**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)

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. - - Piracicaba, 2018.

94 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.....	7
ABSTRACT.....	8
LIST OF FIGURES.....	9
LIST OF TABLES.....	12
1. INTRODUCTION.....	13
REFERENCES.....	13
2. A GENE-RICH FRACTION ANALYSIS OF THE <i>PASSIFLORA EDULIS</i> GENOME....	17
ABSTRACT.....	17
2.1 INTRODUCTION.....	17
2.2. MATERIAL AND METHODS.....	19
2.2.1 BAC Selection and DNA preparation.....	19
2.2.2 DNA Sequencing and Assembly From Long Sequence Reads.....	20
2.2.3 Gene Prediction and Functional Annotation.....	21
2.2.4 Microsatellite prospection.....	22
2.3 RESULTS.....	23
2.3.1 BAC Selection, Sequencing and Assembly.....	23
2.3.2 Gene Representativeness, Structure and Functional Annotation.....	23
2.3.3 Microsatellite prospection.....	30
2.4. DISCUSSION.....	32
2.5. CONCLUSIONS.....	34
REFERENCES.....	34
3. TRANSPOSABLE ELEMENT IDENTIFICATION AND ANALYSIS OF EVOLUTIONARY LINEAGES OF LTR-RETROTRANSPOSONS IN A GENE-RICH FRACTION OF THE <i>PASSIFLORA EDULIS</i> GENOME.....	41
ABSTRACT.....	41
3.1. INTRODUCTION.....	41
3.2. MATERIAL AND METHODS.....	48
3.2.1 Plant Material.....	48
3.2.2 Identification of Transposable Elements.....	48
3.2.3 Analysis of LTR-RTs.....	50
3.2.3.1 Identification and retrieval of LTR-RT internal domains.....	50
3.2.3.2 Phylogenetic analysis of LTR-RT elements.....	50
3.2.3.3 Assignment of LTR-RTs to evolutionary lineages and naming of sequences..	51
3.2.3.4 Structural features of <i>Passiflora edulis</i> LTR-RTs.....	52
3.2.3.5 Estimation of <i>Passiflora edulis</i> LTR-RT insertion time.....	52
3.2.3.6 In silico analysis of LTR-RT expression.....	53
3.2.3.7 RNA extraction and Reverse Transcriptase (RT)-PCR analysis.....	53
3.2.3.8 LTR-RTs from wild <i>Passiflora</i> species.....	54
3.3. RESULTS.....	54
3.3.1 Identification of Transposable Elements.....	54
3.3.2 Analysis of LTR-RTs.....	58

3.3.2.1 Phylogeny and structural features of the LTR-RTs.....	58
3.3.2.2 Estimation of LTR-RT insertion time.....	67
3.3.2.3 LTR-RT in silico transcriptional activity.....	67
3.3.2.4 RNA extraction and Reverse Transcriptase (RT) PCR analysis.....	69
3.3.2.5 LTR-RTs from wild <i>Passiflora</i> species.....	70
3.4. DISCUSSION.....	71
3.5. CONCLUSIONS.....	79
REFERENCES.....	79
APPENDICES.....	89

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

- 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*.....28
- 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* (orange bars) and in coding DNA sequences (CDS, blue bars). (B) Percentage of the most frequent motifs in each class of microsatellites (SSRs) found in all sequences (orange bars) and in coding DNA sequences (CDS, blue bars) of *Passiflora edulis*.....31
- Figure 3: Classification system for transposable elements. Adapted from Wicker et al., 2007.....47
- 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).....57
- 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).....58
- 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.....59
- 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).....61

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.....62

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.....63

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.....64

Figure 11: Number of complete and incomplete copies of LTR-RTs in nine lineages of the *Copia* and *Gypsy* superfamilies.....65

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.....66

Figure 13: Estimated insertion times of 73 full-length LTR-RT lineages of *Passiflora edulis*.....67

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.....68

Figure 15: Number of normalized transcripts assigned to LTR-RT lineages. The transcripts were obtained from three *Passiflora edulis* RNA-seq libraries of shoot apices of juvenile (1), vegetative (2) and reproductive adult plants (3) (Dornelas M.C., unpublished data).....69

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).....70

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).....71

LIST OF TABLES

Table 1. Gene content in a gene-rich fraction of the <i>Passiflora edulis</i> genome (~10.4 Mb) and structural annotation.....	24
Table 2. Most frequent protein signatures (≥ 10) recognized in 1,488 genes of <i>Passiflora edulis</i> according to InterProScan results.....	29
Table 3. Classification and abundance of transposable elements identified in a gene-rich fraction of <i>Passiflora edulis</i> genome.....	56
Table 4. General features of <i>Passiflora edulis</i> LTR-RT lineages in <i>Copia</i> and <i>Gypsy</i> superfamilies.....	66

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 Macdougall, 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., 2015a; 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 lipooxygenase-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.

2. 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.
- Agrianual (2017). *Anuário da agricultura brasileira*. 22 ed. , ed. F. C. & Agroinformativos São Paulo.
- 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.
- 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.
- 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., Coghill, 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.

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

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 nonrepetitive 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.