inra.fr/library/genomic_resource/Ped-B-Flav) of approximately 83,000 clones, which is kept at the National Centre for Plant Genomic Resources (CNRGV: cnrgv.toulouse.inra.fr) at INRA in Toulouse, France. In addition, we were able to provide initial insights into the *P. edulis* genome using BAC-end sequence (BES) data as a major resource (Santos et al., 2014), and describe the structural organization of the plant's chloroplast genome, which differs from that of various Malpighiales species due to rearrangement events (Cauz-Santos et al., 2017).

Although based on small-sized sequences, we were able to map BAC-end sequences to intervals of sequenced related genomes (Huddleston et al., 2014) and identify collinear microsyntenic regions as a preliminary step towards selecting clones for full sequencing, which can be done with high accuracy using the single-molecule real-time (SMRT) sequencing (Pacific Biosciences). This method produces long, unbiased sequences that, in turn, facilitate subsequent assembly (VanBuren et al., 2015), a critical step in plants due to the high proportion of repetitive sequences throughout their genomes (Mayer et al., 2012).

Most of the projects aimed at obtaining a draft or a complete plant genome were performed using large-insert based sequencing methods (Buyyarapu et al., 2013; Li et al.,

2015) to allow estimation of the number of genes, and abundance of transposable elements (TEs) and microsatellites. In the functional part of the genome in particular, the annotation of large-inserts can provide an arsenal of biological information to facilitate comparison against databases and determine the distribution of BAC inserts relative to related genomes in order to examine the degree of synteny between them and gain insights into evolutionary relationships (de Setta et al., 2014; Ming et al., 2015).

In this scenario, we are continuing to study the *P. edulis* genome based on the large-insert BAC library and using the SMRT sequencing platform to completely sequence over 100 inserts of BAC clones. These clones were pre-selected based on BES microsynteny results and probes homologous to transcripts from a subtractive library of *P. edulis* in response to *Xanthomonas axonopodis* infection, which allowed us to obtain a gene-rich fraction of this genome. The repetitive content, predicted genes, and coding sequences were annotated. It is worth noting that part of the results derived from this study was recently published in Munhoz et al. (2018).

2.2. Material and Methods

2.2.1. BAC Selection and DNA preparation

BAC clones were selected from the findings of Santos et al. (2014), which provide an initial overview of the *P. edulis* genome using BAC-end sequence (BES) data as a major resource. The results of comparative mapping between *P. edulis'* BES and the reference genomes of *Arabidopsis thaliana, Populus trichocarpa* and *Vitis vinifera* were also used to choose BAC clones for sequencing. In addition, based on BES functional annotation results, the BAC-inserts with coding sequences (CDS) in one or both BESs were also selected.

A second selection procedure was performed after screening the genomic library using the probes homologous to *P. edulis* transcripts described in Munhoz et al. (2015). Briefly, the authors used suppression subtractive hybridization to construct two cDNA libraries enriched for transcripts induced and repressed by *Xanthomonas axonopodis*, respectively, 24 h after inoculation with a highly virulent bacterial strain.

The homologous probes were prepared via PCR, using as a template the genomic DNA from 'IAPAR-123', the accession used to construct the Ped-B-Flav BAC library. Specific primers were used to generate a single amplicon (200 to 600 bp in size) for each

probe gene sequence. The ‘DecaLabel DNA Labeling Kit’ (Fermentas) was used for radiolabeling the probes. The amplification products were then purified with ‘Illustra ProbeQuant™ G-50 Micro Columns’ (GE Healthcare). The library was previously gridded onto macroarrays in which 41,472 clones were double-spotted on each 22 × 22 cm nylon membrane. These membranes were submerged in a bath of SSC (Saline-Sodium Citrate) solution (6×, 17 min., 50 °C); incubated overnight (68 °C) in hybridization buffer [6× SSC, 5× Denhardt's Solution, 0.5 % (w/v) SDS (Sodium Dodecyl Sulfate)]; hybridized with denatured probes (10 min, 95 °C; 1 min., cooled on ice); and washed twice in buffer 1 [2× SSC, 0.1 % (w/v) SDS] (15 min., 50 °C) and buffer 2 [0.5× SSC, 0.1 % (w/v) SDS] (30 min., 50 °C). Next, the hybridized membranes were placed in a film cassette for 24 h.; radioactive signals were detected using a PhosphorImager™ and Storm 820 scanner (Amersham Biosciences) and analyzed using HDFR3 software, to identify the positive clones. Each positive clone was individually validated by PCR.

In order to estimate insert sizes, the preserved cultures were scraped and a positive single colony of each BAC grown in a 96-well plate (overnight, 37 °C) containing 1200 µL of LB medium with chloramphenicol (12.5 µg/mL) and glycerol (6 %). DNAs were then isolated using a NucleoSpin® 96 Flash (Macherey-Nagel) BAC DNA purification kit, digested with 5 U of FastDigest™ *Not*I enzyme (Fermentas) and size-fractioned by PFGE (6 $V.cm^{-1}$, 5 to 15 s switch time, 16 h run time, 12.5 °C) in a Chef Mapper XA Chiller System 220 V (BioRad), followed by ethidium bromide staining and visualization. The insert sizes were determined by comparison with PFGE (pulsed-field gel electrophoresis) standard size markers.

To prepare the DNA for sequencing, 1 µl of the above cultures was allowed to regrow in 20 mL of LB medium (plus 12.5 µg/mL chloramphenicol at 37 °C overnight) under shaking (250 rpm). The cultures were then mixed in pools, at a maximum of 20 clones per pool. DNA extraction was performed using the Nucleobond Xtra Midi Plus kit (Macherey-Nagel) according to the manufacturer’s instructions.

2.2.2. DNA Sequencing and Assembly From Long Sequence Reads

Approximately 5 µg of each pool was used for the construction of a SMRT library based on the standard Pacific Biosciences (San Francisco, CA, USA) preparation protocol for 10-kb libraries. Each pool was sequenced in one SMRT Cell using P6 polymerase in

combination with C4 chemistry, following the manufacturer's standard operating procedures and using the PacBio RS II long-read sequencer.

Reads were assembled by a hierarchical genome assembly process (HGAP workflow) (Chin et al., 2013), and using the v2.2.0 SMRT® analysis software suite for HGAP implementation. Reads were first aligned by the PacBio long-read aligner or BLASR (Chaisson and Tesler, 2012) against the complete genome of *Escherichia coli*, strain K12, substrain DH10B (GenBank: CP000948.1). The *E. coli* reads, as well as low quality reads (minimum read length of 500 bp and minimum read quality of 0.80) were removed from the data set. Filtered reads were then preassembled to yield long, highly accurate sequences. To perform this step, the smallest and the longest reads were separated from each other to correct errors by mapping single-pass reads to the longest reads (seed reads), which represent the longest portion of the read length distribution. Next, sequences were filtered against vector (BAC) sequences, and the Celera assembler used to assemble data and obtain draft assemblies. The last step was performed in order to significantly reduce the remaining indels and base substitution errors in the draft assembly. The Quiver algorithm was used for this purpose. This quality-aware consensus algorithm uses rich quality scores (Quality Value/QV scores) and QV is a per-base estimate of base accuracy. QV scores over 20 are from very good data with only 1% error probability. Finally, Quiver polishes the assembly for final consensus (Chin et al., 2013).

Once the refined assembly was obtained, each BAC-insert sequence was individualized by matching the end sequences to the pool of assembled sequences using BLAST. Read coverage was assessed by aligning the raw reads on the assembled sequences with BLASR.

2.2.3. Gene Prediction and Functional Annotation

Evidence-driven gene prediction was performed based on gene models of *Arabidopsis thaliana* and *Theobroma cacao* and using the following software: Augustus (Stanke et al., 2004), GlimmerHMM (Majoros et al., 2004), GeneMark.hmm (Borodovsky and Lomsadze, 2011), and SNAP (Korf, 2004). *Ab initio* gene finding was performed with the BRAKER pipeline (Hoff et al., 2016). Protein homology detection and potential intron resolution were detected by Exonerate software (Slater and Birney, 2005) against the annotated genomes of *Populus trichocarpa*, *Salix purpurea*, *Ricinus communis* and *Manihot*

esculenta, downloaded from the Phytozome website (Goodstein et al., 2012). These species are among the plant genomes with the highest number of top hits for *P. edulis* (Santos et al., 2014).

Additionally, a *P. edulis* RNA-seq library (see details below) was used to support gene model predictions. PASA (Haas et al., 2003) was used to produce alignment assemblies based on overlapping transcript alignments from *P. edulis* RNA-seq data. The results were combined by EVidence Modeler software (Haas et al., 2008), and PASA was used to update the EVidence Modeler consensus predictions, adding UTR annotations and models for alternatively spliced isoforms. Exon-intron boundaries were manually examined using GenomeView (Abeel et al., 2012) and adjusted where necessary.

RNA-seq reads (2× 100 bp; Illumina HiSeq 2000) were trimmed based on quality (Phred quality score > 20). Contaminants, remaining adapters, and sequences (< 50 bp) were removed using SeqyClean v1.9.9 (Zhbannikov et al., 2017). Total RNA-seq assembly was implemented by Trinity (Haas et al., 2013). In brief, RNA-seq reads were derived from three libraries (each replicated three times) of shoot apexes of juvenile, vegetative and reproductive adult plants of *P. edulis*, constructed with the aim of performing comparisons of these three developmental stages (Dornelas M.C. et al., unpublished data).

Functional annotation of the predicted gene sequences was performed using Blast2GO v3.2 tools (Conesa et al., 2005) for assigning ontological terms in accordance with BLASTX results (e-value cut-off of 1×10^{-6}). In addition, protein signature recognition was performed using the InterProScan tool (Jones et al., 2014).

2.2.4. Microsatellite prospection

Eukaryotic genomes contain a substantial portion of repetitive elements which are organized into three main classes: dispersed repeats (mostly transposable elements and retrotransposed genes), local repeats (tandem repeats and simple sequence repeats or microsatellites) and segmental duplications (duplicated genomic fragments) (Bao and Eddy, 2002). It is highly recommended to identify and mask repetitive regions before gene prediction. Otherwise, unmasked repeats can produce spurious BLAST alignments, resulting in false evidence for gene annotations (Yandell and Ence, 2012).

MISA (Aggarwal et al., 2007) was used to search for microsatellites based on microsatellite sequences with at least 10 nucleotides in the repeat for mono-, 5 for di -, and 3

for tri-, tetra-, penta- or hexanucleotides. Composite microsatellites were also identified. They are formed by multiple, adjacent, repetitive motifs. Hence, a microsatellite is considered composite if it has a maximum interruption of 10 bp between motifs (Oliveira et al., 2006; revisited by Vieira et al., 2016).

2.3. Results

2.3.1. BAC Selection, Sequencing and Assembly

Sixty-six BAC inserts were selected for complete sequencing based on our previous BAC-end sequencing results (Santos et al., 2014), and 46 were selected using probes homologous to transcripts of *P. edulis* (Munhoz et al., 2015) (Appendix A). Thus, in total, 112 BAC inserts from the *P. edulis* genomic library were sequenced. The sequencing process resulted in 571,565 high quality reads, ranging from 500 to 46,831 bp in length. Assemblies were between 24,316 and 142,456 bp in length, corresponding to their respective band sizes resolved by PFGE. The high quality of the long reads (QV> 47) and high coverage of the contigs (on average 278×) are indications of the reliability of our data (Appendix B), leading to the conclusion that all inserts were completely sequenced and assembled.

The sequencing method was of sufficient quality to provide a single contig per insert, with only two exceptions; in the assembly process, insert sequences Pe101K14 and Pe141H13 had overlapping regions that resulted in a single contig of 172,337 bp; similarly, Pe20N3 and Pe64C12 resulted in a single contig of 114,997 bp. In addition, of the 112 BAC insert sequences, three corresponded to organelle DNA, and therefore these sequences were not included. Thus, 107 sequences were subjected to annotation, totaling 10,401,671 bp (10.4 Mb) corresponding to approximately 1.0 % of the *P. edulis* genome. GC content across this genome fraction was 41.09%, and in the CDS 46.49%.

2.3.2. Gene Representativeness, Structure and Functional Annotation

Structural sequence annotation resulted in the prediction of 1,883 genes ranging from 153 to 24,687 bp in length, with an average of 2,448 bp. These gene sequences represented 44% of the total sequenced nucleotides, corresponding to 4,608,830 bp. Intergenic regions

covered from 0 (overlapped genes) to 92,497 bp, with a mean length of 3,184 bp. Between three and 36 predicted genes were identified per sequenced insert, with an average of 17.6 predicted genes per insert (Table 1). Taking into account the estimated size of the *P. edulis* genome (~1,230 Mb), the high number of genes identified herein (1,833) endorses the efficiency of our strategy in selecting BAC-inserts that were supposedly gene-rich.

Table 1. Gene content in a gene-rich fraction of the *Passiflora edulis* genome (~10.4 Mb) and structural annotation.

(continue)

BAC code	No. of predicted genes	Intronless genes	Exons per gene	Gene length variation (bp)	Average gene length (bp)	Intergenic spacer length variation (bp)	Average intergenic spacer length (bp)	CDS length (bp)	Average CDS length (bp)
Pe101K14	36	17	2 − 17	264 – 11,778	2,720	33 – 6,312	2,070	264 – 6,576	1,187
Pe185D11	36	12	2 − 12	201 – 4,778	1,548	16 – 9,730	1,802	201 – 1,725	689
Pe164B18	29	9	2 − 19	243 – 16,279	2,313	42 – 7,449	1,316	243 – 11,409	1,393
Pe214H11	29	4	2 − 39	799 – 13,956	3,857	194 – 5,728	1,134	174 – 4,572	1,636
Pe164D9	28	13	2 − 11	198 – 5,817	1,868	114 – 5,844	1,600	156 – 2,202	1,066
Pe186E19	28	4	2 − 14	770 – 7,450	2,651	11 – 13,501	1,559	210 – 2,307	1,098
Pe43L2	27	3	2 − 18	339 – 10,097	2,718	162 – 2,768	973	279 – 3,123	1,145
Pe86F9	27	13	2 − 5	201 – 20,501	1,622	147 – 12,507	2,776	201 – 1,740	607
Pe164K17	26	4	2 − 13	436 – 9,502	3,037	11 – 7,775	1,761	204 – 5,334	1,310
Pe215I8	26	5	2 − 18	312 – 8,238	3,007	230 – 13,338	2,168	180 – 3,501	1,253
Pe75K15	25	14	2 − 5	186 – 4,193	857	10 – 11,721	2,951	186 – 2,100	591
Pe84I14	25	6	2 − 12	345 – 8,118	3,014	69 – 4,352	936	198 – 4,275	1,295
Pe84M23	25	5	2 − 13	305 – 8,652	2,753	52 – 5,197	998	177 – 3,018	1,168
Pe93M2	25	5	2 − 16	399 – 7,069	2,274	135 – 11,933	2,170	192 – 2,961	1,109
Pe171P13	25	8	2 − 20	461 – 9,727	2,759	158 – 15,960	2,392	330 – 4,035	1,193
Pe207D11	25	12	2 − 17	213 – 6,756	1,896	5 – 20,551	2,838	213 – 2,730	897
Pe93N7	24	5	2 − 11	921 – 8,889	3,120	18 – 7,588	1,421	387 – 5,085	1,486
Pe108C16	24	8	2 − 14	234 – 6,553	1,892	34 – 9,113	2,209	234 – 3,252	974
Pe173B16	24	6	2 − 32	475 – 15,390	3,079	151 – 15,127	2,134	279 – 6,375	1,523
Pe185J16	24	4	2 − 21	447 – 8,773	2,432	201 – 6,924	2,083	237 – 2,367	1,035
Pe198H23	24	8	2 − 6	180 – 5,279	1,943	1 – 11,008	2,681	180 – 3,510	1,143
Pe212I1	24	5	2 − 35	234 – 12,694	2,715	53 – 15,133	2,607	234 – 3,567	1,080
Pe93J9	23	3	2 − 16	615 – 6,131	2,907	3 – 9,066	1,824	201 – 3,321	1,295
Pe135J12	23	6	2 − 15	162 – 9,543	2,714	81 – 8,758	1,868	162 – 4,433	1,260
Pe195F4	23	2	2 − 20	261 – 8,364	2,843	9 – 11,133	2,208	177 – 5,442	1,192
Pe74I6	22	9	2 − 39	204 – 17,655	3,407	146 – 6,191	1,764	204 – 4,374	1,164
Pe84M18	22	6	2 − 10	321 – 8,124	2,563	22 – 15,224	2,364	321 – 4,356	1,160
Pe101O4	22	6	2 − 19	624 – 9,702	2,678	315 – 10,499	2,782	300 – 2,235	884
Pe141J23	22	6	2 − 15	189 – 9,258	2,567	608 – 12,079	2,407	189 – 2,550	870
Pe201C11	22	11	2 − 17	195 – 5,452	1,865	288 – 17,891	4,128	195 – 2,634	822
Pe69G18	21	3	2 − 22	228 – 8,658	2,958	61 – 19,104	2,304	210 – 3,582	1,192
Pe69H24	21	2	2 − 14	335 – 6,461	2,752	445 – 5705	2,306	234 – 2,559	1,142

BAC code	No. of predicted genes	Intronless genes	Exons per gene	Gene length variation (bp)	Average gene length (bp)	Intergenic spacer length variation (bp)	Average intergenic spacer length (bp)	CDS length (bp)	Average CDS length (bp)
Pe93K19	21	3	2 − 12	792 – 10,373	3,523	196 – 6,322	1,422	387 – 4,629	1,593
Pe125I23	21	5	2 − 14	414 – 7,993	2,526	51 – 8,659	2,406	414 – 1,776	1,106
Pe164A12	21	7	2 − 11	384 – 7,964	2,354	26 – 7,406	1,675	228 – 4,503	1,050
Pe168B17	21	3	2 − 11	321 – 6,861	2,509	47 – 16,932	4,619	174 – 4,140	1,234
Pe214A18	21	7	2 − 11	243 – 6,314	1,944	237 – 27,586	3,916	243 – 2,184	924
Pe7M15	20	11	2 − 15	213 – 9,031	2,388	12 – 17,420	3,676	213 – 3,495	1,046
Pe28D11	20	16	2 − 4	189 – 2,430	780	22 – 28,073	5,567	189 – 1,410	558
Pe60G10	20	6	2 − 24	351 – 9,925	2,513	91 – 10,947	2,767	261 – 3,378	1,291
Pe65F7	20	8	2 − 14	306 – 7,081	1,973	12 – 25,539	2,702	213 – 3,252	844
Pe175N8	20	8	2 − 27	219 – 14,245	2,941	15 – 11,237	2,495	219 – 3,663	1,299
Pe214N19	20	9	2 − 13	234 – 5,913	1,594	37 – 15,598	3,485	189 – 2,470	788
Pe43D2	19	3	2 − 8	447 – 7,338	2,601	271 – 19,633	2,158	222 – 4,872	1,120
Pe51C2	19	5	2 − 16	357 – 8,889	3,603	493 – 6,756	2,110	357 – 5,088	1,520
Pe85B19	19	7	2 − 18	372 – 10,115	2,851	42 – 8,103	2,368	183 – 3,228	1,157
Pe101P7	19	3	2 − 20	234 – 8,484	3,742	16 – 2,340	963	234 – 2,712	1,247
Pe134H15	19	8	2 − 11	295 – 7,290	2,527	208 – 5,953	2,351	219 – 1,899	844
Pe216F3	19	2	2 − 37	393 – 14,151	3,198	241 – 3,160	914	393 – 8,943	1,626
Pe216F9	19	5	2 − 13	207 – 9,274	3,547	420 – 5,573	2,107	207 – 3,417	1,180
Pe20N3 + Pe64C12	18	5	2 − 12	441 – 6,941	2,557	266 – 10,519	2,009	276 – 2,364	1,223
Pe24G19	18	12	2 − 6	165 – 3,803	1,054	184 – 22,176	3,639	165 – 1,593	598
Pe69C7	18	7	2 − 22	210 – 8,505	3,745	132 – 18,029	2,165	210 – 4,164	1,450
Pe69O16	18	4	4 − 19	590 – 17,670	4,339	86 – 1,976	767	177 – 14,583	2,292
Pe212D7	18	7	2 − 36	171 – 21,131	2,654	415 – 20,035	4,436	171 – 9,330	1,229
Pe27H17	17	13	2 − 3	177 – 2,134	620	197 – 13,511	4,390	177 – 1,071	464
Pe85I9	17	5	2 − 12	207 – 8,578	2,908	334 – 20,210	2,892	207 – 1,956	1,107
Pe89E10	17	10	2 − 13	183 – 4,327	974	342 – 18,584	5,178	174 – 1,794	509
Pe101P13	17	4	2 − 21	666 –13,552	4,437	90 – 4,941	1,072	210 – 2,307	1,261
Pe209G15	17	3	2 − 14	219 – 8,353	3,108	118 – 17,105	2,754	219 – 3,084	1,416
Pe21O15	16	7	2 − 13	189 – 4,570	1,512	106 – 14,572	3,633	156 – 1,902	595
Pe63J18	16	10	2 − 5	441 – 6,941	2,750	266 – 10,519	2,054	213 – 3,429	970
Pe84K8	16	3	2 − 18	162 – 12,356	3,570	178 – 4,867	1,891	162 – 2,295	1,072
Pe93M4	16	10	2 − 7	216 – 3,063	972	15 – 37,508	4,704	216 – 1,998	640
Pe117C17	16	11	2 − 12	153 – 6,852	979	7 – 18,168	5,302	153 – 1,188	414
Pe138G17	16	10	2 − 10	178 – 6,113	1,395	40 – 13,394	4,513	178 – 2,934	731
Pe141K8	16	4	2 − 24	1,053 – 11,592	4,060	283 – 5,091	2,179	387 – 3,975	1,653
Pe216B22	16	1	2 − 15	1013 – 8,815	3,931	47 – 19,862	3,119	795 – 3,768	1,575
Pe216I5	16	6	4 − 16	201 – 5,929	3,296	462 – 4,563	1,373	201 – 2,862	1,458
Pe61E2	15	4	3 − 12	231 – 8,598	3,100	223 – 18,187	3,244	231 – 2,103	973
Pe99P16	15	9	2 − 33	249 – 15,022	2,441	501 – 9,387	2,582	216 – 4,605	908
Pe123N8	15	5	2 − 22	163 – 10,051	2,938	70 – 13,306	39,979	163 – 2,397	1,028
Pe3F10	14	4	2 − 14	652 – 6,552	2,471	90 – 4,389	1,557	285 – 3,252	1,080
Pe28E22	14	1	2 − 12	379 – 11,107	3,661	13 – 16,073	2,221	261 – 2,718	1,247

BAC code	No. of predicted genes	Intronless genes	Exons per gene	Gene length variation (bp)	Average gene length (bp)	Intergenic spacer length variation (bp)	Average intergenic spacer length (bp)	CDS length (bp)	Average CDS length (bp)
Pe34M7	14	6	2 − 4	225 – 1,298	652	82 – 39,701	6,611	192 – 1,026	459
Pe75F20	14	6	2 − 13	198 – 6,418	1,859	182 – 21,979	5,567	198 – 1,842	541
Pe85H4	14	1	2 − 51	489 – 22,481	3,938	178 – 17,578	2,764	300 – 5,706	1,546
Pe85J23	14	2	2 − 15	760 – 9,631	3,222	362 – 9,609	2,597	492 – 3,066	1,087
Pe101H15	14	10	2 − 5	225 – 24,687	2,257	122 – 15,195	6,521	255 – 1,008	524
Pe69F22	13	0	2 − 14	438 – 6,597	3,680	196 – 26,118	4,433	207 – 1,710	1,029
Pe75A21	13	8	3 − 10	162 – 5,730	1,569	10 – 15,569	4,038	162 – 2,076	630
Pe84M6	13	8	2 − 13	185 – 3,026	1,059	262 – 16,455	4,686	185 – 1,578	792
Pe86H7	13	7	2 − 3	213 – 4,497	1,429	31 – 28,575	6,964	213 – 3,459	875
Pe34H9	12	3	2 − 14	258 – 6,285	1,961	49 – 44,532	6,154	258 – 1,623	748
Pe213C9	12	8	2 − 5	327 – 3,599	1,246	213 – 31,653	7,880	234 – 2,016	749
Pe71E3	11	2	3 − 9	207 – 3,727	2,185	362 – 31,489	6,138	207 – 1,698	1,047
Pe93A7	11	8	2 − 4	162 – 1,374	582	18 – 25,472	7,604	162 – 759	373
Pe93F5	11	2	2 − 8	192 – 11,041	2,745	5 – 24,167	7,152	192 – 1722	707
Pe93O18	11	3	2 − 11	387 – 7,643	2,714	596 – 49,482	9,337	387 – 1,632	1,080
Pe101F21	11	7	2 − 7	243 – 4,835	947	58 – 27,172	8,438	198 – 1,806	534
Pe141B12	11	4	2 − 15	288 – 6,769	2,412	251 – 24,611	5,214	282 – 3,417	1,142
Pe75D12	10	6	2 − 5	219 – 3,255	778	109 – 39,945	8,052	216 – 1,224	456
Pe75N15	10	8	2	204 – 714	444	78 – 32,243	7,353	204 – 714	402
Pe9E4	9	4	2 − 14	342 – 6,100	2,099	654 – 13,925	6,177	342 – 2,898	1,171
Pe15E1	9	4	2 − 13	270 – 2,896	1,153	700 – 33,021	9,014	270 – 1,578	714
Pe20E10	9	4	2 − 2	159 – 1,578	605	278 – 35,112	9,958	159 – 1,578	496
Pe212M5	9	4	2 − 6	267 – 3,170	1,020	851 – 10,468	4,056	267 – 1,566	727
Pe103M2	8	2	2 − 17	222 – 12,656	3,122	418 – 32,453	6,547	222 – 2,010	807
Pe28I20	7	5	2 − 2	237 – 881	467	67 – 30,516	11,363	237 – 762	437
Pe75F13	7	4	2 − 3	180 – 1,636	654	16,743 – 92,497	58,535	180 – 1,245	519
Pe85O9	7	1	2 − 8	441 – 3,324	2,079	515 – 6,447	1,784	441 – 1,329	765
Pe1M17	6	1	2 − 4	312 – 2,473	1,099	256 – 10,848	5,311	312 – 1,404	784
Pe212J12	6	2	2 − 15	405 – 4,357	1,377	81 – 12,708	3,133	381 – 1,644	692
Pe216B2	6	1	2 − 24	218 – 15,969	5,097	830 – 4,575	2,306	218 – 3,819	1,605
Pe113A7	5	3	2	156 – 2254	1,206	3472 – 26,026	13,464	156 – 681	503
Pe1K19	3	0	2 − 9	958 – 4,737	3,111	287 – 37,487	18,877	840 – 897	869
Pe33M2	3	2	3	210 – 2,037	824	4,001 – 69,199	36,600	210 – 697	377

Approximately one third of the genes (631) had no introns. The remaining (1,252) had up to 50 introns. A total of 6,122 introns (ranging from 26 to 7,869 bp in length) and 8,005 exons (ranging from 3 to 6,249 bp) were recognized. CDS ranged from 153 to 14,583 bp in length, totaling 1,985,892 bp, with a mean of 1,054 bp. Sixty-one were insert end sequences and therefore incomplete gene sequences. According to the RNA-seq read

alignment results, 252 genes exhibited more than one transcript (Table 1), including glutamine synthetase leaf enzyme, chloroplastic (6 transcripts), ultraviolet-B receptor UVR8, a protein responsive to UV-B (5), the auxin response factor (2), an abscisic acid insensitive protein (2) and an ethylene receptor protein (2).

Of the 1,883 predicted genes, 1,502 showed significant levels of similarity (e-values $< 1\times 10^{-6}$) to plant proteins according to the Blast2GO results. The top hits for this large fraction of genes (~80%) were from *Jatropha curcas* (298), *Populus trichocarpa* (275), *Populus euphratica* (232) and *Ricinus communis* (212). These results were expected, since among all available plant genomes, these species are phylogenetically close to *P. edulis*, and all belong to the Malpighiales order. Functional annotation resulted in 3,178 ontological terms assigned to 1,191 genes. These GO terms were related to several processes, and are usually classified into three broad categories (known as level 1): biological process, molecular function and cellular component. The distribution of level 2 terms within each of these major categories is shown in Figure 1 and matches the results of BES annotation (Santos et al., 2014).

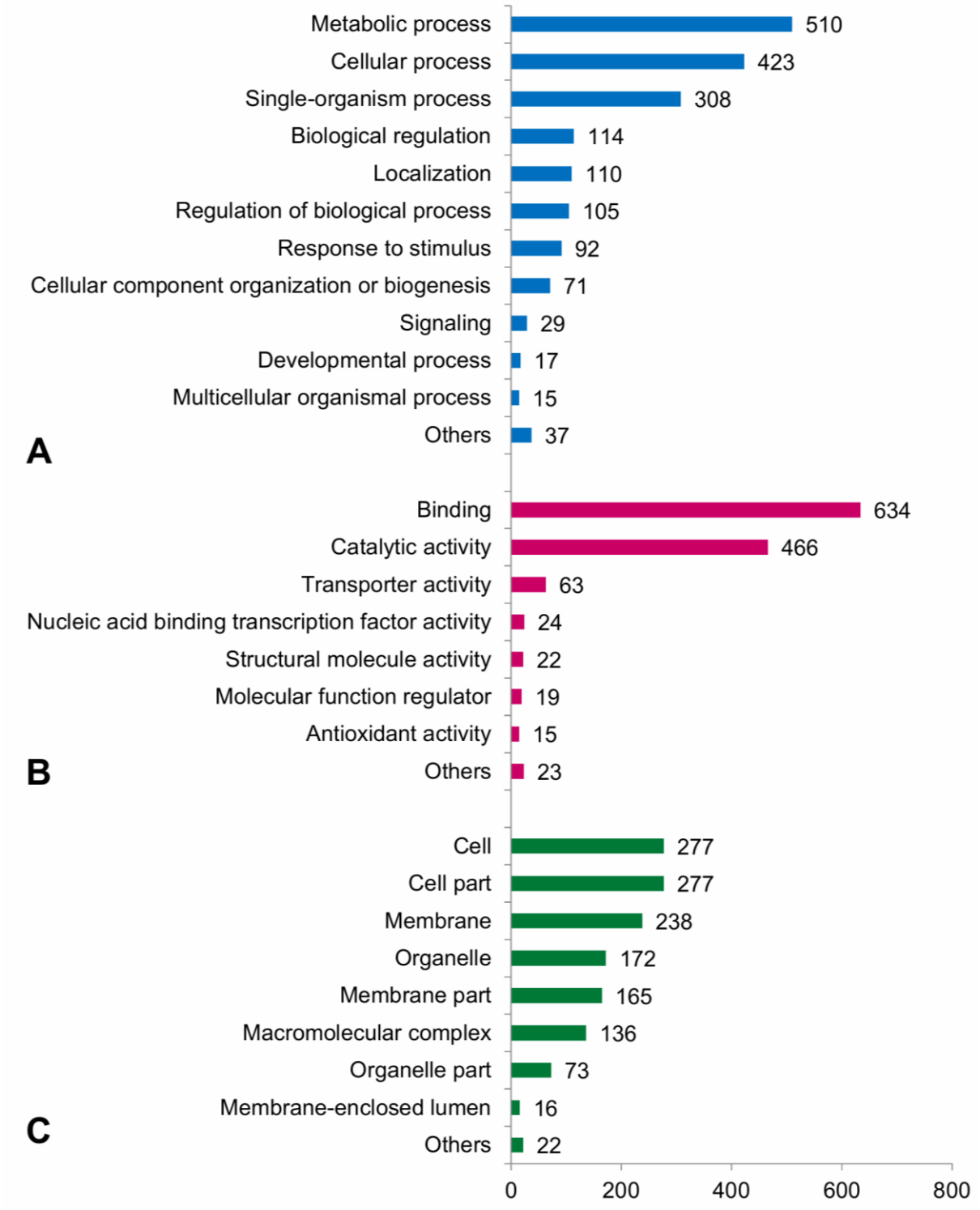

Figure 1. Distribution of GO annotations assigned to gene products in ontological categories: (A) Biological process, (B) Molecular function and (C) Cellular component. GO annotations were extracted from all sequences (10.4 Mb) of *Passiflora edulis*.

Regarding the 46 regions selected using probes homologous to transcripts induced and repressed by *X. axonopodis* infection, none of the functional categories related to plant defense were found to be overrepresented. However, protein signatures related to plant

immunity and defense functions were identified. The serine/threonine-protein kinase active site (32 genes), and the leucine-rich repeat domain, L domain-like (27 genes) were among the most represented signatures (Table 2). In total, automated searches for protein signatures recognized 1,383 signatures in 1,488 genes of *P. edulis*: 783 domains, 453 protein families, 125 sites and 22 replicates (Table 2). Most of these signatures (769) were taken from the Pfam database (Finn et al., 2014), and the remainder from SuperFamily (239) (Gough et al., 2001) and Smart (223) (Letunic and Bork, 2017).

Table 2. Most frequent protein signatures (≥ 10) recognized in 1,488 genes of *Passiflora edulis* according to InterProScan results.

(continue)

InterProScan ID	Number of genes
IPR005162 [Domain]: Retrotransposon gag domain	58
IPR011009 [Domain]: Protein kinase-like domain	51
IPR000719 [Domain]: Protein kinase domain	49
IPR027417 [Domain]: P-loop containing nucleoside triphosphate hydrolase	39
IPR001878 [Domain]: Zinc finger, CCHC-type	36
IPR011990 [Domain]: Tetratricopeptide-like helical domain	34
IPR008271 [Active_Site]: Serine/threonine-protein kinase, active site	32
IPR013083 [Domain]: Zinc finger, RING/FYVE/PHD-type	31
IPR029058 [Domain]: Alpha/Beta hydrolase fold	30
IPR017441 [Binding_Site]: Protein kinase, ATP binding site	30
IPR016024 [Domain]: Armadillo-type fold	27
IPR032675 [Domain]: Leucine-rich repeat domain, L domain-like	27
IPR013320 [Domain]: Concanavalin A-like lectin/glucanase domain	25
IPR009057 [Domain]: Homeodomain-like	25
IPR002885 [Repeat]: Pentatricopeptide repeat	25
IPR011989 [Domain]: Armadillo-like helical	22
IPR016040 [Domain]: NAD(P)-binding domain	19
IPR013242 [Domain]: Retroviral aspartyl protease	19
IPR001841 [Domain]: Zinc finger, RING-type	19
IPR017986 [Domain]: WD40-repeat-containing domain	18
IPR012337 [Domain]: Ribonuclease H-like domain	18
IPR015943 [Domain]: WD40/YVTN repeat-like-containing domain	18
IPR001128 [Family]: Cytochrome P450	17
IPR001611 [REPEAT] - Leucine-rich repeat	17
IPR012677 [Domain]: Nucleotide-binding alpha-beta plait domain	16
IPR001680 [Repeat]: WD40 repeat	16
IPR001005 [Domain]: SANT/Myb domain	15
IPR029044 [Domain]: Nucleotide-diphospho-sugar transferases	15
IPR026960 [Domain]: Reverse transcriptase zinc-binding domain	15
IPR017853 [Domain]: Glycoside hydrolase superfamily	15
IPR000504 [Domain]: RNA recognition motif domain	14
IPR013210 [Domain]: Leucine-rich repeat-containing N-terminal, plant-type	14
IPR001245 [Domain]: Serine-threonine/tyrosine-protein kinase catalytic domain	14

InterProScan ID	Number of genes
IPR018247 [Binding_Site]: EF-Hand 1, calcium-binding site	13
IPR005135 [Domain]: Endonuclease/exonuclease/phosphatase	13
IPR011598 [Domain]: Myc-type, basic helix-loop-helix (bHLH) domain	13
IPR011992 [Domain]: EF-hand domain pair	13
IPR002401 [Family]: Cytochrome P450, E-class, group I	13
IPR005123 [Domain]: Oxoglutarate/iron-dependent dioxygenase	12
IPR002048 [Domain]: EF-hand domain	12
IPR012334 [Domain]: Pectin lyase fold	11
IPR013781 [Domain]: Glycoside hydrolase, catalytic domain	11
IPR011050 [Domain]: Pectin lyase fold/virulence factor	11
IPR017930 [Domain]: Myb domain	11
IPR017972 [Conserved_Site]: Cytochrome P450, conserved site	11
IPR006121 [Domain]: Heavy metal-associated domain, HMA	10
IPR001810 [Domain]: F-box domain	10
IPR000620 [Domain]: EamA domain	10
IPR012336 [Domain]: Thioredoxin-like fold	10
IPR016140 [Domain]: Bifunctional inhibitor/plant lipid transfer protein/seed	10
IPR025558 [Domain]: Domain of unknown function DUF4283	10

2.3.3. Microsatellite prospection

The search for microsatellites resulted in the identification of 11,020 simple sequence repeats (SSR), representing 1.05% of all sequence data (109,695 bp/10,401,671 bp). In CDS (1,985,806 bp) there were 1,762 SSRs (~16% of the total). Taking into account all sequence data, 106 SSRs were found every 100 kb (one SSR every 0.94 kb). Analyzing the CDS region, 89 SSRs were found every 100 kb (one SSR every 1.12 kb); hence, the frequency of SSRs was slightly lower in the CDS region (~1.2×, 1.12 kb/0.94 kb). Our estimates were 10× lower than those reported in Santos et al. (2014) using *P. edulis* BES data as a major resource (10.8 SSRs every 100 kb or one SSR every 9.25 kb).

Microsatellite sequences were grouped according to motif, and all possible classes of repeats were found, with trinucleotides the most prevalent in both data sources. Compound SSRs accounted for 17.4% (1,919/11,020) of all SSRs, and 15.7% (278/1,762) of these SSRs were found in CDS (Figure 2A). Among the mononucleotides, the A/T motif far surpassed the number of G/C motifs. The most frequent dinucleotides were AT/AT (49.3%), followed by AG/CT (35.4%), which were prevalent in CDS (74%). Among the trinucleotides, AAG/CTT were the most frequent in both data sources (~23%). Other occurrences (tetra-, penta- and hexanucleotides) are shown in Figure 2B.

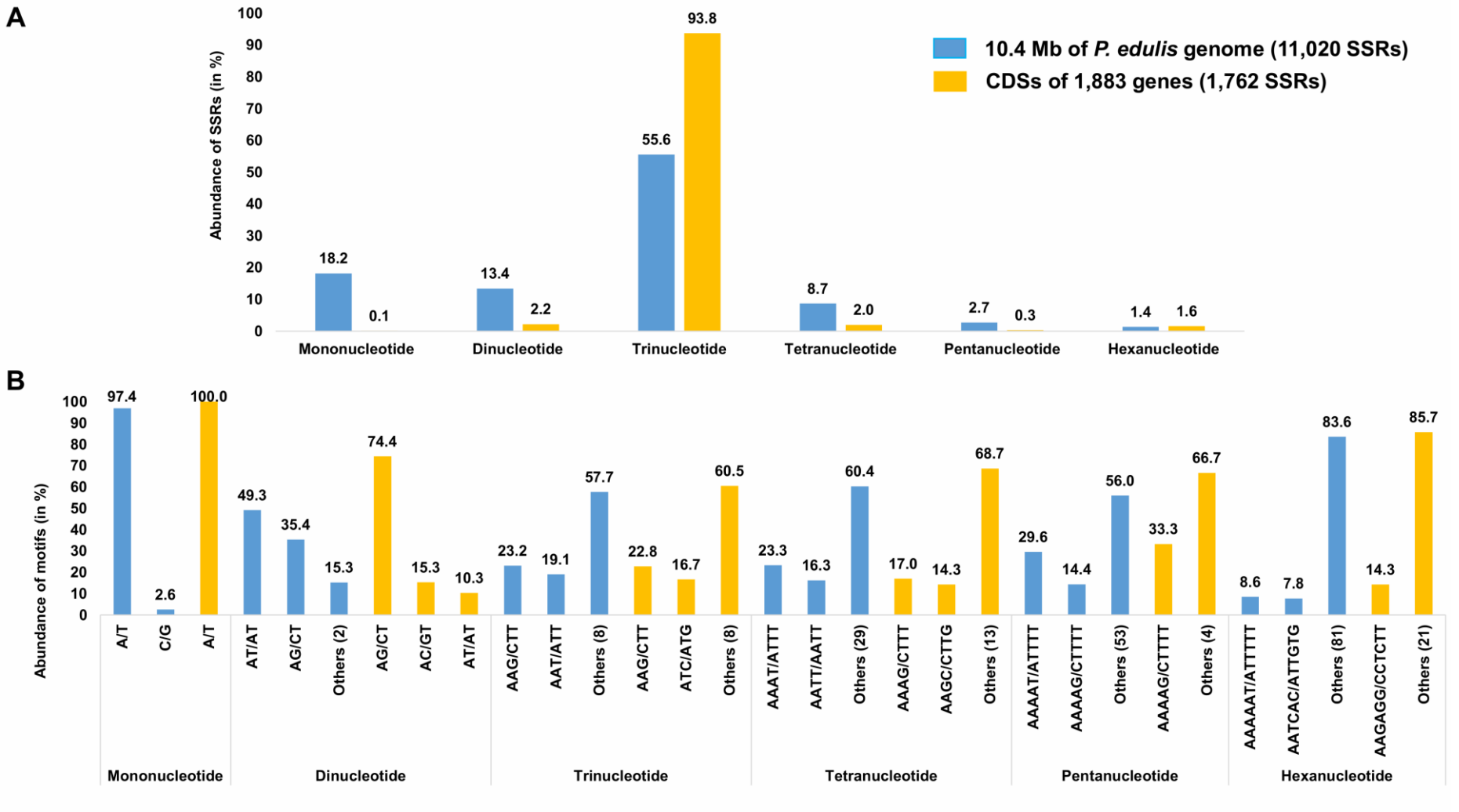

Figure 2. (A) Percentage of mono-, di-, tri-, tetra-, penta- and hexanucleotides in microsatellites (SSRs) found in all sequences (10.4 Mb) of *Passiflora edulis* (blue bars) and in coding DNA sequences (CDS, orange bars). (B) Percentage of the most frequent motifs in each class of microsatellites (SSRs) found in all sequences (blue bars) and in coding DNA sequences (CDS, orange bars) of *Passiflora edulis*.

2.4. Discussion

Despite great advances in genome sequencing, the process of sequencing a plant genome is still laborious, due primarily to the size and complexity of genome regions which pose a challenge when it comes to sequencing and assembly. For instance, *Passiflora* species are extensively diversified in morphological terms, with genome sizes ranging from 207 Mb to 2.15 Gb (Yotoko et al., 2011) and there are no draft genomes for any passionfruits, even the most cultivated species, *P. edulis*. In this study, a gene-rich fraction of the *P. edulis* genome was sequenced and assembled from long sequence reads, allowing us to obtain 10.4 Mb of highly curated data.

About half of all sequences (44%) matched *P. edulis* gene sequences and annotation revealed several functional categories and protein domains. Interestingly, the most frequent domain was retrotransposon gag, associated with transcripts of the LTR retrotransposon, followed by the kinase domains. This abundance was to be expected, since kinases belong to a superfamily of proteins with copies in the hundreds or thousands and are components of all cellular functions. These proteins use ATP γ-phosphate to phosphorylate serine and threonine or tyrosine residues from other proteins (Lehti-Shiu and Shiu, 2012). Note that to date there is an enormous scarcity of information on *Passiflora* nuclear genes in databases. This means that obtaining gene-based probes for selecting new regions for whole sequencing is practically impossible. Our structural and functional annotation of approximately 1,900 genes provides a significant set of high quality gene sequences that can be used in many other studies on *Passiflora*.

The GC content (41.09%) found in this genomic region is high, likewise that reported previously using BES data as the major resource (41.9%, Santos et al., 2014). The GC content of other Malpighiales related species is lower (*Populus trichocarpa*, 33.7%, Tuskan et al., 2006; *Ricinus communis*, 32.5%, Chan et al., 2010; *Hevea brasiliensis*, 34.84%, Tang et al., 2016; *Manihot esculenta*, 33.94%, Wang et al., 2014), with the exception of *Linum usitatissimun* (40%, Wang et al., 2012). In plants, grass genomes contain higher GC content compared with that of other angiosperms (Singh et al., 2016). Interestingly, the GC content of *P. edulis* is similar to those of some monocot species.

In terms of microsatellite abundance, we found that ~1.0% of all *P. edulis* sequences consisted of SSRs, with trinucleotide repeats prevalent (55.6%), even in CDS (93.8%).

Microsatellite abundance generally varies from one genome region to another, but trinucleotides are usually overrepresented in coding sequences, due to selection pressures against mutations that may alter the reading frames (Xu et al., 2013). Our results corroborate the findings of a pioneer study reported by Morgante et al. (2002) to the effect that trinucleotide repeats are significantly more abundant in the expressed regions of plant genomes. Very recently, Araya et al. (2017) reported a total of 1,300 perfect microsatellite sites in *P. edulis* genomic regions (with minimum 15× coverage as a cut off; Illumina paired-end reads) that were selected for marker development and *Passiflora* diversity analysis. In this significant sample, the prevalence of tri-, tetra- and dinucleotides was found to be 41.0%, 36.4% and 22.6%, respectively.

In the *P. trichocarpa* genome, the predominance of mono- (69.8%), di- (19.5%) and trinucleotides (9.0%) decreased stepwise as the motif length increased (mono- to hexanucleotide repeats); 98% of *P. trichocarpa* mononucleotides consist of A/T motifs and only 2% of C/G motifs. The same applies to *P. edulis* (Figure 2B). For di- and trinucleotides, the most frequent motifs were AT/AT (60.5%) and AAT/ATT (48.2%). In terms of coding sequences, 90.3% and 76.6% of the mono- and dinucleotides consist respectively of A/T and AG/CT motifs. Trinucleotides consist mainly of AAG/CTT, ACC/GGT and AGG/CCT motifs (~20% of each), and the frequencies of tetra-, penta- and hexanucleotides were very low (Sonah et al., 2011).

In *M. esculenta*, 37.4% of all SSRs corresponded to dinucleotides, and tri- and pentanucleotides were found in the same proportion (~24%); within the coding sequences, tri- and hexanucleotides accounted for 95.6%. AT/AT and AAT/ATT were the most common di- and trinucleotide motifs (~23% and ~12%, respectively) and, as in *P. edulis*, AG/CT and AAG/CTT were the most prevalent in coding sequences (~4% and ~23%, respectively) (Vásquez and López, 2014). In the *R. communis* genome, most of the SSRs found were also dinucleotides (70.4%), followed by trinucleotides (24.9%). AT/TA was the most frequent motif among dinucleotides (75.3%) and AAT/TTA among trinucleotides (71%) (Tan et al., 2014).

Clearly, the particular occurrence of certain motifs in plant genomes and in different genome regions is due to selection pressure during evolution (Ellegren, 2004; Hancock, 1999), and structural and functional genome attributes, like GC content and codon usage bias, may be responsible for the unique content and distribution patterns of microsatellites (Chakraborty et al., 1997; Whittaker et al., 2003).

This is the first report with good quality data about the identification of genes and identification and distribution of microsatellites in *P. edulis* genomic sequences. Our research group has been worked with the aim of improving knowledge about this species. Our data contributed for understanding a little more about this species and other studies are in progress, like the comparative mapping of some gene-rich regions described here that are been compared to genomic regions of *Populus trichocarpa* and *Manihot esculenta*, two Malpighiales species with complete available genomes.

Other issue is the identification and characterization of transposable elements (next chapter). Many studies have reported the contribution of TEs to genome evolution. This study will facilitate the understanding of the evolution of the *P. edulis* genome, and provides the first report on the dynamics of TEs in *Passiflora*. Despite the availability of sequencing technologies suitable for the generation of high amount of quality data, untangling a large genome like that of *P. edulis* (~1,230 Mb; Yotoko et al., 2011) is still a challenge and the cost remains a limitation. The strategy adopted herein represents an important step towards obtaining the entire genome of *P. edulis*.

2.5. Conclusions

The outcome of this research was a unique set of high quality sequence data on a gene-rich fraction of the *Passiflora edulis* genome, describing gene content and abundance of microsatellites. The structural and functional annotations of some 1,880 genes of *P. edulis* are detailed. Obtaining the complete genome of *P. edulis* is still a challenge. In this study the first steps have been taken and we could generate some information about the gene-rich regions of *P. edulis*, providing genomic data for other studies. For instance, further studies are required to elucidate the organization of gene-poor and large repetitive regions to contribute to our understanding of the evolutionary issues that this genome underwent.

References

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012). GenomeView: a next-generation genome browser. *Nucleic Acids Res.* 40, e12–e12. doi:10.1093/nar/gkr995.

Abreu, P. P., Souza, M. M., Santos, E. A., Pires, M. V., Pires, M. M., and de Almeida, A.-A. F. (2009). Passion flower hybrids and their use in the ornamental plant market: perspectives for sustainable development with emphasis on Brazil. *Euphytica* 166, 307–315. doi:10.1007/s10681-008-9835-x.

Aggarwal, R. K., Hendre, P. S., Varshney, R. K., Bhat, P. R., Krishnakumar, V., and Singh, L. (2007). Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor. Appl. Genet.* 114, 359–72. doi:10.1007/s00122-006-0440-x.

Araya, S., Martins, A. M., Junqueira, N. T. V, Costa, A. M., Faleiro, F. G., and Ferreira, M. E. (2017). Microsatellite marker development by partial sequencing of the sour passion fruit genome (*Passiflora edulis* Sims). *BMC Genomics* 18, 549. doi:10.1186/s12864-017-3881-5.

Bao, Z., and Eddy, S. R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276. doi:10.1101/gr.88502.

Bernacci, L. C., Cervi, A. C., Milward-de-Azevedo, M. A., Nunes, T. S., Imig, D. C., and Mezzonato, A. (2015). Passifloraceae. *List. espécies da flora do Bras.* Available at: http://reflora.jbrj.gov.br/jabot/floradobrasil/FB182 [Accessed November 15, 2017].

Borodovsky, M., and Lomsadze, A. (2011). Eukaryotic Gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr. Protoc. Bioinformatics* CHAPTER, Unit-4.610. doi:10.1002/0471250953.bi0406s35.

Buyyarapu, R., Kantety, R. V, Yu, J. Z., Xu, Z., Kohel, R. J., Percy, R. G., et al. (2013). BAC-pool sequencing and analysis of large segments of A12 and D12 homoeologous chromosomes in upland cotton. *PLoS One* 8, e76757. doi:10.1371/journal.pone.0076757.

Carneiro, M. S., Camargo, L. E. A., Coelho, A. S. G., Vencovsky, R., Rui, P. L. J., Stenzel, N. M. C., et al. (2002). RAPD-based genetic linkage maps of yellow passion fruit (*Passiflora edulis* Sims. f. flavicarpa Deg.). *Genome* 45, 670–678. doi:10.1139/g02-035.

Cauz-Santos, L. A., Munhoz, C. F., Rodde, N., Cauet, S., Santos, A. A., Penha, H. A., et al. (2017). The chloroplast genome of *Passiflora edulis* (Passifloraceae) Assembled from long sequence reads: structural organization and phylogenomic studies in Malpighiales. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.00334.

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13, 238. doi:10.1186/1471-2105-13-238.

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J., and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1041–1046.

Chan, A. P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* 28, 951–956. doi:10.1038/nbt.1674.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth* 10, 563–569. Available at: http://dx.doi.org/10.1038/nmeth.2474.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi:10.1093/bioinformatics/bti610.

Cuco, S. M., Vieira, M. L. C., Mondin, M., and Aguiar-Perecin, M. L. R. (2005). Comparative karyotype analysis of three *Passiflora* L. species and cytogenetic characterization of somatic hybrids. *Caryologia* 58, 220–228. doi:10.1080/00087114.2005.10589454.

de Setta, N., Monteiro-Vitorello, C. B., Metcalfe, C. J., Cruz, G. M. Q., Del Bem, L. E., Vicentini, R., et al. (2014). Building the sugarcane genome for biotechnology and identifying evolutionary trends. *BMC Genomics* 15, 540. doi:10.1186/1471-2164-15-540.

Deng, J., Zhou, Y., Bai, M., Li, H., and Li, L. (2010). Anxiolytic and sedative activities of *Passiflora edulis* f. flavicarpa. *J. Ethnopharmacol.* 128, 148–153. doi:10.1016/j.jep.2009.12.043.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435–45. doi:10.1038/nrg1348.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi:10.1093/nar/gkt1223.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi:10.1093/nar/gkr944.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313, 903–919. doi:10.1006/jmbi.2001.5080.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. J., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi:10.1038/nprot.2013.084.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7–R7. doi:10.1186/gb-2008-9-1-r7.

Hancock, J. M. J. (1999). “Microsatellites and other simple sequences: genomic context and mutational mechanisms,” in *Microsatellites: evolution and applications 1*, eds. D. B. Goldstein and C. Schlötterer (Oxford: Oxford University Press), 3–9. doi:10.1038/mt.2008.186.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32, 767–769. doi:10.1093/bioinformatics/btv661.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* 24, 688–696. doi:10.1101/gr.168450.113.

IBGE (2015). *Produção Agrícola Municipal: culturas temporárias e permanentes*. Rio de Janeiro.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5, 59. doi:10.1186/1471-2105-5-59.

Lehti-Shiu, M. D., and Shiu, S.-H. (2012). Diversity, classification and function of the plant protein kinase superfamily. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 2619–2639. doi:10.1098/rstb.2012.0003.

Letunic, I., and Bork, P. (2017). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, gkx922-gkx922. Available at: http://dx.doi.org/10.1093/nar/gkx922.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi:10.1038/nbt.3208.

Lopes, R., Lopes, M. T. G., Carneiro, M. S., Matta, F. D. P., Camargo, L. E. A., and Vieira, M. L. C. (2006). Linkage and mapping of resistance genes to *Xanthomonas axonopodis* pv. passiflorae in yellow passion fruit. *Genome* 49, 17–29. doi:10.1139/G05-081.

Madureira, H. C., Pereira, T. N. S., Da Cunha, M., and Klein, D. E. (2012). Histological analysis of pollen-pistil interactions in sour passion fruit plants (*Passiflora edulis* Sims). *Biocell* 36, 83–90.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi:10.1093/bioinformatics/bth315.

Mayer, K. F. X., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., Graner, A., et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491, 711–716. doi:10.1038/nature11543.

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., et al. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* 47, 1435. Available at: http://dx.doi.org/10.1038/ng.3435.

Moraes, M. C., Gerald, I. O., Matta, F. P., and Vieira, M. L. C. (2005). Genetic and phenotypic parameter estimates for yield and fruit quality traits from a single wide cross in yellow passion fruit. *HortScience* 40, 1978–1981.

Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with non-repetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200. doi:10.1038/ng822.

Munhoz, C. F., Santos, A. A., Arenhart, R. A., Santini, L., Monteiro-Vitorello, C. B., and Vieira, M. L. C. (2015). Analysis of plant gene expression during passion fruit-*Xanthomonas axonopodis* interaction implicates lipoxygenase 2 in host defence. *Ann. Appl. Biol.* 167, 135–155. doi:10.1111/aab.12215.

Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R., and Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genet. Mol. Biol.* 29, 294–307. doi:10.1590/S1415-47572006000200018.

Oliveira, E. J., Vieira, M. L. C., Garcia, A. A. F., Munhoz, C. F., Margarido, G. R. A., Consoli, L., et al. (2008). An integrated molecular map of yellow passion fruit based on simultaneous maximum-likelihood estimation of linkage and linkage phases. *J. Am. Soc. Hortic. Sci.* 133, 35–41. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-37849007543&partnerID=40&md5=92f8347ed25bb95a14157a5b4912095d.

Santos, A., Penha, H., Bellec, A., Munhoz, C. De, Pedrosa-Harand, A., Bergès, H., et al. (2014). Begin at the beginning: A BAC-end view of the passion fruit (*Passiflora*) genome. *BMC Genomics* 15, 816. doi:10.1186/1471-2164-15-816.

Silvia, M. C., Vieira, M. L. C., Mondin, M., and Aguiar-Perecin, M. L. R. (2005). Comparative karyotype analysis of three passiflora l. species and cytogenetic characterization of somatic hybrids. *Caryologia* 58, 220–228. doi:10.1080/00087114.2005.10589454.

Singh, R., Ming, R., and Yu, Q. (2016). Comparative analysis of GC content variations in plant genomes. *Trop. Plant Biol.* 9, 136–149. doi:10.1007/s12042-016-9165-4.

Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31. doi:10.1186/1471-2105-6-31.

Sonah, H., Deshmukh, R. K., Sharma, A., Singh, V. P., Gupta, D. K., Gacche, R. N., et al. (2011). Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. *PLoS One* 6. doi:10.1371/journal.pone.0021298.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32, W309–W312. doi:10.1093/nar/gkh379.

Suassuna, T. de M. F., Bruckner, H., de Carvalho, R., and Borem, A. (2003). Self-incompatibility in passionfruit: evidence of gametophytic-sporophytic control. *Theor. Appl. Genet.* 106, 298–302. doi:10.1007/s00122-002-1103-1.

Tan, M., Wu, K., Wang, L., Yan, M., Zhao, Z., Xu, J., et al. (2014). Developing and characterising *Ricinus communis* SSR markers by data mining of whole-genome sequences. *Mol. Breed.* 34, 893–904. doi:10.1007/s11032-014-0083-6.

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat. Plants* 2, 1–10. doi:10.1038/NPLANTS.2016.73.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science (80-.).* 313, 1596–1604. doi:10.1126/science.1128691.

ULMER, T., and MacDougal, J. M. J. M. (2004). *Passiflora*: passionflowers of the world. *Timber Press*, 430p.

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* 527, 508–511. doi:10.1038/nature15714.

Vásquez, A., and López, C. (2014). *In silico* genome comparison and distribution analysis of simple sequences repeats in cassava. *Int. J. Genomics* 2014. doi:10.1155/2014/471461.

Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., et al. (2014). Cassava genome from a wild ancestor to cultivated varieties. *Nat. Commun.* 5, 5110. doi:10.1038/ncomms6110.

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., et al. (2012). The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. *Plant J.* 72, 461–473. doi:10.1111/j.1365-313X.2012.05093.x.

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R. M. (2003). Likelihood-based estimation of microsatellite mutation rates. *Genetics* 164, 781–787.

Xu, J., Liu, L., Xu, Y., Chen, C., Rong, T., Ali, F., et al. (2013). Development and characterization of simple sequence repeat markers providing genome-wide coverage and high resolution in maize. *DNA Res.* 20, 497–509. doi:10.1093/dnares/dst026.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13, 329–342. doi:10.1038/nrg3174.

Yotoko, K. S. C., Dornelas, M. C., Togni, P. D., Fonseca, T. C., Salzano, F. M., Bonatto, S. L., et al. (2011). Does variation in genome sizes reflect adaptive or neutral processes? New clues from *Passiflora*. *PLoS One* 6, e18212. doi:10.1371/journal.pone.0018212.

Zhbannikov, I. Y., Hunter, S. S., Foster, J. A., and Settles, M. L. (2017). SeqyClean. *Proc. 8th ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics - ACM-BCB '17*, 407–416. doi:10.1145/3107411.3107446.

3. TRANSPOSABLE ELEMENT IDENTIFICATION AND ANALYSIS OF EVOLUTIONARY LINEAGES OF LTR-RETROTRANSPOSONS IN A GENE-RICH FRACTION OF THE *PASSIFLORA EDULIS* GENOME

ABSTRACT

A large portion of plant genomes is composed of repetitive elements, especially LTR retrotransposons (LTR-RTs), hence, their identification and characterization is an important part of a genome description. The aim of the present study was to identify and characterize transposable elements (TEs) and analyze the LTR-RTs found in ~10 Mb of DNA sequences of *P. edulis*. TEs were identified and classified using REPET/PASTEC. Internal coding domains of LTR-RTs were identified using Repeat Explorer. LTR-RTs were assigned to evolutionary lineages based on the phylogenetic inference estimated from the complete aminoacid sequence of the RT or GAG domains, applying the Maximum Likelihood method with 1,000 bootstraps. Insertion time was estimated based on LTRs divergence. *In silico* transcriptional activity of LTR-RTs was tested based on their association to transcripts (80% coverage and 80% identity). Transcriptional activity was confirmed through amplification of the RT domain in cDNA of LTR-RTs. The presence of LTR-RTs in other eight species of *Passiflora* was investigated through PCR amplification of the RT domain. The search for transposable elements resulted in the identification of 250 TEs, classified in Class I (11 DIRSs, 7 LINEs, 181 LTRs, 2 SINEs, 36 LARDs, 4 TRIMs,), and Class II (2 HELITRONs, 6 TIRs, 1 MITE). TE sequences corresponded to 17.6% of the data. TEs were preferentially located in intergenic spaces (70.4%), but some were observed overlapping genes (30.6%). LTR-RTs corresponded to 13.6% of the data and 44 and 137 were assigned to *Copia* and *Gypsy* superfamilies, respectively. Four plant evolutionary *Copia* lineages (*Angela*, *Ale*, *Tork* and *Sire*) and five *Gypsy* (*Del*, *Athila*, *Reina*, *CRM* and *Galadriel*) were identified. ~40% and ~55% of *Gypsy* and *Copia* elements, respectively, have five internal domains. *Copia* elements were more conserved regarding each domain separately. The majority (95.9%) of the full-length LTR-RTs were recently inserted into *P. edulis* genome (< 2.0 Mya). 2,821 transcripts were associated to full-length *Copia* and *Gypsy* LTR-RTs. Transcriptional activity was confirmed for all elements, except the element from *Athila*. *Angela*, *Del*, *CRM* and *Tork* are possibly conserved in *Passiflora* genus, since representative elements were identified in all species. *Athila* and *Galadriel* lineages were found only in the *Passiflora* subgenus. This is the first analysis on the structure and content of TEs, particularly LTR-RTs, in *P. edulis* genome.

Keywords: *Passiflora edulis*; Passion fruit genome; Transposable elements; LTR-RT

3.1. Introduction

Transposable elements (TEs) are DNA segments that have the ability to move within the genome (Lisch, 2012). They are found in all eukaryotic genomes, with little-known exceptions (Garcia-Perez, 2010; Wicker et al., 2007), and were first described in maize by Barbara McClintock in the middle of the 20th century (see Ravindran, 2012). For many years, TEs were considered ‘junk DNA’ or ‘selfish DNA parasites’ until they were discovered to have many important roles. Genome-scale studies revealed and confirmed that TEs play a key

role in genome function, chromosome evolution, speciation and diversity (Garcia-Perez, 2010; Klein and O'Neill, 2018).

TEs are well-represented in all eukaryote genomes. In animals, for instance, TEs account for 5.3% of the genome in *Drosophila melanogaster*, 18.3% in *Caenorhabditis elegans*, and up to 42.8% in *Homo sapiens* (Canapa et al., 2016). In angiosperms, TEs account for 18.5% of the genome in *Arabidopsis thaliana*, 28.1% in *Brachypodium distachyon*, 63.2% in *Solanum lycopersicum*, and up to 84.2% in *Zea mays* (Oliver et al., 2013).

TEs are subdivided into two major classes based on the transposition mechanism. Class I elements are known as retrotransposons and use a 'copy-and-paste' mechanism with an RNA intermediate stage. Class II elements are known as DNA transposons and move through the genome using a straight-forward 'cut-and-paste' transposition mechanism. Each class is subdivided based on the presence of typical domains and other structures, such terminal repeats (Mita and Boeke, 2016; Padeken et al., 2015; Wicker et al., 2007). Based on their coding ability, transposable elements are referred to as autonomous if they express all the proteins required for their transposition, or non-autonomous if they express only part of the machinery needed for transposition. Alternatively, they may have no expression at all and transposition is then reliant on protein expression of other TEs (Padeken et al., 2015; Zhao et al., 2016).

With the aim coming up with a unified system for classifying TEs, Wicker et al. (2007) proposed hierarchical classification based on the transposition mechanism, sequence similarities and TE structural relationships. The hierarchy includes class, subclass, order, superfamily, family, and subfamily (Figure 3).

As mentioned above, the first level (Class) is based on the presence or absence of an RNA intermediate stage in the transposition process. Class II is subdivided into two subclasses: Subclass 1 includes the 'cut-and-paste' DNA transposons that copy and insert themselves into another position; Subclass 2 includes DNA transposons that are copied from the genome ('rolling circle' and 'self-replicating'), involving the transposition of only one strand. Orders are based on presence, structure and organization of protein or non-coding domains. Superfamilies share the replication strategy, but are differentiated by features, such as protein structure or non-coding domains. Superfamilies also differ in terms of the presence and size of the Target Site Duplication (TSD), a short direct repeat that flanks the TE and is generated upon insertion. Family and subfamily levels share DNA sequence conservation (Figure 3) (Mita and Boeke, 2016; Padeken et al., 2015; Wicker et al., 2007).

Class I elements include the following orders: LTR (Long Terminal Repeat), DIRS (*Dictyostelium* Intermediate Repeat Sequence), PLE (*Penelope*-like Elements), LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements) (Figure 3).

LTR-RTs (LTR retrotransposons) have long terminal repeats at both 5' and 3' ends. Autonomous LTR-RTs also have domains that encode for proteins related to their transposition: GAG (Group-specific Antigen), POL (synthesized as a polyprotein containing PROT, Aspartic Proteinase; RT, Reverse Transcriptase; RH, RNAse H; and INT, Integrase genes). In some cases they even include functional ENV (Envelope) proteins. The GAG domain is a structural protein responsible for packing retrotransposon RNAs and proteins, forming the VLP (Virus-like Particle), inside which reverse transcription occurs; PROT cleaves the Pol polyprotein; INT allows insertion into a new location; RT and RH catalyze the formation of the cDNA from the retrotransposon RNA (Havecker et al., 2004; Padeken et al., 2015; Wicker et al., 2007; Zhao et al., 2016). Additionally, some LTR-RTs can have a chromodomain (Chromatin Organization Modifier Domain, CHD), which is a protein domain of 40−50 amino acids involved in chromatin remodeling and gene expression regulation in eukaryotes. These proteins perform a wide range of functions, including chromatin targeting and interaction between different proteins, RNA and DNA (Novikova, 2009; Novikova et al., 2012).

LTR-RTs are divided into five superfamilies; three are found only in Metazoans (*Bel-Pao*, *Retrovirus* and *ERV*), and the other two (*Copia* and *Gypsy*) are also found in plants. *Copia* and *Gypsy* elements differ only in the organization of their internal domains (Bennetzen and Wang, 2014; Wicker et al., 2007). Viridiplantae LTR-RTs from the *Copia* superfamily are further subdivided into families or evolutionary lineages (*Angela*, *Ale*, *Bianca*, *Ivana*, *Oryco*, *Retrofit*, *Sire*, *Tork* and *Maximus*), as are those in the *Gypsy* superfamily (*Athila*, *CRM*, *Del*, *Galadriel* and *Reina*). Evolutionary lineages share relationships in both sequence similarity and genome structure, and their identification and characterization are central to understanding the evolutionary history of the LTR-RT system. Plant genomes have many copies of LTR-RT evolutionary lineages, but only a few proliferated to high-copy numbers and this changes from one genome to another. The assignment of an LTR-RT to one specific evolutionary lineage is based on phylogenetic inference according to the domains of the internal region (Llorens et al., 2009, 2011).

LTR-RTs are the predominant order in plants. They form a large fraction of all flowering plant genomes investigated so far (Vicient and Casacuberta, 2017). For example, LTR-RTs account for a respective 61.8% and 41.7% of the genomes of *Solanum lycopersicum*

(The Tomato Consortium, 2012) and *Pinus taeda*, the loblolly pine; (Wegrzyn et al., 2014). In addition, LTR-RTs are responsible for genome expansion in some species, such as *Capsicum annuum*, with the accumulation of LTR-RTs and their derivatives (Park et al., 2012; Vicient and Casacuberta, 2017).

LTR-RTs can be found in heterochromatic regions and also inside or close to the genes, influencing their expression and evolution. LTR-RTs can influence alternative splicing, epigenetic control, transduction, duplication, recombination and many other cellular processes (Galindo-González et al., 2017; Oliver et al., 2013). It is therefore crucial to understand the contribution of LTR-RTs to genome structure and function, but only a few detailed studies are available.

For instance, in sugarcane, an initial detailed study of LTR-RTs was carried out by Domingues et al. (2012). Sixty complete LTR-RT elements were classified into 35 families within four *Copia* (*Ale*, *Angela*, *Ivana* and *Maximus*) and three *Gypsy* (*DEL*, *Reina* and *TAT*) lineages. Elements were structurally similar within lineages, but there were large size differences between lineages. By using FISH (Fluorescence *in situ* hybridization), the authors showed that the *Gypsy* elements were found in heterochromatic regions, and *Copia* elements in the euchromatin. However, two lineages (*Ale* and *TAT*) showed a localized clustering pattern on some chromosomes. In addition, individual families had distinct transcript profiles, suggesting that they are differentially expressed. These findings indicate that LTR-RT families can potentially affect the genome in diverse ways.

DIRS elements have a tyrosine recombinase (YR) domain instead of an INT, which is typically involved in site-specific recombination. Some DIRS families exhibit a conserved methyltransferase (MET) domain at the C-terminus. These elements are very diverse in structural terms (Poulter and Butler, 2015; Poulter and Goodwin, 2005). TEs from the PLE order encode an RT and an endonuclease. They also exhibit terminal repeats that can be directly or inversely orientated (Wicker et al., 2007). LINE elements can be several kilobases in length and are found in all eukaryotic kingdoms. SINE elements are small (500−800 bp) and depend on LINE machinery for transposition (Mita and Boeke, 2016; Wicker et al., 2007).

Class II includes DNA transposons with TIRs (Terminal Inverted Repeats). Superfamilies of Class II elements are based on the relatedness of the transposase and shared structural features, including the TIR sequence and the length of the TSD. The TIR order has nine superfamilies, seven of which are found in plants (*Tc1-Mariner*, *hAT*, *Mutator*, *P*, *PIF-Harbinger* and *CACTA*). The order *Helintron* is found in plants too. The other two orders

(*Crypton* and *Maverick*) do not exist in plants (Figure 3) (Mita and Boeke, 2016; Wicker et al., 2007).

Non-autonomous elements include LARDs (Large Retrotransposon Derivatives), which can reach up to 4 kb in length and have an internal domain with no coding capacity; MITEs (Miniature Inverted-Repeat Transposable Elements), which are flanked by TIRs and are often found within or close to genes; SNACs (Small Non-Autonomous CACTA Transposons) and TRIMs (Terminal Repeat Retrotransposons in Miniature), which are less than 4 kb in length and, like LARDs, have an internal domain with no coding capacity (Wicker et al., 2007).

Naming TEs from different classes and orders is no longer the complicated task it used to be, thanks to the guidelines proposed by Wicker et al. (2007). The system uses a three-letter code (class, order and superfamily, respectively); the family (or subfamily) name; the ID (database accession) of the sequence in which the element was identified and the number of the element in the sequence. For example, RLG_*CRM*_Pe33M2-1 belongs to Class I, order LTR-RT, superfamily *Gypsy*, family *CRM*, and is the first element in the Pe33M2 sequence of *Passiflora edulis* (see Figure 3).

Transposable elements are very diverse within and among genomes, posing some identification and annotation problems. As stated above, TEs differ in many aspects, including structure, transposition mechanism, sequence, length, chromosomal location, etc. They are known to have been present in almost all species for a long time, and therefore have been subjected to evolutionary mechanisms, like the overtime process of insertion into new genomic sites, which can generate point mutations, rearrangements and indels. This results in fragmented, divergent and mosaic copies that are difficult to annotate (Hoen et al., 2015; SanMiguel et al., 1996).

There are two main approaches to TE annotation: knowledge-based and *de novo* TE detection methods. Knowledge-based approaches use sequence similarity to known TEs, comparing newly generated sequences to a library of known repeats, like Repbase (http://www.girinst.org/repbase/). This approach is efficient for detecting and classifying TEs, even single-copy TEs, due to the large number of previously reported TE sequences. However, it is biased towards TEs that have already been described, and is therefore not suitable for the discovery of new TEs or elements resulting from rapidly diverging repeat sequences (Arensburger et al., 2016; Bergman and Quesneville, 2007; Hoen et al., 2015). In contrast, *de novo* approaches attempt to discover TEs without using prior information. The first step is self-alignment: the genome is compared with itself, aiming to detect similar

sequences at different sites, resulting in pairwise alignments of similar sequences. The resulting pairwise alignments are then clustered into order to obtain repeat families. A consensus for each family is then constructed and classified. Lastly, all copies within the genome are retrieved (or annotated). However, the *de novo* approach does have some disadvantages, such as the fact that all kinds of repeated sequences are mined and it is not possible to detect TE families occurring at low copy number (Arensburger et al., 2016; Bergman and Quesneville, 2007; Hoen et al., 2015).

There are many tools currently available for TE detection using either a knowledge-based or *de novo* approach. The premises on which these tools are based can differ and, consequently, the results can be quite divergent. There is no consensus regarding TE annotation, and this lack of standardization prevents comparative analyses across species or genomes, limiting the capacity to provide insights into TE evolution. Because TEs are relevant to genome organization and evolution, attempts are being made in the literature to standardize TE annotation (Hoen et al., 2015; Ragupathy et al., 2013). One suggestion is to base annotation on combined methods (knowledge-based and *de novo*) by using several complementary tools (Bergman and Quesneville, 2007; Flutre et al., 2011).

The REPET pipeline is a good option as a first step towards standardization. In fact, it consists of two pipelines: TEdenovo and TEannot. TEdenovo implements the following stages: all-by-all alignment of contigs; clustering of identified repeat segments and finally, the construction of consensus sequences. TEannot classifies TEs by searching for homology and comparing structural features. In the final stage, TE copies found in the entire genome are annotated (Flutre et al., 2011; Quesneville et al., 2005; Ragupathy et al., 2013).

The aims of this study were to identify transposable elements in a set of fully sequenced BAC inserts (~10 Mb) of *Passiflora edulis* (Munhoz et al., 2018) using the REPET pipeline, and to analyze the LTR-RT order at lineage level. This is the first analysis of mobile elements in Passifloraceae, the passionflower family.

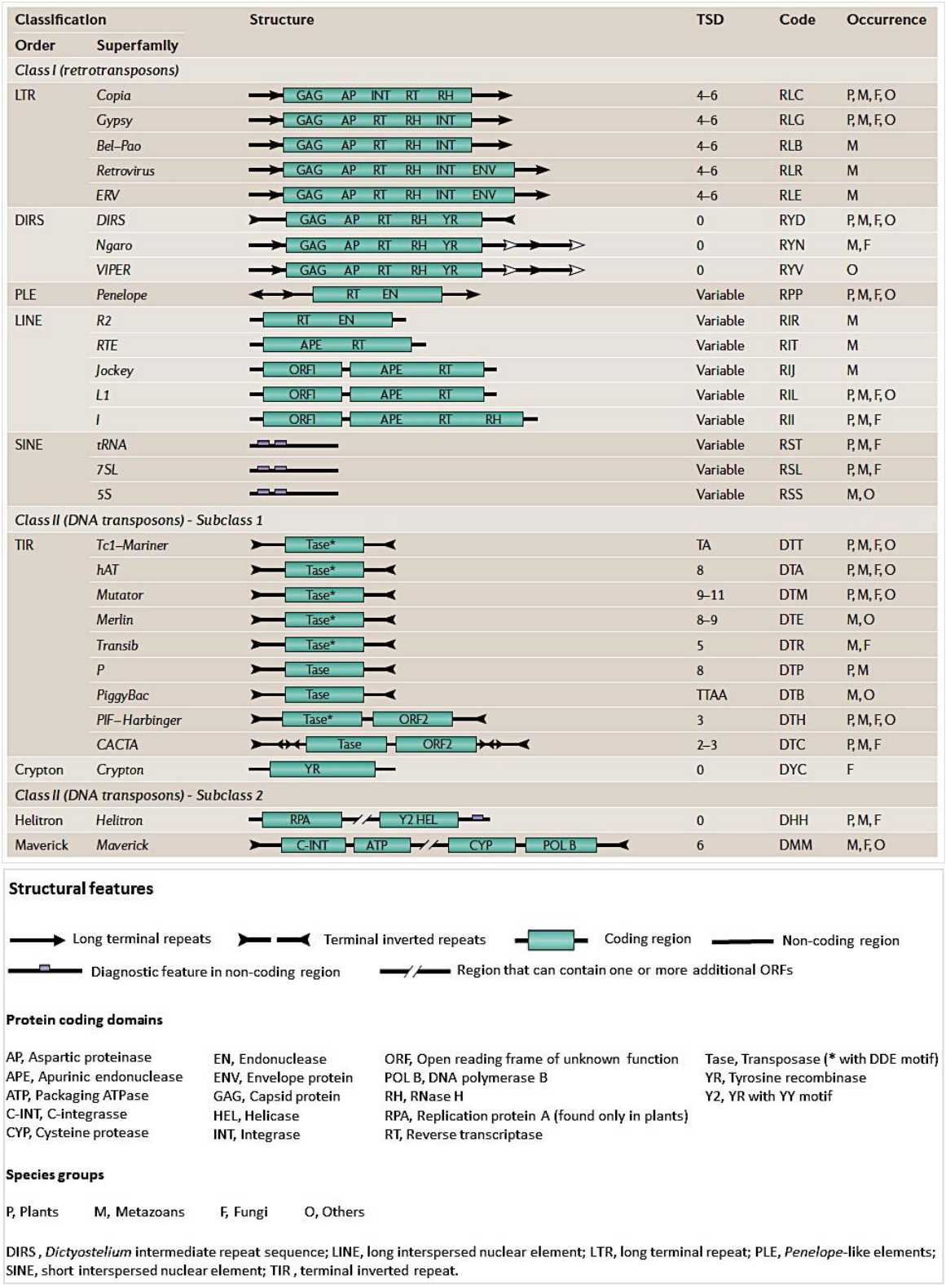

Figure 3. Classification system for transposable elements. Adapted from Wicker et al., 2007.

3.2. Material and Methods

3.2.1. Plant Material

Passiflora edulis accession 'IAPAR-123' is a selection from the Brazilian commercial population used for industrialized juice production, and belongs to the germplasm collection of The Agronomic Institute of Paraná (IAPAR), Paraná state, Brazil. It has favorable characteristics regarding total soluble solids, juice acidity and yield. In addition, this genotype was used as the female parent in a bi-parental cross to generate the first genetic map for the species (Carneiro et al., 2002). Hence, this genotype was selected to build a large-insert genomic BAC (Bacterial Artificial Chromosomes) library denoted 'Ped-B-Flav'. It contains approximately 83,000 clones, and is kept at the French Plant Genomic Resources Center (CNRGV: cnrgv.toulouse.inra.fr)/INRA, Toulouse, France. A set of BAC clones was then selected for sequencing using the PacBio RS® platform, and these sequence data were used herein for searching and identifying *P. edulis* TEs.

Briefly, BAC clones were selected from the findings of Santos et al. (2014), which provide an initial overview of the *P. edulis* genome using BAC-end sequence (BES) data as a major resource. Based on BES functional annotation results, the BAC-inserts with coding sequences (CDS) in one or both BES were selected. A second selection procedure was performed after screening the library using the probes homologous to *P. edulis* transcripts described in Munhoz et al. (2015). Briefly, the authors used suppression subtractive hybridization to construct two cDNA libraries enriched for transcripts respectively induced and repressed by *Xanthomonas axonopodis* 24 h after inoculation with a highly virulent bacterial strain. One hundred and seven BAC clones were sequenced, totaling 10,401,671 bp (10.4 Mb), corresponding to approximately 1.0 % of the *P. edulis* genome. These sequence data were subjected to gene prediction and annotation and constitute a gene-rich fraction of the genome (see Munhoz et al., 2018).

3.2.2. Identification of Transposable Elements

As mentioned above, in this study the same set of *P. edulis* BAC sequences (~10 Mb) was screened to identify transposable elements (TEs). The REPET pipeline was used to

identify, classify and annotate TEs (Flutre et al., 2011; Quesneville et al., 2005). Knowledge-based TE classification was based on the Repbase library (Jurka et al., 2005) .

Transposable elements were classified into the Class I (Retrotransposons) or Class II (DNA transposons) orders using PASTEC (Pseudo Agent System for Transposable Element Classification, Hoede et al., 2014), included in the REPET package. PASTEC classifies TEs by searching for structural features and similarities. It is the only tool that provides classification to order level in the Wicker hierarchical TE classification system. The output of this tool facilitates manual curation by providing all the evidence accumulated for each TE consensus.

TEs were named as follows according to this classification (see Figure 3): RLC or RLG (LTR Retrotransposons from the *Copia* and *Gypsy* superfamilies, respectively), RYD (DIRS order, superfamily *DIRS*), RIX (LINE), RSX (SINE), RXX (unclassified or non-autonomous retrotransposons, like LARDs), for Class I elements; and DTX (TIR Transposon), DHX (Helitron), DXX (MITE) for Class II elements. LTR-RT elements were further characterized and subdivided into evolutionary lineages based on their domains (described below in section 3.2.3).

NCBI's CDD (Conserved Domain Database) is a live search service with a comprehensive collection of protein domain and protein family models tracked by NCBI's Entrez database (Marchler-Bauer et al., 2017). Domains from previously characterized DIRS, LINE, SINE, LARD and TRIM elements from Class I, and from Helitron, TIR and MITE elements from Class II were identified by searching for similarity within the NCBI's CDD. The coordinates of all features of each element were recorded in an Excel sheet and the information used to create a schematic representation of each element in IBS (Illustrator for Biological Sequences) (Liu et al., 2015).

Terminal inverted repeats (from DIRS and TIR elements) were identified using the Einvert tool in the EMBOSS open software suite (Rice et al., 2000). TSDs (from LINE, SINE, LARD, TRIM and TIR elements) were identified by comparing the 5' and 3' flanked sequences of each element using GenomeView (Abeel et al., 2012) in order to visualize the element location and its flanking sequences.

To define their genomic location, the TE sequences were mapped against the annotated BAC sequences using BWA-MEM (Li, 2013), which maps sequences of 70 bp to 1 M bp against a reference. TE locations (within intergenic or genic regions) were examined using GenomeView, a stand-alone browser designed to visualize and manipulate genomic data (Abeel et al., 2012).

3.2.3. Analysis of LTR-RTs

The PASTEC tool enabled us to classify TEs automatically to order level. The LTR-RT elements were then further classified at lineage level. All domains of each LTR-RT were identified, and subsequently used for phylogenetic inferences and assignment to evolutionary lineages. All the steps are described below.

3.2.3.1. Identification and retrieval of LTR-RT internal domains

LTR-RT coding domains GAG, PROT, INT, RT, RH, CHDII, and CHDCR chromodomains were identified using the Repeat Explorer web server, which contains a collection of software tools for the characterization of eukaryotic repetitive elements from next-generation sequence reads. The protein-domain based tools identify, extract and analyze sequence regions corresponding to conserved domains of retrotransposon proteins. The domains are identified based on their similarity to a representative set of reference sequences from the RepeatExplorer database of Viridiplantae TE protein domains. The comparison parameters were minimum similarity 60% and minimum identity 40%. The proportion of hit length to database sequence length was set to 0.8. The output from this analysis is a dataset of protein sequences translated from query DNA and best matching sequences from the protein database. Two output datasets were created: one for *Copia* related sequences and one for *Gypsy* sequences (Novák et al., 2010, 2013).

3.2.3.2. Phylogenetic analysis of LTR-RT elements

Phylogenetic analysis was based preferentially on the translated RT domain. For incomplete LTR-RT elements with no RT domain, the GAG domain was used as an alternative. Domains from previously described lineages were retrieved from the Gypsy Database (GyDB) (Llorens et al., 2011) and used in phylogenetic analysis to assign *P. edulis* elements to lineages. Domains from previously characterized LTR-RT lineages from sugar cane were also used (Domingues et al., 2012). A total of nine *Copia* and six *Gypsy* lineages

were used to initially classify the LTR-RTs from *P. edulis* into *Copia* or *Gypsy* superfamilies. Phylogenetic inferences were then made for each superfamily.

Translated domains were aligned using the MUSCLE program (Edgar, 2004) implemented in MEGA 7.0 (Kumar et al., 2016), with default parameters, and then manually edited. Model Generator was used to find the best amino acid substitution model (Keane et al., 2006). All phylogenetic inferences were drawn using the highest-ranked substitution model available. Phylogenetic trees were constructed using Raxml (Stamatakis, 2014), applying the Maximum Likelihood method with 1,000 bootstrap replicates. Trees were visualized using FigTree v1.4 (Rambaut, 2012) and edited using Dendroscope (Huson et al., 2007).

3.2.3.3. Assignment of LTR-RTs to evolutionary lineages and naming of sequences

P. edulis LTR-RTs were assigned to evolutionary lineages based on phylogenetic inferences. Only groups supported with a high bootstrap value (> 50) were considered. On the basis of a proposed universal classification of TEs (Wicker et al., 2007), we were able to assign names to LTR sequences. We standardized the name of *P. edulis* LTR-RT sequences, following the example of Domingues et al. (2012) for sugarcane sequences. Sequences were named 'RLC' (*Copia*) or 'RLG' (*Gypsy*), 'pe' for '*Passiflora edulis*' and the lineage name, e.g. '*Angela*'. The BAC sequence within the element was located (e.g. 'Pe1K19'); then each sequence of the same lineage and within the same BAC clone was numbered sequentially. For instance, 'RLC_*peAngela*_Pe1K19-1' is the first *P. edulis Angela* element found in the BAC Pe1K19, from the superfamily *Copia*.

3.2.3.4. Structural features of *Passiflora edulis* LTR-RTs

Entire sequences of LTR-RTs were submitted as both 'query' and 'subject' to the blast2seq tool (Sayers et al., 2011). In addition, full lengths were self-aligned using the

MUSCLE program (Edgar, 2004), implemented in MEGA 7.0 (Kumar et al., 2016), to confirm the presence and boundaries of terminal repeats.

TSDs were identified by comparing the 4 to 6 bp of the 5' and the 3' flanked sequences of each element using GenomeView (Abeel et al., 2012) to visualize each LTR-RT location and its flanking sequences. Most of the LTR-RTs found to date are incomplete copies. Therefore, an LTR-RT is considered complete only if it exhibits all the coding domains necessary for its transposition (GAG, PROT, INT, RT and RH) and the two long terminal repeats.

Coordinates of all features of each of the complete LTR-RTs were recorded in an Excel sheet and the information used to create a schematic representation in IBS (Liu et al., 2015).

3.2.3.5. Estimation of *Passiflora edulis* LTR-RT insertion time

When a TE creates a new copy, the two copies are often identical at the time of transposition. In this study, the insertion time of intact LTR-RT lineages was calculated based on the assumption that they are identical at the time of integration (SanMiguel et al., 1998; Zhao et al., 2016). For each element, we aligned the 5' and 3' LTR sequences using the MUSCLE program (Edgar, 2004) set to default parameters. Divergence between LTRs (K) was calculated using MEGA 7.0 (Kumar et al., 2016) and the Kimura-2-parameter distance (Kimura, 1980). The insertion time (T) for each intact element was calculated using the formula: $T = K/2r$ (Yin et al., 2015). A value of 1.5×10^{-8} substitutions per site per year (r) was used to calculate the insertion time (Koch et al., 2000). The age of the LTR-RTs was expressed in Mya (million years ago).

3.2.3.6. *In silico* analysis of LTR-RT expression

A *P. edulis* transcriptome derived from three RNA-seq libraries of shoot apexes of juvenile, vegetative and reproductive adult plants (details provided in Munhoz et al., 2018)

were mapped against the full lengths of our LTR-RTs. Mapping was done using the BWA-MEM package (Li, 2013), and transcripts similar to LTR-RTs were assigned to a lineage according to criteria based on Wicker et al. (2007): 80 % coverage and 80 % nucleotide identity. The number of hits for each library was normalized using the Cufflinks package (Trapnell et al., 2010) available in the Galaxy suite (Afgan et al., 2016).

3.2.3.7. RNA extraction and Reverse Transcriptase (RT)-PCR analysis

Leaf tissues were collected from 4-month-old plants cultivated under greenhouse conditions. Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions.

Primers were manually designed within the reverse transcriptase (RT) domain to amplify full-length LTR-RT elements. Gene Runner (Spruyt and Buquicchio, 2004) was used to verify the quality of the primer sequences as follows: primer length from 18 to 22 bases, 50 to 60% GC-content, absence of secondary structures (such as dimers and hairpins), and melting temperature from 58 to 62° C.

The total RNA preparation was treated with DNAse (Promega, Madison, WI, USA) to remove residual genomic DNA. One µg of RNA was used to generate the first cDNA strand using the SMARTer™ PCR cDNA Synthesis Kit (Clontech Laboratories, Mountain View, CA, USA) according to the manufacturer's instructions. The reaction mixture contained onefold of M-MLV reverse transcriptase (RT) buffer, 0.6 mM of each dNTP, 25 U of RNasin® (Promega), 200 U of M-MLV RT (Promega), 0.25 µM of primer CDS (5' AAG CAG TGG TAT CAA CGC AGA GTA CTT TTT TTT TTT TTT TTT TTT TTT TTT TTN V 3'). Diethylpyrocarbonate (DEPC)-treated water (0.01%) was added to make up the volume to 25 µL, and the reaction was incubated at 42° C for 1 h.

cDNA dilutions were used in PCR reactions as follows: 3.0 µL of cDNA, 1× PCR buffer, 0.25 mM dNTP, 0.3 µM of each primer, 2 mM MgCl2, and 1.2 U Taq DNA polymerase (Promega). Ultrapure water was added to make up the volume to 20 µL. Cycling consisted of an initial denaturing step of 5 min at 95 °C, 30 cycles of 1 min at 95 °C, 1 min at 55 °C, 1 min at 72 °C and a final extension of 8 min at 72 °C. PCR products were electrophoresed in 1.2% agarose gels stained with SYBR SAFE® (Invitrogen), with 1× TBE as the running buffer, and compared to the 100 bp DNA Ladder (Invitrogen).

3.2.3.8. LTR-RTs from wild *Passiflora* species

We investigated the presence of the LTR-RTs identified herein in other species of the *Passiflora* genera. *Passiflora* (Passifloraceae) is subdivided in four subgenera (*Passiflora*, *Decaloba*, *Deidamioides* and *Astrophea*) (Muschner et al., 2003), and the following wild species were screened: *P. edimundoi*, *P. setacea* and *P. alata* (*Passiflora*); *P. organenesis* and *P. capsularis* (*Decaloba*); *P. deidamiodes* and *P. contracta* (*Deidamiodes*) and *P. rhamnifolia* (*Astrophea*). These species belong to the *Passiflora* collection kept in our laboratory (Genetics Department, 'Luiz de Queiroz' College of Agriculture, University of São Paulo, Brazil).

The primers described in section 3.2.3.7 were used to amplify the Reverse Transcriptase domain using the total DNA of the wild species as a template; it was extracted from young leaf tissues using the acetyltrimethyl ammonium bromide method adapted from Murray and Thompson (Murray and Thompson, 1980). PCR and gel electrophoresis procedures were as described in section 3.2.3.7.

3.3. Results

3.3.1. Identification of Transposable Elements

The search for transposable elements resulted in the identification of 250 TEs, corresponding to 1,830,620 bp or 17.6% of total sequence data. Class I was the most common, representing 96.4% (241/250) of the elements. The most frequent were LTR-RTs, corresponding to 72.4% (181/250), and accounting for 1,418,389 bp or 13.6% of the data (Table 3).

The DIRS order was found to contain 11 incomplete elements that ranged in length from 7,923 to 13,640 bp, totaling 115,640 bp or 1.1% of the data. The *P. edulis* DIRS has the following particular domains: GAG, MET, PROT, RT, RH and INT. The YR and TIR domains were not identified, although they are typically found in plant DIRS (Figure 4).

Seven elements were classified as LINEs with lengths ranging from 4,030 to 18,937 bp, totaling 54,468 bp or 0.5% of the data. All LINEs found herein were incomplete. Four elements contained an RT-like domain and three elements contained a domain of unknown

function of about 300 nucleotides, typical of LINEs. All these elements had TSDs as expected, with 4 to 5 nucleotides. Two SINE elements were identified, corresponding to only 0.01% of the data. No conserved domains were identified, even though TSDs were found in these two elements.

LARDs are non-autonomous retrotransposons derived from LTR-RTs. The LARD order is represented in *P. edulis* sequences by 36 elements with lengths ranging from 1,816 to 12,505 bp. These LARDs accounted for 189,161 bp or 1.8% of the data, and included 22 with LTRs, and 129 up to 1,964 bp in length. Thirty-five elements were found to contain TSDs 4 to 6 bp long. Additionally, four non-autonomous TRIM elements were identified, corresponding to only 0.05% of the data. No conserved domains were identified, but TSDs 4 to 6 bp in length were found in all elements. Other Class I orders were poorly represented.

Only 3.6% (9/250) of the TEs were classified as belonging Class II, the majority (6) in the TIR order (Table 3). A GT-1 domain was found in two TIRs and DDE motifs in three TIRs. The GT-1 protein has DNA-binding transcription factor activity and DDE is an endonuclease motif involved in DNA transposition. One TIR element exhibited peptidase motifs, and the other had no conserved domains. Only two elements formed TSDs with 4 bp. We were also able to identify two Helitrons and only one non-autonomous MITE. Helitron elements have a TIR-domain (Toll-Interleukin-1 Receptor) that is involved in signaling processes and Leucine-rich Repeats (LRR).

The majority of elements (70.8%, 177/250) were incomplete. In addition, 22.4% (56/250) were potentially chimeric. TEs were preferentially located in intergenic spaces (70.4%, 176/250), but 74 were observed overlapping genes (70 exonic and only four intronic sequences). Additionally, some elements were found within another transposable element, thus creating a TE cluster (Figure 5).

Table 3. Classification and abundance of transposable elements identified in a gene-rich fraction of *Passiflora edulis* genome.

Class	Order	Number of copies	Percentage	Sequence length (bp)	Percentage	Percentage of total sequence data
Class I (Retrotransposons)	DIRS total (RYX)	11	4.4	115,640	6.32	1.11
	DIRS incomplete	11				
	DIRS potential chimeric	11				
	LINE total (RIX)	7	2.8	54,468	2.98	0.52
	LINE incomplete	7				
	LINE potential chimeric	6				
	LTR total (RLX)	181	72.4	1,418,389	77.48	13.64
	LTR complete	73				
	LTR incomplete	108				
	LTR potential chimeric	36				
	SINE total (RSX)	2	0.8	614	0.03	0.01
	SINE incomplete	2				
	LARD total (RXX-LARD)	36	14.4	189,161	10.33	1.82
	LARD potential chimeric	2				
	TRIM total (RXX-TRIM)	4	1.6	4,981	0.27	0.05
Class I total		**241**	**96.4**	**1,783,253**	**97.41**	**17.15**
Class II (DNA transposons)	Helitron total (DHX)	2	0.8	13,827	0.76	0.13
	Helitron incomplete	2				
	TIR total (DTX)	6	2.4	32,748	1.79	0.31
	TIR incomplete	6				
	TIR potential chimeric	1				
	MITE total (DXX-MITE)	1	0.4	792	0.04	0.01
Classe II total		**9**	**3.6**	**47,367**	**2.59**	**0.45**
TOTAL		**250**	**100**	**1,830,620**	**100**	**17.60**

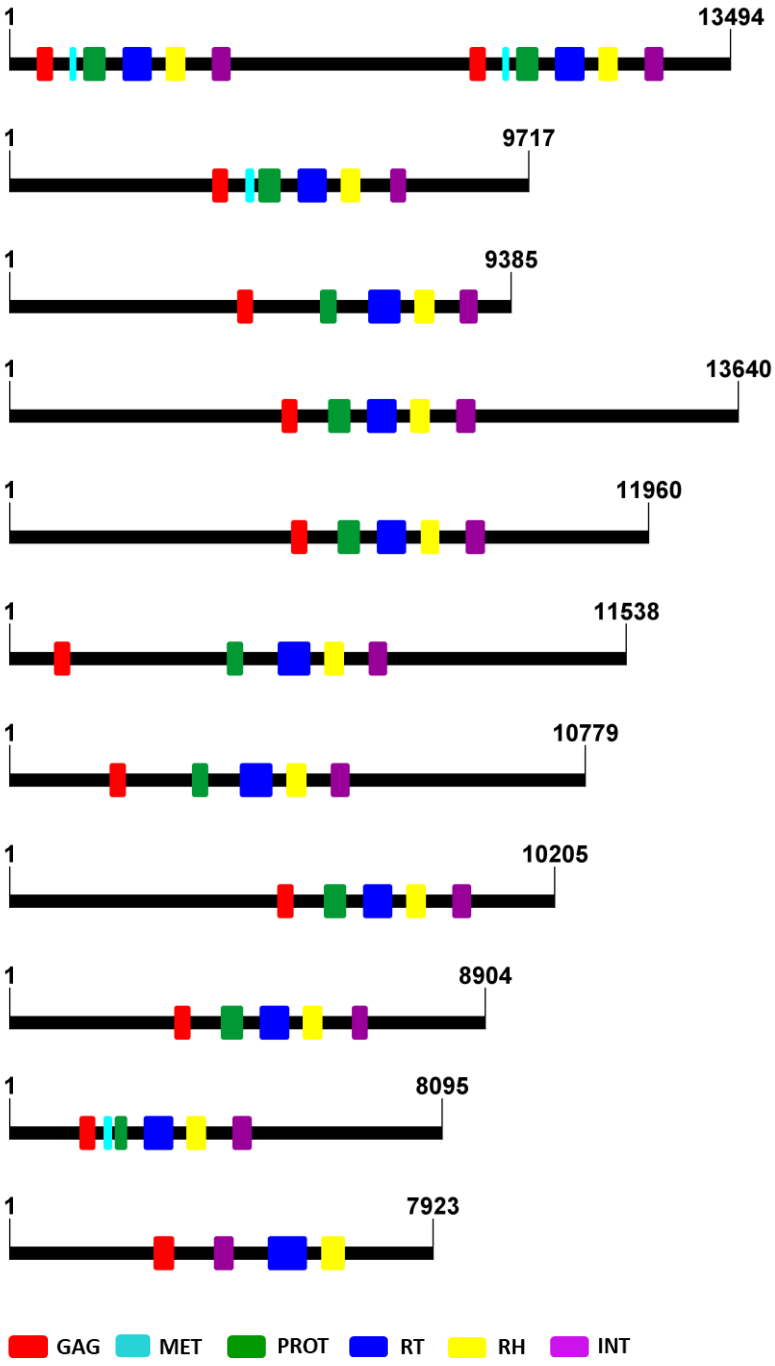

Figure 4. Schematic representation of domains found in 11 DIRS elements from *Passiflora edulis*. Domain abbreviations and color-coding: GAG = gag (red); MET = methyltransferase (light blue); PROT = protease (green); RT = reverse transcriptase (blue); RH = ribonuclease H (yellow); INT = integrase (purple). Numbers on the right are element lengths (in bp).

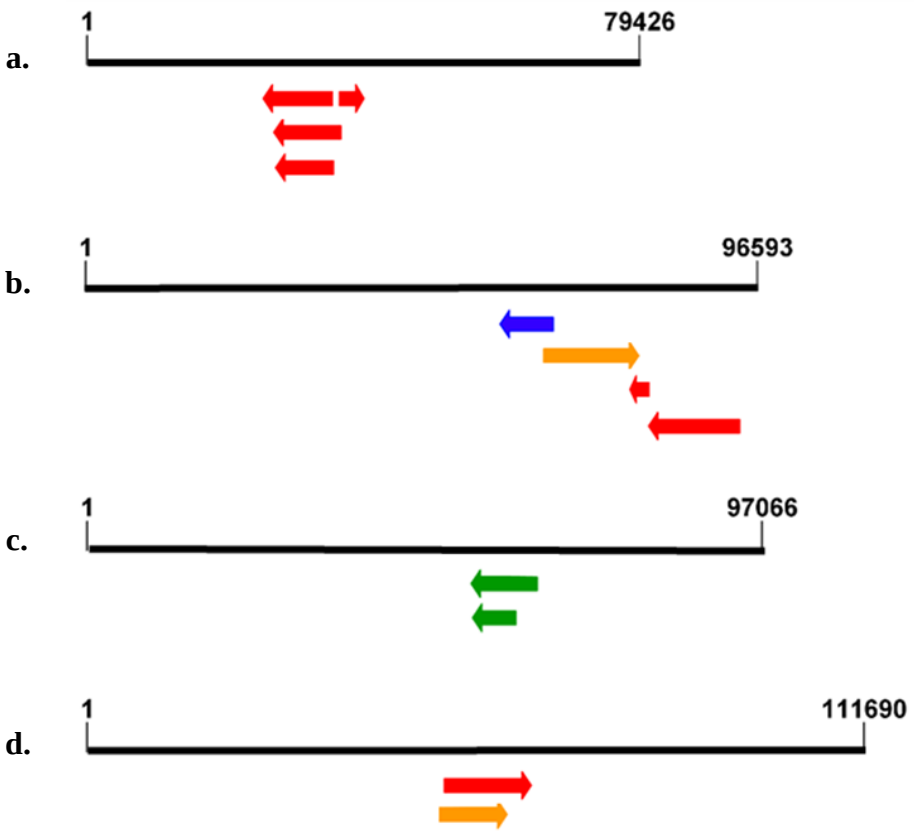

Figure 5. Schematic representation of some TE clusters (arrowed) found in four genomic regions of *Passiflora edulis*. Numbers on the right are genomic region lengths (in bp). Colors indicate different orders, as follows: red (LTR-RT), blue (LINE), orange (DIRS), and green (LARD).

3.3.2. Analysis of LTR-RTs

3.3.2.1. Phylogeny and structural features of the LTR-RTs

A total of 181 LTR-RTs were identified using the REPET pipeline (Flutre et al., 2011; Quesneville et al., 2005). Forty-four were assigned to the *Copia* and 137 to the *Gypsy* superfamilies. These elements were input to Repeat Explorer (Novák et al., 2013) and all their internal domains identified (Figure 6).

Copia and *Gypsy* elements had similar proportions of internal recognizable domains. The majority of *Gypsy* elements (~60%) had up to 4 internal domains. Six *Gypsy* elements had none of the main internal domains, but contained putative sequences of chromodomains, used to classify them as *Gypsy*. All *Copia* elements had at least one domain; the majority (~50%) exhibited five domains (Figure 6a). Taking each domain separately, their frequencies were higher in *Copia* elements than in *Gypsy*, except for the GAG domain, which is prevalent

in *Gypsy* (Figure 6b). This is evidence of the higher level of conservation and potential activity of *Copia* compared to *Gypsy* elements, although in absolute terms, *Gypsy* contained over three times more elements. In terms of the absolute number of LTR-RTs, the GAG domain was the most prevalent in all elements. Chromodomains were also identified in some *Gypsy* elements.

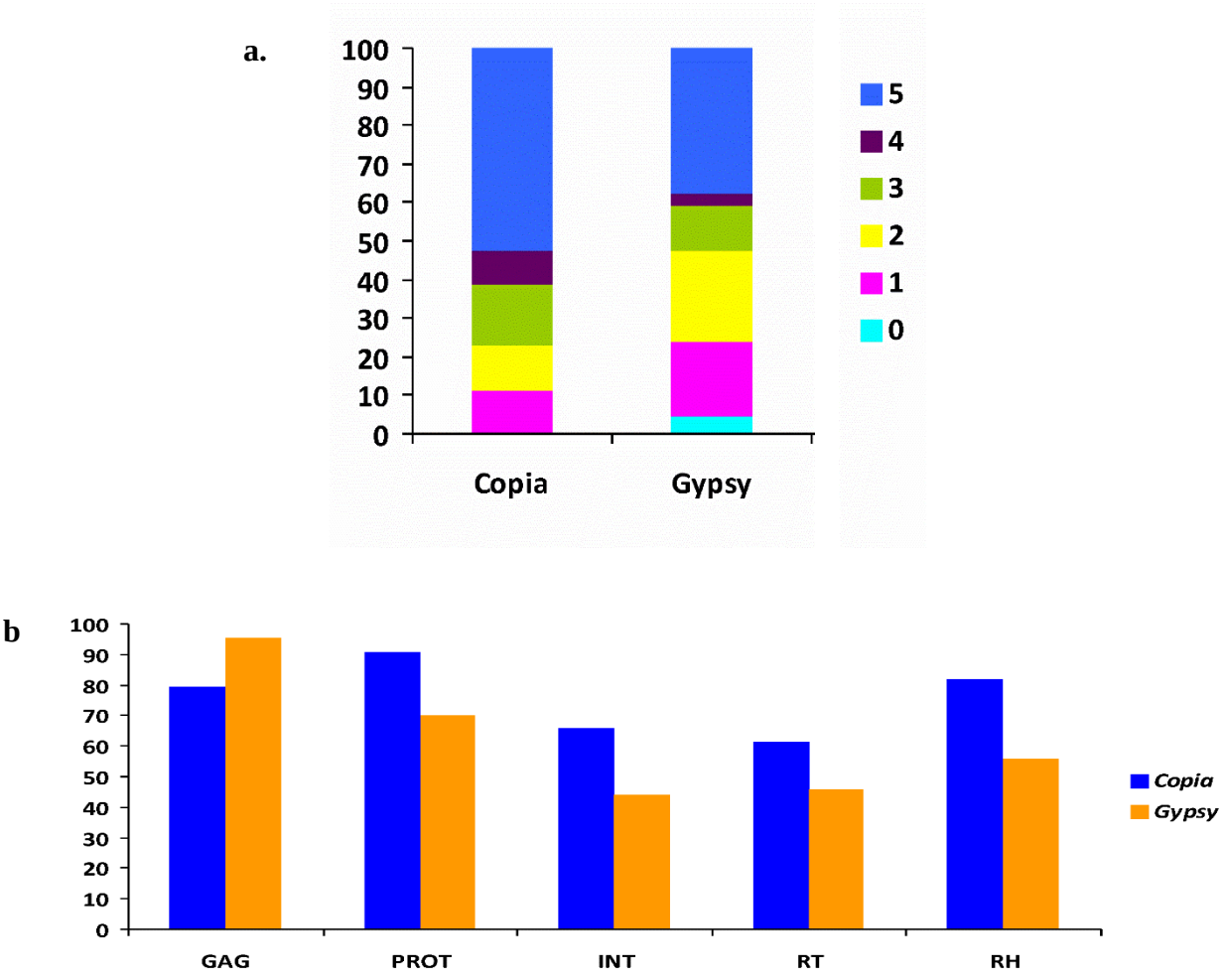

Figure 6. LTR-RT elements of *Passiflora edulis*: a) Percentage of elements from *Copia* and *Gypsy* superfamilies that contain up to 5 domains displayed in different colors; b) Percentage of elements from *Copia* and *Gypsy* superfamilies containing the following internal domains: GAG, PROT, INT, RT and RH.

TSDs were identified in 134 LTR-RTs. TSD length ranged from 4 to 6 bp, although the great majority were 4 bp long. LTRs were recognized in 162 LTR-RTs. In total, 73 LTR-RTs were complete elements (22 from *Copia* and 51 from *Gypsy*), i.e. they contained the five main internal domains, and LTRs at any end of the LTR-RTs.

LTR-RT elements were then classified into *Copia* and *Gypsy* evolutionary lineages, based on the translated RT or GAG domain. Of the total (181), 165 LTR-RTs were classified into evolutionary lineages: 85 were classified using the RT domain and 80 using the GAG domain. Ten LTR-RTs had only one putative domain, half with the PROT domain and half the RH domain. These elements could not be classified into evolutionary lineages, but were classified as *Copia* or *Gypsy*.

Of the resulting 165 LTR-RTs, 37 were assigned to *Copia* and 128 to *Gypsy*. We were able to recognize four plant evolutionary *Copia* lineages (*Angela*, *Ale*, *Tork* and *Sire*) and five *Gypsy* lineages (*Del*, *Athila*, *Reina*, *CRM* and *Galadriel*). All lineage classifications were supported by bootstrap values over 50. For the RT and the GAG domains, the RLG*_peDel* lineage was supported by bootstrap values of 54 (RT) and 57 (GAG). This lineage is highly diverse within *P. edulis* and this is probably the reason why the bootstrap value is low. In addition, the domains from other characterized lineages used to draw this inference were from phylogenetically distant species (*Legolas* and *Tma* from *Arabidopsis thaliana*, and *Peabody* from *Pisum sativum*). The RLG*_peReina* lineage was also supported by a low bootstrap value (54), and the domains from other characterized lineages used to draw this inference were also from distant species (*Gloin* and *Gimli* from *Arabidopsis thaliana*, and *Ifg7* from *Pinus sp)*.

All lineages were monophyletic, except for RLC*_peAle*. This lineage had only one element and we were not able to match it to one specific lineage. This element could belong to RLC*_peTork* or RLC*_peAle*, and was considered to belong to the *Ale/Tork* lineage (Figures 7 to 10; Table 4).

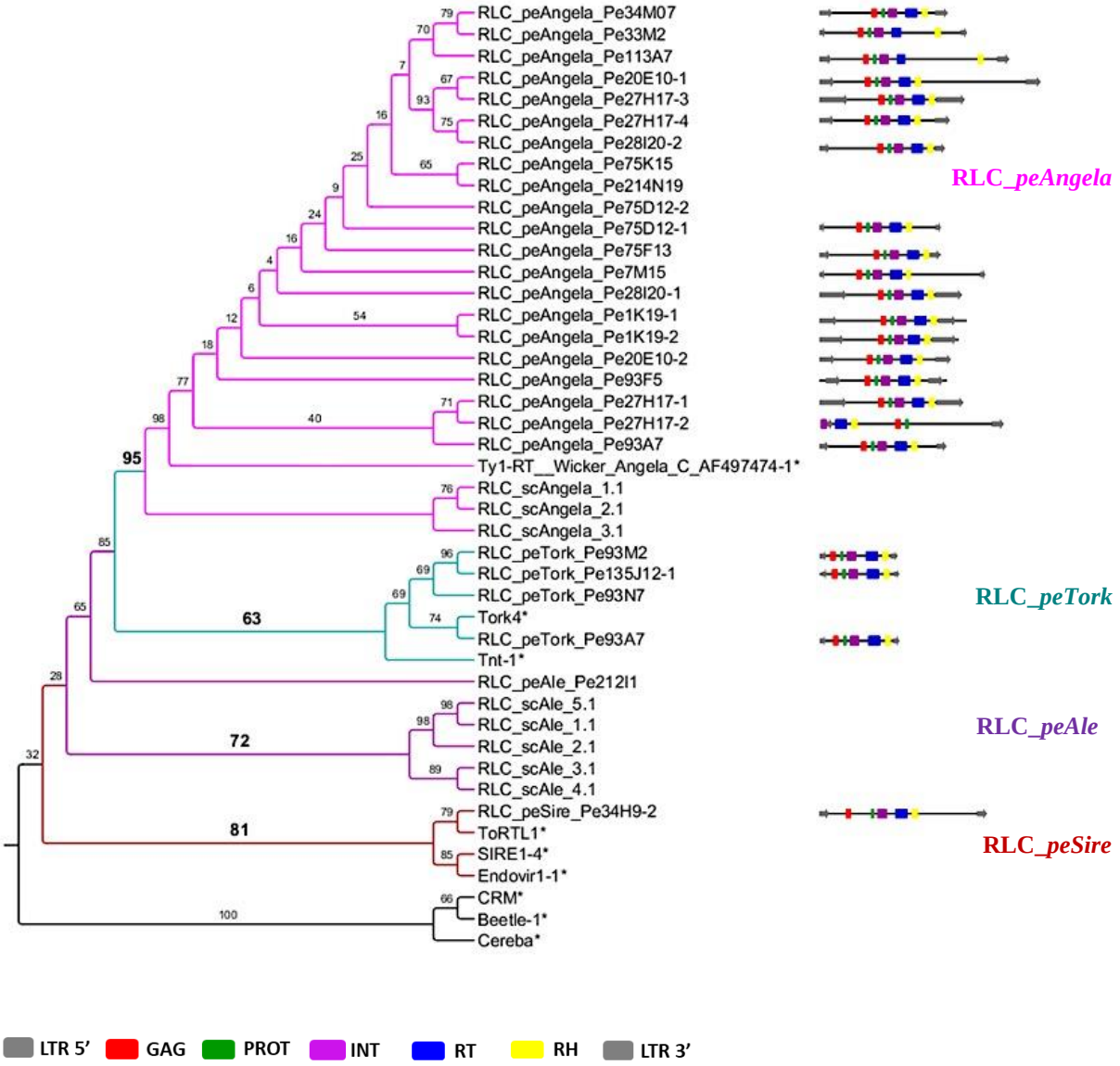

Figure 7. Phylogenetic tree of *Copia* lineages inferred from the complete amino acid sequence of the Reverse Transcriptase domain. Phylogenetic inferences were drawn based on Maximum Likelihood (1,000 bootstraps). The bootstrap values for each node are indicated in bold type. Sequences from the Gypsy database are indicated by an asterisk. *Angela* and *Ale* are sugarcane-characterized lineages. Names and colored blocks indicate lineages. The *CRM* lineage from the *Gypsy* superfamily was used as an outgroup to produce a rooted tree. A schematic representation of full-length elements is shown on the right. Abbreviations and color-coding of domains: LTR = long terminal repeat (grey); GAG = gag (red); PROT = protease (green); INT = integrase (purple); RT = reverse transcriptase (blue); RH = ribonuclease H (yellow).

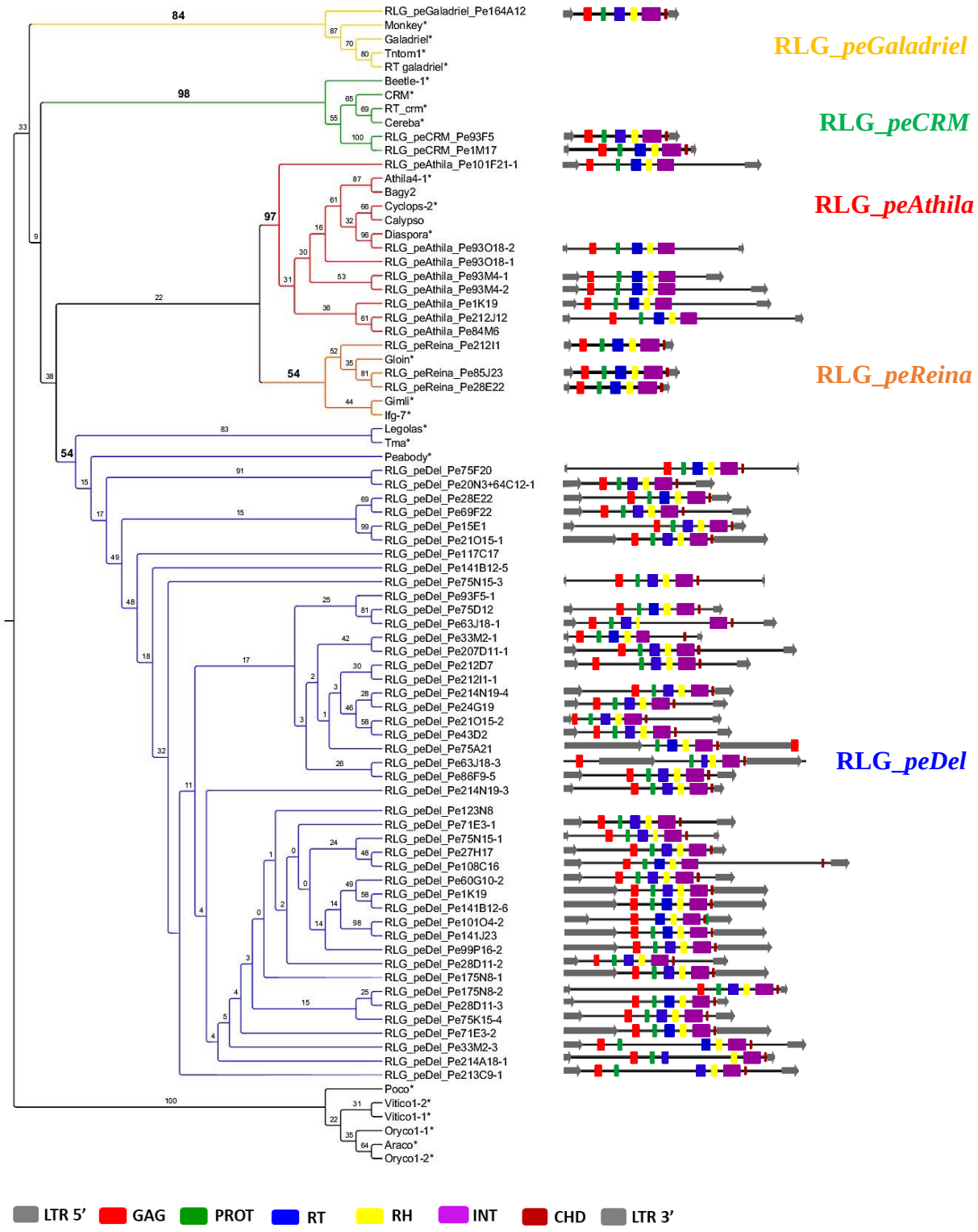

Figure 8. Phylogenetic tree of *Gypsy* lineages inferred from the complete amino acid sequence of the Reverse Transcriptase domain. Phylogenetic inferences were drawn based on Maximum Likelihood (1,000 bootstraps). The bootstrap values for each lineage node are indicated in bold type. Sequences from the Gypsy database are indicated by an asterisk. Names and colored blocks indicate lineages. The *Oryco* lineage from the *Copia* superfamily was used as the outgroup to produce a rooted tree. A schematic representation of full-length elements is shown on the right. Abbreviations and color-coding of domains: LTR = long terminal repeat (grey); GAG = gag (red); PROT = protease (green); RT = reverse transcriptase (blue); RH = ribonuclease H (yellow); INT = integrase (purple); CHD = chromodomain.

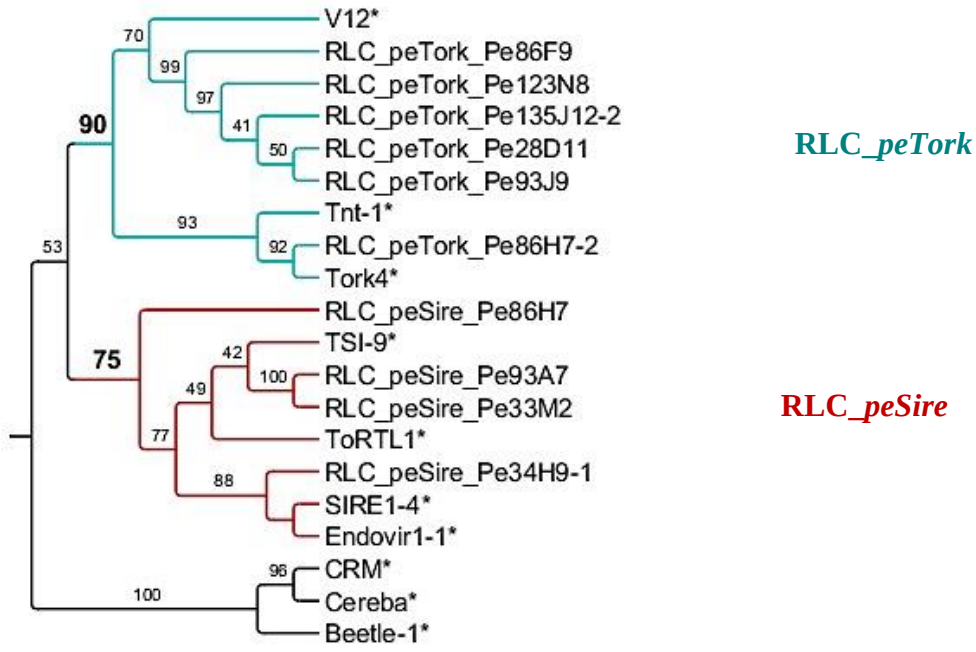

Figure 9. Phylogenetic tree of *Copia* lineages inferred from the complete amino acid sequence of the GAG domain. Phylogenetic inferences were drawn based on Maximum Likelihood (1,000 bootstraps). The bootstrap values for each lineage node are indicated in bold type. Sequences from the Gypsy database are indicated by an asterisk. Names and colored blocks indicate lineages. The *CRM* lineage from the *Gypsy* superfamily was used as the outgroup to produce a rooted tree.

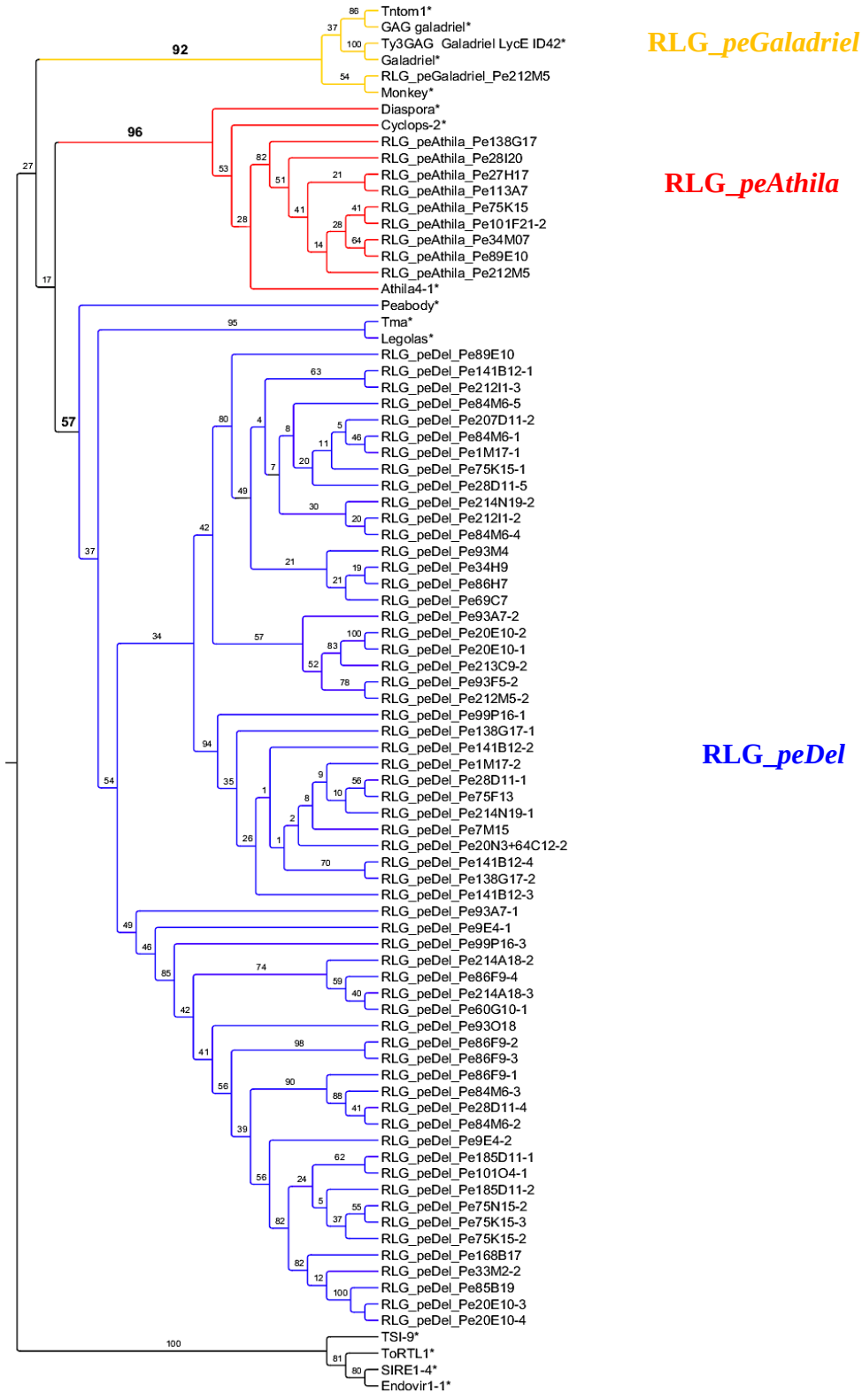

Figure 10. Phylogenetic tree of Gypsy lineages inferred from the complete amino acid sequence of the GAG domain. Phylogenetic inferences were drawn based on Maximum Likelihood (1,000 bootstraps). The bootstrap values for each lineage node are indicated in bold type. Sequences from the Gypsy database are indicated by an asterisk. Names and colored blocks indicate lineages. The Sire lineage from the Copia superfamily was used as the outgroup to produce a rooted tree.

The *Angela* lineage was the most abundant in the *Copia* superfamily and the *Del* lineage in *Gypsy*. RLC_*peAngela* was the only lineage with a higher proportion of complete elements than incomplete elements (Table 4, Figure 11).

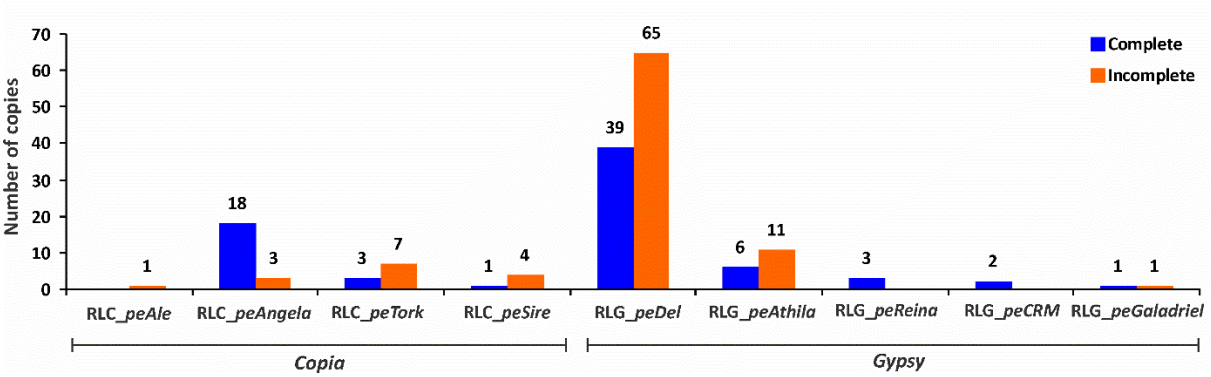

Figure 11. Number of complete and incomplete copies of LTR-RTs in nine lineages of the *Copia* and *Gypsy* superfamilies.

Overall, the lengths of the elements varied within the superfamilies. For *Copia*, RLC_*peAngela* element length ranged from 5.5 to 14.3 kb, with an average of ~9.7 kb. RLC_*peTork* exhibited elements with median lengths from 2.2 to 6.2 kb, with an average of ~4.2 kb. For *Gypsy*, the larger lineage (RLG_*peDel*) exhibited elements ranging in length from 0.6 to 13.3 kb, with an average of ~8.3 kb. RLG_*peAthila* element length ranged from 1.0 to 13.7 kb, with an average of ~7.7 kb (Table 4).

LTR sequence lengths were also variable overall but in terms of the lineages, there was no association between the lengths of elements and LTRs. Differences in total lengths were probably due to differences in the spacer regions between the internal coding domains and LTRs.

Table 4. General features of *Passiflora edulis* LTR-RT lineages in *Copia* and *Gypsy* superfamilies.

Superfamily/Lineage	Element length range (bp)	LTR length range (bp)	Number of sequences
Copia			
RLC_*peAngela*	5,688−14,300	316−1831	21
RLC_*peAle*	4640	285−288	1
RLC_*peTork*	2,267−6,289	147−987	10
RLC_*peSire*	2,108−13,073	102−1270	5
Total *Copia*			37
Gypsy			
RLG_*peDel*	669−15,362	103−4096	104
RLG_*peAthila*	1,033−13,706	290−1401	17
RLG_*peReina*	5,224−5,684	293−475	3
RLG_*peCRM*	6,529−6,888	255−460	2
RLG_*peGaladriel*	1,624−5,708	513	2
Total *Gypsy*			128
TOTAL			**165**

The median lengths of GAG, PROT and RH domains were similar within the superfamilies for the most representative lineages (RLC_*peAngela*, RLC_*peTork* and RLC_*peSire* in *Copia*, and RLG_*peDel* and RLG_*peAthila* in *Gypsy*). However, the INT domain was significantly longer in *Gypsy* elements, while the RT domain was longer in *Copia* elements (Figure 12).

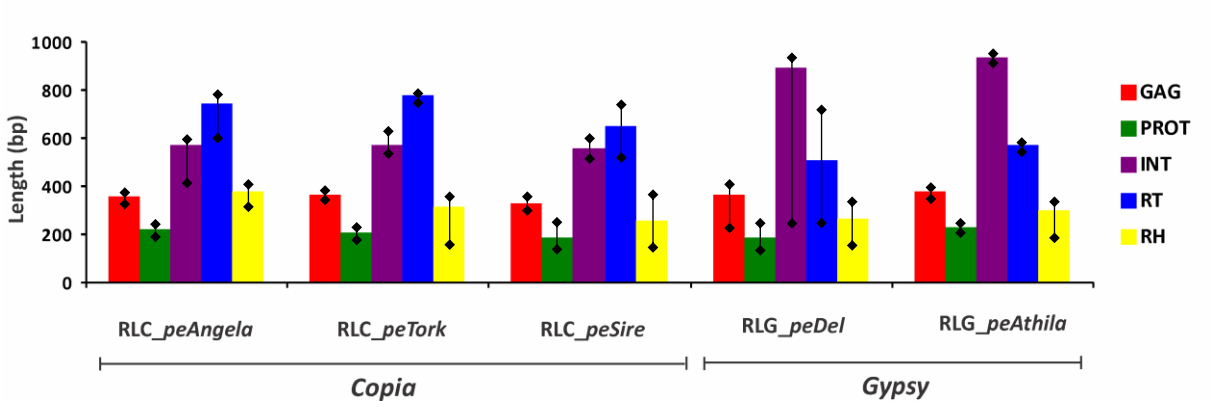

Figure 12. Median sequence lengths of five LTR-RT domains in the most representative lineages of the *Copia* and *Gypsy* superfamilies. Bars indicate the lengths of each domain.

3.3.2.2. Estimation of LTR-RT insertion time

We estimated the insertion time of all 73 LTR-RTs with intact copies e.g. five internal coding domains (GAG, PROT, INT, RT and RH) and both LTRs (Figure 13). The majority (74.3%, 29/39) of the RLG*_peDel* copies were recently inserted into the *P. edulis* genome (< 1.0 Mya), but insertion of three other copies was dated at > 2.0 Mya. RLG*_peReina,* RLG*_peGaladriel* and RLG*_peCRM*, with 3, 1, and 2 copies respectively, had a similar insertion time pattern (< 1.0 Mya); interestingly, one copy of RLG*_peReina* had an estimated age of 0.0 Mya. RLG*_peAthila* copies (6) were inserted less than 1.5 Mya.

All RLC*_peAngela* copies (18) were inserted into the genome < 2.0 Mya, and one of them was found to be dated at 0.0 Mya. One of the two copies of RLC*_peTork* lineage was inserted into the genome up to 0.5 Mya, and the other between 1.5 and 2.0 Mya. The RLC*_peSire* copy had an estimated age of between 1.5 and 2.0 million years.

The full-length LTR-RTs are new. They were found to be recently inserted into *P. edulis* genome, most (95.9%, 70/73) < 2.0 Mya. These observations provide further evidence of their transcriptional activity.

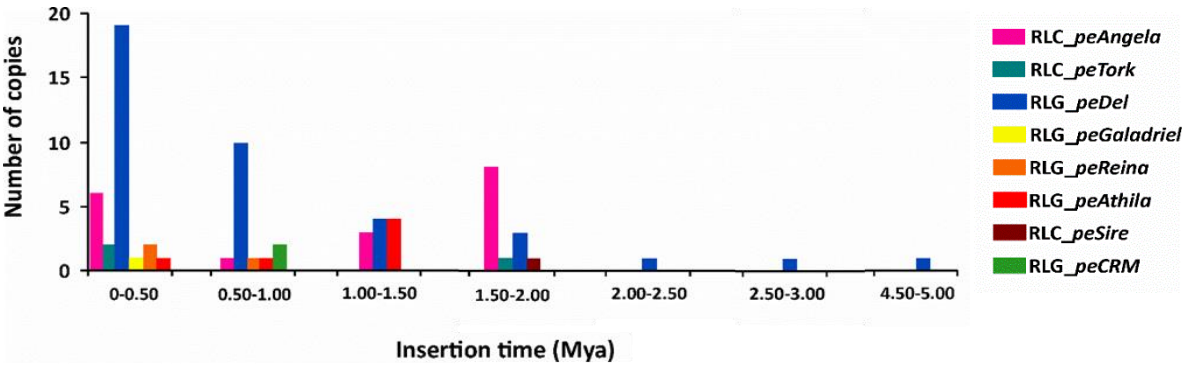

Figure 13. Estimated insertion times of 73 full-length LTR-RT lineages of *Passiflora edulis*.

3.3.2.3. LTR-RT *in silico* transcriptional activity

As mentioned above, our results provide evidence that LRT-RTs are transcriptionally active. All were therefore investigated *in silico*. We were able to associate 2,821 transcripts derived from the RNA-seq libraries (described in section 2.3.6) with full-length LTR-RTs. All

Copia and *Gypsy* lineages were associated with transcripts. The number of associated transcripts was correlated with the number of representative elements from each lineage, i.e. the higher the number of elements representative of the lineage, the higher the number of transcripts associated. Hence, there is no evidence for assuming that lineages have different expression levels. The largest number of transcripts was associated with the RLG_*peDel* lineage, probably because this lineage has more elements (Figure 14). In addition, the highest number of transcripts was identified in the RNA-seq library prepared from vegetative adult plants (Figure 15).

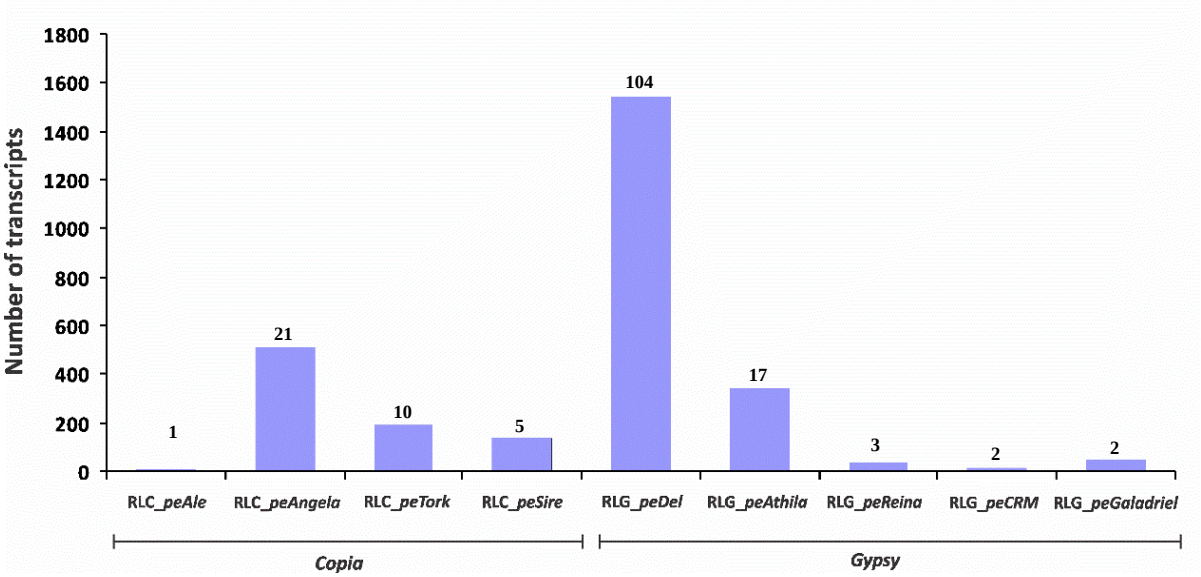

Figure 14. Number of transcripts assigned to LTR-RT lineages of the *Copia* and *Gypsy* superfamilies. The number of representative elements is indicated at the top of each bar.

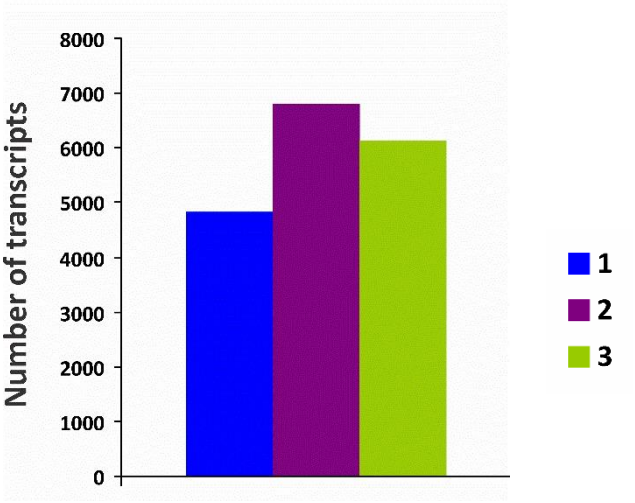

Figure 15. Number of normalized transcripts assigned to LTR-RT lineages. The transcripts were obtained from three *Passiflora edulis* RNA-seq libraries of shoot apexes of juvenile (1), vegetative (2) and reproductive adult plants (3) (Dornelas M.C., unpublished data).

3.3.2.4. RNA extraction and Reverse Transcriptase (RT) PCR analysis

Based on our previous results, an RT-PCR analysis was performed in order to confirm the transcriptional activity of some LTR-RTs. The youngest LTR-RT from each lineage was selected based on estimated insertion time. The following elements were selected:

RLC_*peAngela*_Pe93F5 (0.0 Mya)

RLC_*peTork*_Pe93M2 (0.07 Mya)

RLG_*peAthila*_Pe93M4-1 (0.5 Mya)

RLG_*peCRM*_Pe1M17 (0.53 Mya)

RLG_*peDel*_Pe99P16-2 (0.03 Mya)

RLG_*peGaladriel*_Pe164A12 (0.4 Mya)

RLG_*peReina*_Pe212I1 (0.0 Mya)

RT-PCR analysis confirmed the transcriptional activity of all elements, except RLG_*peAthila*_Pe93M4-1. As expected, strong PCR bands were detected for RLC_*peAngela*_Pe93F5 and RLG_*peDel*_Pe99P16-2 (Figure 16).

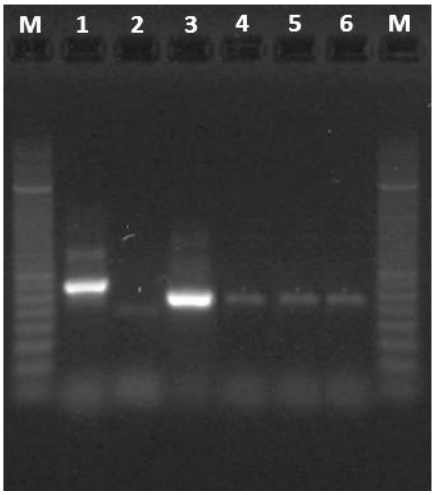

Figure 16. Agarose (2%) gel electrophoresis of RT-PCR products from the selected transposable elements: RLC_*peAngela*_Pe93F5 (1), RLC_*peTork*_Pe93M2 (2), RLG_*peDel*_Pe99P16-2 (3), RLG_*peCRM*_Pe1M17 (4), RLG_*peGaladriel*_Pe164A12 (5) and RLG_*peReina*_Pe212I1 (6), using cDNA templates of *P. edulis*. M, 100 bp ladder (Invitrogen).

3.3.2.5. LTR-RTs from wild *Passiflora* species

The presence of the selected *P. edulis* TEs (RLC_*peAngela*_Pe93F5, RLC_*peTork*_Pe93M2, RLG_*peAthila*_Pe93M4-1, RLG_*peCRM*_Pe1M17, RLG_*peDel*_Pe99P16-2, RLG_*peGaladriel*_Pe164A12 and RLG_*peReina*_Pe212I1) was tested in the following 8 wild species: *P. edimundoi*, *P. setacea*, *P. alata* (subgenus *Passiflora*); *P. organensis*, *P. capsularis* (subgenus *Decaloba*); *P. deidamiodes*, *P. contracta* (subgenus *Deidamiodes*); *P. rhamnifolia* (subgenus *Astrophea*) (Figure 17).

RLC_*peAngela*_Pe93F5, RLC_*peTork*_Pe93M2, RLG_*peCRM*_Pe1M17 and RLG_*peDel*_Pe99P16-2 were identified in all species tested.

RLG_*peAthila*_Pe93M4-1 was identified in *P. alata* and *P. setacea*, and RLG_*peGaladriel*_Pe164A12 was identified only in *P. setacea*.

RLG_*peReina*_Pe212I1 was not present in any of the species.

P. capsularis (*Decaloba*) was the only species where no amplification was detected.

In general, the *Angela*, *Del*, *CRM* and *Tork* lineages may be conserved in *Passiflora* genus, since representative elements were identified in all species. The element from *CRM* lineage showed a weak PCR band, probably because this lineage is less represented in these species compared to the others with stronger PCR bands.

The LTR-RTs from the *Athila* and *Galadriel* lineages were found only in the *Passiflora* subgenus. These lineages are probably less conserved through the *Passiflora* genus and could be less represented in these species.

It seems that elements from *Reina* lineage are not conserved through the *Passiflora* genus.

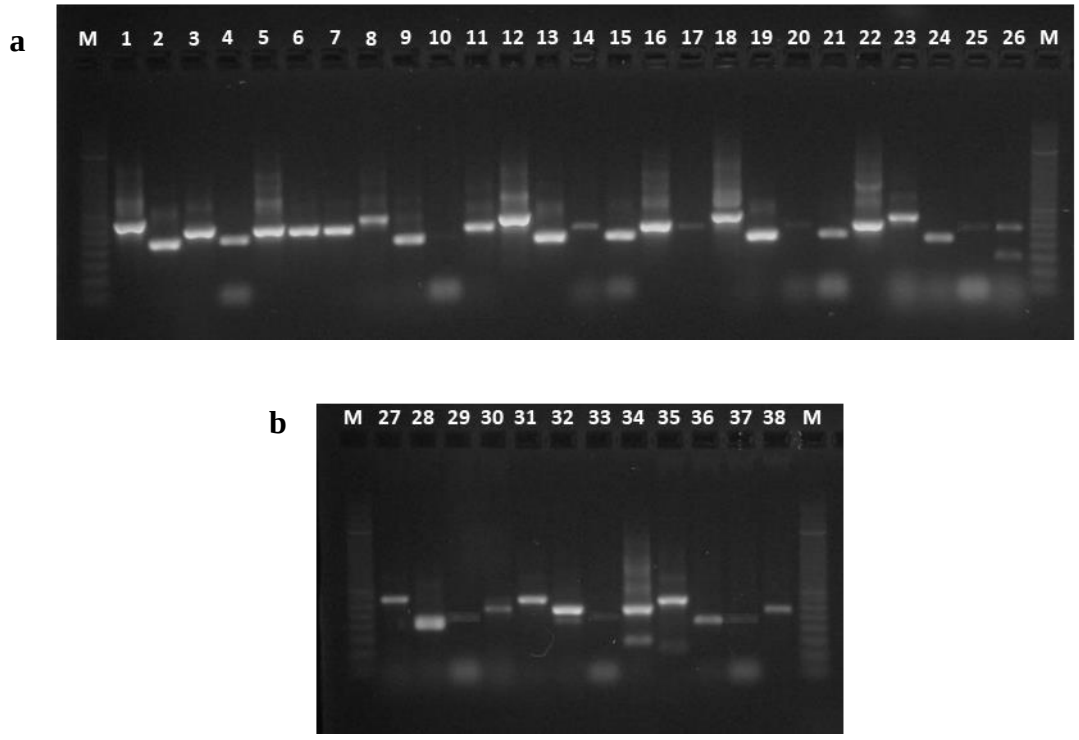

Figure 17. Agarose (2%) gel electrophoresis of PCR products from the following elements: RLC_*peAngela*_Pe93F5 (1), RLC_*peTork*_Pe93M2 (2), RLG_*peAthila*_Pe93M4-1 (3), RLG_*peCRM*_Pe1M17 (4), RLG_*peDel*_Pe99P16-2 (5), RLG_*peGaladriel*_Pe164A12 (6) and RLG_*peReina*_Pe212I1 (7). M: 100 bp ladder (Invitrogen). a. Line 1-7, *P. edulis* (TEs 1-7); 8-11, *P. edimundoi* (TEs 1, 2, 4 and 5); 12-17, *P. setacea* (TEs 1, 2, 3, 4, 5 and 6); 18-22, *P. alata* (TEs 1, 2, 3, 4 and 5); 23-26, *P. organensis* (TEs 1, 2, 4 and 5). b. Line 27-30, *P. deidamioides* (TEs 1, 2, 4 and 5); 31-34, *P. contracta* (TEs 1, 2, 4 and 5); 35-38, *P. rhamnifolia* (TEs 1, 2, 4 and 5).

3.4. Discussion

We analyzed the repetitive mobile elements in a gene-rich fraction of the *Passiflora edulis* genome using large sequences retrieved from a BAC library. This provided the first complete overview of the content and structure of these elements in *Passiflora*.

This gene-rich fraction (~10 Mb or 0.85% of the *P. edulis* genome) consists of 17.6% (~1.8 Mb) corresponding to TEs (transposable elements). Santos et al. (2014) sequenced

10,000 BES (BAC-end sequence) from the same BAC library (~6.2 Mb or 0.39% of the *P. edulis* genome) and were able to identify reads likely to contain TEs in 18.54% of all BES. Although these studies analyzed two different genomic regions, the gene-rich region (44.0% containing coding sequences) and the gene-poor region (9.6% containing coding sequences) both have very similar proportions of TEs.

The abundance of TEs in plant genomes is highly variable, and the most abundant repeats occur in plant families. For example, TEs constitute a respective 18.3%, 21.5% and 42.4% of the entirely sequenced genomes of *Medicago truncatula* (Fabaceae, Young et al., 2011), *Vitis vinifera* (Vitaceae, Jaillon et al., 2007) and *Malus* × *domestica* (Rosaceae, Velasco et al., 2010), all representative dicot species. This wide variation also occurs in representative monocots. For example, a respective 28.1%, 79.8% and 85.0% was reported in the following Poaceae species: *Brachypodium distachyon* (Vogel et al., 2010), *Triticum aestivum* (Mayer et al., 2014) and *Zea mays* (Schnable et al., 2009). In *Musa acuminata* (Musaceae, D'Hont et al., 2012), 43.7% of the genome consists of TEs and in *Ananas comosus* (Bromeliaceae, Ming et al., 2015) the proportion is 44.5%.

In completely sequenced genomes of Malpighiales (the taxonomic order to which *Passiflora* belongs), the TE content is approximately the same as in *Linum usitatissimun* (23.0%, Gonzalez and Deyholos, 2012) and *Manihot esculenta* (24.4%, Wang et al., 2014), but it is much higher in *Hevea brasiliensis* (71%, Tang et al., 2016), *Ricinus communis* (50.3%, Chan et al., 2010), *Jatropha curcas* (45.9%, Wu et al., 2015) and *Populus trichocarpa* (~42.0%, Tuskan et al., 2006). This variation is a consequence of species' genome evolution, which affects genome size (Civáň et al., 2011; Lee and Kim, 2014), species-specific factors that influence TE amplification and repression (Hua-van et al., 2011), and polyploidization events (Galindo-González et al., 2017; Vicient and Casacuberta, 2017). In addition, polyploidization and TE amplification greatly influence one another, boosting their potential to drive plant genome evolution (Vicient and Casacuberta, 2017).

Class I is predominant in flowering plants, the majority belonging to the LTR order (see Oliver et al., 2013). In our data, 97.4% of all TEs are from Class I elements, ~77.5% of which are composed of LTR retroelements. Using BES data as a major resource, Santos et al. (2014) also identified a large proportion of Class I TEs (94.1%) in *P. edulis*, 98.0% of which belong to the LTR order.

Similarly, LTR-RTs are predominant in representative Dicotyledons. For example, these elements constitute a respective 92.3%, 90.2% and 88.7% in *M. truncatula*, *V. vinifera* and *M. domestica*. This also occurs in representative monocots: for example, respective

figures of 97%, 83.2%, 82.9%, 79.8% and 71.1% have been reported in *M. acuminata*, , *Z. mays, B. distachyon, T. aestivum* and *A. comosus*.

In Malpighiales species, LTR-RTs are the most abundant (74.6% in *Linum*, 74.2% in *Jatropha*, 71.7 in *Hevea*, 60% in *Populus* and 45.6% in *Mahihot*). In *Ricinus*, only 32.2% of TEs belong to the LTR order, but a large portion of retrotransposons have yet to be classified.

Eleven DIRS elements were identified in *P. edulis*, corresponding to 1.1% of total sequence data. All these elements were incomplete and surprisingly no YR or TIR domains were identified in any of them. However, MET domains were identified in three elements. DIRS was first described in the slime mold *Dictyostelium discoideum* (Cappello et al., 1985). It seems that DIRS-like retrotransposons are widespread in eukaryotes and have been reported in 61 species, including the green algae *Chlamydomonas reinhardtii* and *Volvox carteri* (Piednoël et al., 2011). Some DIRS elements are well-described only in vertebrates and fungi, (see Poulter and Butler, 2015).

In higher plants, there are only a few reports of the presence of DIRS elements and there are no detailed studies. Only two DIRS were reported in *L. usitatissimun* (Gonzalez and Deyholos, 2012), five in *P. trichocarpa* and three in *P. euphratica* (Yi et al., 2018). In *Elaeis guineensis* and *E. oleifera*, 3% of the TEs were classified as DIRS. DIRS seem to constitute only a minor fraction of TEs in angiosperms. However, there is a considerable unclassified fraction of TEs in some reported genomes, and these elements can be very diverse and mutated, which makes it difficult to identify and properly characterize them. Detailed descriptions of DIRS in plants could help elucidate their roles in eukaryote evolution.

Seven incomplete LINEs were identified, corresponding to only 0.5%. LINEs are found in all eukaryotic kingdoms (Wicker et al., 2007), but are predominant in animal genomes (Lee and Kim, 2014). There are some reports of LINEs in plants. In Salicaceous, 87, 112 and 81 LINEs were identified in *P. trichocarpa*, *Salix shuchowensis* and *P. euphratica* respectively (Yi et al., 2018). Interestingly, *Del-2*, a LINE-like element, is highly abundant in *Lilium* species, with 240,000 copies constituting some 4% of the *L. speciosum* genome (Leeton and Smyth, 1993; Smyth, 1991).

Only two SINEs were identified and they had no conserved domains. SINEs are also predominant in animal genomes and there are very few reports of these elements in plants (Gonzalez and Deyholos, 2012; Lee and Kim, 2014; Oliver et al., 2013). The Large Retrotransposon Derivatives (LARD) order was represented by 36 elements, accounting for 1.8% of *P. edulis* sequence data, and only two TRIMs were identified, with no conserved domains. LARDs were one of the first TEs to be described. They have been characterized in

some species in which well-known TE families are described, such as *Sukkula* in Triticaceae, but mostly in *Hordeum* spp. and *T. aestivum* (Kalendar et al., 2004), and *Dasheng*, *Spit* and *Squid* in rice (Jiang et al., 2002; Vitte et al., 2007). In flax (*L. usitatissimun*), at least some of the larger LTR elements could be classified as LARDs and some of the shorter-than-expected LTR retrotransposons are probably TRIMs (Gonzalez and Deyholos, 2012). Cossu et al. (2012) studied the dynamics of LTR-RTs in *P. trichocarpa* and reported that a large number of unknown elements probably belong to the LARD or TRIM orders.

Only 2.6% of *P. edulis* mobile elements were classified as DNA transposons, which are generally less abundant in plant genomes (2.1% in *M. domestica*, 6.5% in *V. vinifera*, 7.6% in *M. truncatula*, 17.1% in *B. distachyon* and 18.7% in *T. aestivum*) with little-known exceptions, such as *Arabidopsis thaliana* (59.4%, Oliver et al., 2013). In terms of species related to *P. edulis*, the proportion of DNA transposons is 2.7% in *M. esculenta* (Wang et al., 2014), 1.8% in *R. communis* (Chan et al., 2010) and 15.7% in *L. usitatissimun* (Gonzalez and Deyholos, 2012).

We also report herein the occurrence of gene fragments in some *P. edulis* DNA transposons. In fact, although less abundant in plant genomes, they tend to be associated or interact with the host genes, and there are some reports of the acquisition of genes or gene fragments by DNA transposons (Dooner and Weil, 2007; Gupta et al., 2005; Morgante et al., 2005). A lot of adaptive or evolutionary potential in angiosperms is indeed due to the activity of TEs, including DNA-TEs, resulting in an extraordinary array of genetic changes, including gene modifications, duplications and the creation of novel genes (Oliver et al., 2013).

The majority of the *P. edulis* TEs (70.8%) were incomplete, corroborating the findings of Cruz et al. (2014) who stated that most TE copies are either defective or fossilized. In actual face, TEs are frequently recognized as genomic fossils that were once autonomous, but at some point in time experienced a deletion, inversion or other type of mutation that rendered them inactive. However, a non-autonomous TE can remain active using the enzymatic machinery required for transposition provided by an autonomous partner (Piskurek and Jackson, 2012; Zhao and Ma, 2013), such as the *Dasheng* (non-autonomous) and *RIRE2* (autonomous) elements in rice (Jiang et al., 2002).

There were cases of TEs confined to clusters within the genome. When TEs accumulate, even if the insertions tend to occur randomly, this is more likely to occur within another transposable element. Furthermore, these insertions are usually selectively neutral, leaving TEs free to accumulate in clusters (Hua-van et al., 2011). On the other hand, TEs are preferentially located in intergenic spaces (Vicient and Casacuberta, 2017; Zhao and Ma,

2013), but there are reports of transposable elements in the neighborhood of genes or regulatory regions (Contreras et al., 2015) and this is consistent with our findings, with a considerable portion of TEs (17.6%) in the *P. edulis* gene-rich fraction.

There are several reports regarding the influence of transposable elements on genome regulation and evolution. When inserted inside, overlapping or near genes, they can be potent mutagenic and regulatory agents, capable of disrupting gene function and producing alternative splicing and chimeric TE-gene sections (Galindo-González et al., 2017). In addition, genome size and structure is largely determined by TEs (Oliver et al., 2013).

Genome size and TE content varies in Malpighiales species with completely sequenced genomes. For instance, the genome of *Hevea brasiliensis* is estimated to be 1,470 Mb in size and 71% of it is made up of mobile elements (Tang et al., 2016). These figures are 742 Mb and 24.4% in *Manihot esculenta* (Wang et al., 2014); 485 Mb and 42.0% in *Populus trichocarpa* (Tuskan et al., 2006); 416 Mb and 45.9% in *Jatropha curcas* (Wu et al., 2015); 373 Mb and ~23% in *Linum usitatissimun* (Gonzalez and Deyholos, 2012) and 320 Mb and 50.0% in *Ricinus communis* (Chan et al., 2010). The genome size of *P. edulis* is estimated at ~1,230 Mb (Yotoko et al., 2011) and we found that the TE content in a small genome fraction (~1.0%) is almost 18%, very similar to the figures for *L. usitatissimun* and *M. esculenta*.

P. edulis has the second largest genome of these Malpighiales, similar in size to that of *Hevea*, which consists of 71% TEs. There is a positive correlation between genome size and TE content in angiosperms, especially those with large genomes (Civáň et al., 2011; Lee and Kim, 2014). We therefore think that the proportion of TEs is even higher in the *P. edulis* genome as a whole. The same applies to *Hevea*. In addition, compared to other *Passiflora* species, *P. edulis* has one of the largest genomes (Yotoko et al., 2011). We therefore believe that *P. edulis* has undergone a duplication event that could explain its genome size (Munhoz et al., 2018), mediated by TE proliferation along evolutionary time. This hypothesis is in line with many reports regarding the contribution of mobile elements to the expansion of plant genomes (Lee and Kim, 2014; Park et al., 2012).

It is very surprising that this kind of gene-rich fraction has a TE content of approximately 18%, the majority LTR-RTs. TEs in gene-rich fractions of other genomes have been reported. For instance, in *Linum* the location of TEs was not completely random and some regions had equal coverage of genes and TEs, with a predominance of certain superfamilies (Gonzalez and Deyholos, 2012). As far as we know, all mobile elements influence genome structure and function. For instance, TEs close to genes can become positive regulators of gene expression, or may become targets for epigenetic silencing,

affecting adjacent gene regions (Galindo-González et al., 2017). This led us to consider the activity of LTR-RTs in the *P. edulis* genome, and we therefore identified and investigated LTR-RTs in detail.

We have already discussed the predominance of LTR-RTs and their roles in the evolution of plant genomes. It is crucial to identify and characterize these elements. However, only a few studies discuss the structure and function of LTR-RTs in detail. Furthermore, most of the TEs identified herein were LTR-RTs (97.4%) corresponding to 13.6% of the sequence data. At superfamily level, 75.7% of them belonged to *Gypsy* and 24.3% to *Copia*. The *Gypsy* superfamily is the most representative in the majority of the plant genomes adequately characterized, such as the dicots *Malus* × *domestica* (82%), *Solanum tuberosum* (80%), *Arabidopsis thaliana* (78.8%), *S. lycopersicum* (75.7%), *Vitis vinifera* (74.4%), and *Glycine max* (70.2%), as well as the monocots *Oryza sativa* (82.7%), *Sorghum bicolor* (78.5%), *Brachypodium distachyon* (76.6%), *Setaria italica* (75.4%), *Triticum aestivum* (71.7%), *Hordeum vulgare* (67.9%) and *Zea mays* (66.2%) (reviewed in Oliver et al., 2013).

In related species of *P. edulis*, the *Gypsy* superfamily was also the most representative, accounting for 85% in *Hevea*, 70.6% in *Ricinus*, 54.5% in *Jatropha* and 52% in *Populus*. On the other hand, *Copia* elements constitute the majority of LTR-RTs in *Linum* (54.1%). In general, *Gypsy* elements are mostly associated with heterochromatic regions and *Copia* elements with euchromatic regions (Cossu et al., 2012; Domingues et al., 2012; Gonzalez and Deyholos, 2012). Nevertheless, in the euchromatic region we examined, *Gypsy* elements were prevalent.

We were able to identify four *Copia* (*Angela*, *Ale*, *Tork* and *Sire*) and five *Gypsy* (*Del*, *Athila*, *Reina*, *CRM* and *Galadriel*) evolutionary lineages. The RLC_*peAngela* lineage was the most representative within *Copia*, and RLG_*peDel* within *Gypsy*. Considering all lineages, RLG_*peDel* was the most abundant (~63.0%, 104/164). The *Angela* lineage has been described in other plant species. In fact, *Angela* was significantly predominant in the genome of *Setaria italica* (Poaceae), corresponding to 28.2% of all complete LTR-RTs (Domingues et al., 2012; Ochoa Cruz et al., 2016). The RLC_*peAngela* lineage is well-conserved in the genome fraction analyzed herein, since most of the *P. edulis* elements were complete (78.26%, 18/23) and had similar structure in terms of total length and internal domain length and organization (Figure 7).

The RLG_*peDel* lineage dominated the population of TEs in the *P. edulis* genome fraction, representing 48.4% of all TEs and 8.5% of total sequence data. We observed that the

RLG_*peDel* lineage is not well-conserved in terms of the number of complete elements (37.5%, 39/104), total length and internal domain organization (Figure 8).

The prevalence of particular families is highly variable among plant species (Ochoa Cruz et al., 2016; Vicient and Casacuberta, 2017). The predominance of particular families has been reported in some plant genomes. It is well known that the expansion of some plant genomes is due to accumulation of few types of elements, showing a clear correlation between genome size and the activity of these highly repeated families (El Baidouri and Panaud, 2013; Vicient and Casacuberta, 2017). One example is the expansion of the *Capsicum annuum* genome through a massive accumulation of single-type *Ty3/Gypsy*-like elements that belong to the *Del* subgroup (Park et al., 2012).

The observation that retrotransposition occurs in all plant genomes, associated with the prevalence of some elements, leads to the assumption that only few families have undergone transpositional bursts in the recent past (El Baidouri and Panaud, 2013; Vicient and Casacuberta, 2017). Retrotransposon replication has determined the structure of eukaryotic genomes and LTR-RT insertion was a frequent occurrence during evolution (Cossu et al., 2012). However, the scales and periods of activity of LTR-RT amplification vary dramatically among lineages, species and families (Du et al., 2010). In addition, some LTR-RT families are preferentially distributed in gene-rich chromosomal arms, indicating that individual families have specific properties (Zhao and Ma, 2013)

In passion fruit, we found that full-length LTR-RT elements became active quite recently. The majority of elements (95.9%, 70/73) were inserted into the genome < 2.0 Mya. In fact, two of these elements had 100% LTR pair similarity and are probably currently influential in the genome, explaining their abundance in this gene-rich portion.

In *Linum*, *Copia* elements appear to have been increasingly and continuously active over the last 5 million years, while *Gypsy* elements have been active for the last 7−8 million years, but to a lesser extent (Gonzalez and Deyholos, 2012). In *Populus*, both *Gypsy* and *Copia* LTR-RTs have been active during the same period, although the mean insertion date of *Copia* full-length elements predates that of *Gypsy* (9.3 vs. 10.3 Mya) (Cossu et al., 2012).

Analyzing *in silico* transcriptional activity showed that all *Copia* and *Gypsy* lineages were associated with transcripts, which is an indication of LTR-RT activity. Based on the association of transcripts to LTR-RT lineages in *L. usitatissimun*, just a small proportion of TEs might be active, most of which were *Copia* LTR elements (Gonzalez and Deyholos, 2012). In *P. trichocarpa* only a small number of families appear to be transcriptionally active, and they are often composed of one or at most two full-length elements (Cossu et al., 2012).

Reverse Transcriptase PCR analysis confirmed the transcriptional activity of the youngest element of each *P. edulis* LTR-RT lineage, excluding RLG*_peAthila*_Pe93M4-1. RLC*_peAngela*_Pe93F5 and RLG*_peDel*_Pe99P16-2 exhibited stronger PCR bands. Apparently, these elements are more expressed than the others. This is plausible since these elements are the most representative elements of the *Copia* and *Gypsy* superfamilies respectively. Amplifications of RLC*_peTork*_Pe93M2, RLG*_peCRM*_Pe1M17, RLG*_peGaladriel*_Pe164A12 and RLG*_peReina*_Pe212I1 exhibited weak PCR bands. These lineages were poorly represented in the genome fraction studied, possibly because they are less represented in the whole genome.

Our results suggest that the RLC*_peAngela* and RLG*_peDel* lineages dominate the gene-rich region of the *P. edulis* genome and our data confirm their activity in the recent past (2.0 Mya), with some possibly still active. These elements might have important implications, acting as positive regulators of gene expression, via their cis-acting elements (promoters found in LTR sequences), or as epigenetic silencers. On the other hand, some of these elements could be undergoing purifying selection to avoid detrimental effects on genome function. More detailed studies on LTR-RTs and their identification and characterization in the whole genome could help elucidate their roles in *P. edulis* genome evolution.

Finally, some *P. edulis* lineages seem to be conserved in wild species of the *Passiflora* genus. Our results show that the RLC*_peAngela*, RLC*_peTork*, RLG*_peCRM* and RLG*_peDel* lineages seem to be found in all four *Passiflora* subgenera (*Passiflora*, *Decaloba*, *Deidamiodes* and *Astrophea*). This could be a result of their shared evolutionary history, i.e. these lineages could have undergone amplifications before the evolutionary differentiation of these *Passiflora* species.

LRT-RTs seem to be better conserved throughout the subgenus *Passiflora* since RLG*_peAthila* and RLG*_peGaladriel* are present only in species of this subgenus (*P. alata* and *P. setacea*). This is plausible since these species belong to the same subgenus of *P. edulis* (Feuillet and Macdougal, 2003). Another explanation is that these lineages might have undergone relatively recent proliferation in the subgenus, since they are absent from the other species of the genus *Passiflora*.

RLG*_peReina* was the only lineage not found in any of the species analyzed, and is possibly a specific lineage of *P. edulis*, due to recent proliferation in *P. edulis* or decay in the other species.

LTR-RT activity varies widely among species. This variation could be largely responsible for variation in genome size and genomic differentiation among related species

(Zhao and Ma, 2013). Comparisons of closely related species are therefore important to understanding the fine-scale dynamics of retrotransposon evolution and how they shaped the genomic evolution of their hosts.

3.5. Conclusions

Mobile elements need to be identified and characterized. This is an important step towards providing a description of an organism's whole genome, especially in plant species, which are much less represented in the literature than animal species, including humans. This is the first report of the identification and detailed characterization of transposable elements (TEs) in a gene-rich fraction (~10 Mb) of *Passiflora edulis* (Malpighiales: Passifloraceae). We found that TEs occupy 17.6% of this region, and most were found in intergenic spaces, although some TEs overlapped genes. Long terminal repeat retrotransposons dominated the *P. edulis* genome portion analysed herein, consisting mainly of *Gypsy* elements, with over-representation of the RLG*_peDel* and RLC*_peAngela* lineages. *P. edulis* LTR-RTs have been reported as active over the last 2 million years. We were also able to confirm the transcriptional activity of full-length LTR-RTs by means of transcript association and reverse transcriptase PCR analysis. The recent activity and abundance of LTR-RTs in this gene-rich portion is indicative of their potential role in shaping the *P. edulis* genome. Interestingly, some lineages seem to be conserved in wild species of *Passiflora* and further detailed characterization of the repeat portion will contribute to our understanding of their influence on the evolution of this genus. Our study provides a detailed description of how transposable elements are identified and characterized. This applies especially to LTR-RTs. It will certainly facilitate the description of the whole genome of *P. edulis* and contribute to other studies of *Passiflora* species.

References

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012). GenomeView: a next-generation genome browser. *Nucleic Acids Res.* 40, e12–e12. doi:10.1093/nar/gkr995.

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10. Available at: http://dx.doi.org/10.1093/nar/gkw343.

Arensburger, P., Piégu, B., and Bigot, Y. (2016). The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? *Mob. Genet. Elements* 6, e1256852. doi:10.1080/2159256X.2016.1256852.

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi:10.1146/annurev-arplant-050213-035811.

Bergman, C. M., and Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* 8, 382–392. doi:10.1093/bib/bbm048.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., and Olmo, E. (2016). Transposons, genome size, and evolutionary insights in animals. *Cytogenet. Genome Res.* 147, 217–239. doi:10.1159/000444429.

Cappello, J., Handelsman, K., and Lodish, H. F. (1985). Sequence of *Dictyostelium* DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell* 43, 105–115. doi:10.1016/0092-8674(85)90016-9.

Carneiro, M. S., Camargo, L. E. A., Coelho, A. S. G., Vencovsky, R., Rui, P. L. J., Stenzel, N. M. C., et al. (2002). RAPD-based genetic linkage maps of yellow passion fruit (*Passiflora edulis* Sims. f. flavicarpa Deg.). *Genome* 45, 670–678. doi:10.1139/g02-035.

Chan, A. P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* 28, 951–956. doi:10.1038/nbt.1674.

Civáň, P., Švec, M., and Hauptvogel, P. (2011). On the coevolution of transposable elements and plant genomes. *J. Bot.* 2011, 1–9. doi:10.1155/2011/893546.

Contreras, B., Vives, C., Castells, R., and Casacuberta, J. (2015). The impact of transposable elements in the evolution of plant genomes: from selfish elements to key players. In: Pontarotti P, ed. Evolutionary biology: biodiversification from genotype to phenotype. *Cham Springer Int. Publ.*, 93–105.

Cossu, R. M., Buti, M., Giordani, T., Natali, L., and Cavallini, A. (2012). A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. *Tree Genet. Genomes* 8, 61–75. doi:10.1007/s11295-011-0421-3.

Cruz, G. M. Q., Metcalfe, C. J., De Setta, N., Cruz, E. A. O., Prata Vieira, A., Medina, R., et al. (2014). Virus-like attachment sites and plastic CpG Islands: landmarks of diversity in plant del retrotransposons. *PLoS One* 9, 1–14. doi:10.1371/journal.pone.0097099.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488, 213. Available at: http://dx.doi.org/10.1038/nature11241.

Domingues, D. S., Cruz, G. M. Q., Metcalfe, C. J., Nogueira, F. T. S., Vicentini, R., Alves, C. D. S., et al. (2012). Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genomics* 13, 137. doi:10.1186/1471-2164-13-137.

Dooner, H. K., and Weil, C. F. (2007). Give-and-take : interactions between DNA transposons and their host plant genomes. 486–492. doi:10.1016/j.gde.2007.08.010.

Du, J., Tian, Z., Hans, C. S., Laten, H. M., Cannon, S. B., Jackson, S. A., et al. (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J.* 63, 584–598. doi:10.1111/j.1365-313X.2010.04263.x.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340.

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. doi:10.1093/gbe/evt025.

Feuillet, C., and Macdougal, J. M. (2003). *A new infrageneric classification of Passiflora L. (Passifloraceae).*

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* 6, e16526. doi:10.1371/journal.pone.0016526.

Galindo-González, L., Mhiri, C., Deyholos, M. K., and Grandbastien, M.-A. (2017). LTR-retrotransposons in plants: engines of evolution. *Gene* 626, 14–25. doi:https://doi.org/10.1016/j.gene.2017.04.051.

Garcia-Perez, M. M.-L. and J. L. (2010). DNA Transposons: nature and applications in genomics. *Curr. Genomics* 11, 115–128. doi:http://dx.doi.org/10.2174/138920210790886871.

Gonzalez, L. G., and Deyholos, M. K. (2012). Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. *BMC Genomics* 13. doi:10.1186/1471-2164-13-644.

Gupta, S., Gallavotti, A., Stryker, G. A., Schmidt, R. J., and Lal, S. K. (2005). A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol. Biol.* 57, 115–127. doi:10.1007/s11103-004-6636-z.

Havecker, E. R., Gao, X., and Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biol.* 5. doi:10.1186/gb-2004-5-6-225.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., et al. (2014). PASTEC: an automatic transposable element classification tool. *PLoS One* 9, e91929. doi:10.1371/journal.pone.0091929.

Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., et al. (2015). A call for benchmarking transposable element annotation methods. *Mob. DNA* 6, 13. doi:10.1186/s13100-015-0044-6.

Hua-van, A., Rouzic, A. Le, Boutin, T. S., Filée, J., and Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biol. Direct* 6, 19. doi:10.1186/1745-6150-6-19.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8, 460. doi:10.1186/1471-2105-8-460.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463. Available at: http://dx.doi.org/10.1038/nature06148.

Jiang, N., Jordan, I. K., and Wessler, S. R. (2002). Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. *Plant Physiol.* 130, 1697–1705. doi:10.1104/pp.015412.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi:10.1159/000084979.

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450.

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., and Mclnerney, J. O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* 6, 29. doi:10.1186/1471-2148-6-29.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.

Klein, S. J., and O'Neill, R. J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* doi:10.1007/s10577-017-9569-5.

Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* 17, 1483–1498. doi:10.1093/oxfordjournals.molbev.a026248.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054.

Lee, S.-I., and Kim, N.-S. (2014). Transposable elements and genome size variations in plants. *Genomics Inform.* 12, 87. doi:10.5808/GI.2014.12.3.87.

Leeton, P. R., and Smyth, D. R. (1993). An abundant LINE-like element amplified in the genome of *Lilium speciosum*. *Mol. Gen. Genet.* 237, 97–104.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303. Available at: http://arxiv.org/abs/1303.3997 [Accessed October 14, 2017].

Lisch, D. (2012). How important are transposons for plant evolution? *Nat Rev Genet* 14, 49–61. doi:10.1038/nrg3374.

Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., et al. (2015). IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* 31, 3359–3361. Available at: http://dx.doi.org/10.1093/bioinformatics/btv362.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., et al. (2011). The Gypsy database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39, D70–D74. doi:10.1093/nar/gkq1061.

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct* 4. doi:10.1186/1745-6150-4-41.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi:10.1093/nar/gkw1129.

Mayer, K. F. X., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., and Catherine Feuillet (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345, 1251788. doi:10.1126/science.1251788.

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., et al. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* 47, 1435. Available at: http://dx.doi.org/10.1038/ng.3435.

Mita, P., and Boeke, J. D. (2016). How retrotransposons shape genome regulation. *Curr. Opin. Genet. Dev.* 37, 90–100. doi:10.1016/j.gde.2016.01.001.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* 37, 997–1002. doi:10.1038/ng1615.

Munhoz, C. F., Costa, Z. P., Cauz-Santos, L. A., Reátegui, A. C. E., Rodde, N., Cauet, S., et al. (2018). A gene-rich fraction analysis of the *Passiflora edulis* genome reveals highly conserved microsyntenic regions with two related Malpighiales species. *Sci. Rep.* 8, 13024. doi:10.1038/s41598-018-31330-8.

Munhoz, C. F., Santos, A. A., Arenhart, R. A., Santini, L., Monteiro-Vitorello, C. B., and Vieira, M. L. C. (2015). Analysis of plant gene expression during passion fruit-*Xanthomonas axonopodis* interaction implicates lipoxygenase 2 in host defence. *Ann. Appl. Biol.* 167, 135–155. doi:10.1111/aab.12215.

Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–25. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC324241/.

Muschner, V. C., Lorenz, A. P., Cervi, A. C., Bonatto, S. L., Souza-Chies, T. T., Salzano, F. M., et al. (2003). A first molecular phylogenetic analysis of *Passiflora* (Passifloraceae). *Am. J. Bot.* 90, 1229–1238. doi:10.3732/ajb.90.8.1229.

Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11, 378. doi:10.1186/1471-2105-11-378.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29, 792–793. doi:10.1093/bioinformatics/btt054.

Novikova, A., Smyshlyaev, G., and Novikova, O. (2012). Evolutionary history of LTR retrotransposon chromodomains in plants. *Int. J. Plant Genomics* 2012. doi:10.1155/2012/874743.

Novikova, O. (2009). Chromodomains and LTR retrotransposons in plants. *Commun. Integr. Biol.* 2, 158–162. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686373/.

Ochoa Cruz, E. A., Cruz, G. M. Q., Vieira, A. P., and Van Sluys, M.-A. (2016). Virus-like attachment sites as structural landmarks of plants retrotransposons. *Mob. DNA* 7, 14. doi:10.1186/s13100-016-0069-5.

Oliver, K. R., McComb, J. A., and Greene, W. K. (2013). Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol. Evol.* 5, 1886–1901. doi:10.1093/gbe/evt141.

Padeken, J., Zeller, P., and Gasser, S. M. (2015). Repeat DNA in genome organization and stability. *Curr. Opin. Genet. Dev.* 31, 12–19. doi:10.1016/j.gde.2015.03.009.

Park, M., Park, J., Kim, S., Kwon, J.-K., Park, H. M., Bae, I. H., et al. (2012). Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. *Plant J.* 69, 1018–1029. doi:10.1111/j.1365-313X.2011.04851.x.

Piednoël, M., Gonçalves, I. R., Higuet, D., and Bonnivard, E. (2011). Eukaryote DIRS1-like retrotransposons : an overview. 1–18.

Piskurek, O., and Jackson, D. J. (2012). Transposable elements: from DNA parasites to architects of metazoan evolution. *Genes (Basel).* 3, 409–422. doi:10.3390/genes3030409.

Poulter, R. T. M., and Butler, M. I. (2015). Tyrosine recombinase retrotransposons and transposons. *Microbiol. Spectr.* 3, MDNA3-0036-2014. doi:10.1128/microbiolspec.MDNA3-0036-2014.

Poulter, R. T. M., and Goodwin, T. J. D. (2005). DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenet. Genome Res.* 110, 575–588. Available at: https://www.karger.com/DOI/10.1159/000084991.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., et al. (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* 1, e22. doi:10.1371/journal.pcbi.0010022.

Ragupathy, R., You, F. M., and Cloutier, S. (2013). Arguments for standardizing transposable element annotation in plant genomes. *Trends Plant Sci.* 18, 367–376. doi:10.1016/j.tplants.2013.03.005.

Rambaut, A. (2012). FigTree v. 1.4.0. *http://tree.bio.ed.ac.uk/software/figtree/*. Available at: citeulike-article-id:13604993.

Ravindran, S. (2012). Barbara McClintock and the discovery of jumping genes. *Proc. Natl. Acad. Sci.* 109, 20198–20199. doi:10.1073/pnas.1219372109.

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

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* 20, 43–45. doi:10.1038/1695.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765–768.

Santos, A., Penha, H., Bellec, A., Munhoz, C. De, Pedrosa-Harand, A., Bergès, H., et al. (2014). Begin at the beginning: a bac-end view of the passion fruit (*Passiflora*) genome. *BMC Genomics* 15, 816. doi:10.1186/1471-2164-15-816.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., et al. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 39, D38-51. doi:10.1093/nar/gkq1172.

Smyth, D. R. (1991). Dispersed repeats in plant genomes. *Chromosoma* 100, 355–359. doi:10.1007/BF00337513.

Spruyt, M., and Buquicchio, F. (2004). Gene runner for windows. *http://www.generunner.net/*.

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

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat. Plants* 2, 1–10. doi:10.1038/NPLANTS.2016.73.

The Tomato Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi:10.1038/nature11119.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511. Available at: http://dx.doi.org/10.1038/nbt.1621.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science (80-.).* 313, 1596–1604. doi:10.1126/science.1128691.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* 42, 833. Available at: http://dx.doi.org/10.1038/ng.654.

Vicient, C. M., and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* 120, 195–207. doi:10.1093/aob/mcx078.

Vitte, C., Chaparro, C., Quesneville, H., and Panaud, O. (2007). Spip and Squiq, two novel rice non-autonomous LTR retro-element families related to RIRE3 and RIRE8. *Plant Sci.* 172, 8–19. doi:https://doi.org/10.1016/j.plantsci.2006.07.008.

Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., and Bevan, M. W. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463, 763. Available at: http://dx.doi.org/10.1038/nature08747.

Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., et al. (2014). Cassava genome from a wild ancestor to cultivated varieties. *Nat. Commun.* 5, 5110. doi:10.1038/ncomms6110.

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., et al. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* 196, 891–909. doi:10.1534/genetics.113.159996.

Wicker, T., Sabot, F. F., Hua-Van, A. A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. doi:10.1038/nrg2165.

Wu, P., Zhou, C., Cheng, S., Wu, Z., Lu, W., Han, J., et al. (2015). Integrated genome sequence and linkage map of physic nut (*Jatropha curcas* L.), a biodiesel plant. *Plant J.* 81, 810–821. doi:10.1111/tpj.12761.

Yi, F., Jia, Z., Xiao, Y., Ma, W., and Wang, J. (2018). SPTEdb: a database for transposable elements in salicaceous plants. *Database J. Biol. Databases Curation* 2018, bay024. doi:10.1093/database/bay024.

Yin, H., Du, J., Wu, J., Wei, S., Xu, Y., Tao, S., et al. (2015). Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. *Sci. Rep.* 5, 1–15. doi:10.1038/srep17644.

Yotoko, K. S. C., Dornelas, M. C., Togni, P. D., Fonseca, T. C., Salzano, F. M., Bonatto, S. L., et al. (2011). Does variation in genome sizes reflect adaptive or neutral processes? New clues from *Passiflora*. *PLoS One* 6, e18212. doi:10.1371/journal.pone.0018212.

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., et al. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480, 520. Available at: http://dx.doi.org/10.1038/nature10625.

Zhao, D., Ferguson, A. A., and Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1859, 366–380. doi:10.1016/j.bbagrm.2015.12.005.

Zhao, M., and Ma, J. (2013). Co-evolution of plant LTR-retrotransposons and their host genomes. *Protein Cell* 4, 493–501. doi:10.1007/s13238-013-3037-6.

APPENDICES

APPENDIX A.

BAC inserts selected from the *Passilfora edulis* genomic library for complete sequencing.

(continue)

BAC code	BAC selection criterion
Pe1K19	Presence of gene in the forward BAC-end sequence (BES)
Pe1M17	Potential non co-localized microsyntenic region with *Populus trichocarpa*
Pe3F10	Probe complementary to the gene cellulose synthase
Pe7A23	Probe complementary to mitochondrial gene
Pe7M15	Probe complementary to the gene stromal cell-derived factor 2-like protein precursor
Pe9E4	Probe complementary to the gene auxin response factor 2
Pe15E1	Probe complementary to the gene (+)-neomenthol dehydrogenase
Pe20E10	Probe complementary to the gene cytochrome C oxidase subunit 2
Pe20N3	Probe complementary to the gene lipoxygenase 2
Pe21O15	Probe complementary to the gene glutamine synthetase
Pe24G19	Probe complementary to the gene sugar transport protein 13
Pe27H17	Probe complementary to the gene kinesin-like protein
Pe28D11	Probe complementary to the gene heat stress transcription factor C-1
Pe28E22	Probe complementary to the gene glutamate receptor
Pe28I20	Probe complementary to the gene glycolate oxidase
Pe33M2	Presence of genes in both BES
Pe34H9	Probe complementary to the gene copper/zinc-superoxide dismutase 1a
Pe34M7	Presence of genes in both BES
Pe43D2	Presence of genes in both BES
Pe43L2	Presence of genes in both BES
Pe51C2	Probe complementary to the gene F-box/LRR-repeat protein 4
Pe60G10	Probe complementary to the gene serine/threonine-protein kinase-like protein ACR4
Pe61E2	Probe complementary to the gene mitochondrial outer membrane protein porin of 36 kDa
Pe63J18	Probe complementary to the gene chlorophyll a b binding
Pe64C12	Probe complementary to the gene basic endochitinase B
Pe65F7	Probe complementary to the gene celulose sintase
Pe69C7	Potential non co-localized microsyntenic region with *Populus trichocarpa*
Pe69F22	Presence of genes in both BES
Pe69G18	Presence of genes in both BES
Pe69H24	Presence of genes in both BES
Pe69N18	Potential collinear microsyntenic region with *Arabidopsis thaliana*
Pe69O16	Presence of genes in both BES
Pe71E3	Probe complementary to the gene stearoyl-acp desaturase
Pe74I6	Probe complementary to the gene beta-amylase 1
Pe75A21	Presence of genes in both BES
Pe75D12	Potential collinear microsyntenic region with *Populus trichocarpa*
Pe75F13	Presence of gene in the reverse BES
Pe75F20	Presence of gene in the forward BES
Pe75K15	Presence of genes in both BES

BAC code	BAC selection criterion
Pe75N15	Presence of genes in both BES
Pe84I14	Presence of genes in both BES
Pe84K8	Presence of genes in both BES
Pe84M18	Presence of gene in the forward BES
Pe84M23	Presence of genes in both BES
Pe84M6	Presence of gene in the reverse BES
Pe85B19	Presence of genes in both BES
Pe85H4	Presence of genes in both BES
Pe85I9	Presence of genes in both BES
Pe85J23	Presence of genes in both BES
Pe85L8	Potential collinear microsyntenic region with *Arabidopsis thaliana*
Pe85O9	Presence of genes in both BES
Pe86F9	Probe complementary to the gene ATP-citrate synthase beta chain protein 2
Pe86H07	Probe complementary to the gene cytochrome C oxidase
Pe89E10	Probe complementary to the gene homeobox-leucine zipper athb-6
Pe93A7	Presence of gene in the forward BES
Pe93F5	Presence of gene in the reverse BES
Pe93J9	Presence of genes in both BES
Pe93K19	Presence of genes in both BES
Pe93M2	Presence of genes in both BES
Pe93M4	Presence of gene in the reverse BES
Pe93N7	Presence of gene in the forward BES
Pe93O18	Presence of genes in both BES
Pe99P16	Probe complementary to the gene glutamate-cysteine ligase
Pe101F21	Presence of genes in both BES
Pe101H15	Presence of genes in both BES
Pe101K14	Presence of genes in both BES
Pe101O4	Presence of genes in both BES
Pe101P13	Presence of genes in both BES
Pe101P7	Presence of gene in the reverse BES
Pe103M2	Probe complementary to the gene alpha-L-arabinofuranosidase 1
Pe108C16	Probe complementary to the gene glutamate-cysteine ligase
Pe113A7	Probe complementary to the gene harpin-induced protein/protein YLS9-like
Pe117C17	Probe complementary to the gene inactive beta-amylase 9
Pe123N8	Probe complementary to the gene phosphoribulokinase
Pe125I23	Probe complementary to the gene cyclin-dependent protein kinase regulator
Pe134H15	Probe complementary to the gene 1-aminocyclopropane-1-carboxylate oxidase
Pe135J12	Probe complementary to the gene ubiquitin-conjugating enzyme E2-23 kDa-like
Pe138G17	Probe complementary to the gene putative serine proteinase
Pe141B12	Presence of genes in both BES
Pe141H13	Presence of genes in both BES
Pe141J23	Presence of genes in both BES
Pe141K8	Potential non co-localized microsyntenic region with *Populus trichocarpa*

BAC code	BAC selection criterion
Pe164A12	Potential collinear microsyntenic region with *Vitis vinifera*
Pe164B18	Presence of genes in both BES
Pe164D9	Presence of genes in both BES
Pe164K17	Potential collinear microsyntenic region with *Populus trichocarpa*
Pe168B17	Probe complementary to the gene nucleoredoxin 1
Pe171P13	Presence of genes in both BES
Pe173B16	Potential rearranged microsyntenic region with *Populus trichocarpa*
Pe175N8	Probe complementary to the gene disease resistance family protein
Pe185D11	Presence of genes in both BES
Pe185J16	Presence of genes in both BES
Pe186E19	Probe complementary to the gene brassinosteroid insensitive 1
Pe195F4	Probe complementary to the gene glycerate dehydrogenase
Pe198H23	Probe complementary to the gene ethylene response sensor
Pe201C11	Probe complementary to the gene maltose excess protein 1
Pe207D11	Probe complementary to the gene kinesin-like protein
Pe209G15	Probe complementary to the gene lipoxygenase 2
Pe212D7	Probe complementary to the gene oxygen-evolving enhancer protein 1-1
Pe212I1	Presence of genes in both BES
Pe212J12	Presence of gene in the reverse BES
Pe212M5	Presence of gene in the reverse BES
Pe213C9	Probe complementary to the gene F-box protein PP2-A12 isoform X2
Pe214A18	Potential non co-localized microsyntenic region with *Populus trichocarpa/Vitis vinifera*
Pe214H11	Potential non co-localized microsyntenic region with *Populus trichocarpa*
Pe214N19	Presence of genes in both BES
Pe215I8	Probe complementary to the gene glyceraldehyde-3-phosphate dehydrogenase
Pe216B2	Presence of genes in both BES
Pe216B22	Potential collinear microsyntenic region with *Populus trichocarpa*
Pe216F3	Presence of genes in both BES
Pe216F9	Presence of genes in both BES
Pe216I5	Potential non co-localized microsyntenic region with *Populus trichocarpa*

APPENDIX B. Sequencing results of the 112 BAC inserts selected from the *Passiflora edulis* genomic library.

(continue)

BAC code	Gel-estimated insert size (bp)	Reads number	Size range of reads (bp)	% GC	Mean coverage (X)	Mean QV*	Contig size (pb)
Pe1K19	52,000	809	599-35,075	42	122	48.51	48,154
Pe1M17	50,000	628	500-23,379	43	40	48.49	54,675
Pe3F10	48,000	572	518-23,463	40	40	48.50	50,560
Pe7A23	30,000	206	509-16,142	43	86	48.40	22,590
Pe7M15	122,000	5,667	500-31,277	42	186	48.55	117,775
Pe9E4	90,000	1,921	501-42,009	42	87	48.54	88,822
Pe15E1	90,000	4,981	500-39,121	42	212	48.55	88,169
Pe20E10	105,000	10,632	500-42,297	41	437	48.52	107,924
Pe20N3	100,000	5,322	503-39,062	40	410	48.55	96,493
Pe21O15	80,000	3,202	510-36,046	40	238	48.52	86,369
Pe24G19	80,000	2,652	501-23,230	44	118	48.57	83,926
Pe27H17	95,000	2,280	504-36,741	43	222	48.77	95,184
Pe28D11	125,000	9,650	500-42,658	45	350	48.54	122,897
Pe28E22	97,000	4,226	501-22,839	41	137	48.58	97,719
Pe28I20	85,000	3,212	502-38,789	41	216	48.51	91,911
Pe33M2	100,000	1,621	500-38,247	41	154	48.81	97,233
Pe34H9	80,000	4,859	502-35,035	40	191	48.56	92,133
Pe34M7	130,000	5,824	500-29,279	42	136	48.57	130,776
Pe43D2	80,000	3,912	504-42,904	40	322	48.55	88,638
Pe43L2	75,000	5,002	520-38,476	39	398	48.56	93,150
Pe51C2	100,000	11,496	501-46,831	40	554	48.56	93,290
Pe60G10	100,000	6,392	500-37,191	42	307	48.55	96,564
Pe61E2	100,000	14,523	500-45,838	38	670	48.56	95,754
Pe63J18	90,000	3,575	501-22,529	43	119	48.56	91,586
Pe64C12	90,000	2,738	500-23,673	42	101	48.58	97,338
Pe65F7	90,000	3,340	500-21,851	41	121	48.58	89,394
Pe69C7	97,000	3,068	500-25,108	40	104	48.59	103,741
Pe69F22	90,000	4,436	501-40,778	39	380	48.56	87,349
Pe69G18	95,000	4,517	500-35,983	38	371	48.57	91,665
Pe69H24	110,000	5,497	517-38,176	39	398	48.57	104,185
Pe69N18	95,000	2,340	500-22,458	37	87	48.48	94,052
Pe69O16	90,000	4,555	500-41,504	40	394	48.55	90,157
Pe71E3	90,000	4,437	500-35,709	41	382	48.55	85,814
Pe74I6	120,000	14,131	501-39,105	39	582	48.56	112,038
Pe75A21	110,000	5,760	501-37,523	42	386	48.56	109,110
Pe75D12	90,000	9,698	500-25,329	42	344	48.54	92,130
Pe75F13	115,000	3,251	500-42,063	41	221	48.53	112,223
Pe75F20	108,000	4,175	500-34,745	39	319	48.52	101,327
Pe75K15	105,000	4,519	503-37,147	44	327	48.55	100,781
Pe75N15	90,000	4,020	503-42,177	43	322	48.54	94,339

BAC code	Gel-estimated insert size (bp)	Reads number	Size range of reads (bp)	% GC	Mean coverage (X)	Mean QV*	Contig size (pb)
Pe84I14	100,000	5,268	504-39,183	40	327	48.55	97,848
Pe84K8	85,000	4,069	503-34,441	39	289	48.55	85,616
Pe84M6	96,000	2,859	501-44,938	42	240	48.51	92,167
Pe84M18	111,000	5,102	509-39,425	40	383	48.55	103,941
Pe84M23	100,000	5,081	502-43,845	38	397	48.56	93,217
Pe85L8	90,000	4,972	500-26,035	38	173	48.57	91,155
Pe85B19	100,000	4,916	508-35,487	40	306	48.54	96,953
Pe85H4	90,000	4,388	500-30,909	39	297	48.55	88,298
Pe85I9	95,000	4,870	500-39,620	41	301	48.55	95,720
Pe85J23	85,000	3,943	501-36,770	39	278	48.54	85,126
Pe85O9	30,000	391	573-33,837	39	78	48.52	25,363
Pe86F9	110,000	6,693	502-34,245	46	234	48.55	103,497
Pe86H7	95,000	4,691	503-37,269	43	284	48.54	96,593
Pe89E10	100,000	12,007	500-41,450	42	529	48.55	99,734
Pe93A7	104,000	3,503	509-38,522	41	272	48.53	99,578
Pe93F5	108,000	3,694	509-38,440	40	278	48.53	102,639
Pe93J9	110,000	6,340	501-35,344	40	362	48.55	107,024
Pe93K19	100,000	5,460	504-38,289	40	335	48.55	99,992
Pe93M2	100,000	5,170	500-38,116	41	311	48.55	100,436
Pe93M4	92,000	1,909	500-37,458	42	170	48.53	87,142
Pe93N7	119,000	5,922	501-39,176	40	423	48.54	106,968
Pe93O18	105,000	1,836	500-40,738	40	173	48.79	98,251
Pe99P16	100,000	12,248	500-42,116	41	545	48.55	99,641
Pe101F21	95,000	6,810	505-36,921	44	456	48.52	96,919
Pe101H15	80,000	6,099	509-35,269	42	335	47.51	88,834
Pe101K14	80,000/85,000	12,454	501-37,335	39	502	48.52	172,337
Pe101O4	110,000	6,498	503-39,352	38	376	48.53	117,581
Pe101P7	98,000	2,642	520-36,389	38	232	48.52	90,691
Pe101P13	85,000	8,236	501-37,124	37	620	48.52	91,580
Pe103M2	75,000	2,859	501-34,151	40	132	48.53	73,357
Pe108C16	105,000	2,800	500-40,287	41	277	48.79	96,753
Pe113A7	110,000	9,784	500-37,691	41	424	48.55	106,440
Pe117C17	95,000	6,484	501-39,316	42	232	48.53	103,905
Pe123N8	100,000	7,516	500-42,993	41	350	48.55	96,994
Pe125I23	95,000	4,004	500-39,630	39	243	48.55	96,568
Pe134H15	90,000	4,993	500-26,558	40	177	48.55	91,362
Pe135J12	105,000	6,446	503-40,401	40	238	48.55	103,564
Pe138G17	90,000	5,931	503-40,512	43	283	48.54	93,983
Pe141B12	75,000	4,423	502-33,544	43	360	48.53	79,426
Pe141J23	90,000	8,374	508-42,274	41	590	48.51	95,795
Pe141K8	90,000	2,619	500-23,918	40	99	48.58	97,973
Pe164A12	85,000	4,493	501-26,430	41	174	48.57	82,998

BAC code	Gel-estimated insert size (bp)	Reads number	Size range of reads (bp)	% GC	Mean coverage (X)	Mean QV*	Contig size (pb)
Pe164B18	100,000	9,006	501-37,510	40	574	48.52	104,102
Pe164D9	85,000	8,089	500-41,095	40	581	48.53	93,527
Pe164K17	112,000	3,334	502-25,218	40	106	48.58	113,504
Pe168B17	135,000	7,467	500-30,781	41	209	48.55	137,256
Pe171P13	110,000	2,710	500-43,329	40	234	48.79	111,123
Pe173B16	110,000	3,638	501-33,974	40	122	48.59	109,801
Pe175N8	115,000	9,084	500-38,499	41	402	48.55	106,381
Pe185D11	120,000	3,404	500-41,999	40	283	48.79	119,061
Pe185J16	105,000	2,343	506-40,202	39	221	48.78	103,095
Pe186E19	110,000	6,647	500-39,704	39	222	48.55	115,218
Pe195F4	110,000	6,995	500-38,598	40	235	48.55	113,443
Pe198H23	112,000	8,595	500-24,834	42	231	48.56	108,433
Pe201C11	140,000	8,471	500-39,592	40	226	48.56	140,216
Pe207D11	120,000	11,469	500-40,071	41	470	48.56	111,690
Pe209G15	90,000	6,971	500-40,057	39	336	48.56	94,376
Pe212D7	125,000	8,350	500-36,802	39	255	48.55	123,561
Pe212I1	120,000	3,365	502-41,492	40	265	48.79	121,384
Pe212J12	28,000	98	739-30,328	41	26	47.35	24,316
Pe212M5	48,000	549	500-36,767	41	91	48.5	43,763
Pe213C9	110,000	8,311	502-40,348	42	350	48.52	106,552
Pe214A18	97,000	3,152	500-24,957	43	110	48.55	106,977
Pe214H11	140,000	5,952	500-24,825	40	157	48.59	142,456
Pe214N19	100,000	1,871	509-38,049	44	173	48.78	98,343
Pe215I8	130,000	6,262	500-22,437	40	150	48.58	129,737
Pe216B2	50,000	153	504-37,581	36	32	48.57	42,359
Pe216B22	114,000	7,389	500-23,820	40	218	48.57	111,836
Pe216F3	80,000	907	504-41,986	40	111	48.79	79,451
Pe216F9	110,000	2,974	504-43,753	36	276	48.82	105,476
Pe216I5	68,000	1,526	500-18,384	40	77	48.58	72,701

*QV: Quality Value (Probability of incorrect base call: QV40= 1 in 10,000, QV50= 1 in 100,000.