**UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO**

**CENTRO DE CIÊNCIAS AGRÁRIAS E ENGENHARIAS**

**PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO**

DRIELLI CANAL

# The draft genome assembly of *Psidium guajava* L.

**ALEGRE, ES**

**2018**

DRIELLI CANAL

# The draft genome assembly of *Psidium guajava* L.

Dissertation presented to the UniversidadeFederal do Espírito Santo, as part of therequirements of the Genetics and BreedingGraduate Program, for the attainment of the titleMagister Scientiae.

Advisor: DSc. Dr. Adésio Ferreira

**ALEGRE, ES**

**2018**

Modelo de ficha catalográfica fornecido pelo Sistema Integrado de Bibliotecas da Ufes para ser confeccionada pelo autor

DRIELLI CANAL

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

APROVADA: 26 de fevereiro de 2019

Comissão Examinadora:

| | |
|---|---|
| _____ | _____ |
| Dr. Adésio Ferreira<br>Universidade Federal do Espírito<br>Orientador | Dra. Marcia Flores da Silva Ferreira<br>Universidade Federal do Espírito Santo<br>Coorientadora |
| _____ | _____ |
| Dr. Otavio J. B. Brustolini<br>Universidade Federal do Espírito Santo<br>Coorientador | Dra. Paola Carpinetti-Oliveira<br>Universidade Federal do Espírito Santo<br>Coorientadora |
| _____ | _____ |
| Dr. Pedro Henrique Dias Dos Santos<br>Universidade Estadual do Norte Fluminens<br>Membroexterno | Dr. Wellington Ronildo Clarindo<br>Universidade Federal do Espírito Santo<br>Membroexterno |

*dedicated to my beloved family*

## Acknowledgements

First, I would like to thank Godfor grace, health and mercy with me. Second, to my advisors Dr. Adésio Ferreira and Marcia Flores for their patience, help, encouragementand guidance during my dissertation work. I would also like to express my sincere gratitude to my friends OtávioBrustolinifor hismentorshipand support andto Miquéias Fernandes for his patience, teachingsand advice.I also sincerely appreciate the contributions of Pedro, Paola and Well in this work.I'm very grateful to all my colleagues at the Biometry Lab and Laboratório de Genética e Melhoramentofor their time, the jokes, respect, affection, confidence,always supporting and assisting in working hours, even on weekends and late at night.I would like to thank the Laboratório de Biotecnologia and Laboratório de Citogenética, for the reagents, time, space provided and support.The mainsourcesoffundingConselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), also, Fundação de Apoio à Pesquisa do Espírito Santo (FAPES), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)and Universidade Federal do Espírito Santo(UFES) for theopportunities. Finally, and most importantly, my utmost gratitude goes to my family, none of this would have been possible without their unconditionalsupport, trust, sacrifice and love.

## List of Abbreviations

**A**Adenine

**BLAST** Basic Local Alignment Search Tool

**bp** Base pair

**BUSCO** Benchmarking Universal Single-Copy Orthologs

**C** Cytosine

**CDS** Coding sequence

**CTAB** Cetyl Trimethyl Ammonium Bromide protocol

**DAPI** 40 ,60 - diamidino-2-phenylindole

***de novo***starting from the beginning; anew

**DBG** *de Brujin* graphs

**DIRS** *Dictyostelium* intermediate repeat sequence

**DNA** deoxyribonucleic acid

**DTT**dithiothreitol

**FastQ** format for short-read sequence data along with quality

**FI** fluorescence intensity

**FMC**Flow cytometry analysis

**G** Guanine

**GB** giga bytes

**GC** Guanine-Cytosine, used as shorthand for GC-content

**GFF3** Genome Feature Format, version 3

**GO** Gene ontology

**HSPs** High-scoring segment pairs

**kbp**kilo base pairs

**LARDs** Large Retrotransposon Derivates

**LINE** Long interspersed nuclear element

**LTR** Long Terminal Repeat

**N50** Size of contig or scaffold in an assembly such that 50% of the assembly is incontigs of that size or larger

**NCBI** National Center for Biotechnology Information

**NGS** Next-Generation Sequencing

**Mbp**Megabase pairs ($\times 10^6$ bp)

**MITE** Miniature inverted-repeat transposable elements

**MYA** Million years ago

**NGS** Next generation of sequencing

**OLC** Overlap-Layout-Consensus

**PE** Paired-End

**PI** propidiumiodide

**pg**picograms

**RNA r**ibonucleicacid

**SINE** Short interspersed nuclear element

**SMRT** Single molecule real time (sequencing)

**SSR** Simple sequence repeats

**SRA** Sequence Read Archive database

**T** Thymine

**TR tandem repeat**

**TE** Transposable element

**TIR** Terminal Inverted Repeats

**TPS** terpene synthase genes

**WGS** Whole genome sequencing

**ZMWs** zero-mode waveguide

## List of Figures

## List of Tables

## Table of Contents

**ABSTRACT**

*Psidium guajava* is a major commercial fruit of Myrtaceae family, mainly cropped for fresh fruit consumption, valued for its taste and high C vitamin content. Also, is used for industrialized production and has medicinal applications. Is widely distributed in all tropical and subtropical regions. However, guava conventional breeding could be a difficult and slow process and itshampered by limited molecular resources.This dissertation focuses on establishing the genome sequence for the *Psidium guajava*, that comprises all inheritable traits of an organism and provides essential information required to determine phenotypic traits and accelerate breeding programns. In this dissertation, I describe the process of creating a *de novo* assembly of *Psidium guajava* that was generated using a hybrid assembly strategy with Illumina short reads and long reads produced by Pacbio platforms.The genome assembly comprised 2,881 scaffolds with a total length of 412.5 Mb and a scaffold N50 of 312 Kb. Genome completeness indicated that 95,15% of the expected eudicotyledons genes were present in the genome assembly and cytogenetic analysis confirmed the genome size of 465 Mband GC content estimation around 40%. To detect the genomic features, an automatic annotation was performed, which revealed a large proportion of repetitive elements (57%), which suggests that they have played important roles in the evolution of *P. guajava.*The richness and distribution of microsatellites was also evaluated, revealing 210,037 potential SSR s allowing the fast and cost□effective development of molecular markersfor genetic research.A structural sequence annotation resulted in the prediction of about 61,089 putative genes, that are possible related to protein coding regions. The next stage of this project included the annotation of these proteins.

Keywords: whole genome sequencing, repetitive elements, microsatellites, PacBio sequencing, Illumina, bioinformatics.