

# Minicurso

## Exploração de dados de sequenciamento Illumina com ênfase em análise de genomas cloroplastidiais

Dr. Luiz Augusto Cauz dos Santos – [luizcauz@usp.br](mailto:luizcauz@usp.br)

Dra. Zirlane Portugal da Costa – [zirlane@usp.br](mailto:zirlane@usp.br)

13 de Julho de 2020

# PROGRAMA DO CURSO

## ➤ Parte I

- Introdução a genômica
- Introdução a montagem e anotação de genomas
- Análise dos dados Illumina para identificar repeats no genoma nuclear

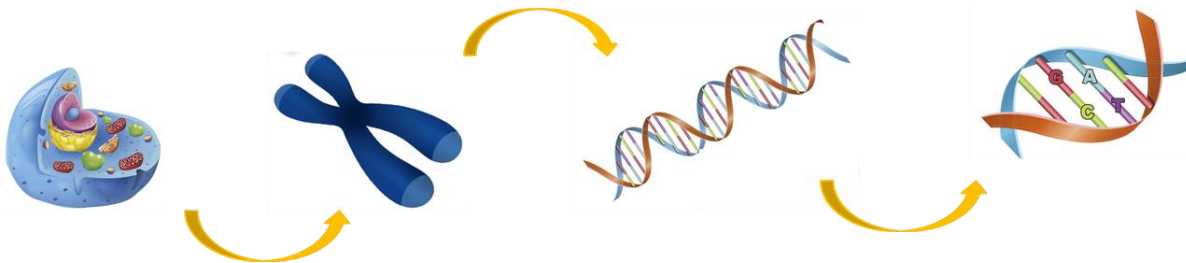
## ➤ Parte II

- Genomas extranucleares
- Montagem de genomas cloroplastidiais com dados Illumina no programa NOVOPLASTY
- Anotação e curadoria manual de genomas cloroplastidiais
- Genômica comparativa

# INTRODUÇÃO À GENÔMICA

- **Genômica** é um ramo da genética que estuda o genoma completo de um organismo
- A genômica revolucionou a maneira de fazer a **análise genética** e abriu caminhos para investigações que não eram concebíveis até alguns anos atrás
- As **análises de genomas inteiros** hoje contribuem para todos os aspectos das pesquisas biológicas

## ✓ Genética humana



# INTRODUÇÃO À GENÔMICA

## Genomas de plantas sequenciados

FGS (*First Generation Sequencing*)

- Sanger



NGS (*Next Generation Sequencing*)

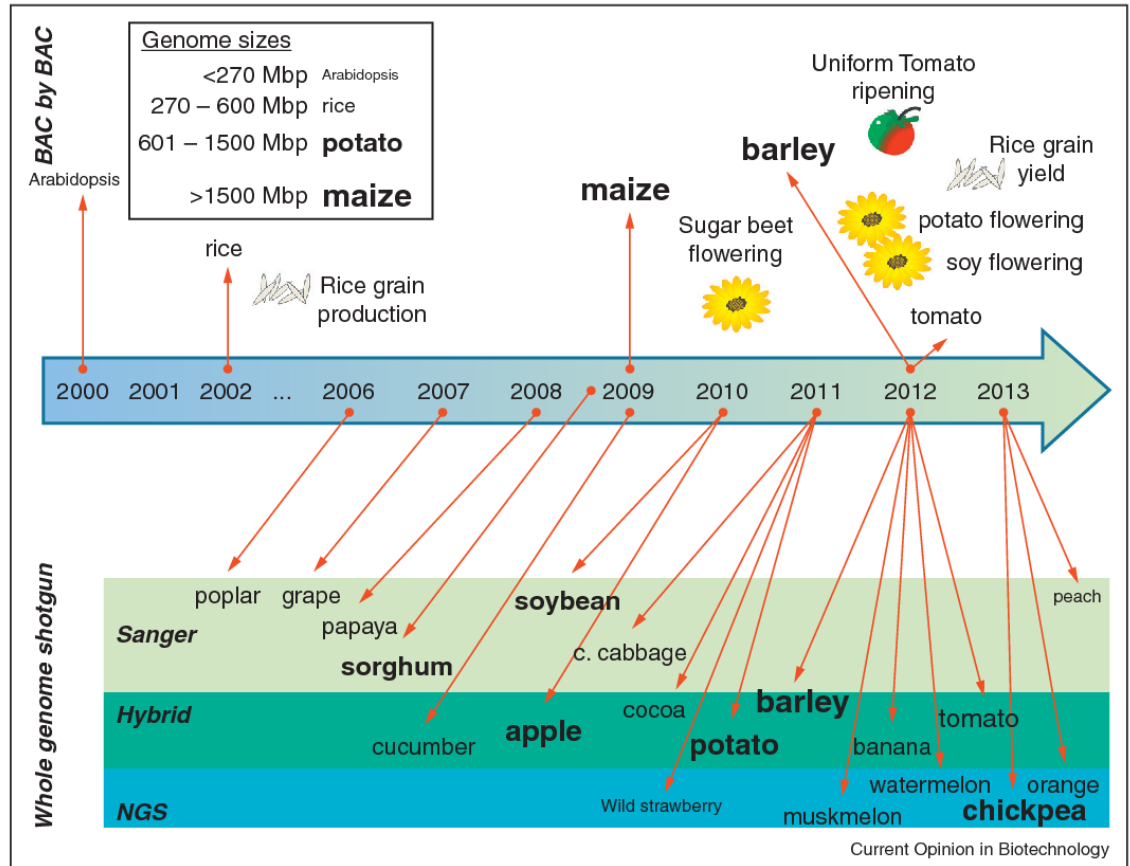
- 454
- Illumina
- SOLiD



TGS (*Third Generation Sequencing*)

- Plataforma PacBio
- Oxford nanopore

Evolução dos métodos de sequenciamento de DNA





# INTRODUÇÃO À GENÔMICA

illumina®



## MiniSeq System

Power and simplicity for targeted sequencing.



## MiSeq Series

Small genome and targeted sequencing.



## NextSeq Series

Everyday genome, exome transcriptome sequencing, and more.



## HiSeq Series

Production-scale genome, exome, transcriptome sequencing, and more.



## HiSeq X Series

Population- and production-scale human whole-genome sequencing.



## NovaSeq Series

Population- and production-scale genome, exome, transcriptome sequencing, and more.

7.5 Gb

15 Gb

120 Gb

1.5 Tb

1.8 Tb

1 Tb - 6 Tb<sup>1</sup>

MAX READ LENGTH  
2x150 bp

MAX READ LENGTH  
2x300 bp

MAX READ LENGTH  
2x150 bp

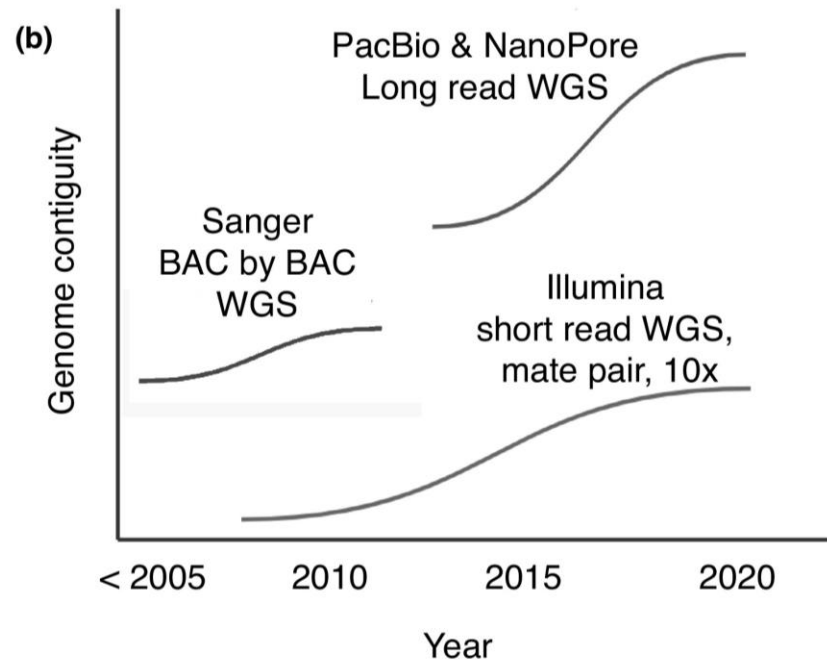
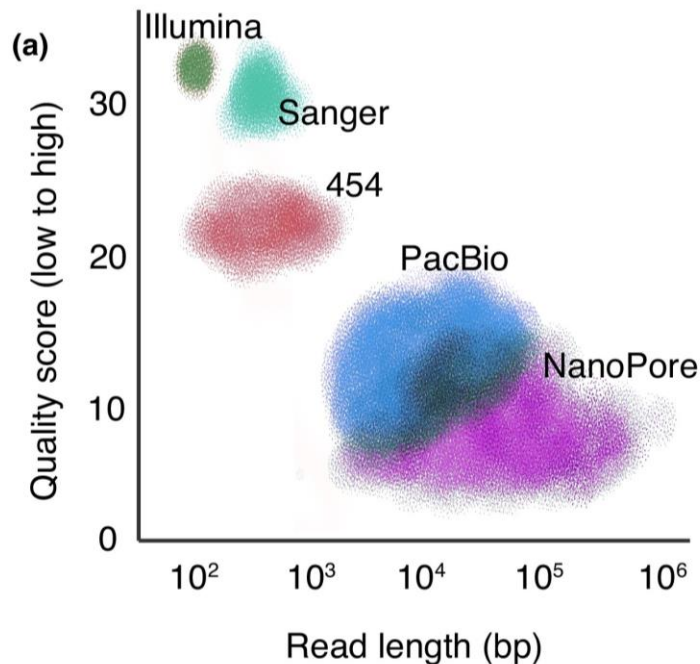
MAX READ LENGTH  
2x150 bp

MAX READ LENGTH  
2x150 bp

MAX READ LENGTH  
2x150 bp

# INTRODUÇÃO À GENÔMICA

- Sequenciamento de molécula única (SMRT – *Single Molecule Real Time*)
- *Reads* longas



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

ScienceDirect



Building near-complete plant genomes

Todd P Michael<sup>1</sup> and Robert VanBuren<sup>2,3</sup>

# INTRODUÇÃO À GENÔMICA

## Genomas de plantas sequenciados

<https://www.plabipd.de/index.ep>

plaBiPD

~400 genomas  
333 Angiospermas  
15 Gimnospermas  
2 Chlorophyta  
44 Algas Verdes

- IMPRESSUM INFORMATION
- GDPR PRIVACY NOTICE

Published plant genomes

- Chronology (timeline)
- Phylogeny (cladogram)

Protein function annotation

- About Mercator
- Mercator4 (v.2.0)
- Mercator (v.3.6)

Plant genome projects

- *Solanum pennellii*
- *Cuscuta campestris*

PLANT 2030

- plaBi-PD
- gaBi-PD



Oldenlandia  
affinis



Passiflora



Pea



Peach tree



Pepper



Pine  
processionary



Rapeseed



Sugarcane



Sunflower



Tomato



Vanilla

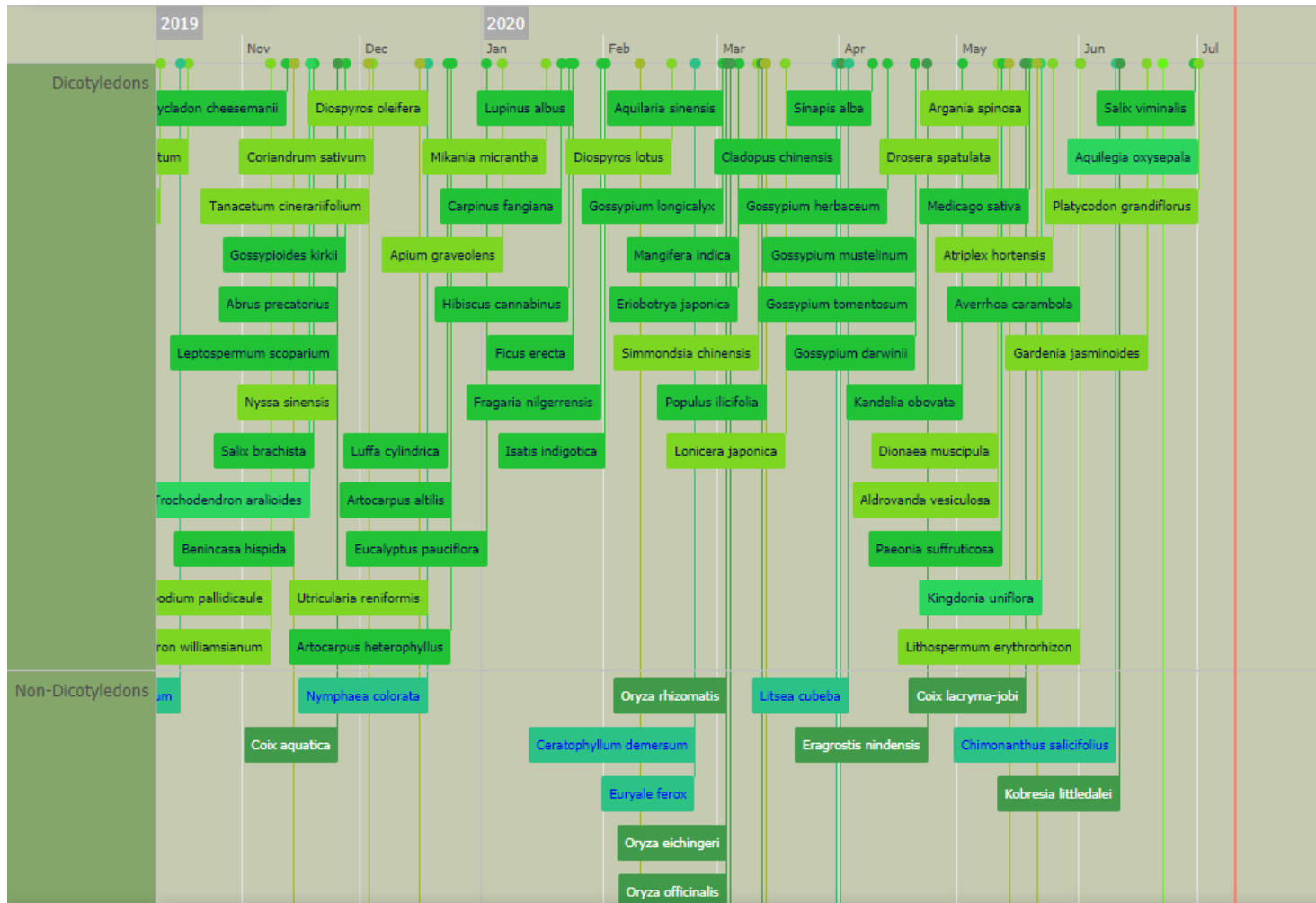


Wheat

# INTRODUÇÃO À GENÔMICA

## Genomas de plantas sequenciados

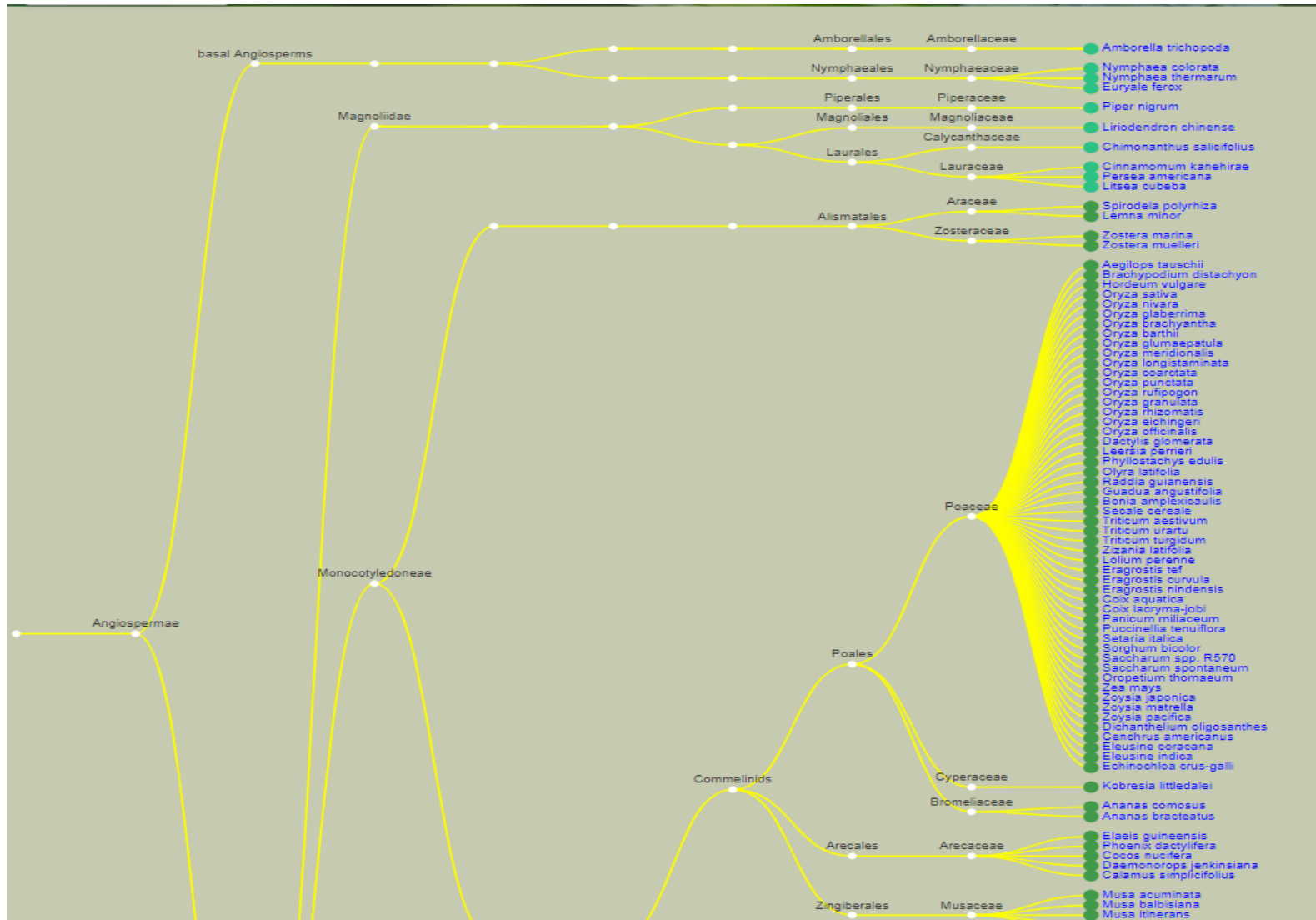
<https://www.plabipd.de/index.ep> - apresentação cronológica das espécies com genomas publicados



# INTRODUÇÃO À GENÔMICA

## Genomas de plantas sequenciados

<https://www.plabipd.de/index.ep> - apresentação filogenética das espécies com genomas publicados



# INTRODUÇÃO À GENÔMICA

- Complexidade dos genomas eucariotos

## Variação no tamanho de genomas

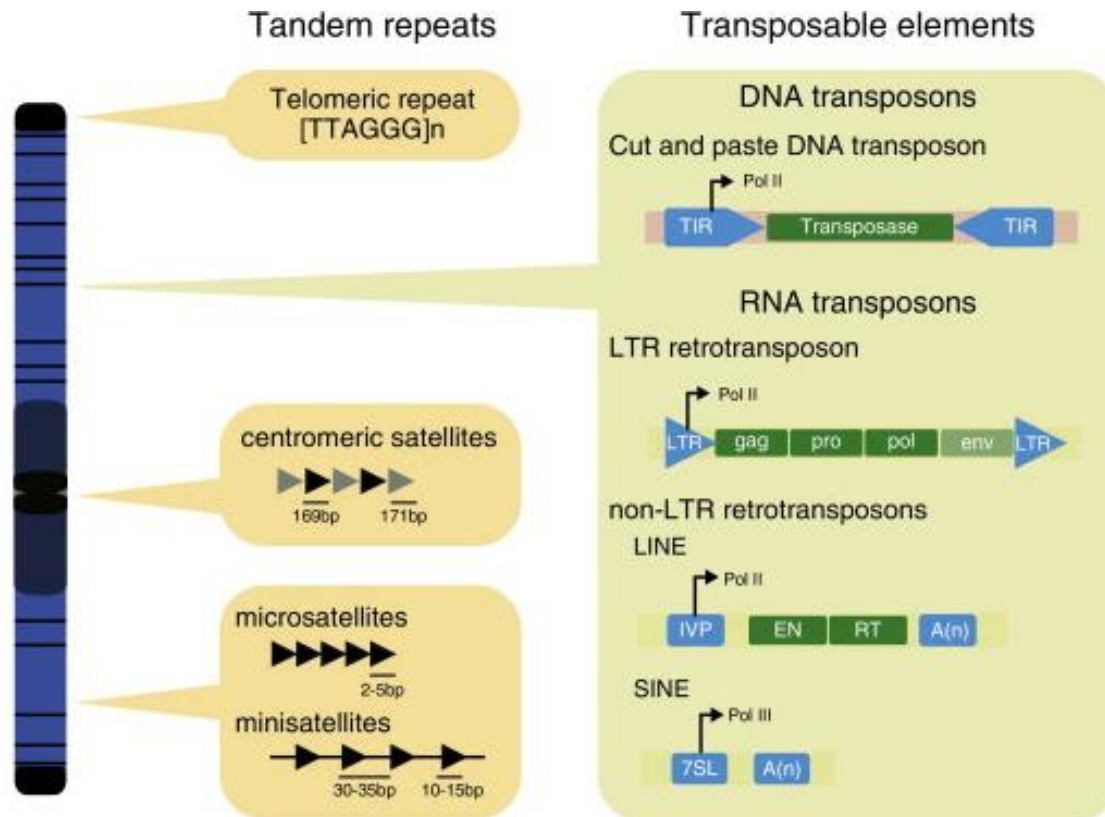
Organismo	Tamanho do genoma (pb)
$\lambda$ (bacteriophage)	50,000
<i>E. coli</i> (bacterium)	4,640,000
<i>Saccharomyces cerevisiae</i> (yeast)	12,000,000
<i>Arabidopsis thaliana</i> (plant)	167,000,000
<i>Drosophila melanogaster</i> (insect)	180,000,000
<i>Homo sapiens</i> (human)	3,400,000,000
<i>Zea mays</i> (corn)	4,500,000,000
<i>Amphiuma</i> (salamander)	765,000,000,000



# INTRODUÇÃO À GENÔMICA

- Complexidade dos genomas eucariotos

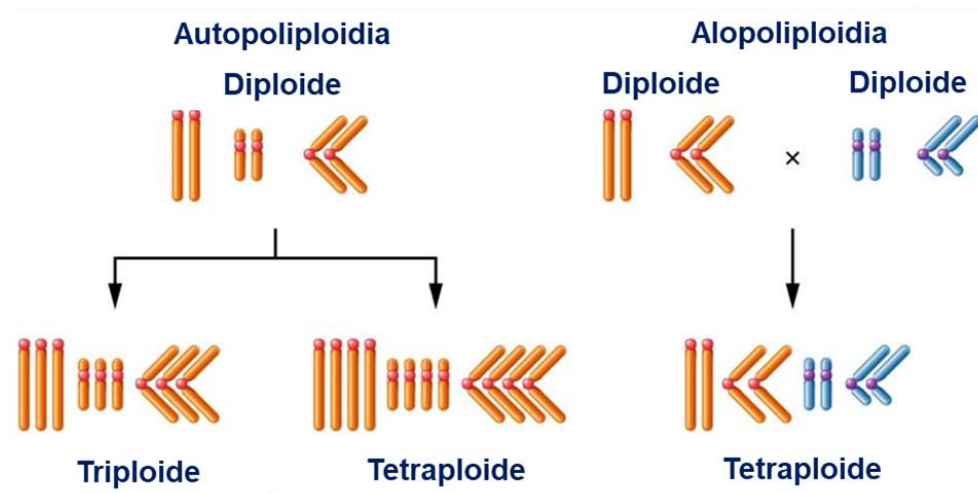
## Porção repetitiva do genoma



Representação esquemática das classes de repetição, sua distribuição e suas características estruturais

# INTRODUÇÃO À GENÔMICA

- Complexidade dos genomas eucariotos
- **Poliploidia: condição na qual uma célula ou organismo adquire um ou mais conjuntos adicionais de cromossomos**
- Comum entre plantas, bem como entre certos grupos de animais
- Autopoliploide: conjunto cromossômico derivado de uma única espécie
- Alopoliploide: conjunto cromossômico derivado de diferentes espécies





# INTRODUÇÃO À GENÔMICA

- Obtenção de genomas

## OVERVIEW

Sequenciamento

Pré-processamento das  
sequências

Montagem

Anotação estrutural

Anotação funcional



# Pré-processamento das sequências

## ➤ Arquivo Fastq

Nome

Sequência

+

Qualidade

```
@D3NZ4HQ1:111:D2DM2ACXX:1:1101:1243:2110 2:N:0:TGACCA
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC
+
!' '*((( (***) ) %%%++) (%%%) .1***-+*'') **55CCF>>>>>CC
```

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99,9%
40	1 in 10,000	99,99%
50	1 in 100.000	99,999%
60	1 in 1.000.000	99,9999%

# Pré-processamento das sequências



- Controle de qualidade dos dados brutos

Qua 8 jul 2020  
phaema1r1.fastq

## FastQC Report

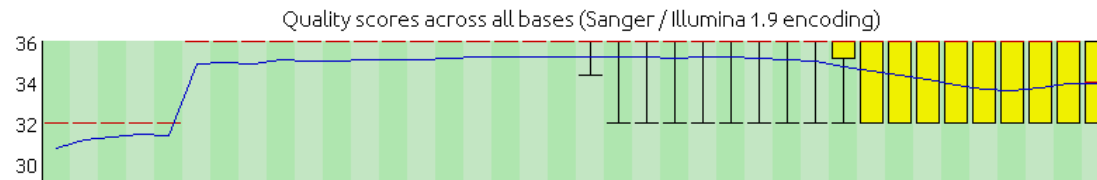
### Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✗ [Per base sequence content](#)
- ! [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

### ✓ Basic Statistics

Measure	Value
Filename	phaema1r1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2214862
Sequences flagged as poor quality	0
Sequence length	1-151
%GC	38

### ✓ Per base sequence quality



# Pré-processamento das sequências

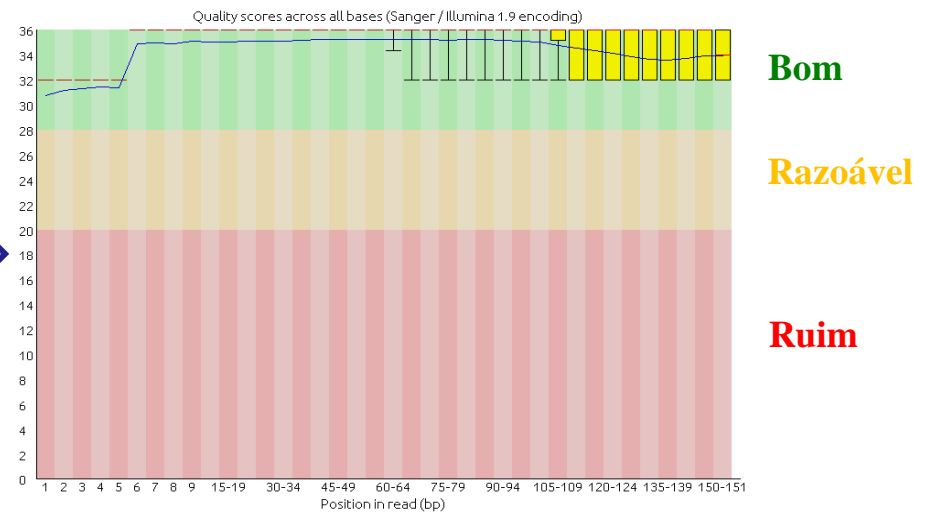
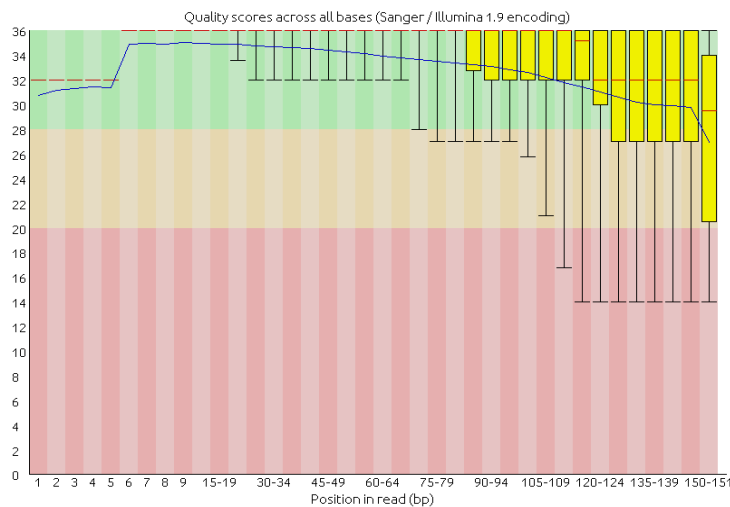
- Controle de qualidade dos dados brutos

## 1. Qualidade das bases

- *Phred* 34 - 40;
- *Phred* +64 (ASCII 59 a 126) / *Phred* +33 (ASCII 0 a 62) - atentar para escala de *phred* usada no software;

- Retirar (*trimming*) bases de baixa qualidade do final das reads;  
`./reformat.sh qtrim=r trimq=30 in=8-_S13_L001_R1_001.fastq out=L1R1.fastq`

`./reformat.sh in=L1R1.fastq out=L1R1.fasta`



# Pré-processamento das sequências

- Controle de qualidade dos dados brutos

## ➤ Armazenamento dos dados

- Renomear as sequências e transformar o arquivo fastq em fasta garantem economia de espaço



<b>L1R1_1.fastq</b>	<b>L1R1_2.fastq</b>	<b>L1R1.fasta</b>
815,6 Mb	546,8 Mb	216,7 Mb

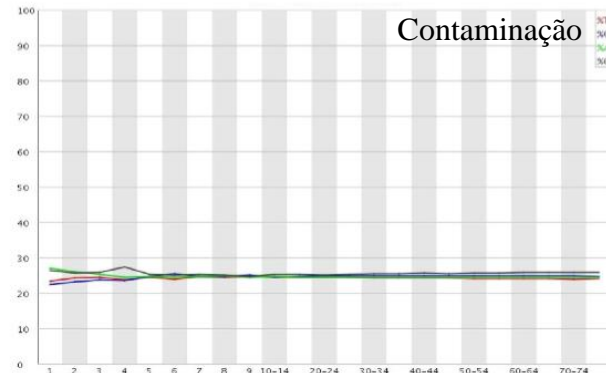
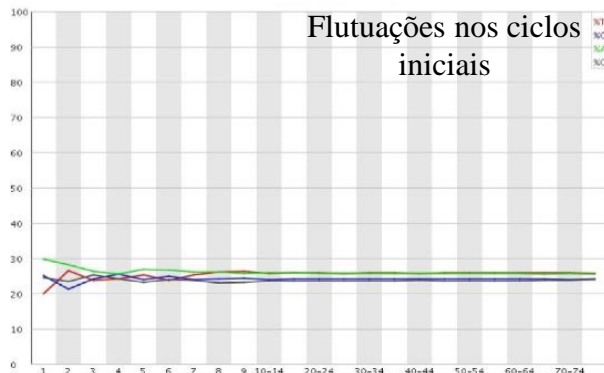
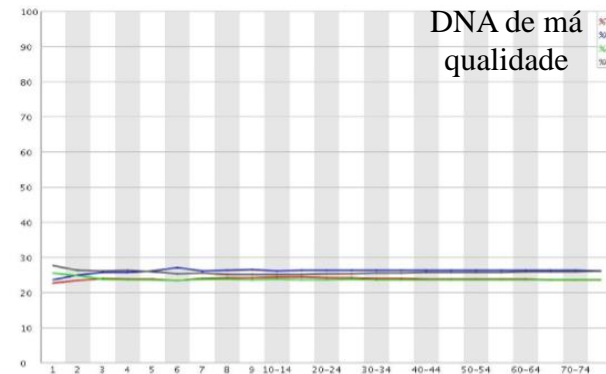
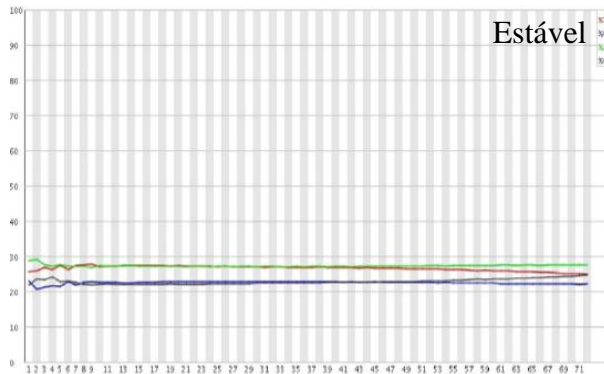
# Pré-processamento das sequências

- Controle de qualidade dos dados brutos

## 2. Distribuição nucleotídica

- Deve permanecer estável ao longo das reads;
- Associação com a qualidade das bases;

**%T**  
**%C**  
**%A**  
**%G**

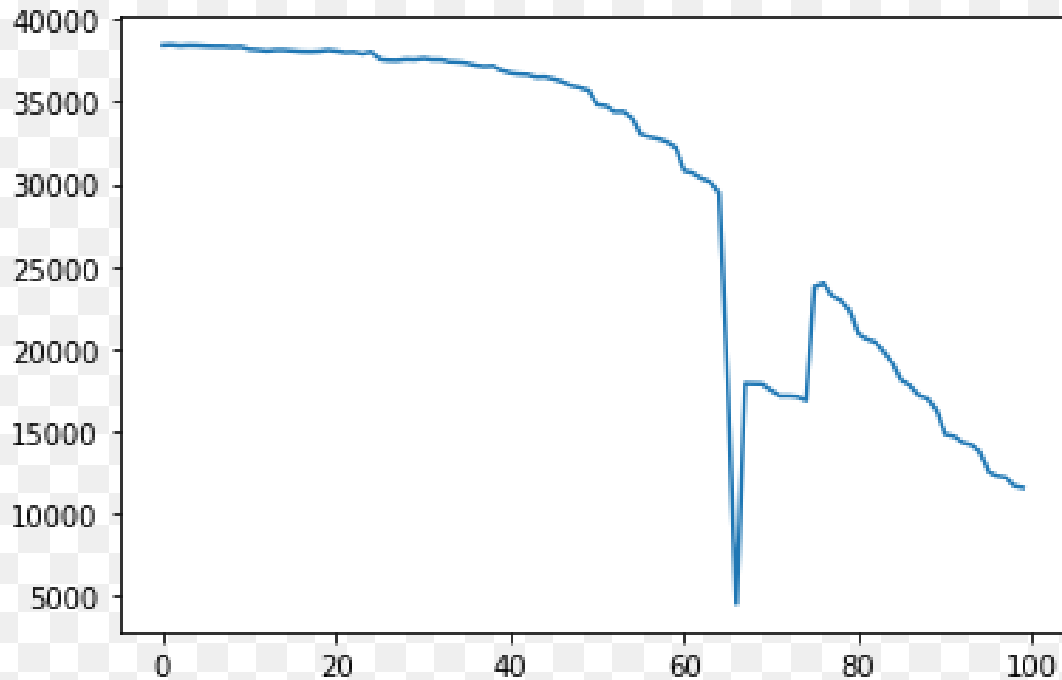


# Pré-processamento das sequências

- Controle de qualidade dos dados brutos

## 3. Distribuição do conteúdo GC

- % GC varia entre espécies e entre regiões do genoma;
- WES: % GC 49-51 - WGS: % GC 38-39 (humanos);
- % GC 38-42: *Saccharomyces cerevisiae* e *Mycobacterium tuberculosis*;
- >10% de desvio pode indicar contaminação;



# Pré-processamento das sequências

- Controle de qualidade dos dados brutos

## SOFTWARES

- **FastQC package: visualização** - qualidade média das bases por read, distribuição do conteúdo GC, identificação das reads duplicadas, etc;
- **FASTX-Toolkit**: qualidade das bases, distribuição nucleotídica;
- **PRINSEQ**
- **BBTools - reformat.sh**



# Galaxy: recurso gratuito, público e acessível na internet



**Galaxy** Analyze Data Workflow Visualize Shared Data Help User

All tools should be functioning normally with the exception of RNA STAR.

**Tools** ☆ ⬆

search tools ✕

- Get Data
- Collection Operations
- GENERAL TEXT TOOLS
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Datamash
- GENOMIC FILE MANIPULATION**
- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore
- Convert Formats
- Lift-Over
- COMMON GENOMICS TOOLS
- Operate on Genomic Intervals
- Fetch
- Sequences/Alignments
- GENOMICS ANALYSIS
- Assembly
- Annotation

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

An illustration of two sneakers with a galaxy-themed pattern on the sides and soles. The design is credited to Rebekka Paisner.

Design by Rebekka Paisner

**James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.**

To ensure implementation of this vision, the Galaxy community has established a foundation—Junior Training and Educational Connections Hotspot (JTech). JTech's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting contributions. More details can be found on [the @jtx page in the Galaxy Hub](#). Please, help us continue what James has started.

[Donate now](#)

**i** Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at [covid19.galaxyproject.org](https://covid19.galaxyproject.org)

The logo for Penn State University, featuring a lion's head and the text 'PennState'.

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at Oregon Health & Science University.

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

**Galaxy:** recurso gratuito, público e acessível na internet



coursera

Explorar ▾

O que você deseja aprender?



Navegar > Ciência de Dados > Análise de Dados

## Análise de Dados Genômicos através do Projeto Galaxy

★★★★★ 3.7 724 classificações • 210 avaliações



James Taylor, PhD

Parte da **Série de cursos 8 Programa de cursos integrados Análise de Dados Genômicos**  
**Análise Genômica**

Ir para o curso

Já inscrito

18.620 já se inscreveram

oferecido por



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

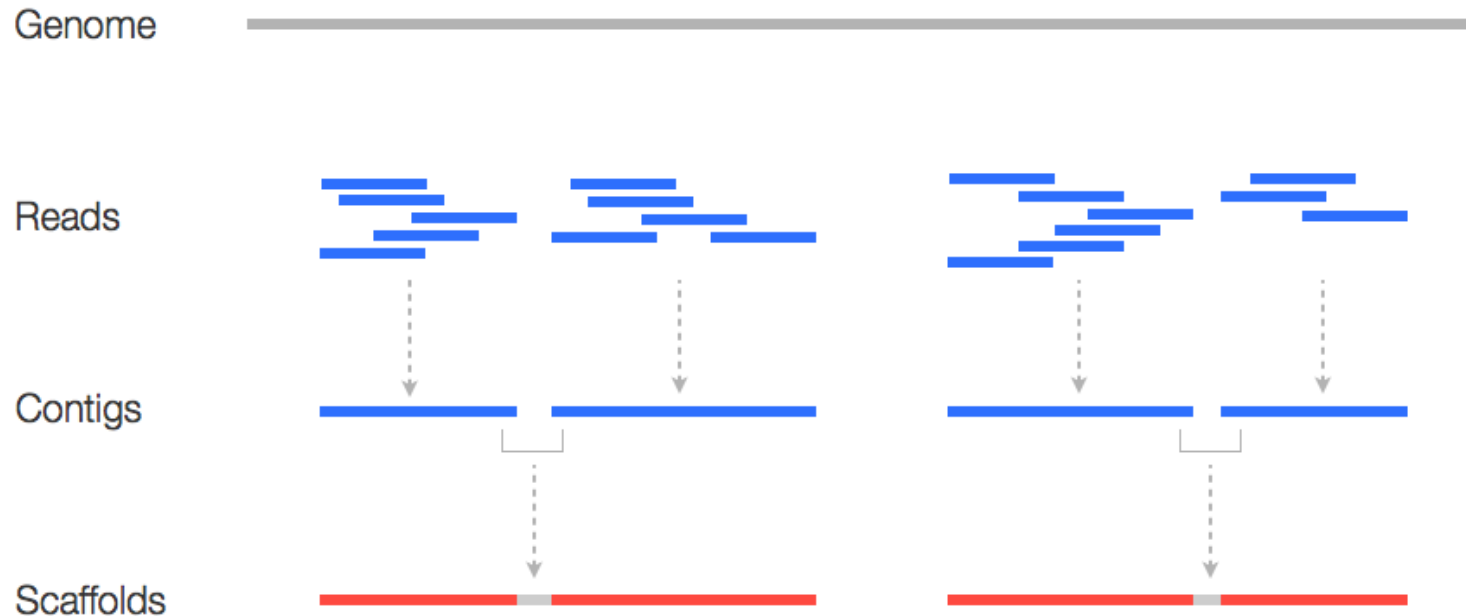
## ➤ Montagem

Processo de reconstrução da sequência de DNA original a partir das *reads* obtidas com o sequenciamento

**Read:** Fragmento que foi sequenciado

**Contigs:** Peçaço contíguo de sequência formado a partir da sobreposição de *reads*

**Scaffolds:** Resultado da conexão entre *contigs*



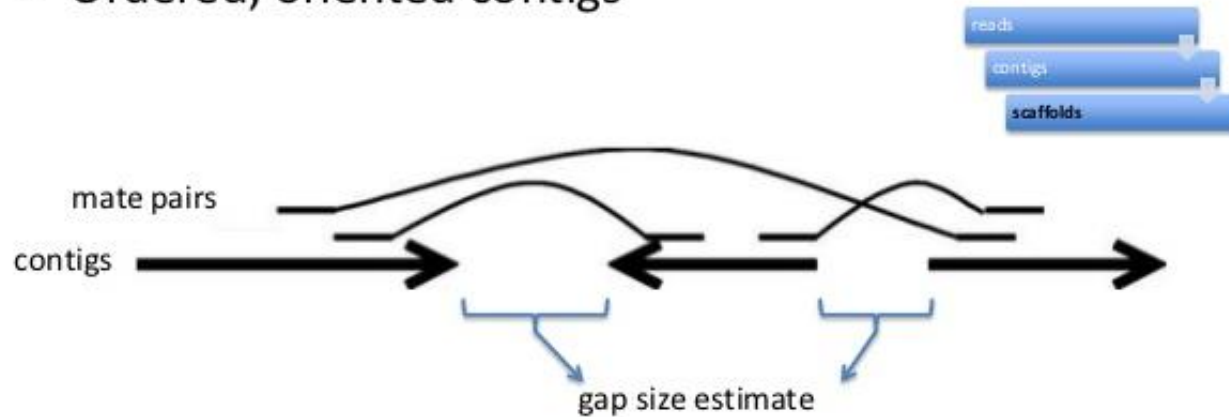
# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Montagem

### Obtenção dos *scaffolds*

#### Scaffolds

- Ordered, oriented contigs



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Montagem



- De-novo



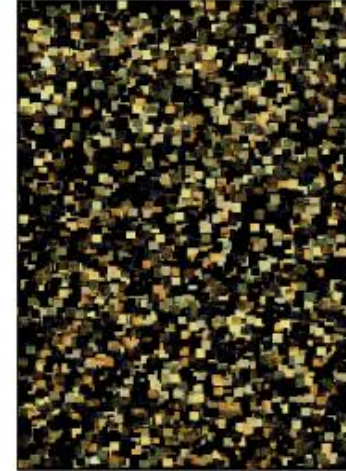
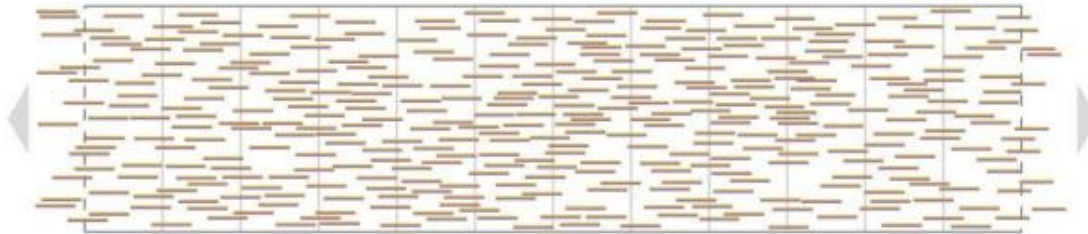
- Reference



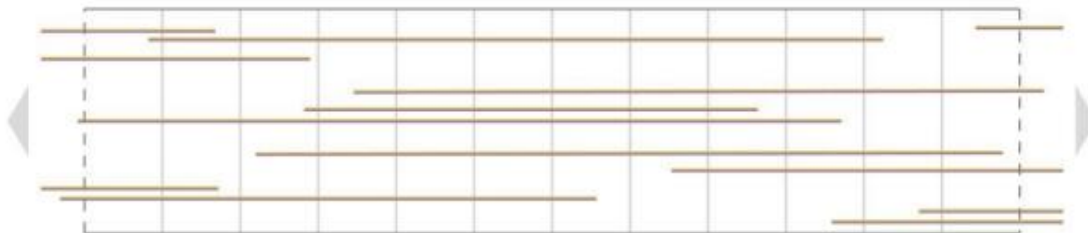
# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Montagem

Short Reads



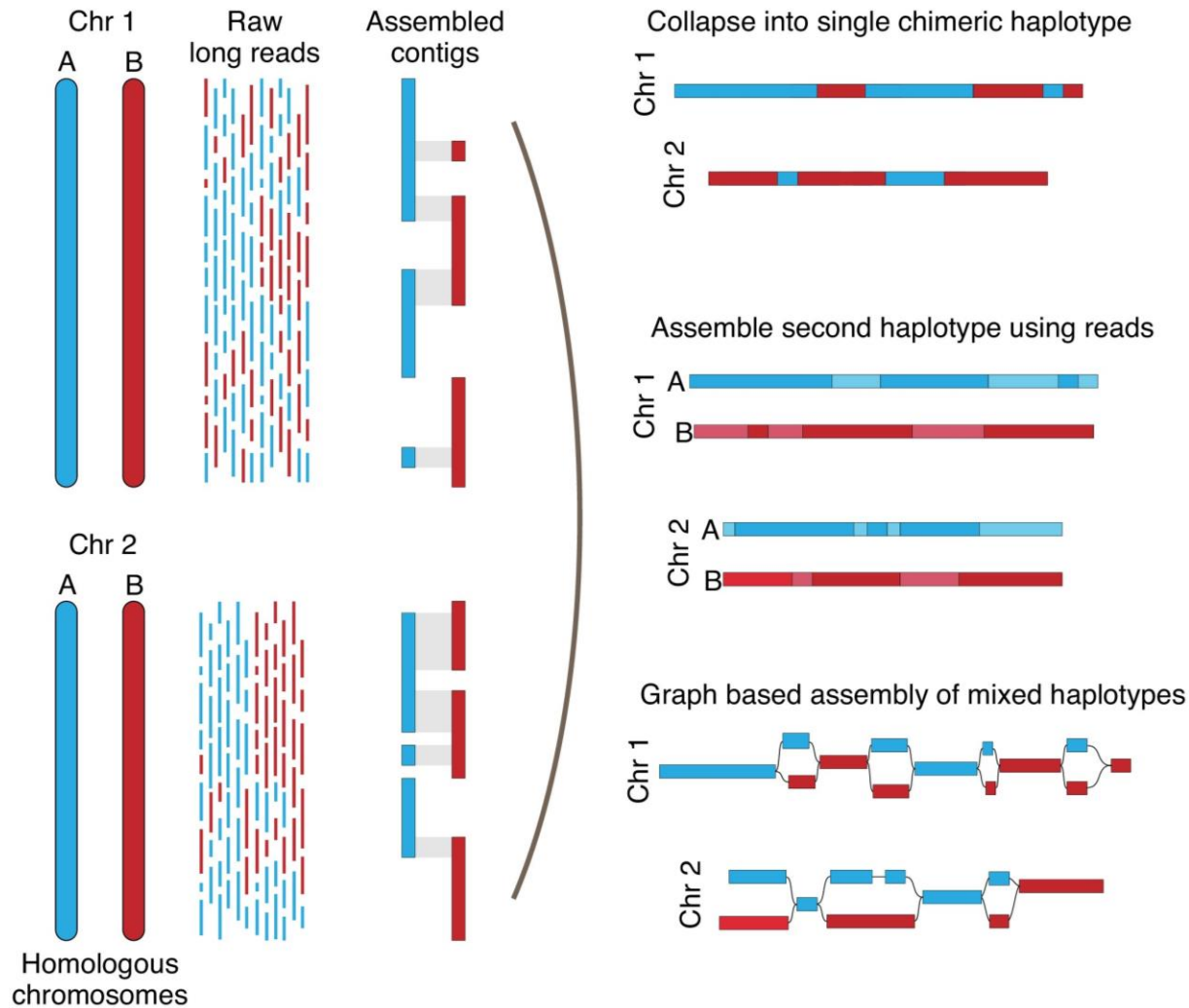
Long Reads





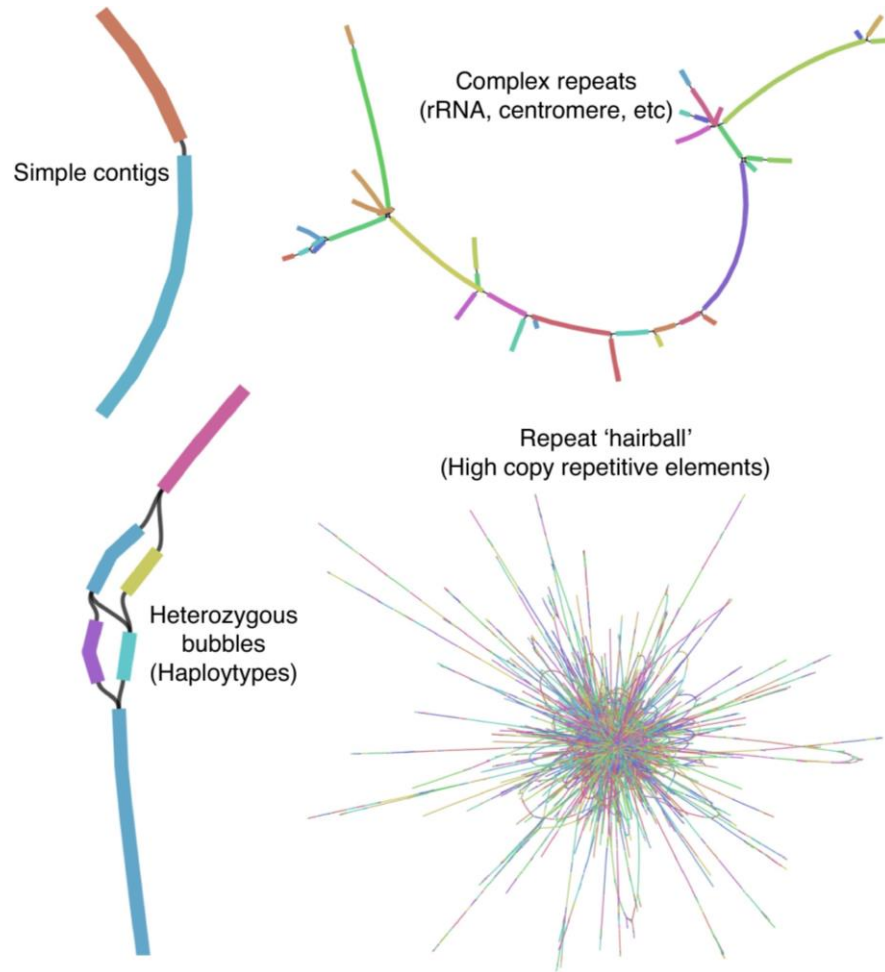
# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

- Principais desafios em projetos de sequenciamento de genomas de plantas
  - ✓ Espécies altamente heterozigóticas: muitas diferenças nos cromossomos homólogos



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Gráfico de Montagem



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

**ScienceDirect**

**Building near-complete plant genomes**  
Todd P Michael<sup>1</sup> and Robert VanBuren<sup>2,3</sup>

*Bioinformatics*, 31(20), 2015, 3350–3352  
doi: 10.1093/bioinformatics/btv383  
Advance Access Publication Date: 22 June 2015  
Applications Note

Genome analysis

**Bandage: interactive visualization of *de novo* genome assemblies**

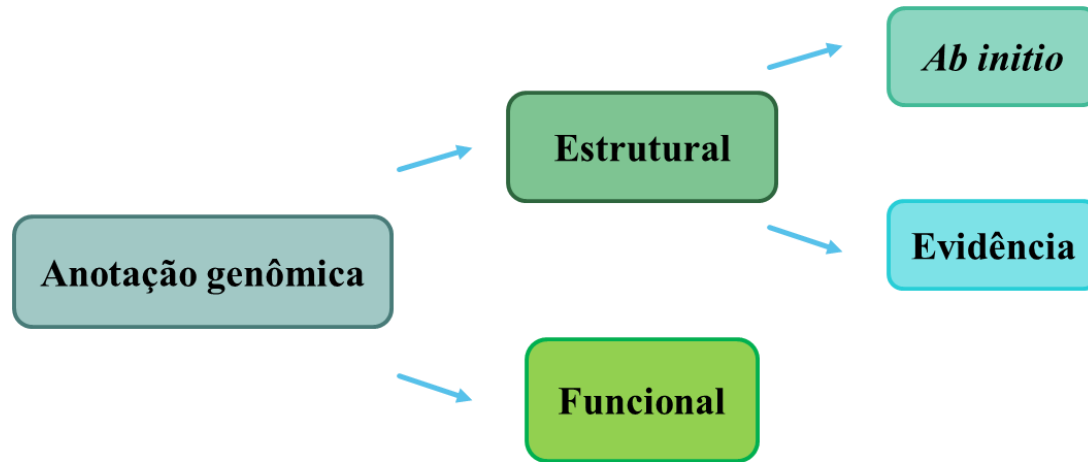
Ryan R. Wick<sup>1,\*</sup>, Mark B. Schultz<sup>1</sup>, Justin Zobel<sup>2</sup> and Kathryn E. Holt<sup>1</sup>



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação de genomas

- atribuição de informação às sequências genômicas realizada através de processos computacionais



- ✓ **Anotação estrutural:** processo de identificação de genes e seus íntrons e éxons
- ✓ **Anotação funcional:** processo de atribuir informações, como termos de ontologia gênica, à anotação estrutural

# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

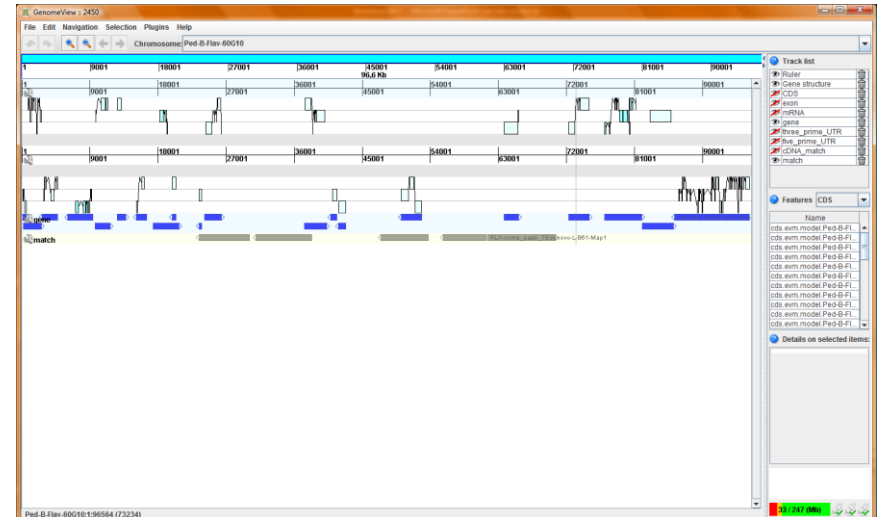
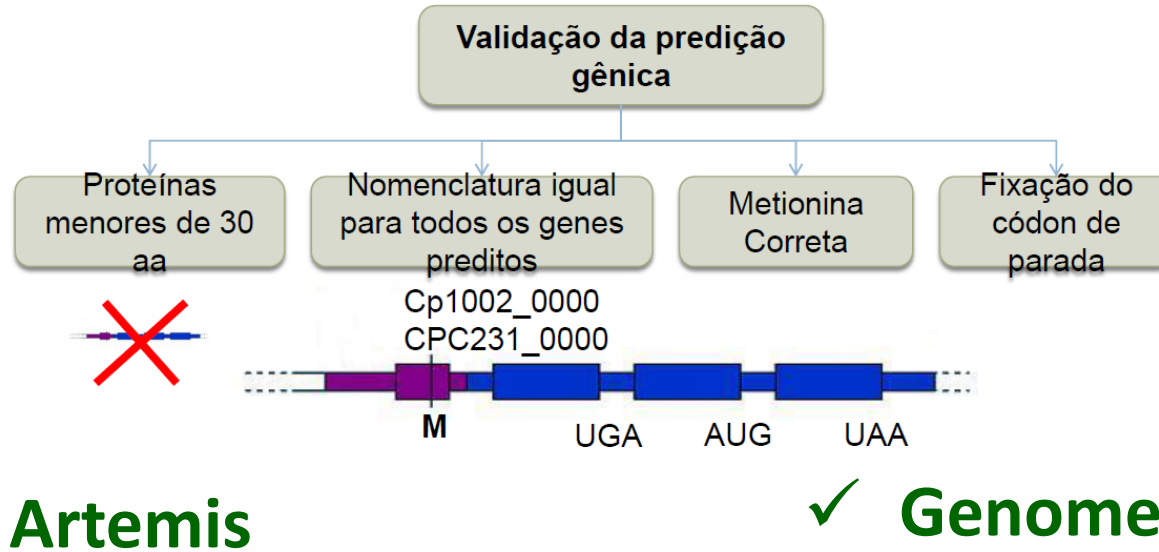
## ➤ Anotação estrutural

aatgcatgCGGctatgctaataatgcatgCGGctatgctaagctGGGatccgatgacaatgcatgCGGctatgctaa  
tgcatgCGGctatgcaagctGGGatccgatgactatgctaagctGGGatccgatgacaatgcatgCGGctatgc  
taatgaatGGTcttGGGatttaccttGgaatgctaagctGGGatccgatgacaatgcatgCGGctatgctaatga  
atGGTcttGGGatttaccttGgaatgctaataatgcatgCGGctatgctaagctGGGatccgatgacaatgcatgc  
GGctatgctaataatgcatgCGGctatgcaagctGGGatccgatgactatgctaagctgCGGctatgctaataatgcatg  
CGGctatgctaagctGGGatccgatgacaatgcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatc  
ctgCGGctatgctaataatgaatGGTcttGGGatttaccttGgaatgctaagctGGGatccgatgacaatgcatgCGG  
ctatgctaataatgaatGGTcttGGGatccgatgacaatgcatgCGGctatgctaagctGGGaatgcatg  
CGGctatgctaagctGGGatccgatgacaatgcatgCGGctatgcaagctGGGatc  
cgatgactatgctaagctgCGGctatgctaataatgcatgCGGctatgctaagctcatgCGGctatgctaagctGGG  
aatgcatgCGGctatgctaagctGGGatccgatgacaatgcatgCGGctatgctaataatgcatgCGGctatgcaag  
ctGGGatccgatgactatgctaagctgCGGctatgctaataatgcatgCGGctatgctaagctCGGctatgctaataatga  
atGGTcttGGGatttaccttGgaatgctaagctGGGatccgatgacaatgcatgCGGctatgctaataatgaatGGTc  
ttGGGatttaccttGgaatgctaataatgcatgCGGctatgctaagctGGGaatgcatgCGGctatgctaagctGG  
gatccgatgacaatgcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatccgatgactatgctaagc  
tgCGGctatgctaataatgcatgCGGctatgctaagctcatgCGG

Gene!

# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Curadoria Manual



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação Funcional

✓ BLAST: Basic Local Alignment Search Tool

- BLASTX : verificar similaridade com proteínas nos bancos de dados
- BLASTN: verificar similaridade com ESTs e RNA ribossomal

## NCBI: National Center for Biotechnology Information

### Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

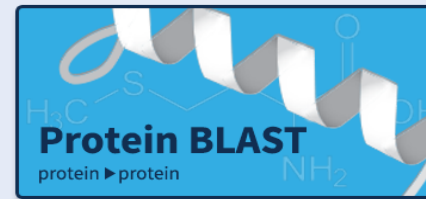
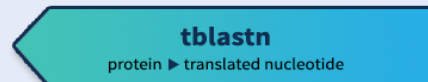
#### QuickBLASTP

Try [QuickBLASTP](#) for a fast protein search of nr.

Tue, 23 May 2017 13:00:00 EST

[More BLAST news...](#)

### Web BLAST



### BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

[Human](#)

[Mouse](#)

[Rat](#)

[Microbes](#)

# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação Funcional

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

### Overview

Entry	Protein names	Match hit	Identity
Q9FI46	Chromophore lyase CRL, chloroplastic ( <i>Arabidopsis thaliana</i> )		77.6%
B0JGA0	Chromophore lyase CpcT/CpeT 3 ( <i>Microcystis aeruginosa</i> (stra..))		33.5%
Q7NNH3	Chromophore lyase CpcT/CpeT 1 ( <i>Gloeobacter violaceus</i> (strai..))		30.0%



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação Funcional

### Gene Ontology

- ✓ Primeira ontologia criada em biologia molecular, 2000
- ✓ Consórcio para a padronização da anotação gênica
- ✓ Vocabulário padrão para a descrição de genes em três categorias
  - Processo biológico
  - Função molecular
  - Localização celular



#### Example GO Term

```
id:          GO:0000016
name:        lactase activity
namespace:   molecular function
def:         "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]
synonym:     "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym:     "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref:        EC:3.2.1.108
xref:        MetaCyc:LACTASE-RXN
xref:        Reactome:20536
is_a:        GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

OPEN

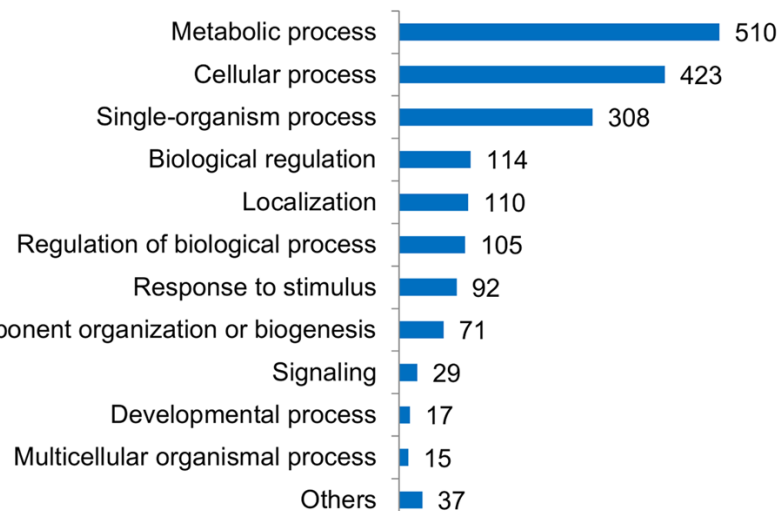
## A gene-rich fraction analysis of the *Passiflora edulis* genome reveals highly conserved microsyntenic regions with two related Malpighiales species

Received: 23 April 2018  
 Accepted: 14 August 2018  
 Published online: 29 August 2018

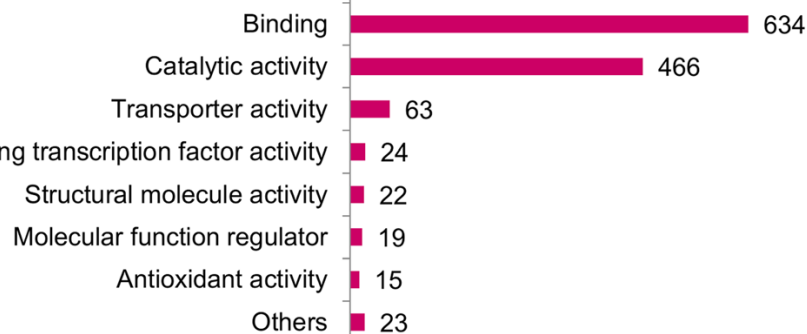
Carla Freitas Munhoz<sup>1</sup>, Zirlane Portugal Costa<sup>1</sup>, Luiz Augusto Cauz-Santos<sup>2</sup>, Alina Carmen Egoavil Reátegui<sup>2</sup>, Nathalie Rodde<sup>2</sup>, Stéphane Cauet<sup>2</sup>, Marcelo Carnier Dornelas<sup>3</sup>, Philippe Leroy<sup>4</sup>, Alessandro de Mello Varani<sup>5</sup>, Hélène Bergès<sup>2</sup> & Maria Lúcia Cameiro Vieira<sup>2</sup>

*Passiflora edulis* is the most widely cultivated species of passionflowers, cropped mainly for industrialized juice production and fresh fruit consumption. Despite its commercial importance, little is known about the genome structure of *P. edulis*. To fill in this gap in our knowledge, a genomic library was built, and now completely sequenced over 100 large-inserts. Sequencing data were assembled from long sequence reads, and structural sequence annotation resulted in the prediction of about 1,900 genes, providing data for subsequent functional analysis. The richness of repetitive elements was also evaluated. Microsyntenic regions of *P. edulis* common to *Papulus trichocarpa* and *Manihot esculenta*, two related Malpighiales species with available fully sequenced genomes were examined. Overall, gene order was well conserved, with some disruptions of collinearity identified as rearrangements, such as inversion and translocation events. The microsynteny level observed between the *P. edulis* sequences and the compared genomes is surprising, given the long divergence time that separates them from the common ancestor. *P. edulis* gene-rich segments are more compact than those of the other two species, even though its genome is much larger. This study provides a first accurate gene set for *P. edulis*, opening the way for new studies on the evolutionary issues in Malpighiales genomes.

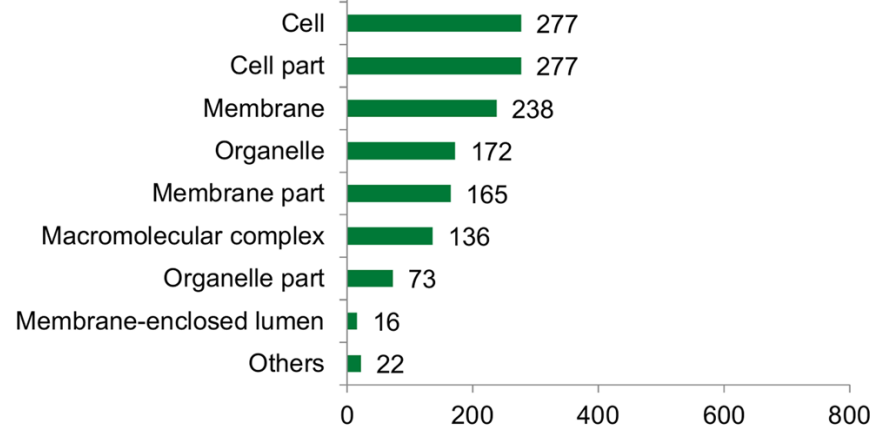
A



B



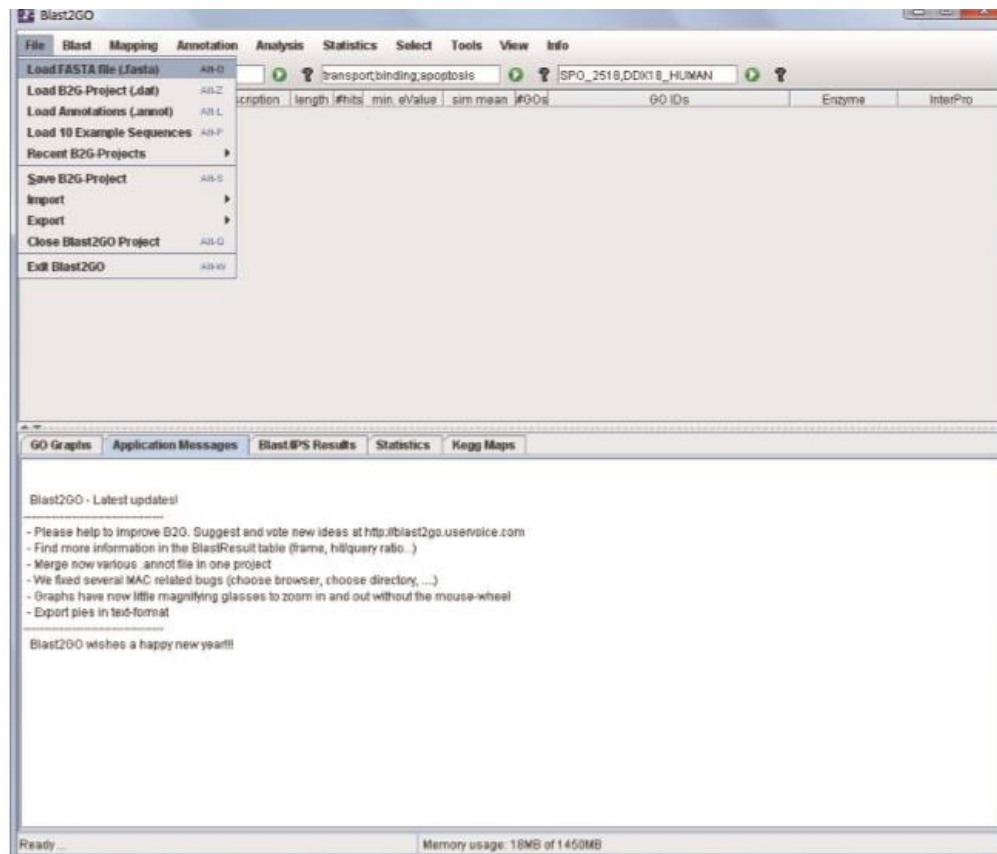
C



# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação Funcional

✓ Anotação automática

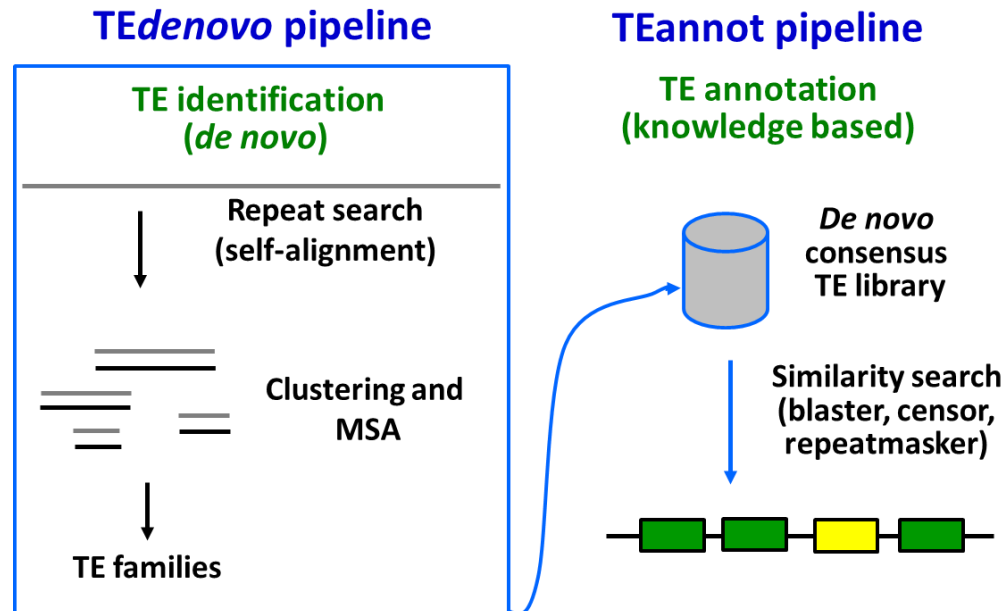




# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação de elementos de transposição

- Identificação e Classificação: REPET/PASTEC

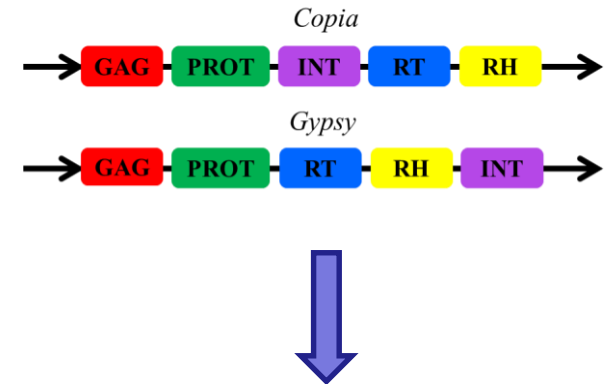


Métodos utilizados para a identificação de elementos de transposição

# INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

## ➤ Anotação de elementos de transposição

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
<b>Class I (retrotransposons)</b>					
LTR	<i>Copia</i>	→ GAG AP INT RT RH →	4-6	RLC	P, M, F, O
	<i>Gypsy</i>	→ GAG AP RT RH INT →	4-6	RLG	P, M, F, O
	<i>Bel-Pao</i>	→ GAG AP RT RH INT →	4-6	RLB	M
	<i>Retrovirus</i>	→ GAG AP RT RH INT ENV →	4-6	RLR	M
	<i>ERV</i>	→ GAG AP RT RH INT ENV →	4-6	RLE	M
DIRS	<i>DIRS</i>	→ GAG AP RT RH YR →	0	RYD	P, M, F, O
	<i>Ngaro</i>	→ GAG AP RT RH YR →	0	RYN	M, F
	<i>VIPER</i>	→ GAG AP RT RH YR →	0	RYV	O
PLE	<i>Penelope</i>	← RT EN →	Variable	RPP	P, M, F, O
LINE	<i>R2</i>	— RT EN —	Variable	RIR	M
	<i>RTE</i>	— APE RT —	Variable	RIT	M
	<i>Jockey</i>	— ORF1 — APE RT —	Variable	RIJ	M
	<i>L1</i>	— ORF1 — APE RT —	Variable	RIL	P, M, F, O
	<i>I</i>	— ORF1 — APE RT RH —	Variable	RII	P, M, F
SINE	<i>tRNA</i>	— — —	Variable	RST	P, M, F
	<i>7SL</i>	— — —	Variable	RSL	P, M, F
	<i>5S</i>	— — —	Variable	RSS	M, O
<b>Class II (DNA transposons) - Subclass 1</b>					
TIR	<i>Tc1-Mariner</i>	→ Tase* ←	TA	DTT	P, M, F, O
	<i>hAT</i>	→ Tase* ←	8	DTA	P, M, F, O
	<i>Mutator</i>	→ Tase* ←	9-11	DTM	P, M, F, O
	<i>Merlin</i>	→ Tase* ←	8-9	DTE	M, O
	<i>Transib</i>	→ Tase* ←	5	DTR	M, F
	<i>P</i>	→ Tase ←	8	DTP	P, M
	<i>PiggyBac</i>	→ Tase ←	TTAA	DTB	M, O
	<i>PIF-Harbinger</i>	→ Tase* — ORF2 ←	3	DTH	P, M, F, O
	<i>CACTA</i>	→ Tase — ORF2 ←	2-3	DTC	P, M, F
	<i>Crypton</i>	→ YR ←	0	DYC	F
<b>Class II (DNA transposons) - Subclass 2</b>					
<i>Helitron</i>	<i>Helitron</i>	→ RPA — Y2 HEL ←	0	DHI	P, M, F
<i>Maverick</i>	<i>Maverick</i>	→ C-INT — ATP — CYP — POL B ←	6	DMM	M, F, O

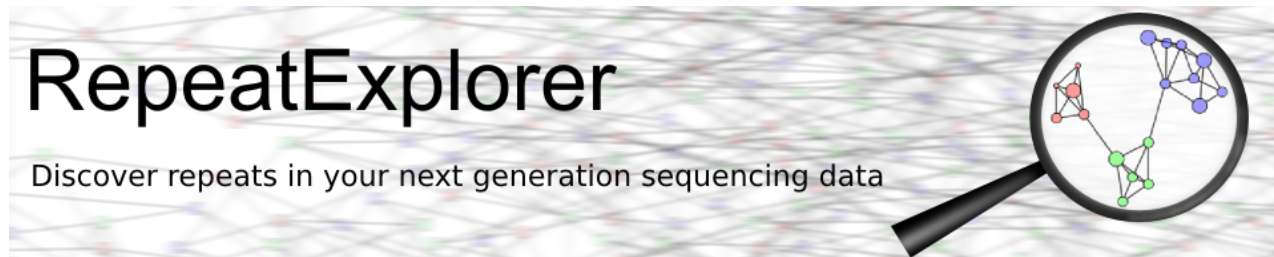


Linhagens evolutivas

Classificação dos elementos de transposição segundo Wicker et al. (2007)

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ Anotação de elementos de transposição



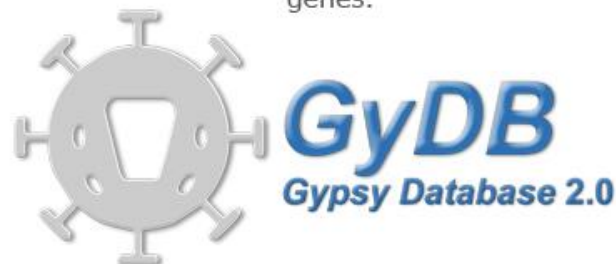
### ➤ Identificação e classificação de elementos de transposição em sequências curtas de DNA

- Busca por domínios de LTRs
- Análise filogenética de domínios

- Bancos de dados



A cooperative scientific network devoted to the evolutionary dynamics of Mobile Genetic Elements, Viruses and related host genes.



# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

RESEARCH

Open Access



Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification

Pavel Neumann<sup>1</sup>, Petr Novák, Nina Hošťáková and Jiří Macas

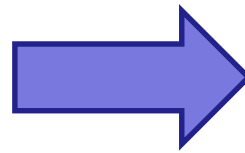
## Abstract

**Background:** Plant LTR-retrotransposons are classified into two superfamilies, Ty1/copia and Ty3/gypsy. They are further divided into an enormous number of families which are, due to the high diversity of their nucleotide sequences, usually specific to a single or a group of closely related species. Previous attempts to group these families into broader categories reflecting their phylogenetic relationships were limited either to analyzing a narrow range of plant species or to analyzing a small number of elements. Furthermore, there is no reference database that allows for similarity based classification of LTR-retrotransposons.

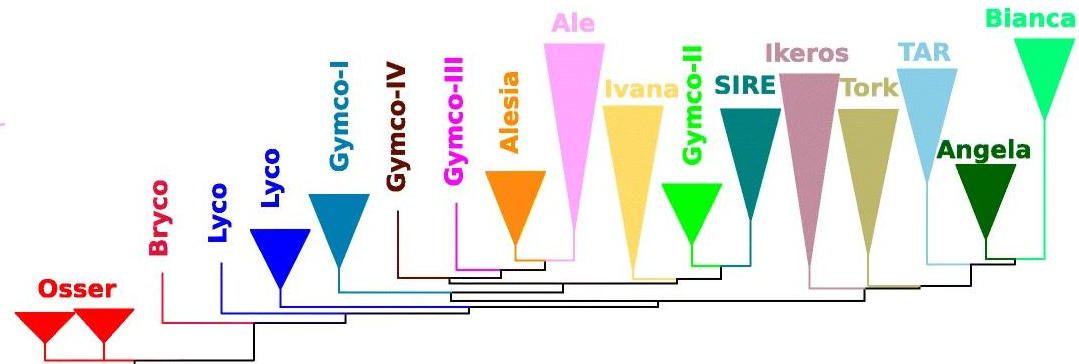
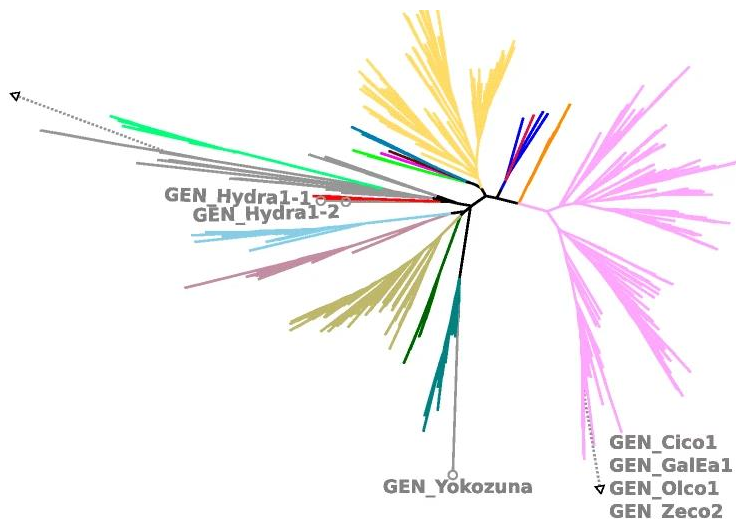
**Results:** We have assembled a database of retrotransposon encoded polyprotein domains sequences extracted from 5410 Ty1/copia elements and 8453 Ty3/gypsy elements sampled from 80 species representing major groups of green plants (Viridiplantae). Phylogenetic analysis of the three most conserved polyprotein domains (RT, RH and INT) led to dividing Ty1/copia and Ty3/gypsy retrotransposons into 16 and 14 lineages respectively. We also characterized various features of LTR-retrotransposon sequences including additional polyprotein domains, extra open reading frames and primer binding sites, and found that the occurrence and/or type of these features correlates with phylogenies inferred from the three protein domains.

**Conclusions:** We have established an improved classification system applicable to LTR-retrotransposons from a wide range of plant species. This system reflects phylogenetic relationships as well as distinct sequence and structural features of the elements. A comprehensive database of retrotransposon protein domains (REXdb) that reflects this classification provides a reference for efficient and unified annotation of LTR-retrotransposons in plant genomes. Access to REXdb related tools is implemented in the RepeatExplorer web server (<https://repeatexplorer.elmir.cerit-sc.cz/>) or using a standalone version of REXdb that can be downloaded separately from RepeatExplorer web page (<http://repeatexplorer.org/>).

**Keywords:** LTR-retrotransposons, Transposable elements, Polyprotein domains, Primer binding site, RepeatExplorer



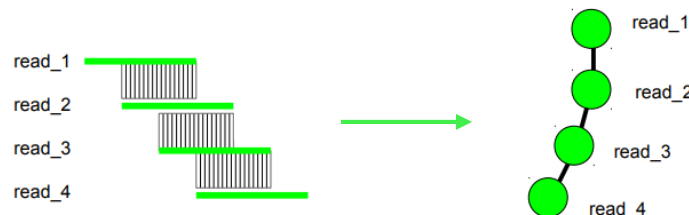
**REXdb:** a reference database of transposable element protein domains



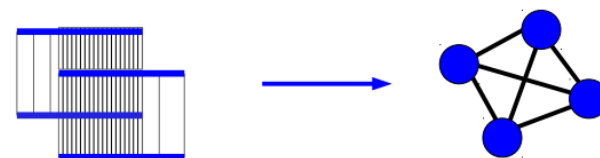
# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

1. Comparação par a par de todas as leituras;



2. Agrupamento das leituras que compartilham semelhanças significativas de sequência em clusters;



- ✓ Esses clusters representam principalmente repetições, porque apenas as leituras derivadas de sequências presentes no genoma várias vezes podem produzir um número suficiente de ocorrências de similaridade nos dados de sequenciamento de baixa cobertura (0,01 a 0,50x);
- ✓ Em princípio, o número de leituras em cada cluster é proporcional à abundância genômica da repetição correspondente, possibilitando sua quantificação;

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR



## ➤ RepeatExplorer

### 1. Arquivo na forma interlaçada



```
>0001_f CGTAATATACATACTTGCTAGCTAGT  
>0001_r GATTTGACGGACACACTAAGCTA  
>0002_f ACTCATTGGACTTAACTTTGATAAT  
>0002_r TATGTTGAAAAATTGAATTCGGGAC  
>0003_f TGACATTTGTGAACGTTAATGTTCAA  
>0003_r TATTGAAATACTGGACACAAATTGGA
```

### 2. Cálculo da quantidade de sequências para a cobertura desejada

$$\text{Genome Coverage} = \frac{(N^\circ \text{ de reads} \times \text{Tamanho das reads})}{\text{Tamanho do genoma}}$$

Exemplo: genoma de 1.000 Mb  
Cobertura desejada: 0,25  
Reads: 100 pb



$$0,25 = \frac{(N^\circ \text{ de reads} \times 100)}{1.000 \times 10^6}$$

$$N^\circ \text{ de reads} = 2.500.000$$



$$0,25 \times 1.000 \times 10^6 = 250.000.000 \text{ pb}$$

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

- ✓ Inserir as sequências processadas
- ✓ Processá-las dentro do RepeatExplorer

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆ ⬇  
search tools ✕

Get Data  
RepeatExplorer2  
TAREAN  
ChIP-Seq Mapper  
**RepeatExplorer Utilities**  
DANTE  
PROFREP  
Text Manipulation  
**NGS: QC and manipulation**  
JBrowse  
FASTA manipulation  
Experimental Tools  
Obsolete tools  
Workflows

## RepeatExplorer

Discover repeats in your next generation sequencing data

Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic

This RepeatExplorer Galaxy portal is a part of services provided by ELIXIR (European research infrastructure for biological information). Please acknowledge this fact in your publications by adding a statement: "Computational resources for RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure."

If **issue to open tools/history** (e.g. `[history_id=None] Failed to retrieve history.`) then please try to clean your browser cache and delete cookies or try to open new browser window in anonymous mode.

### Resources

- Documentation and training information
- The impact of genome coverage and sequence read sampling on reproducibility of repeat identification
- Official Galaxy Project website with information how to use the Galaxy platform
- For command line version of the RepeatExplorer tools, see the source code repository

Follow @RepeatExplorer

**History** ⌕ + ⌵ ⚙  
search datasets ✕

**Unnamed history**  
4 shown, 46 deleted  
62.83 GB ✓ ⌵ ⌶

- 50: RepeatExplorer2 - HTML report from data 47 ✓ ⌵ ✕
- 49: RepeatExplorer2 - Archive with HTML report from data 47 ✓ ⌵ ✕
- 48: RepeatExplorer2 - logfile ✓ ⌵ ✕
- 47: Pd\_11\_25.fa ✓ ⌵ ✕

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆

search tools

**Get Data**

RepeatExplorer2

TAREAN

ChIP-Seq Mapper

RepeatExplorer Utilities

DANTE

PROFREP

Text Manipulation

NGS: QC and manipulation

JBrowse

FASTA manipulation

Experimental Tools


Obsolete tools

Workflows

All workflows

# RepeatExplorer

Discover repeats in your next generation sequencing data



Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic

This RepeatExplorer Galaxy portal is a part of services provided by ELIXIR (European research infrastructure for biological information). Please acknowledge this fact in your publications by adding a statement: "Computational resources for RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure."

If **issue to open tools/history** (e.g. [history\_id=None] Failed to retrieve history..) then please try to clean your browser cache and delete cookies or try to open new browser window in anonymous mode.

## Resources

- Documentation and training information
- The impact of genome coverage and sequence read sampling on reproducibility of repeat identification
- Official Galaxy Project website with information how to use the Galaxy platform
- For command line version of the RepeatExplorer tools, see the source code repository

[Follow @RepeatExplorer](#)

## News

**History**

search datasets

**Unnamed history**

4 shown, 46 deleted

62.83 GB

50: RepeatExplorer2 - HTML report from data 47

49: RepeatExplorer2 - Archive with HTML report from data 47

48: RepeatExplorer2 - logfile

47: Pd\_I1\_25.fa

Caixa de entrada x NCGR Summer B... x Inbox - zirlane@... x Bbmap - Anaconi... x BBTools - DOE Jo... x Galaxy x Galaxy Platform... x Como usar o Rep... x Galaxy x

repeatexplorer-elixir.cerit-sc.cz/galaxy/

Apps Gmail e-mail do G... ESALQ - USP Google Tradutor Google Fundação de Amp... Oportunidades de... Whole Genome Se... Coursera | Online C... Published plant ge... Galaxy

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆

search tools

**Get Data**

Upload File from your computer

EBI SRA ENA SRA

RepeatExplorer2

TAREAN

ChIP-Seq Mapper

RepeatExplorer Utilities

DANTE

PROFREP

Text Manipulation

NGS: QC and manipulation

JBrowse

FASTA manipulation

Experimental Tools

Obsolete tools

Workflows

All workflows

**Download from web or upload from disk**

Regular Composite Collection Rule-based

Drop files here

Type (set all): Auto-detect Genome (set all): Additional S...

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

**History**

search datasets

**Unnamed history**

4 shown, 46 deleted

62.83 GB

50: RepeatExplorer2 - HTML report from data 47

49: RepeatExplorer2 - Archive with HTML report from data 47

48: RepeatExplorer2 - logfile

47: Pd\_I1\_25.fa

## News

Jan 13, 2020 Galaxy server was updated to version 19.09

RepeatExplorer2 and TAREAN tools were updated to version 2.3.7

Jan 13, 2020 See changelog for detailed description of included changes. Previous versions can be found in Obsolete tools section

RepeatExplorer tools are now available in Galaxy toolshed

Packages repeatexplorer2, dante and re\_utils can be installed from Galaxy toolshed to custom Galaxy instance if you prefer analysis using RepeatExplorer on your server.

Jan 8, 2020

Data quota increased to 200G

Nov 10, 2019 Data quota for registered users was increased from 50G to 200G to enable easier analysis of larger datasets

Changelog



# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## RepeatExplorer

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆ ⬇ search tools ✕

**Get Data**

**RepeatExplorer2**

**RepeatExplorer2 clustering:** improved version or repeat discovery and characterization using graph-based sequence clustering

**TAREAN**

**ChIP-Seq Mapper**

**RepeatExplorer Utilities**

**DANTE**

**PROFREP**

**Text Manipulation**

**NGS: QC and manipulation**

**JBrowse**

**FASTA manipulation**

**Experimental Tools**

**Obsolete tools**

**Workflows**

All workflows

## RepeatExplorer

Discover repeats in your next generation sequencing data

Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic

This RepeatExplorer Galaxy portal is a part of services provided by ELIXIR (European research infrastructure for biological information). Please acknowledge this fact in your publications by adding a statement: "Computational resources for RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure."

If **issue to open tools/history** (e.g. [history\_id=None] Failed to retrieve history..) then please try to clean your browser cache and delete cookies or try to open new browser window in anonymous mode.

### Resources

- Documentation and training information
- The impact of genome coverage and sequence read sampling on reproducibility of repeat identification
- Official Galaxy Project website with information how to use the Galaxy platform
- For command line version of the RepeatExplorer tools, see the source code repository

[Follow @RepeatExplorer](#)

### News

Jan 13, 2020 **Galaxy server was updated to version 19.09**

**RepeatExplorer2 and TAREAN tools were updated to version 2.3.7**

**History** ↻ + 🗄 ⚙ search datasets ✕

**Unnamed history**

4 shown, 46 deleted

62.83 GB

**50: RepeatExplorer2 - HTML report from data 47** 👁 ✎ ✕

**49: RepeatExplorer2 - Archive with HTML report from data 47** 👁 ✎ ✕

**48: RepeatExplorer2 - log file** 👁 ✎ ✕

**47: Pd\_I1\_25.fa** 👁 ✎ ✕

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## RepeatExplorer

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** search tools

**Get Data**

**RepeatExplorer2**

**RepeatExplorer2 clustering:** Improved version or repeat discovery and characterization using graph-based sequence clustering (Galaxy Version 2.3.8) Favorite Options

**NGS reads** 47: Pd\_I1\_25.fa

Input file must contain FASTA-formatted NGS reads. Illumina paired-end reads are recommended.

**Paired-end reads** Yes No

If paired-end reads are used, left- and right-hand reads must be interlaced and all pairs must be complete. Example of the correct format is provided in the help below.

**Read sampling** Yes No

Use this option if you want to analyze only a part of the reads

**Select taxon and protein domain database version (REXdb)** Viridiplantae version 3.0

Reference database of transposable element protein domains - REXdb - is used for annotation of repeats

**Advanced options** Yes No

**Select queue** long (max runtime 2 weeks, 64 GB RAM)

**Modify parameters (optional)** -l select=1:ncpus=16:mem=112gbs:scratch\_local=50gb -l walltime=336:00:00 -q elixirre@pbs.elixir-czech.cz -v TAREAN\_MAX\_MEM=64000000,TAREAN\_CPU=15

Execute RepeatExplorer2 HELIX clustering; (2.3.8)

RepeatExplorer2 clustering is a computational pipeline for unsupervised identification of repeats from unassembled sequence reads. The pipeline uses low-pass whole genome sequence reads and performs graph-based clustering. Resulting clusters, representing all types of repeats, are then examined to identify and classify into repeats groups.

**Input data**

**History** search datasets

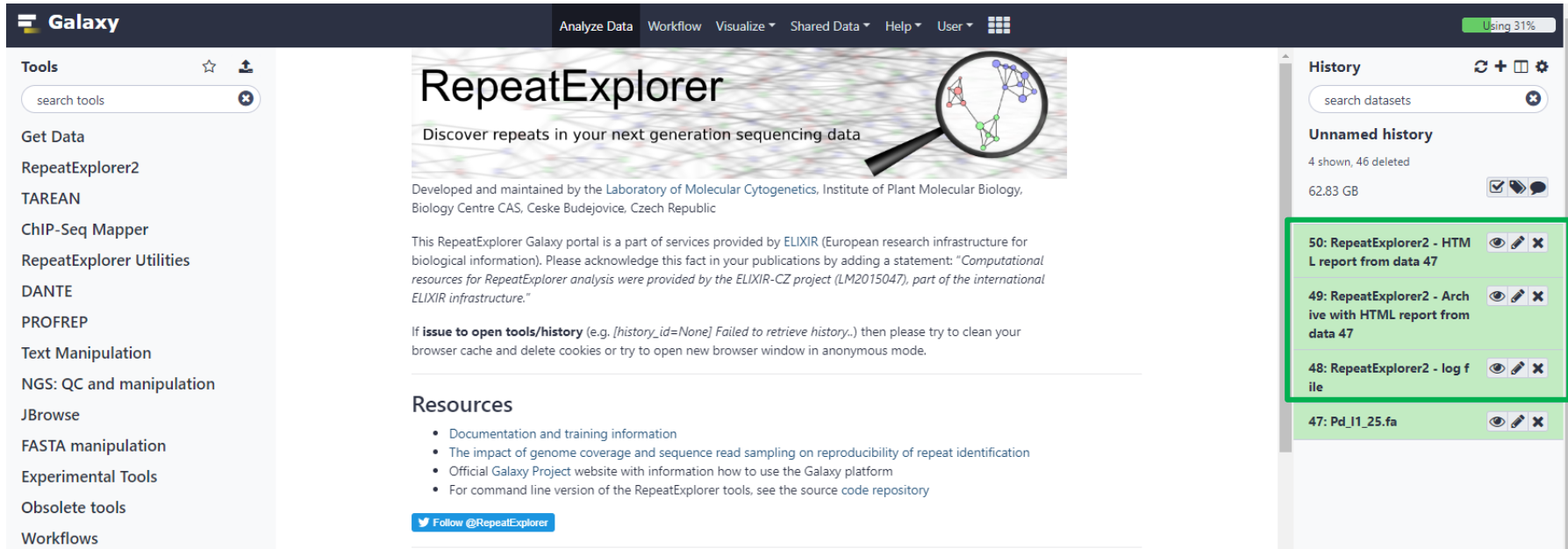
**Unnamed history** 4 shown, 46 deleted 62.83 GB

- 50: RepeatExplorer2 - HTML report from data 47
- 49: RepeatExplorer2 - Archive with HTML report from data 47
- 48: RepeatExplorer2 - log file
- 47: Pd\_I1\_25.fa

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

### ✓ Resultados



**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆ ⬇

search tools ✕

Get Data

RepeatExplorer2

TAREAN

ChIP-Seq Mapper

RepeatExplorer Utilities

DANTE

PROFREP

Text Manipulation

NGS: QC and manipulation

JBrowse

FASTA manipulation

Experimental Tools

Obsolete tools

Workflows

## RepeatExplorer

Discover repeats in your next generation sequencing data

Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic

This RepeatExplorer Galaxy portal is a part of services provided by ELIXIR (European research infrastructure for biological information). Please acknowledge this fact in your publications by adding a statement: "Computational resources for RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure."

If **issue to open tools/history** (e.g. [history\_id=None] Failed to retrieve history.) then please try to clean your browser cache and delete cookies or try to open new browser window in anonymous mode.

### Resources

- Documentation and training information
- The impact of genome coverage and sequence read sampling on reproducibility of repeat identification
- Official Galaxy Project website with information how to use the Galaxy platform
- For command line version of the RepeatExplorer tools, see the source code repository

Follow @RepeatExplorer

**History** ↻ + 🗑 ⚙

search datasets ✕

**Unnamed history**

4 shown, 46 deleted

62.83 GB

- 50: RepeatExplorer2 - HTML report from data 47
- 49: RepeatExplorer2 - Archive with HTML report from data 47
- 48: RepeatExplorer2 - log file
- 47: Pd\_I1\_25.fa

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

### ✓ Resultados

The screenshot displays the RepeatExplorer Galaxy interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and a 'Using 31%' indicator. The left sidebar lists various tools under 'Tools', including RepeatExplorer2, TAREAN, ChIP-Seq Mapper, RepeatExplorer Utilities, DANTE, PROFREP, Text Manipulation, NGS: QC and manipulation, JBrowse, FASTA manipulation, Experimental Tools, Obsolete tools, and Workflows. The main content area features the RepeatExplorer logo and a magnifying glass over a network graph. Below the logo, it states: 'Discover repeats in your next generation sequencing data'. The text describes the tool's development by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic. It also mentions that the RepeatExplorer Galaxy portal is part of services provided by ELIXIR. A note indicates an issue with opening tools/history and suggests cleaning the browser cache. The 'Resources' section lists documentation, genome coverage impact, the Galaxy Project website, and the source code repository. The 'News' section contains three entries: 'Galaxy server was updated to version 19.09' (Jan 13, 2020), 'RepeatExplorer2 and TAREAN tools were updated to version 2.3.7' (Jan 13, 2020), and 'RepeatExplorer tools are now available in Galaxy toolshed' (Jan 8, 2020). The right sidebar shows the 'History' section with a search bar and a list of datasets. The highlighted dataset is '49: RepeatExplorer2 - Archive with HTML report from data 47', which is 22.6 GB and in zip format. The preview shows a list of files being added to the archive, including CLUSTER\_TABLE.csv (deflated 79%), HOW\_TO\_CITE.html (deflated 49%), PROFREP\_CLASSIFICATION\_TEMPLATE.csv (deflated 82%), SUPERCLUSTER\_TABLE.csv (deflated 92%), and TAREAN\_consensus\_rank\_... The bottom of the history list shows '48: RepeatExplorer2 - log file' and '47: Pd\_11\_25.fa'.

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

### Tools

search tools

Get Data

- RepeatExplorer2
- TAREAN
- ChIP-Seq Mapper
- RepeatExplorer Utilities
- DANTE
- PROFREP
- Text Manipulation
- NGS: QC and manipulation
- JBrowse
- FASTA manipulation
- Experimental Tools
- Obsolete tools
- Workflows
- All workflows

## RepeatExplorer

Discover repeats in your next generation sequencing data

Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic

This RepeatExplorer Galaxy portal is a part of services provided by ELIXIR (European research infrastructure for biological information). Please acknowledge this fact in your publications by adding a statement: "Computational resources for RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure."

If **issue to open tools/history** (e.g. [history\_id=None] Failed to retrieve history..) then please try to clean your browser cache and delete cookies or try to open new browser window in anonymous mode.

### Resources

- Documentation and training information
- The impact of genome coverage and sequence read sampling on reproducibility of repeat identification
- Official Galaxy Project website with information how to use the Galaxy platform
- For command line version of the RepeatExplorer tools, see the source code repository

[Follow @RepeatExplorer](#)

### News

Jan 13, 2020 **Galaxy server was updated to version 19.09**

Jan 13, 2020 **RepeatExplorer2 and TAREAN tools were updated to version 2.3.7**  
See changelog for detailed description of included changes. Previous versions can be found in *Obsolete tools* section

Jan 8, 2020 **RepeatExplorer tools are now available in Galaxy toolshed**  
Packages repeatexplorer2, dante and re\_utils can be installed from Galaxy toolshed to custom Galaxy instance if you prefer analysis using RepeatExplorer on on your server.

### History

search datasets

**Unnamed history**  
4 shown, 46 deleted  
62.83 GB

**50: RepeatExplorer2 - HTML report from data 47**

**49: RepeatExplorer2 - Archive with HTML report from data 47**  
22.6 GB  
format: zip, database: ?

Scanning files .....  
adding: CLUSTER\_TABLE.csv (deflated 79%)  
adding: HOW\_TO\_CITE.html (deflated 49%)  
adding: PROFREP\_CLASSIFICATION\_TEMPLATE.csv (deflated 82%)  
adding: SUPERCLUSTER\_TABLE.csv (deflated 92%)  
adding: TAREAN\_consensus\_rank\_...

Compressed zip file

**48: RepeatExplorer2 - log file**

**47: Pd\_11\_25.fa**

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

### ✓ Resultados

Galaxy Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆ ⬇

search tools ⊕

- Get Data
- RepeatExplorer2
- TAREAN
- ChIP-Seq Mapper
- RepeatExplorer Utilities
- DANTE
- PROFREP
- Text Manipulation
- NGS: QC and manipulation
- JBrowse
- FASTA manipulation
- Experimental Tools
- Obsolete tools
- Workflows
- All workflows

### Clustering Summary

**2661222 reads total**

2071799 reads in 74865 superclusters (74932 clusters)

589423 singlets

Number of reads

Proportion of reads [%]

**Graphical summary of the clustering results.** Bars represent superclusters, with their heights and widths corresponding to the numbers of reads in the superclusters (y-axis) and to their proportions in all analyzed reads (x-axis), respectively. Rectangles inside the supercluster bars represent individual clusters. If the filtering of abundant satellites was performed, the affected clusters are shown in green, and their sizes correspond to the adjusted values. Blue and pink background panels show proportions of reads that were clustered and remained single, respectively. Top clusters are on the left of the dotted line.

**Run information:**

- Number of input reads: 2661222
- Number of analyzed reads: 2661222
- Proportion of reads in top clusters : 68 %
- Cluster merging: No
- Paired-end reads: Yes

**History** ↺ + ⊞ ⚙

search datasets ⊕

**Unnamed history**

4 shown, 46 deleted

62.83 GB ☑ 🗨

- 50: RepeatExplorer2 - HT ML report from data 47 👁 ✎ ✕
- 49: RepeatExplorer2 - Archive with HTML report from data 47 👁 ✎ ✕
- 48: RepeatExplorer2 - log file 👁 ✎ ✕
- 47: Pd\_I1\_25.fa 👁 ✎ ✕

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## RepeatExplorer

### ✓ Resultados

**Galaxy** Analyze Data Workflow Visualize Shared Data Help User Using 31%

**Tools** ☆ ⬆

search tools ✕

**Get Data**

- RepeatExplorer2
- TAREAN
- ChIP-Seq Mapper
- RepeatExplorer Utilities
- DANTE
- PROFREP
- Text Manipulation
- NGS: QC and manipulation
- JBrowse
- FASTA manipulation
- Experimental Tools
- Obsolete tools
- Workflows
- All workflows

**Run information:**

- Number of input reads: 2661222
- Number of analyzed reads: 2661222
- Proportion of reads in top clusters : 68 %
- Cluster merging: No
- Paired-end reads: Yes

**Available analyses:**

- [Tandem repeat analysis](#)
- [Cluster annotation](#)
- [Supercluster annotation](#)
- [Repeat annotation summary](#)

**How to cite**

Novak, P., Neumann, P., Pech, J., Steinhaisl, J., 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.

*Classification of repetitive elements using REXdb:*

Neumann, P., Novak, P., Hostakova, N., Macas, J. (2019) - [Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification](#). *Mobile DNA* **10**:1.

*The principle of repeat identification implemented in the RepeatExplorer:*

Novak, P., Neumann, P., Macas, J. (2010) - [Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data](#). *BMC Bioinformatics* **11**:378.

*Using TAREAN for satellite repeat detection and characterization:*

Novak, P., Robledillo, L.A., Koblizkova, A., Vrbova, I., Neumann, P., Macas, J. (2017) - [TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads](#). *Nucleic Acid Research* **45**:e111

**History** ⬆ ⬇ ⚙

search datasets ✕

**Unnamed history**

4 shown, 46 deleted

62.83 GB

- 50: RepeatExplorer2 - HT ML report from data 47
- 49: RepeatExplorer2 - Archive with HTML report from data 47
- 48: RepeatExplorer2 - log file
- 47: Pd\_11\_25.fa

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

✓ Resultados: Sumário dos repeats anotados

The screenshot displays the Galaxy web interface. At the top, there is a navigation bar with 'Galaxy' logo, 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and a 'Using 31%' indicator. On the left, a 'Tools' sidebar lists various tools like RepeatExplorer2, TAREAN, ChIP-Seq Mapper, etc. The main content area is titled 'Repeat annotation summary' and contains a table with the following columns: 'Proportion[%]', 'Nsuperclusters', 'Nclusters', and 'Nreads'. The table lists various repeat elements such as rDNA, 45S\_rDNA, 18S\_rDNA, 25S\_rDNA, 5.8S\_rDNA, 5S\_rDNA, satellite, mobile\_element, Class\_I, SINE, LTR, Ty1\_copia, Ale, Alesia, Angela, Bianca, Bryco, Lyco, Gymco-III, Gymco-I, Gymco-II, Ikeros, Ivana, Gymco-IV, Osser, SIRE, TAR, Tork, Ty1-outgroup, Ty3\_gypsy, non-chromovirus, non-chromo-outgroup, Phygy, Selgy, OTA, Athila, Tat, TatI, TatII, TatIII, Ogre, Retand, chromovirus, Chlamyvir, Tcn1, chromo-outgroup, CRM, Galadriel, and Tekay. The right sidebar shows a 'History' panel with a search bar and a list of recent jobs, including '50: RepeatExplorer2 - HT ML report from data 47'.

	Proportion[%]	Nsuperclusters	Nclusters	Nreads
Unclassified_repeat (conflicting evidences)	0	0	0	0
--rDNA	0	0	0	0
--45S_rDNA	0.36	1	1	9599
--18S_rDNA	0	0	0	0
--25S_rDNA	0.18	1	1	4858
--5.8S_rDNA	0	0	0	0
--5S_rDNA	0.03	1	1	809
--satellite	0	0	0	0
--mobile_element	0	0	0	0
--Class_I	0.46	1	1	12243
--SINE	0	0	0	0
--LTR	17.73	6	30	471735
--Ty1_copia	0	0	0	0
--Ale	0	0	0	0
--Alesia	0	0	0	0
--Angela	10.6	3	16	282190
--Bianca	0.13	1	1	3573
--Bryco	0	0	0	0
--Lyco	0	0	0	0
--Gymco-III	0	0	0	0
--Gymco-I	0	0	0	0
--Gymco-II	0	0	0	0
--Ikeros	0	0	0	0
--Ivana	0	0	0	0
--Gymco-IV	0	0	0	0
--Osser	0	0	0	0
--SIRE	0	0	0	0
--TAR	0	0	0	0
--Tork	0.29	2	2	7823
--Ty1-outgroup	0	0	0	0
--Ty3_gypsy	0	0	0	0
--non-chromovirus	0	0	0	0
--non-chromo-outgroup	0	0	0	0
--Phygy	0	0	0	0
--Selgy	0	0	0	0
--OTA	0	0	0	0
--Athila	1.85	3	3	49132
--Tat	0	0	0	0
--TatI	0	0	0	0
--TatII	0	0	0	0
--TatIII	0	0	0	0
--Ogre	0	0	0	0
--Retand	0	0	0	0
--chromovirus	0	0	0	0
--Chlamyvir	0	0	0	0
--Tcn1	0	0	0	0
--chromo-outgroup	0	0	0	0
--CRM	0	0	0	0
--Galadriel	0.08	2	2	2088
--Tekay	5.91	5	6	157340

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

✓ Resultados: Sumário dos repeats anotados

The screenshot shows the Galaxy web interface with the RepeatExplorer tool results. The main panel displays a tree view of repeat classifications and a summary table. The table highlights 'organelle' and 'contamination' categories.

	Proportion[%]	Nsuperclusters	Nclusters	Nreads
organelle	0	0	0	0
--plastid	1.81	35	32	48283
'--mitochondria	0.16	3	3	4143
Unclassified repeat (No evidence)	28.41	91	99	756071
contamination	0	0	0	0

The right sidebar shows a history of datasets, including '50: RepeatExplorer2 - HT ML report from data 47', '49: RepeatExplorer2 - Archive with HTML report from data 47', '48: RepeatExplorer2 - log file', and '47: Pd\_I1\_25.fa'.



# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## RepeatExplorer

✓ Resultados: anotação dos clusters

Galaxy Analyze Data Workflow Visualize Shared Data Help User Using 31%

Tools search tools

Get Data

RepeatExplorer2

TAREAN

ChIP-Seq Mapper

RepeatExplorer Utilities

DANTE

PROFREP

Text Manipulation

NGS: QC and manipulation

JBrowse

FASTA manipulation

Experimental Tools

Obsolete tools

Workflows

All workflows

### Cluster annotation

[For table legend see documentation.](#)

Show  entries Search:

Cluster	Super cluster	Proportion[%]	Proportion adjusted[%]	Number of reads	Graph layout	Similarity hits [above 0.1%]
1	1	8	1.8	1.8	47139	68.08% Class_ILTR/Ty3_gypsy/chromovirus/Tekay:Ty3-INT 0.81% Class_ILTR/Ty3_gypsy/chromovirus/Tekay:Ty3-CHD
2	2	10	1.4	1.4	37033	
3	3	5	1.3	1.3	33793	55.81% 0.54% 0.38% 0.18% 0.16% 0.13%
4	4	11	1.2	1.2	33225	
5	5	12	1.2	1.2	31234	

read\_1  
read\_2  
read\_3  
read\_4

read\_1  
read\_2  
read\_3  
read\_4

History search datasets

Unnamed history

4 shown, 46 deleted

62.83 GB

50: RepeatExplorer2 - HT ML report from data 47

49: RepeatExplorer2 - Archive with HTML report from data 47

48: RepeatExplorer2 - log file

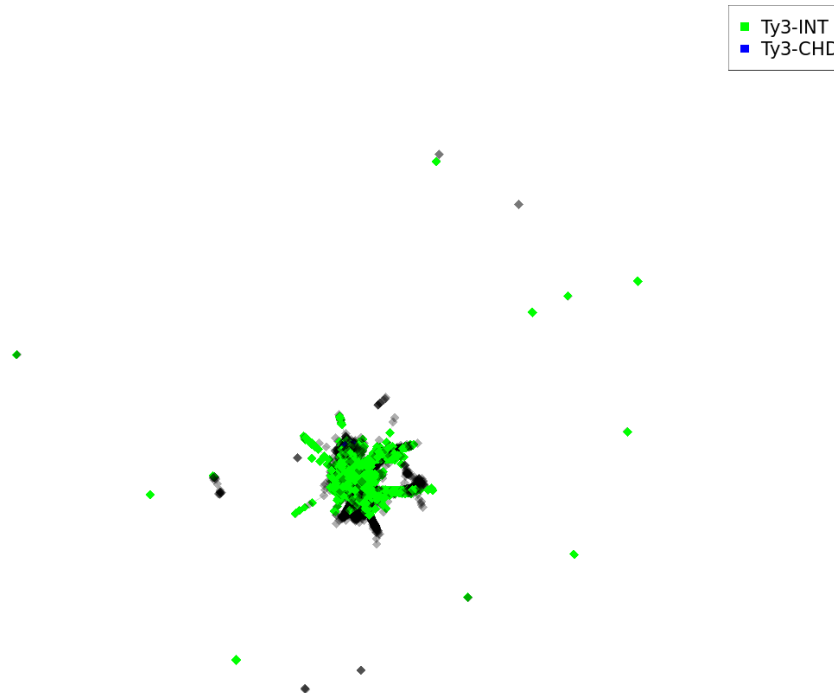
47: Pd\_I1\_25.fa

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

### ✓ Resultados: anotação dos clusters

Similarity based annotation: 68.08% Class\_I/LTR/Ty3\_gypsy/chromovirus/Tekay:Ty3-INT  
0.81% Class\_I/LTR/Ty3\_gypsy/chromovirus/Tekay:Ty3-CHD



# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

### ✓ Resultados

libdir	28/04/2020 13:09	Pasta de arquivos	
seqclust	28/04/2020 13:10	Pasta de arquivos	
cluster_report	28/04/2020 13:09	Chrome HTML Do...	105 KB
CLUSTER_TABLE	28/04/2020 13:10	Microsoft Excel C...	11 KB
contigs	28/04/2020 10:13	Arquivo FASTA	13.421 KB
documentation	28/04/2020 13:09	Chrome HTML Do...	18 KB
HOW_TO_CITE	27/04/2020 12:24	Chrome HTML Do...	2 KB
index	28/04/2020 13:10	Chrome HTML Do...	4 KB
logfile	28/04/2020 13:10	Arquivo TXT	107 KB
PROFREP_CLASSIFICATION_TEMPLATE	28/04/2020 13:09	Microsoft Excel C...	6 KB
style1	28/04/2020 13:09	Documento de fol...	5 KB
summarized_annotation	28/04/2020 10:39	Chrome HTML Do...	10 KB
summary_histogram	28/04/2020 13:09	Arquivo PNG	42 KB
supercluster_report	28/04/2020 10:39	Chrome HTML Do...	118 KB
SUPERCLUSTER_TABLE	28/04/2020 10:39	Microsoft Excel C...	86 KB
TAREAN_consensus_rank_1	28/04/2020 10:32	Arquivo FASTA	0 KB
TAREAN_consensus_rank_2	28/04/2020 10:32	Arquivo FASTA	0 KB
TAREAN_consensus_rank_3	28/04/2020 10:32	Arquivo FASTA	0 KB
TAREAN_consensus_rank_4	28/04/2020 10:32	Arquivo FASTA	0 KB
tarean_report	28/04/2020 13:09	Chrome HTML Do...	173 KB

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## RepeatExplorer

### ✓ Resultados

CLUSTER\_TABLE.csv - LibreOffice Calc

Arquivo Editar Exibir Inserir Formatar Estilos Planilha Dados Ferramentas Janela Ajuda

Liberation Sans 10 N I S A

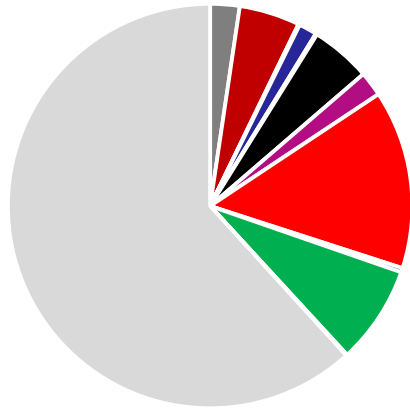
	A	B	C	D	E	F	G	H	I
7	Cluster	Supercluster	Size	Size adjusted	Automatic annotation	TAREAN_annotation	Final_annotation		
8	1	8	47139	47139	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	Other			
9	2	10	37033	37033	All	Other			
10	3	5	33793	33793	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	Other			
11	4	11	33225	33225	All/repeat/mobile_element/Class_I/LTR	Other			
12	5	12	31234	31234	All	Other			
13	6	13	29738	29738	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	Other			
14	7	14	29736	29736	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/non-chromovirus/OTA/Athila	Other			
15	8	3	28387	28387	All/repeat/mobile_element/Class_I/LTR	Other			
16	9	15	26462	26462	All	Other			
17	10	5	25920	25920	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	Other			
18	11	2	25777	25777	All/repeat/mobile_element/Class_I/LTR	Other			
19	12	2	25549	25549	All/repeat/mobile_element/Class_I/LTR	Other			
20	13	1	25485	25485	All/repeat/mobile_element/Class_I/LTR/Ty1_copia/Angela	Other			
21	14	16	23821	23821	All	Other			
22	15	17	23338	23338	All/repeat/mobile_element/Class_I/LTR/Ty1_copia/Angela	Other			
23	16	18	23212	23212	All	Other			
24	17	1	23144	23144	All/repeat/mobile_element/Class_I/LTR/Ty1_copia/Angela	Other			
25	18	20	22270	22270	All	Other			
26	19	7	22126	22126	All	Other			
27	20	1	21489	21489	All/repeat/mobile_element/Class_I/LTR/Ty1_copia/Angela	Other			
28	21	1	21155	21155	All/repeat/mobile_element/Class_I/LTR/Ty1_copia/Angela	Other			
29	22	21	21035	21035	All	Other			

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

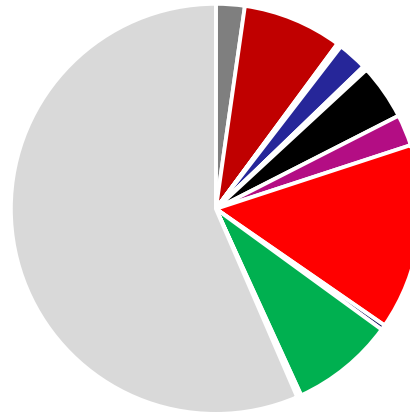
## ➤ RepeatExplorer

✓ Testes com diferentes coberturas

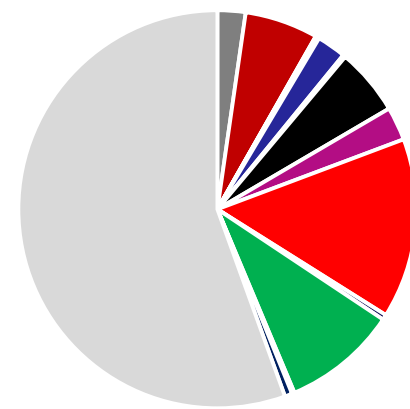
Cobertura de **0,1**: 35,30% de TEs



Cobertura de **0,25**: 38,01% de TEs



Cobertura de **0,5**: 38,17% de TEs



# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

✓ **Teste: *Passiflora loefgrenii***



	rDNA	45S_rDNA			0,20
			18S_rDNA		0,16
			25S_rDNA		0,18
		5S_rDNA			0,03
Mobile_element	Class_I				2,47
		LTR			
			Ty1_copia		
				<b>Angela</b>	<b>11,64</b>
				Bianca	0,19
				Ivana	0,02
				SIRE	1,62
				Tork	0,16
			Ty3_gypsy		
				<b>Athila</b>	<b>9,92</b>
				<b>Tekay</b>	<b>5,96</b>
Unclassified repeat					42,23
					<b>74,78</b>

✓ **2n = 18**

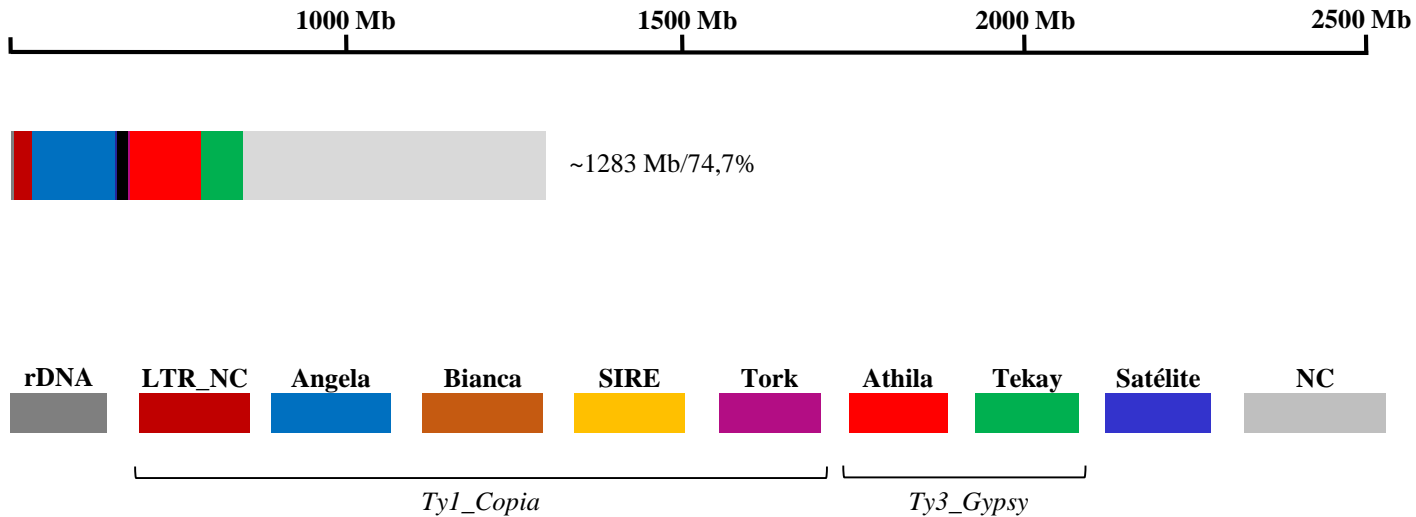
✓ **Genoma: 1283,80 Mb**

✓ **pb para 0,25x = 321.000.000 pb**

# ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

## ➤ RepeatExplorer

✓ *Passiflora loefgrenii*





# Obrigada!

[zirlane@usp.br](mailto:zirlane@usp.br)



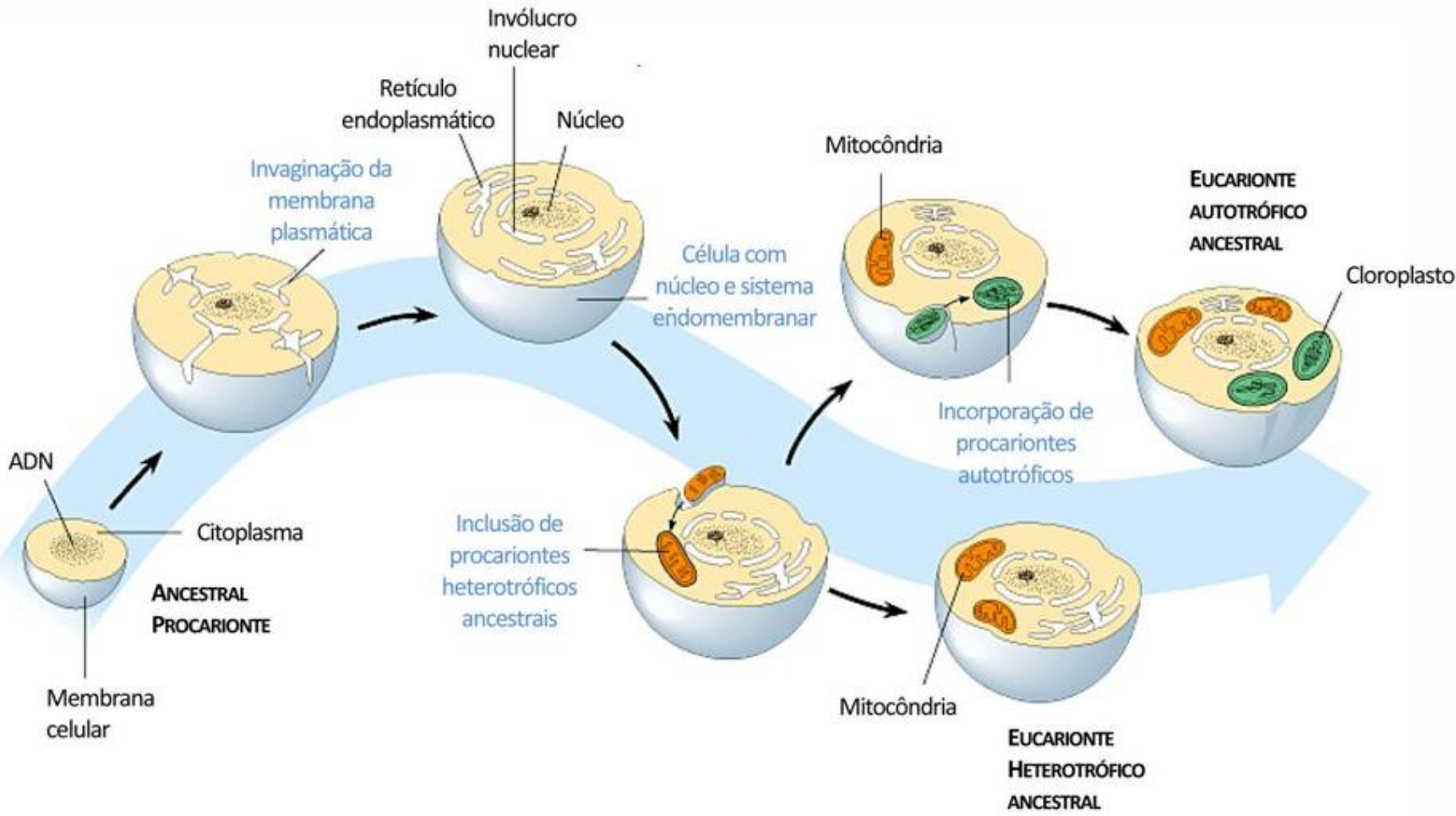
# PARTE II

# Análise de genomas cloroplastidiais

Luiz Augusto Cauz dos Santos

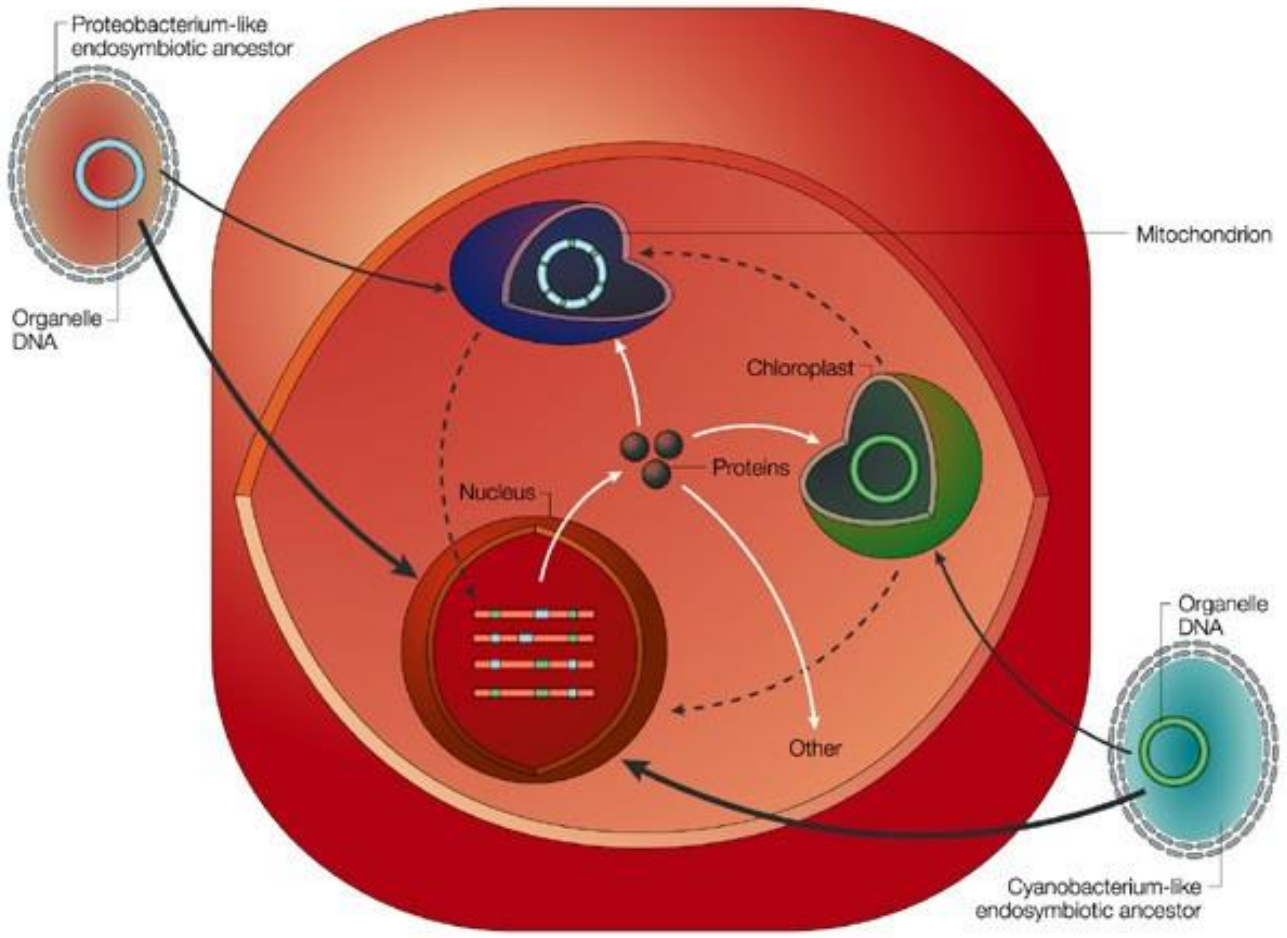
# ➤ Origem dos genomas extranucleares

A teoria da endossimbiose



# ➤ Origem dos genomas extranucleares

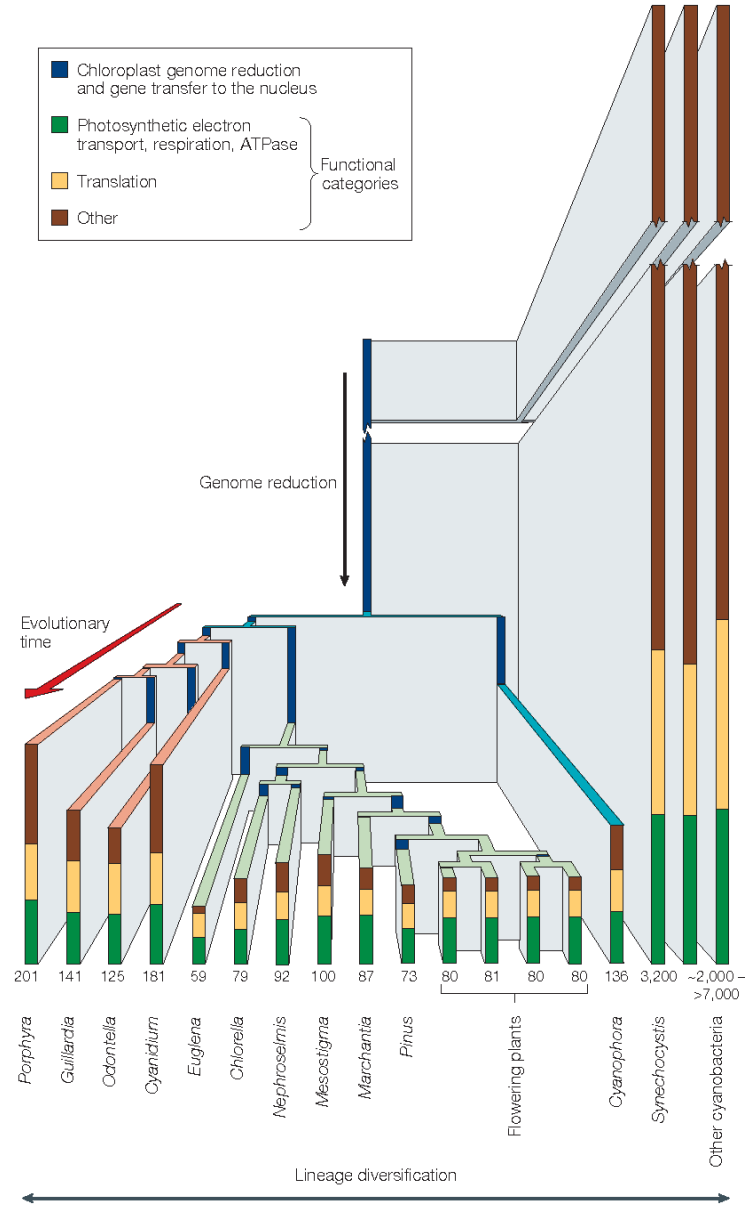
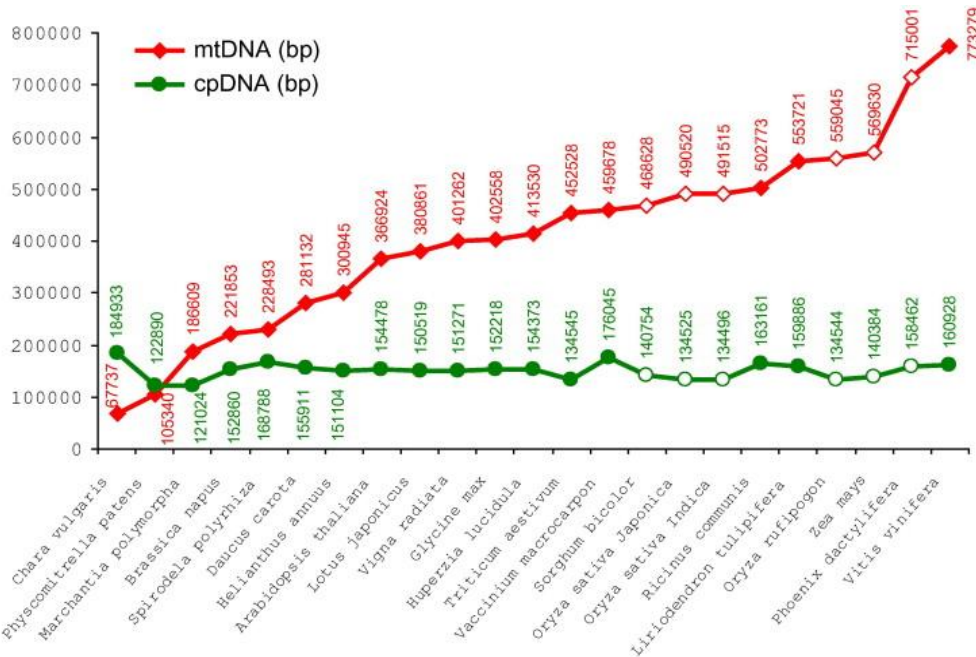
A transferência de material genético entre os genomas



# ➤ Origem dos genomas extranucleares

Redução do genoma cloroplastial

Variação de tamanho dos genomas mitocondriais



# GENOMA MITOCONDRIAL

Moléculas circulares

Tamanho entre 69.2kb a 521kb (Algas)  
200 a 750kb (Angiospermas)

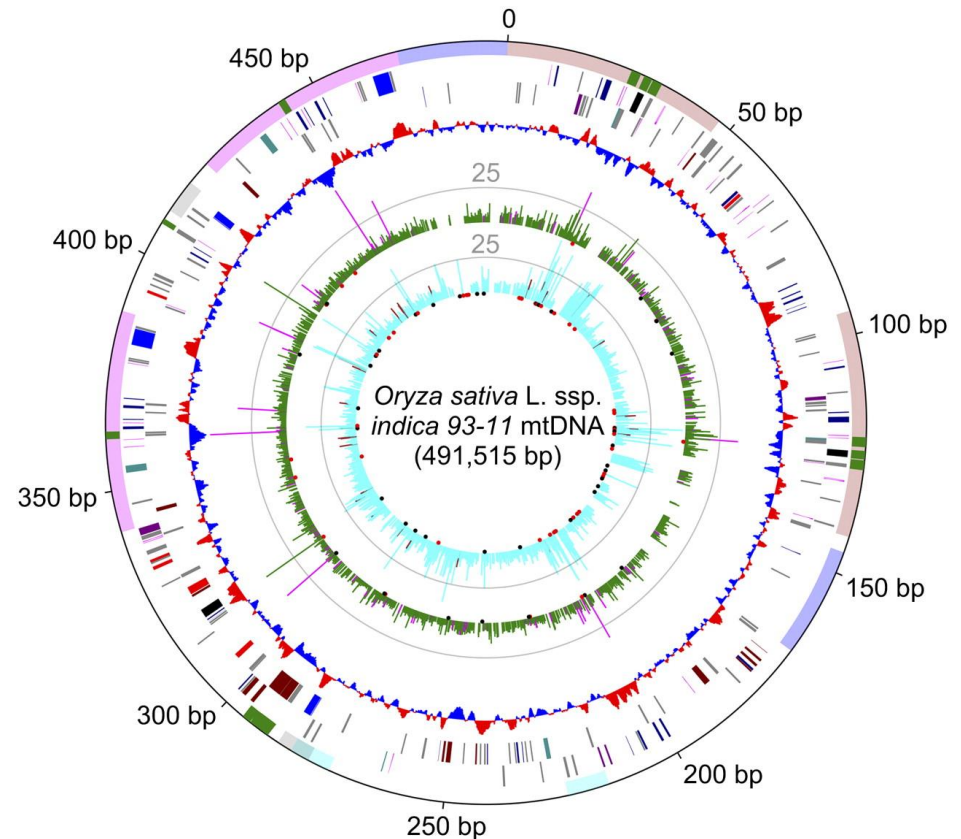
Múltiplas cópias em uma célula

Em angiospermas:

**Genes:** 50 – 60 genes

Genes que codificam para:

- tRNA
- rRNA
- Citocromo oxidase
- NDH-desidrogenase
- Subunidades de ATPase





# GENOMA CLOROPLASTIDIAL

Genoma cp, plastomas, plastídios

Moléculas circulares ou lineares

Tamanho entre 107 a 218Kb

**Estrutura Quadripartida**

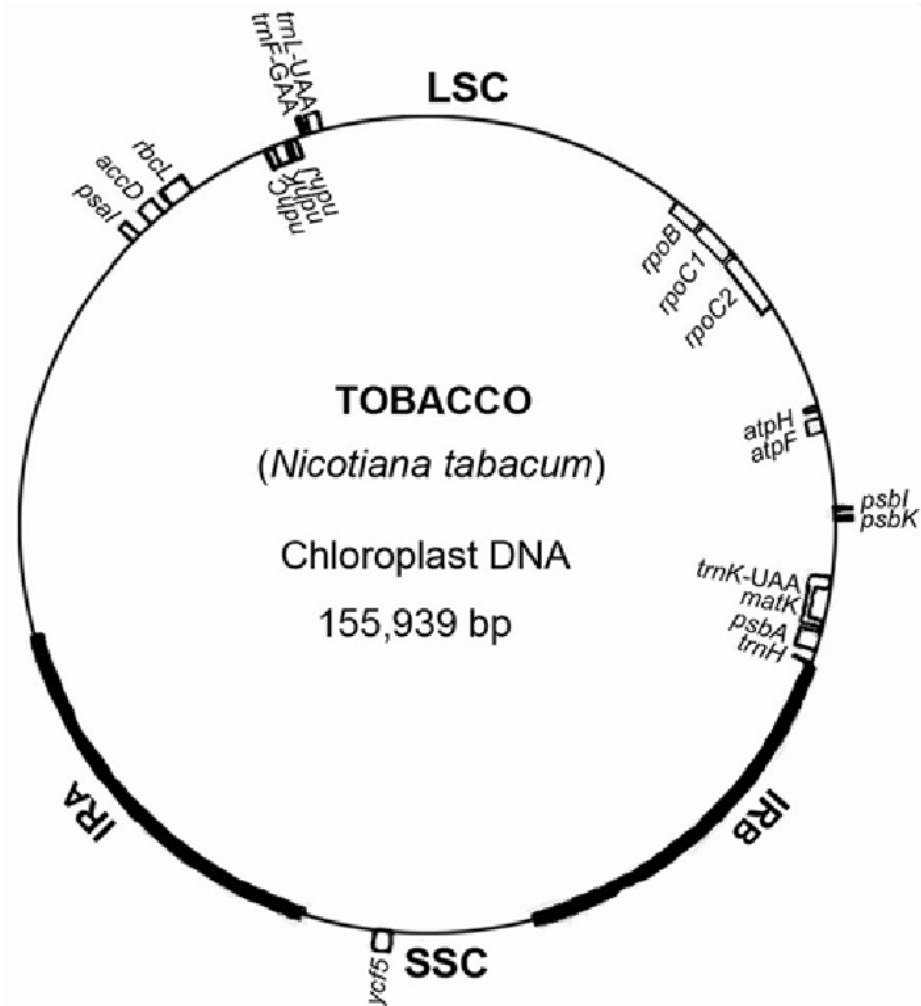
Múltiplas cópias em uma célula  
(média de 600/célula em *Arabidopsis*)

**Genes:** ~120 genes

Genes codificadores de proteínas

~ 30 genes **únicos** de tRNA

~ 4 genes **únicos** de rRNA  
(16S, 23S, 4.5S e 5S)



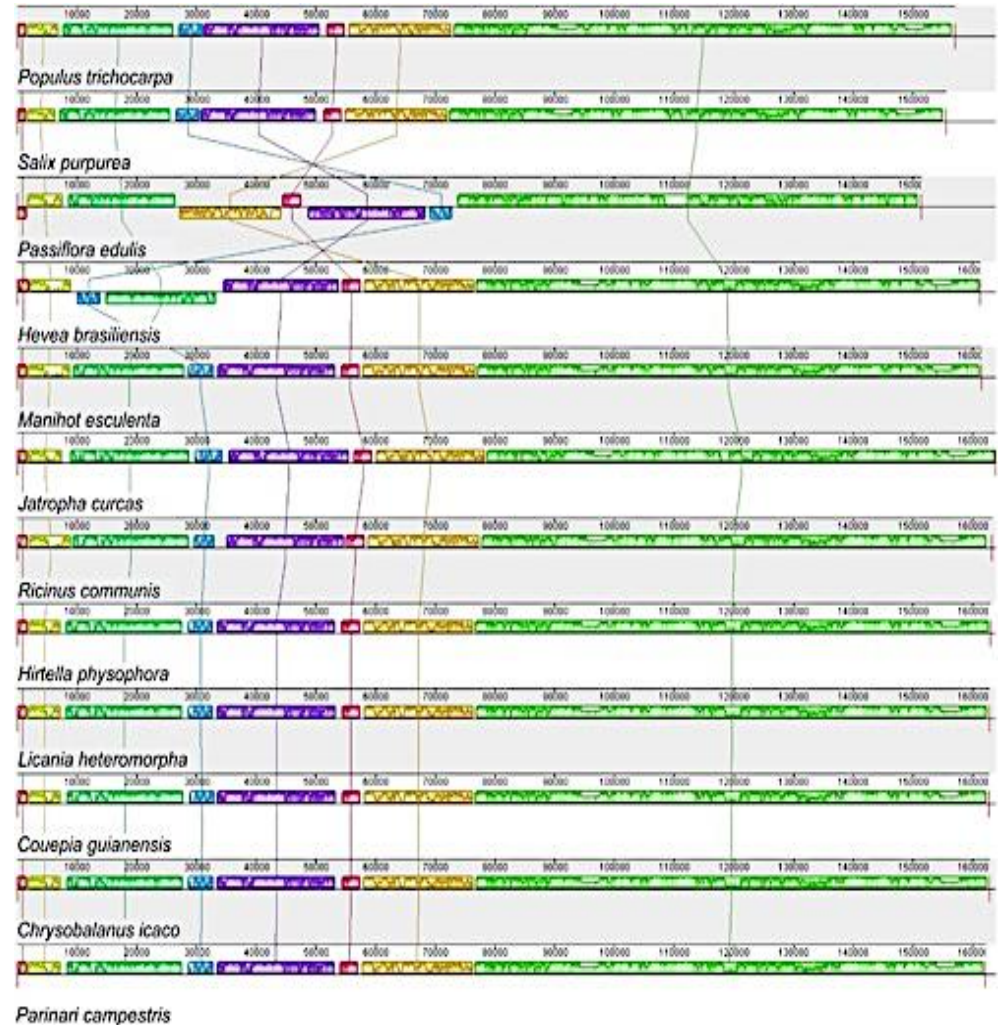
## ➤ Genomas cloroplastidiais em Passifloraceae

Características de 20 genomas cloroplastidiais de espécies da família Passifloraceae

Subgenus	Species	Cp genome size (bp)	LSC (bp)	SSC (bp)	IR (bp)	Total GC %	Total unique genes
<b>Astrophea</b>	<i>Passiflora cerradensis</i>	164,276	84,004	12,854	33,709	36,6	108
	<i>Passiflora haematostigna</i>	161,755	88,759	12,912	30,042	36,4	109
	<i>Passiflora rhamnifolia</i>	162,601	89,512	12,921	30,084	36,4	109
<b>Decaloba</b>	<i>Passiflora candollei</i>	138,081	72,565	13,506	26,005	37,2	104
	<i>Passiflora capsularis</i>	113,152	*	*	*	36,0	104
	<i>Passiflora costaricensis</i>	114,211	*	*	*	36,1	104
	<i>Passiflora suberosa</i>	156,355	56,445	13,033	43,496	37,3	103
	<i>Passiflora vespertilio</i>	138,456	72,902	13,158	26,196	37,1	104
<b>Deidamioides</b>	<i>Passiflora contracta</i>	167,141	87,972	13,731	32,714	36,7	107
	<i>Passiflora deidamioides</i>	167,827	82,408	13,739	35,840	36,8	107
<b>Passiflora</b>	<i>Passiflora alata</i>	148,113	86,001	13,494	24,309	36,9	107
	<i>Passiflora cristalina</i>	145,053	85,661	13,530	22,931	36,8	107
	<i>Passiflora edmundoi</i>	142,646	85,476	13,314	21,928	37,2	107
	<i>Passiflora loefgrenii</i>	146,320	86,153	13,267	23,450	37,0	107
	<i>Passiflora miniata</i>	151,758	85,737	13,477	26,272	37,0	107
	<i>Passiflora mucronata</i>	151,259	85,261	13,552	26,223	36,9	107
	<i>Passiflora recurva</i>	151,976	86,146	13,504	26,163	37,0	107
	<i>Passiflora watsoniana</i>	146,782	87,381	13,355	23,023	37,0	107
<b>Dilkea</b>	<i>Dilkea retusa</i>	161,923	88,575	12,686	30,331	36,2	109
<b>Mitostemma</b>	<i>Mitostemma brevifilis</i>	162,350	88,837	12,695	30,409	36,1	108

➤ Os genomas cp são altamente conservados em Angiospermas

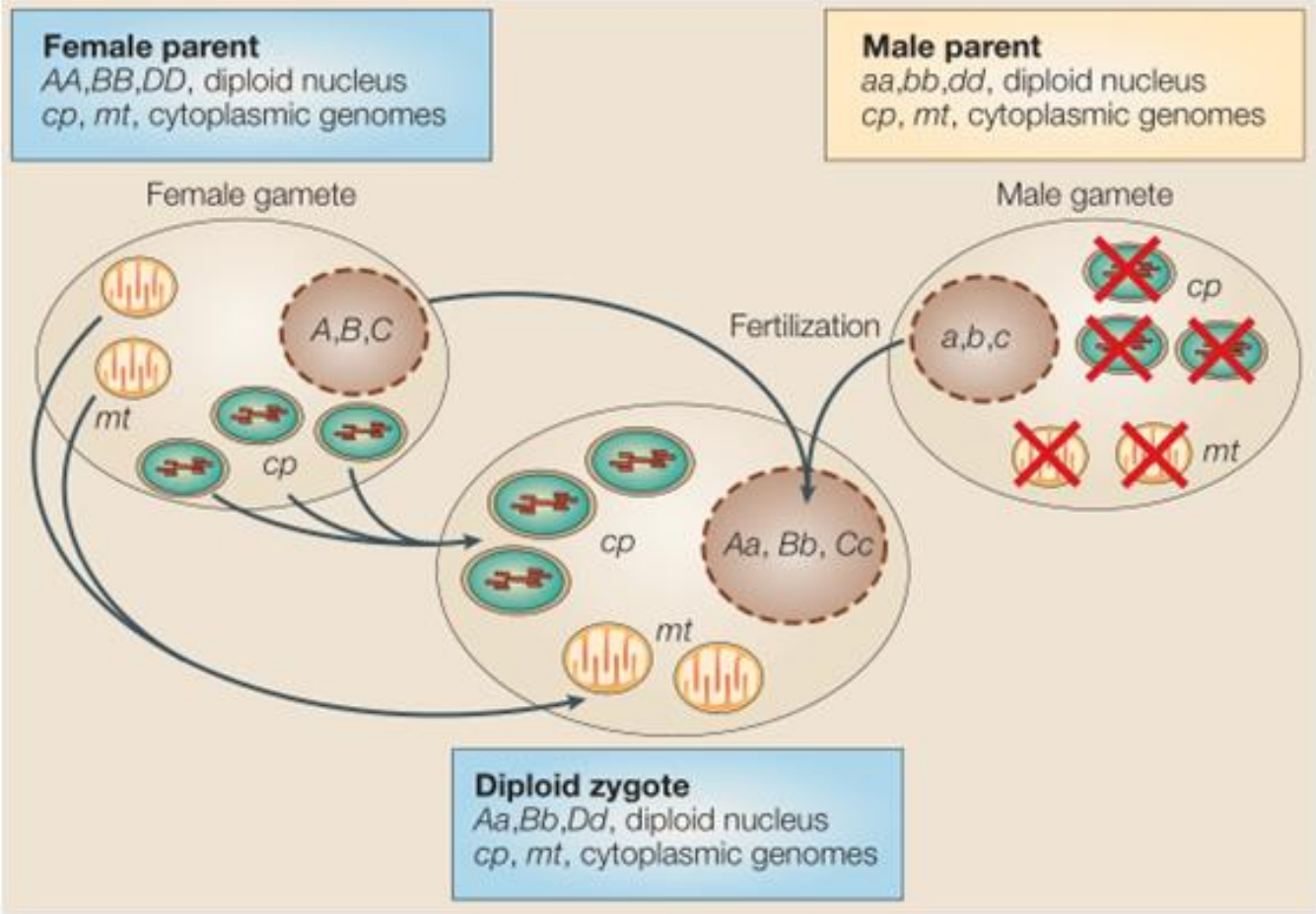
- Exemplo na ordem Malpighiales: Oito regiões adjacentes de sintonia nas sequências dos genomas cp: as regiões incluem genes, tRNAs e rRNAs



Alinhamento de 12 genomas cp completos de Malpighiales

# ➤ Herança citoplasmática

Herança uniparental de cloroplastos e mitocôndrias

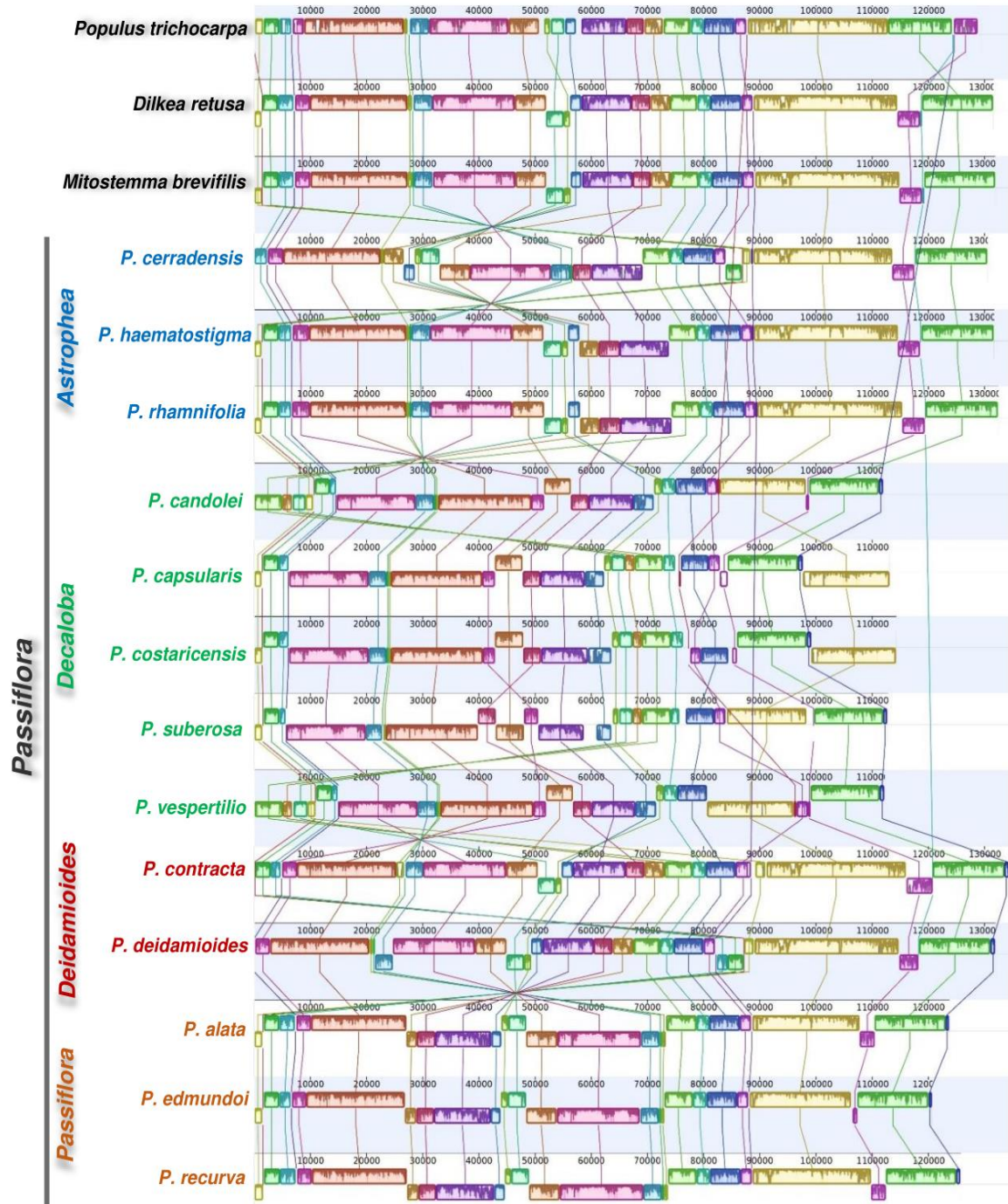


Cerca de 20% das angiospermas poderiam apresentar herança biparental



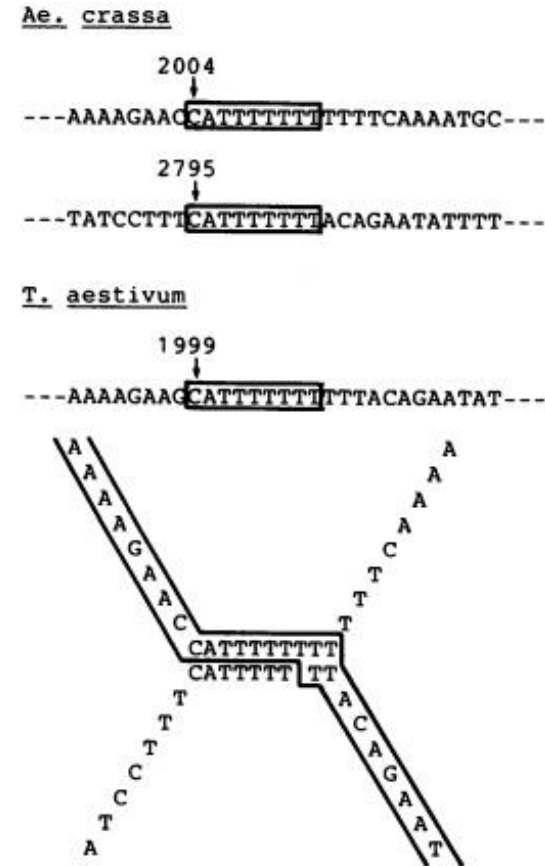
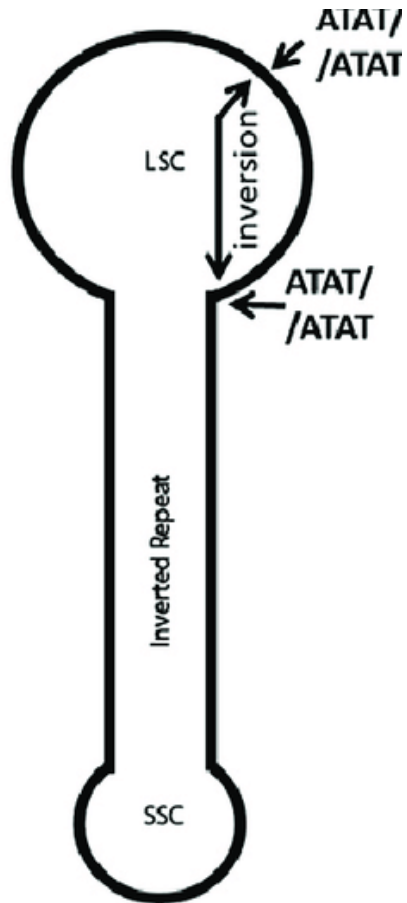
## ➤ Rearranjos

Recombinação intramolecular entre pequenas sequencias repetidas dispersas no genoma, ou genes de tRNA



## ➤ Rearranjos

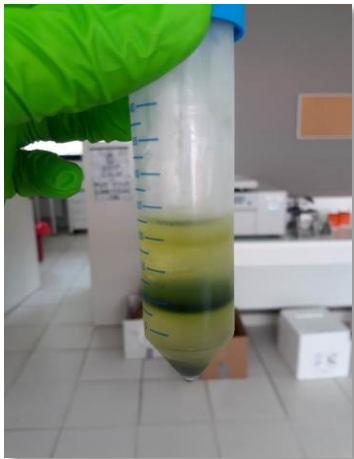
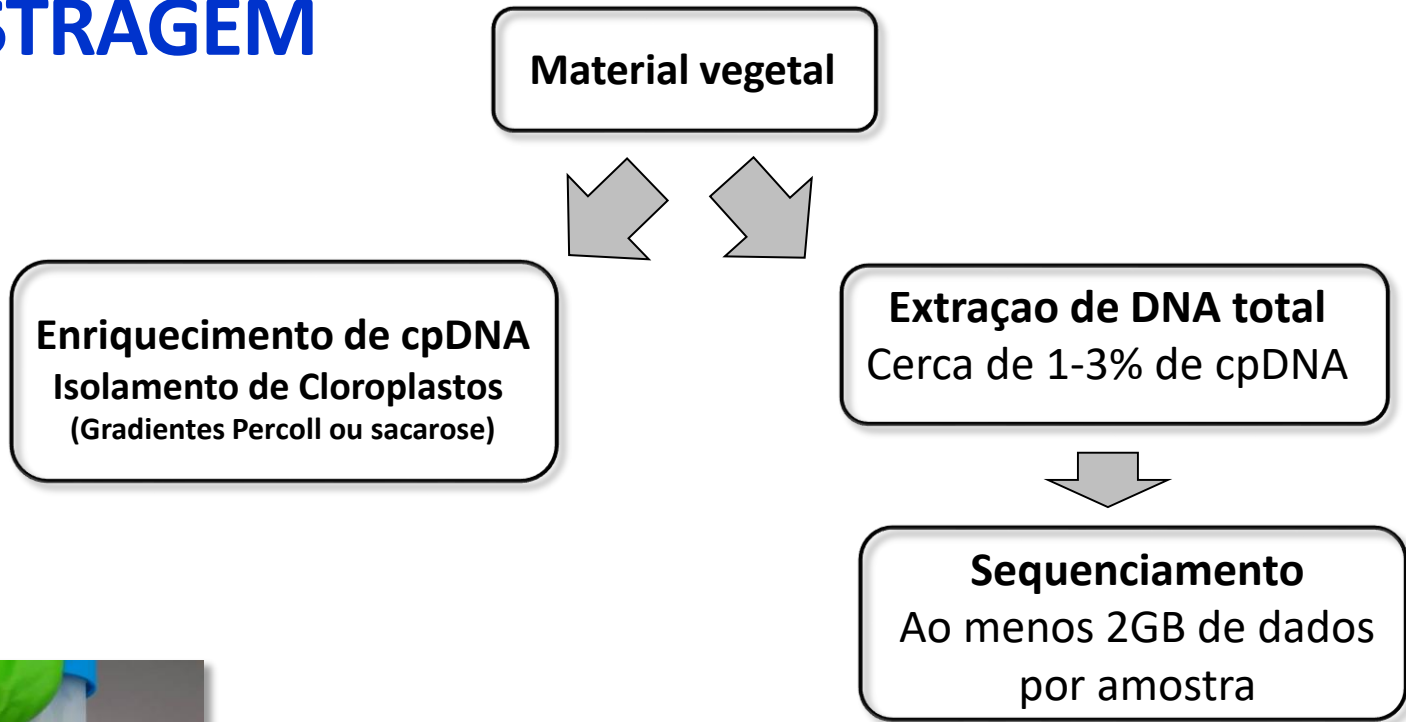
Recombinação homóloga intramolecular



Sinn *et al.*, 2018. American Journal of Botany

Ogihara *et al.*, 1988. PNAS

# AMOSTRAGEM



**Isolamento de cloroplastos  
em gradiente de sacarose**

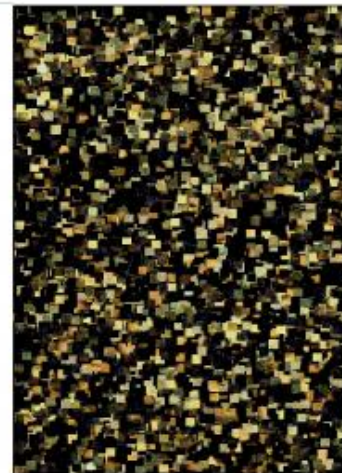
NextSeq 500 System Performance Parameters<sup>§</sup>

Flow Cell Configuration	Read Length (bp)	Output (Gb)	Run Time
High Output Flow Cell	2 × 150	100–120	29 hours
Up to 400 M single reads	2 × 75	50–60	18 hours
Up to 800 M paired-end reads	1 × 75	25–30	11 hours
Mid Output Flow Cell	2 × 150	32–39	26 hours
Up to 130 M single reads	2 × 75	16–19	15 hours

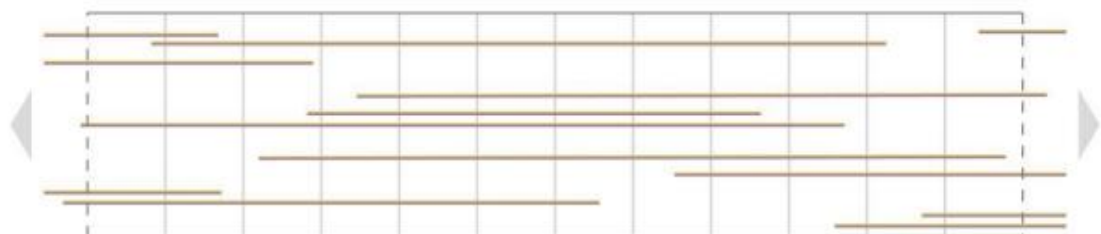
**Exemplo:** 2Gb 1% cpDNA= 20Mb, para um genoma cp 150 Kb = cobertura de 133x

## ➤ Montagem de Genomas

### Short Reads



### Long Reads

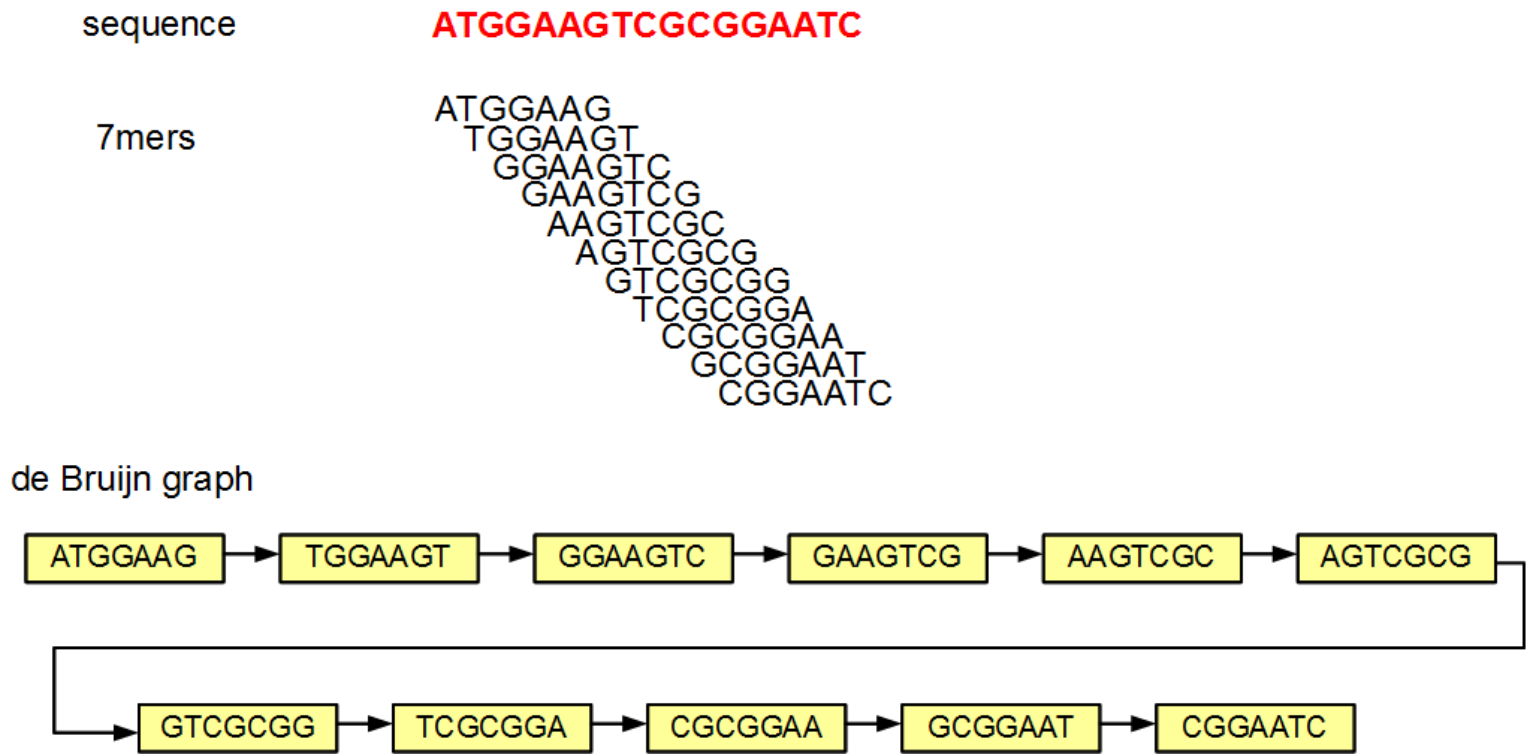




## ➤ Montagem de Genomas

- Ajuste do tamanho do k-mer

**K-mer:** Subsequência de tamanho k



**K-mer alto:** mais especificidade

**K-mer baixo:** mais sensibilidade

## ➤ Montagem de Genomas

Avaliação do número e tamanho dos contigs:

- 1) Tamanho do maior contig
- 2) N50: 50% do total de pb esteja contida em contigs

Ex. cálculo N50 em um genoma de 300 Mb  
8 Contigs: 3Mb, 3Mb, 15Mb, 24Mb, 39Mb, 45Mb, 54Mb  
e 117Mb  
**N50 = 54Mb**

# NOVOPLASTY

**Montador de genomas organelares** (Dierckxsens *et al.*, 2017. Nucleic Acids Research)

A montagem se baseia na extensão a partir de uma sequência semente

**<https://github.com/ndierckx/NOVOPlasty>**

NOVOPlasty - The organelle assembler and heteroplasmy caller

organelle-assembler mitochondria chloroplast-genome-assembly heteroplasmy

406 commits 1 branch 0 packages 31 releases 1 contributor View license

Branch: master

New pull request

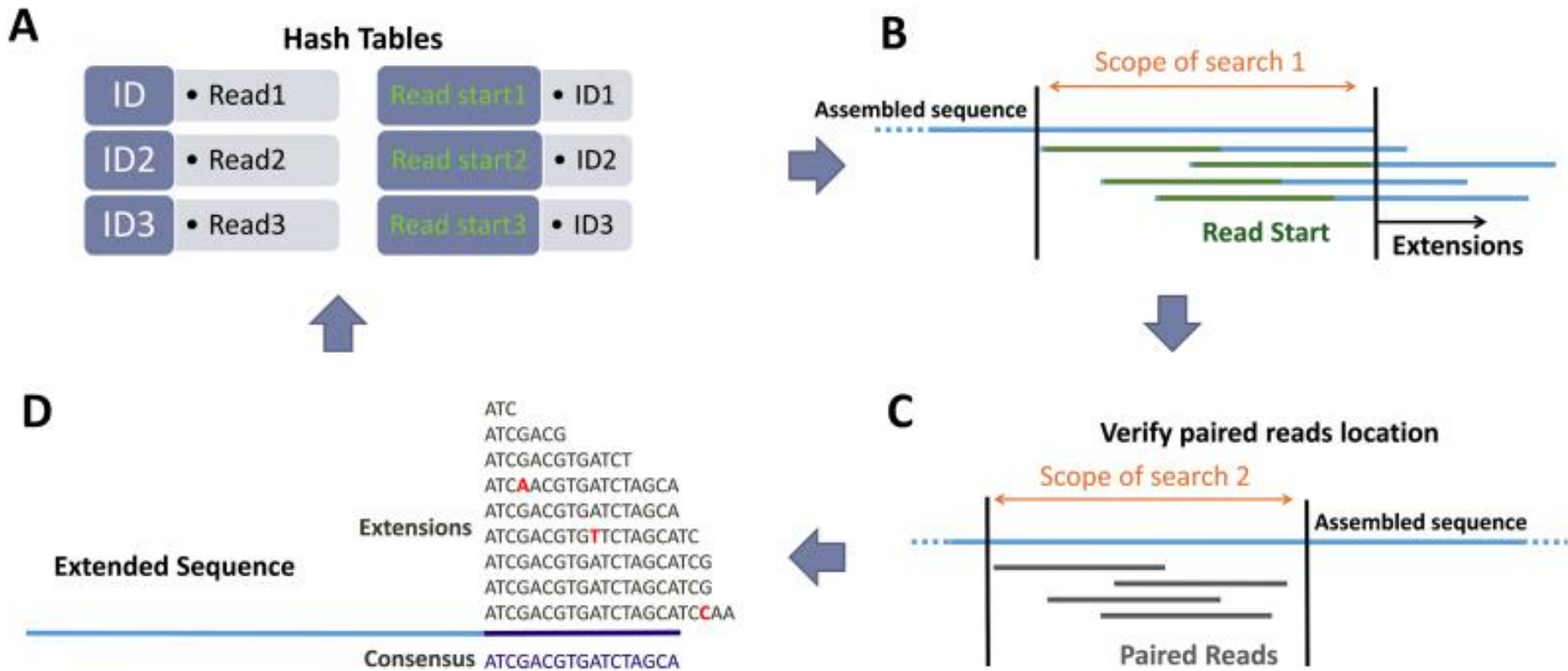
Find file

Clone or download

ndierckx Create NOVOPlasty4.1.pl		Latest commit d804316 4 days ago
Circos	Create Circos.pl	9 months ago
Test datasets	Rename log_mtDNA-Server_1:100_v3.7.txt to log_mtDNA-Server_1_100_v3.7...	2 months ago
.gitignore	initial commit	5 years ago
LICENSE	Update LICENSE	4 years ago
NOVOPlasty4.1.pl	Create NOVOPlasty4.1.pl	4 days ago
README.md	Update README.md	12 days ago

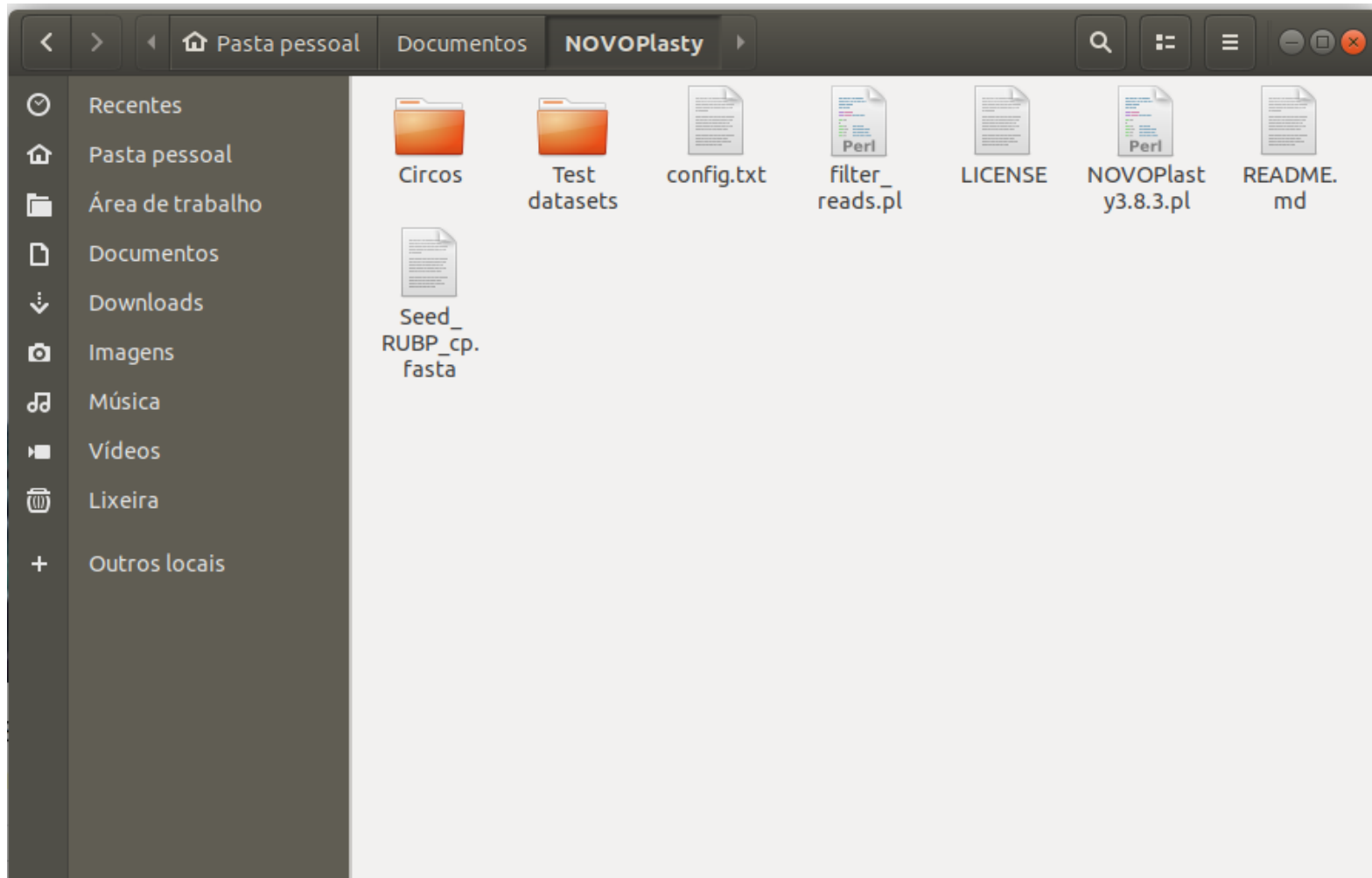
# NOVOPLASTY

A montagem se baseia na extensão a partir de uma sequência semente






# NOVOPLASTY

## Arquivos que compõem o programa de montagem



# NOVOPLASTY

O arquivo config.txt deve ser utilizado para determinar os parâmetros a serem utilizados na análise

```
Abrir ▾  *config.txt ~/Documentos/NOVOPlasty Salvar  
```

```
Project:
-----
Project name      = Test
Type              = mito
Genome Range      = 12000-22000
K-mer             = 39
Max memory        =
Extended log      = 0
Save assembled reads = no
Seed Input        = /path/to/seed_file/Seed.fasta
Extend seed directly = no
Reference sequence = /path/to/reference_file/reference.fasta (optional)
Variance detection =
Chloroplast sequence = /path/to/chloroplast_file/chloroplast.fasta (only for "mito_plant" option)

Dataset 1:
-----
Read Length       = 151
Insert size       = 300
Platform          = illumina
Single/Paired     = PE
Combined reads    =
Forward reads     = /path/to/reads/reads_1.fastq
Reverse reads     = /path/to/reads/reads_2.fastq

Heteroplasmy:
-----
MAF               =
HP exclude list   =
PCR-free          =

Optional:
-----
```

# NOVOPLASTY

## Parâmetros a serem especificados no arquivo config.txt

**Project name** = Nome do projeto a ser utilizado para os arquivos de saída

**Type** = chloro (Dados de genoma cloroplastidial)  
          mito (Dados de genoma mitocondrial)  
          mito\_plant (Dados de genoma mitocondrial de plantas)

**Genome Range** = 120000-200000 (Genoma cloroplastidial)  
                  12000-22000 (Genoma mitocondrial)  
                  45000-800000 (Genoma mitocondrial de plantas)

Ajustar o valor caso se tenha uma referência pois auxilia na correta estimação da circularização do genoma

**K-mer** = 39 (Default)

Para dados com baixa cobertura ou com *reads* menores de 90 pb diminuir o K-mer para 23

# NOVOPLASTY

## Parâmetros a serem especificados no arquivo config.txt

**Max memory** = Determinar um valor máximo a ser utilizado caso a capacidade de memória seja limitada. Exemplo: Para um computador 8 GB de RAM determinar no arquivo Max memory = 7

### Max memory = 7

```
Subsampled fraction: 37.06 %  
Forward reads without pair: 502077  
Reverse reads without pair: 212157  
  
Retrieve Seed.....OK  
  
Initial read retrieved successfully: GATTTTACCATGACTGCAATTTTAGAGAGACGCGAAAGCGAAAGCCTATGGGGTCGTTTCTGTAAGTGGATAACC  
  
Start Assembly...
```

### Max memory = 4

```
Subsampled fraction: 20.99 %  
Forward reads without pair: 286623  
Reverse reads without pair: 118187  
  
Retrieve Seed.....OK  
  
Initial read retrieved successfully: GATTTTACCATGACTGCAATTTTAGAGAGACGCGAAAGCGAAAGCCTATGGGGTCGTTTCTGTAAGTGGATAACC  
  
Start Assembly...
```



# NOVOPLASTY

## Parâmetros a serem especificados no arquivo config.txt

**Extended log** = 0 ou 1

Utilizar 1 para gerar um log.txt final com mais informações de todo o processo de montagem

**Save assembled reads** = yes ou no

Utilizar yes para gerar um arquivo separado com todas as *reads* utilizadas na montagem

**Seed Input** = Caminho para o arquivo (Exemplo: ./nomedoarquivo.fasta)

**Obs:** De preferência utilizar uma sequência pequena, por exemplo, um gene cloroplastidial

**Extend seed directly** = yes ou no

Utilizar yes apenas se a sequência *seed* for proveniente do mesmo conjunto de dados

**Reference sequence** = Caminho para o arquivo (Exemplo: ./nomedoarquivo.fasta)

**Variance detection** = yes ou no

Utilizar yes apenas se possuir uma sequência referência

**Chloroplast sequence** = Caminho para o arquivo (Exemplo: ./nomedoarquivo.fasta)

Obs: utilizar esta opção apenas se for realizar montagem de genomas mitocondriais de plantas

# NOVOPLASTY

**Parâmetros a serem especificados no arquivo config.txt.**

**Read Length** = Exemplo: 76 ou 151

**Insert size** = Exemplo: 350 ou 550 para NextSeq

**Platform** = illumina ou ion

**Single/Paired** = PE

Obs: O programa aceita apenas Paired-end

**Combined reads** = Caminho para o arquivo (Exemplo: ./nomedoarquivo.fq ou .fastq)

**Forward reads** = Caminho para o arquivo (Exemplo: ./nomedoarquivo.fq ou .fastq)

**Reverse reads** = Caminho para o arquivo (Exemplo: ./nomedoarquivo.fq ou .fastq)

# NOVOPLASTY

**Parâmetros a serem especificados no arquivo config.txt.**

## **Heteroplasmy:**

**MAF** = (0.007-0.49)

Obs: Proceder a montagem do genoma cp previamente sem esta opção, e depois utilize a sequência montada como referência e como *seed*.

**PCR-free** = yes ou no

Utilizar em casos de bibliotecas construídas com o kit TruSeq.

## **Optional:**

**Insert size auto** = yes ou no

**Insert Range** = 1.9 (Default)

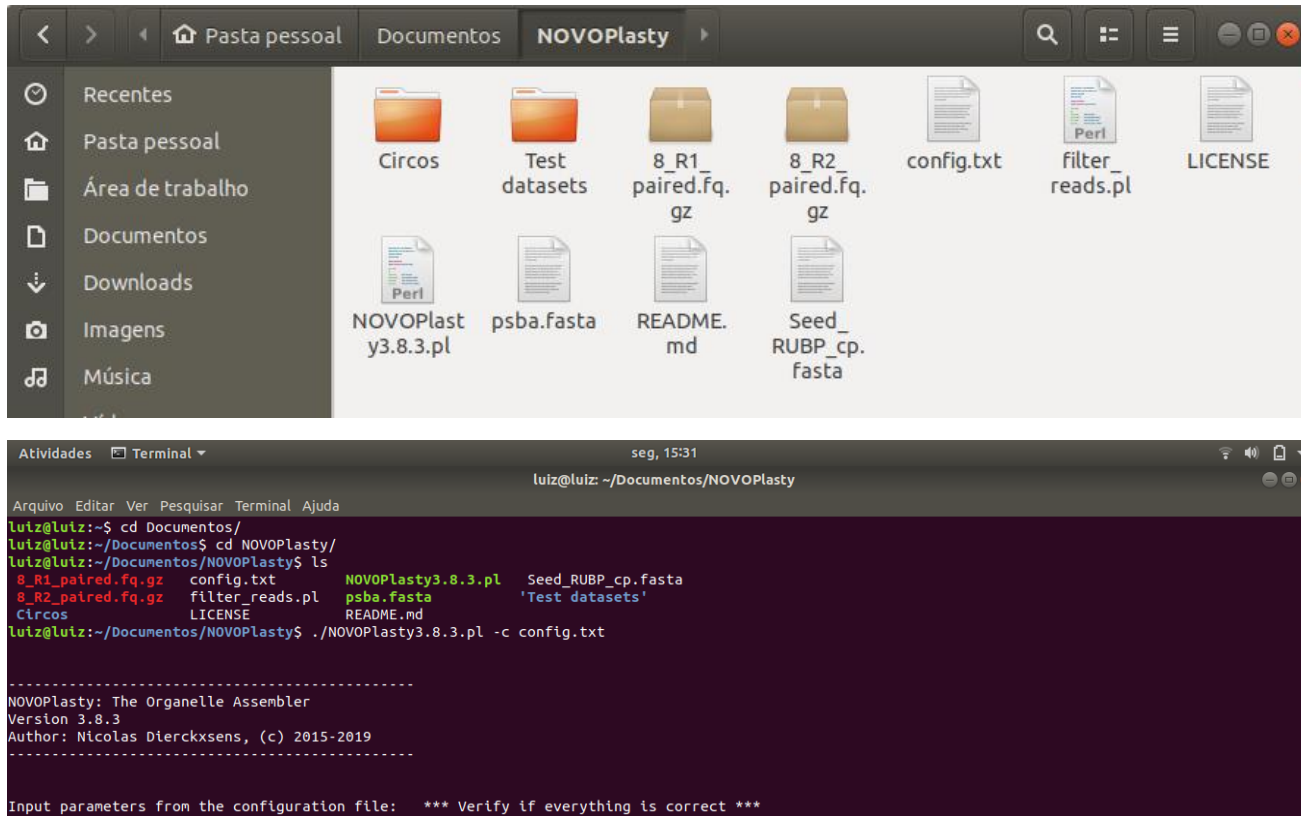
**Insert Range strict** = 1.3 (Default)

**Use Quality Scores** = yes ou no

Utilizar esta opção quando as *reads* apresentarem baixa qualidade

# NOVOPLASTY

Rodando o programa:



The image shows a file manager window displaying the contents of the NOVOPlasty directory. The files and folders are:

- Circos
- Test datasets
- 8\_R1\_paired.Fq.gz
- 8\_R2\_paired.Fq.gz
- config.txt
- filter\_reads.pl
- LICENSE
- NOVOPlasty3.8.3.pl
- psba.fasta
- README.md
- Seed\_RUBP\_cp.fasta

Below the file manager is a terminal window showing the execution of the program. The terminal output is as follows:

```
luiz@luiz: ~/Documentos/NOVOPlasty
luiz@luiz:~/Documentos$ cd NOVOPlasty/
luiz@luiz:~/Documentos/NOVOPlasty$ ls
8_R1_paired.fq.gz  config.txt      NOVOPlasty3.8.3.pl  Seed_RUBP_cp.fasta
8_R2_paired.fq.gz  filter_reads.pl psba.fasta          'Test datasets'
Circos             LICENSE        README.md
luiz@luiz:~/Documentos/NOVOPlasty$ ./NOVOPlasty3.8.3.pl -c config.txt

-----
NOVOPlasty: The Organelle Assembler
Version 3.8.3
Author: Nicolas Dierckxsens, (c) 2015-2019
-----
Input parameters from the configuration file:  *** Verify if everything is correct ***
```

Após abrir o terminal e acessar a pasta com os arquivos do Programa, digitar o comando:  
**./NOVOPlasty3.8.3.pl -c config.txt**

# NOVOPLASTY

## Rodando o programa:

```
luiz@luiz: ~/Documentos/NOVOPlasty
Arquivo Editar Ver Pesquisar Terminal Ajuda
Single/Paired      = PE
Combined reads    =
Forward reads     = ./8_R1_paired.fq.gz
Reverse reads     = ./8_R2_paired.fq.gz

Heteroplasmy:
-----
Heteroplasmy      =
HP exclude list  =
PCR-free         =

Optional:
-----
Insert size auto  = yes
Insert range     = 1.9
Insert range strict = 1.3
Use Quality Scores =

Reading Input.....OK

Building Hash Table.....OK

Subsampled fraction: 20.99 %
Forward reads without pair: 286623
Reverse reads without pair: 118187

Retrieve Seed.....OK

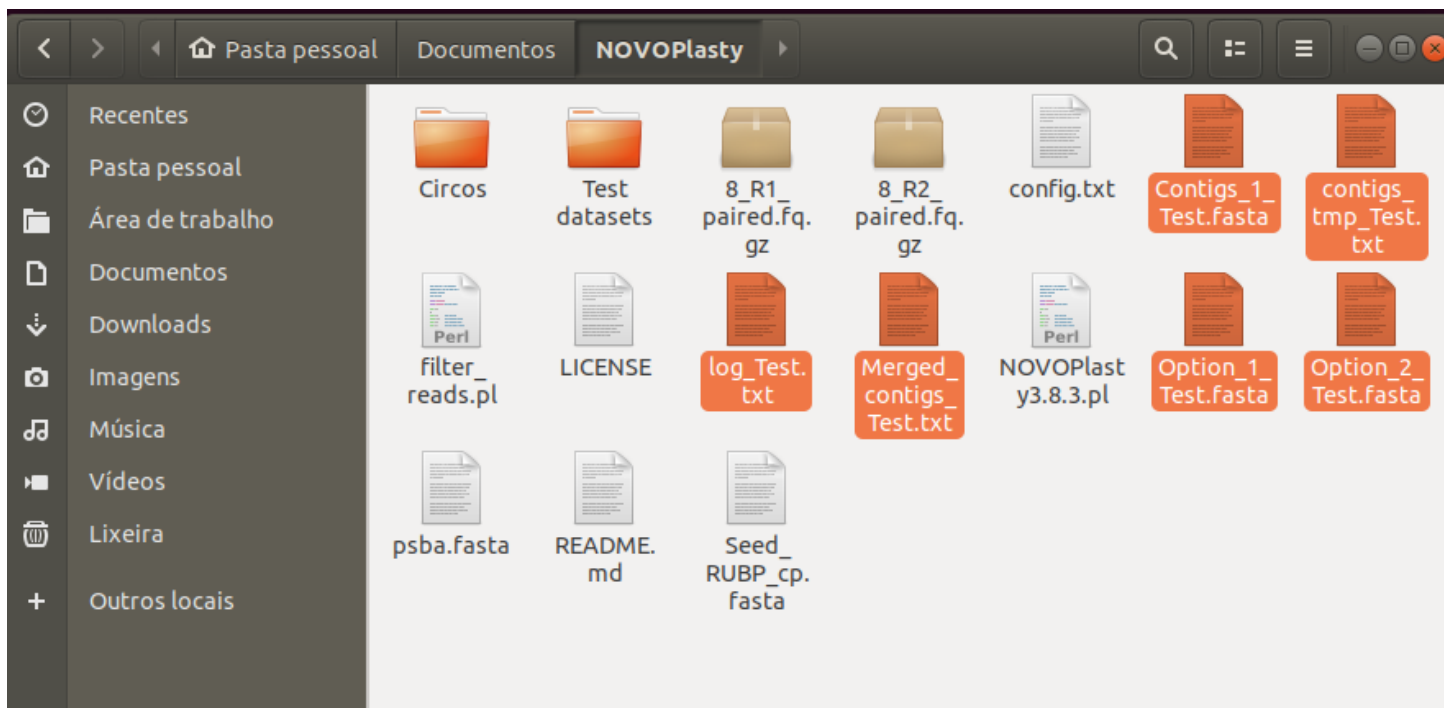
Initial read retrieved successfully: GATTTTACCATGACTGCAATTTTAGAGAGACGCCAAAGCGAAAGCCTATGGGGTCGTTTCTGTAAGTGGATAACC

Start Assembly...

61752 bp assembled
```

# NOVOPLASTY

Arquivos de saída do programa:



**Contigs\_1\_Test.fasta:** Arquivo contendo os contigs que foram gerados durante a montagem

**Contigs\_tmp\_Test.txt:** Arquivo backup

# NOVOPLASTY

## Arquivos de saída do programa:

**log\_Test.txt:** Arquivo contendo todas as informações da análise.

```
log_Test.txt
~/Documentos/NOVOPlasty
Abrir Salvar
Initial read retrieved successfully: GATTTTACCATGACTGCAATTTTAGAGAGACGCGAAAGCGAAAGCCTATGGGGTCGTTTCTGTAAGTGGATAACC
Start Assembly...

-----Assembly 1 finished: Contigs are automatically merged in Merged_contigs file-----

Contig 01          : 64935 bp
Contig 02          : 127 bp
Contig 03          : 806 bp
Contig 04          : 597 bp
Contig 05          : 6578 bp
  (Check manually if the two contigs overlap to merge them together!)
Contig 05          : 75377 bp
Contig 06          : 6578 bp
  (Check manually if the two contigs overlap to merge them together!)
Contig 06          : 75377 bp

Total contigs      : 8
Largest contig     : 75377 bp
Smallest contig    : 127 bp
Average insert size : 312 bp

-----Input data metrics-----

Total reads        : 7497130
Aligned reads      : 1202058
Assembled reads    : 1095808
Organelle genome % : 16.03 %
Average organelle coverage : 623

-----

Texto sem formatação  Largura da tabulação: 8  Lin 1, Col 1  INS
```

# NOVOPLASTY

## Arquivos de saída do programa:

**Option\_Test.fasta:** Arquivo contendo sequencia gerada a partir da junção dos contigs. Os diferentes arquivos Option\_Test.fasta podem diferir na orientação das regiões IRs

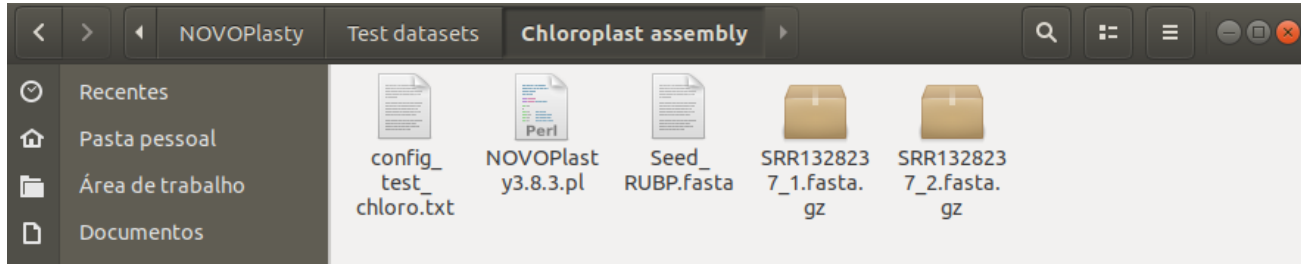
```
Option_1_Test.fasta
~/Documentos/NOVOPlasty

>Contig1
TTTGAATAAAGATGGGTGAATTTTTTACCATAAAATGAAATCCTCCTCCTACCGCCATTTTTAAAATAGTTTTATTGATCAGGAATAATAACAAAAATGGAATAAGAATAATAAATAAGAATGTCAATTCATTTCT
CAAAAATCTTCAATTCGTTAGCAATCCGAATTTTGTCAAATAGAATTTCTTATTAATTAATACAATTTTTTTTTCTTTTTTTTAGTTGGCTAGAAATCAAAAAGGATAGTATATCTCTTCACTTACAAAAGAGAAA
CAAAGTAAATTCATGGCTTCTTTCTAGATAGAATTGATAGGCAATCGAATTCAGGTATCAGAATAGAATATAGAATGGAAAACAAGGAATGTGTCTATATCCTTGTTTTCTATATTAGATCAGATATGCTATAGA
TATGACAAACGTACCTTTATTTCAATTTGGATACATACGTATCCTTAACATACTGAACCGACTGGAATTTATTGATTAACCAAAAAGCGGTTACATAAACTAAATCTCGAATCGAAAATTTGGCCAAAG
ATTGAAGTAGTTTTCTTTCTCTATCAAATATTGACTTTTTTCTATAATGTCATATTGACTTTTTTCTATAATGTCAAAGCGGCTGCAGTTTTTCTATTACTTTTGAAAAGTTCTGTATTCATATGAATATA
TTGAATTGAAAGAGATTAGAATTTCAACTCTCTACGACTTCTAGTGGATAAAAAGATTTTCAGGAACAAGGAAATCTTTTTCTTTCTAAATCCAATTAGACTTTTAAGTCTTTCAATAGATTTGAAAAATTATCTTT
TTTTTCTCCGCATCCTTGATTAATTTGTTGTTGGTCTTATGAACCAATAAAATAGCTATGGTTTGATACATAAGAAATTTCCATTAAGCTTCGCTCTGTTTTTTTTCGGTACACGCGGAAACTCAACAATCAATTT
TCAATTTTAGATCTCCCCTAAGCCTATCCCCTAAGCCATCTCTTTCTTTTGACCCAAAACCAAAAAGGTCCACATTTAATTTCTTTCTGTCGAATCAAGGATTTAATCCATAGATTTTTCTTTGTTTGTCTTTTGA
GTCTTTTACCAGGTAAGTATATGTTGGATATTGACGAAATGCTTCGTCTACATCCTTAGCATCTATGGTCGCTCGAATGGCTGGTGAGCTCGCCGCTCGAGTGGCCTCAAAAAGAAAATTTCCCAACTGGGTTT
GGTGTTTCTGTTCTCCTGGTGTTTTTTGGTCTACTGGCGTTTTTGGTCTACTGATGATTATTCTGCTCCTGGTTTGGTCTACTGATGATTATTCTTCTCTGGCGTTTTTGGTTTTTGGTCTACTGTGGATTACT
GATTACTCTTCTACTGGTGTTTTTTGGTCTTTTCTATTTCGGAGAATTGATTGAAGATTCACTGGGTTCTGAAATCGAGTTTTCTTTATAATTGAAAAGAAGTAATTTGGGTTGGTATTACCCACGGTCTTCTTTATAT
TCTCTCGAATTTGAAAAAAGAACCCTCAATTTCAAATTTGATACGAGACAGTGTACTATTTCTGTCATTCATTCCCATCCAATCTAACTCAGGTGGGTGCTTTCTACAAATCTTTACCTTTTGACAAGCATAGGA
TTTGTCAAAAAACGATCTAGGTCACCGTCTTTCTCAATTTTATGATTTTATTCTCTTACTCTCAGTTTCAGTATTTTCTTTGTTTTTTCAATTAATCCAGAATTCATTTTTTCTTTGTTGTAAGACAAAAA
Texto sem formatação ▾ Largura da tabulação: 8 ▾ Lin 1, Col 1 ▾ INS
```



# NOVOPLASTY

Rodando o programa com os arquivos de Teste:



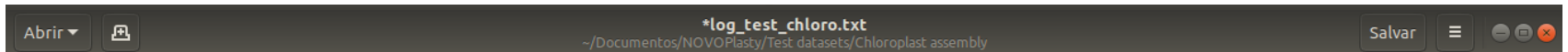
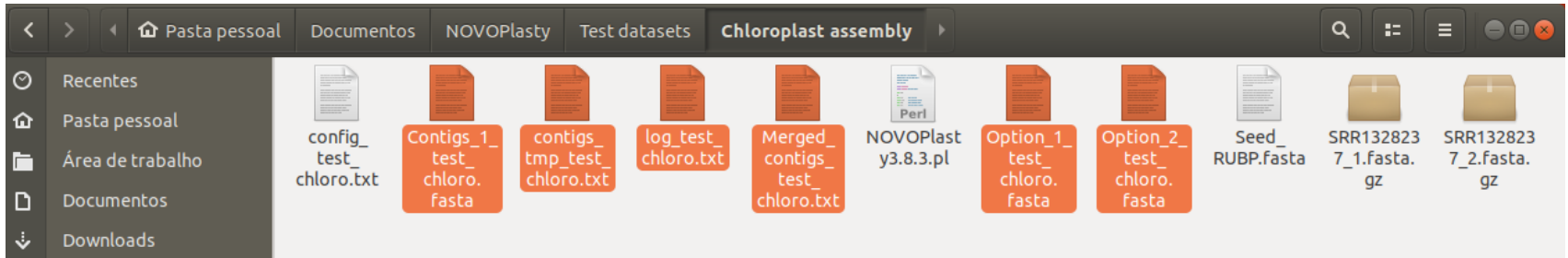
```
luiz@luiz: ~/Documentos/NOVOPlasty/Test datasets/Chloroplast assembly
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~$ cd Documentos/
luiz@luiz:~/Documentos$ cd NOVOPlasty/
luiz@luiz:~/Documentos/NOVOPlasty$ ls
 8_R1_paired.fq.gz      filter_reads.pl          Option_2_Test.fasta
 8_R2_paired.fq.gz      LICENSE                  psba.fasta
 Circos                 log_Test.txt             README.md
 config.txt             Merged_contigs_Test.txt Seed_RUBP_cp.fasta
 Contigs_1_Test.fasta   NOVOPlasty3.8.3.pl      'Test datasets'
 contigs_tmp_Test.txt   Option_1_Test.fasta
luiz@luiz:~/Documentos/NOVOPlasty$ cd Test\ datasets/
luiz@luiz:~/Documentos/NOVOPlasty/Test datasets$ ls
'Chloroplast assembly'  Heteroplasmy  'Mitochondrial assembly'
luiz@luiz:~/Documentos/NOVOPlasty/Test datasets$ cd Chloroplast\ assembly/
luiz@luiz:~/Documentos/NOVOPlasty/Test datasets/Chloroplast assembly$ ls
config_test_chloro.txt  Seed_RUBP.fasta          SRR1328237_2.fasta.gz
NOVOPlasty3.8.3.pl     SRR1328237_1.fasta.gz
luiz@luiz:~/Documentos/NOVOPlasty/Test datasets/Chloroplast assembly$ ./NOVOPlasty3.8.3.pl -c config_test_chloro.txt
```

**Para usar o arquivo de Teste:** No terminal, acesse a pasta do NOVOPlasty, abra a pasta Test datasets e procure pela pasta chloroplast assembly. Para rodar a análise teste utilize o comando:

**`./NOVOPlasty3.8.3.pl -c config_test_chloro.txt`**

# NOVOPLASTY

## Arquivos de saída do programa:



Start Assembly...

-----Assembly 1 finished: Contigs are automatically merged in Merged\_contigs file-----

```
Contig 01      : 122155 bp
Contig 02      : 13669 bp
Contig 03      : 13375 bp
```

```
Total contigs      : 3
Largest contig      : 122155 bp
Smallest contig     : 13375 bp
Average insert size : 479 bp
```

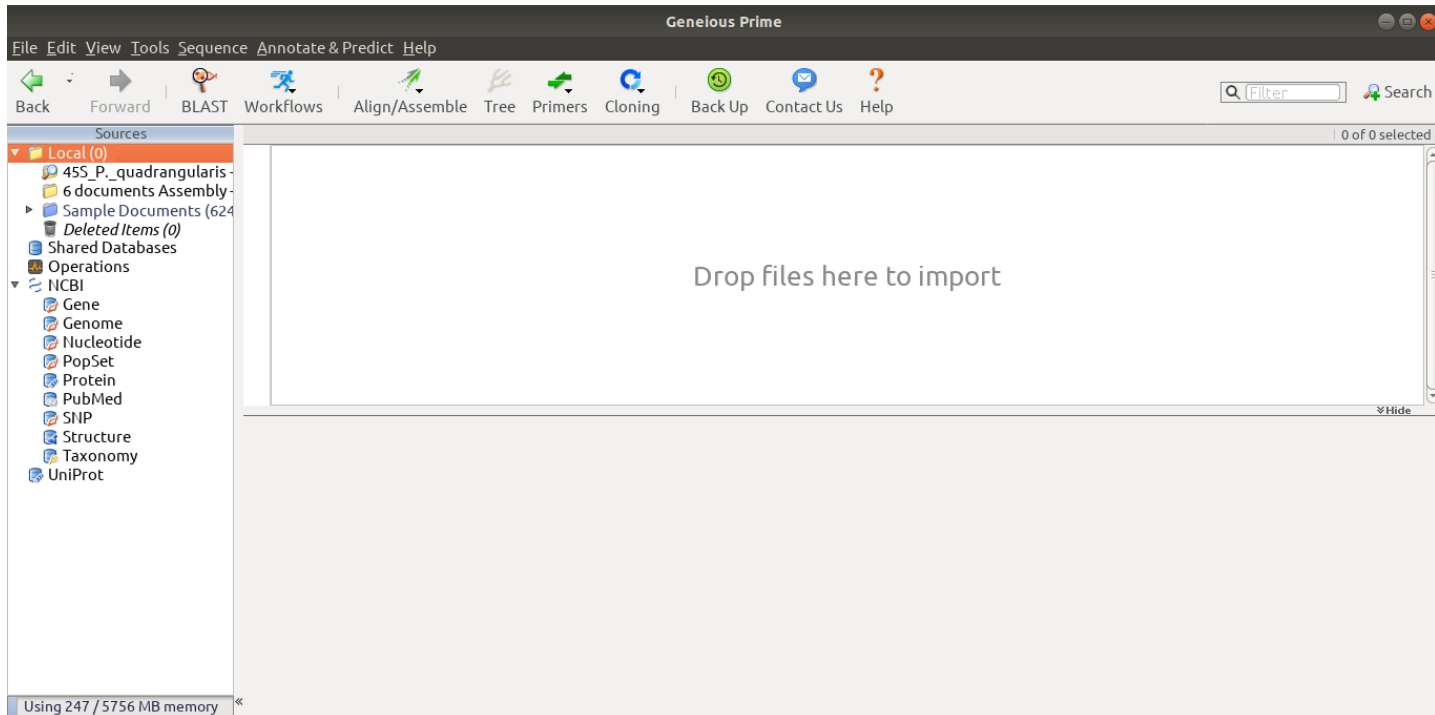
-----Input data metrics-----

```
Total reads      : 279568
Aligned reads     : 271382
Assembled reads   : 246460
Organelle genome % : 97.07 %
Average organelle coverage : 305
```

# COMO AVALIAR A MONTAGEM?

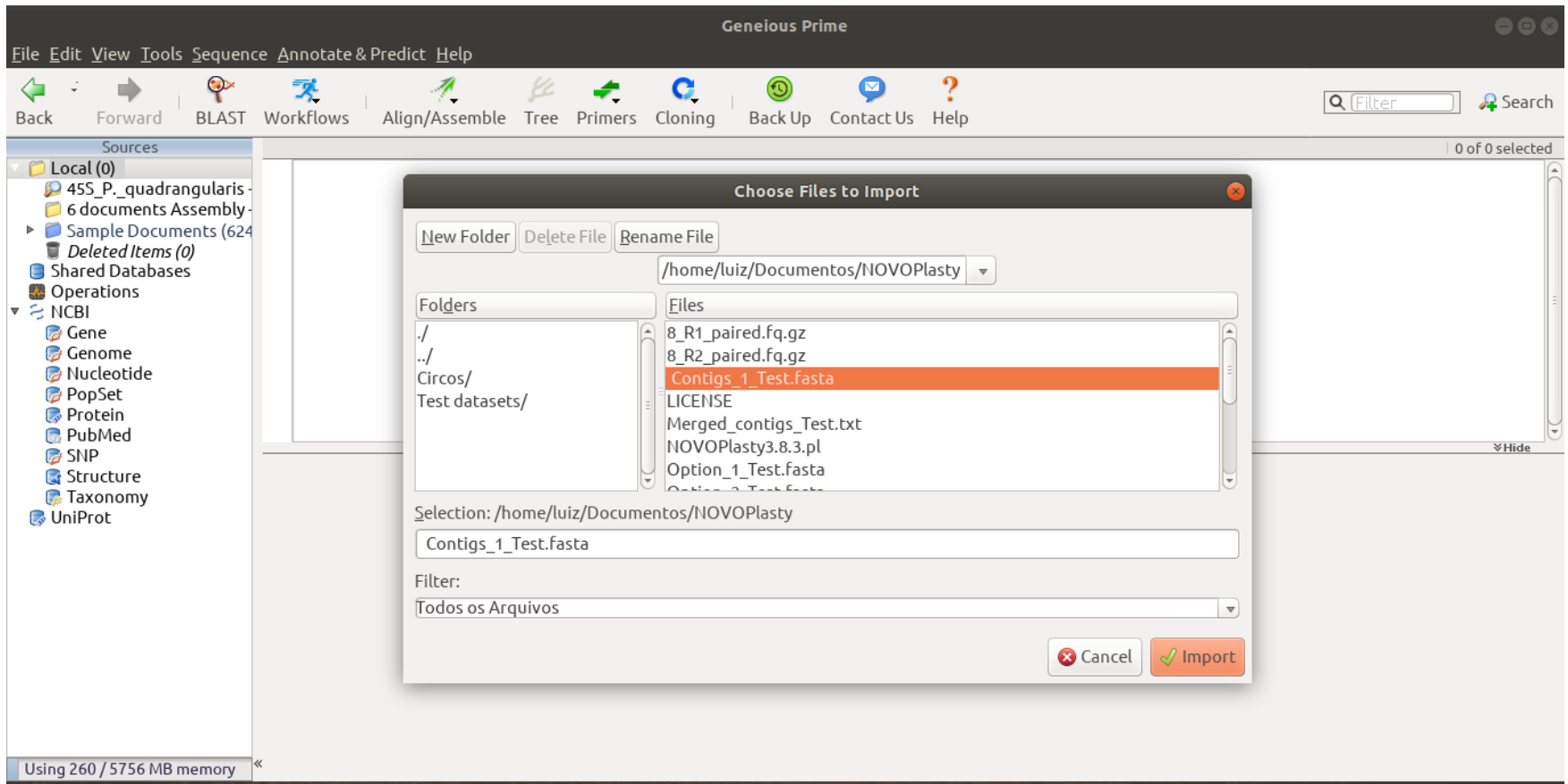
- Tamanho esperado do genoma
- Quantidade de contigs
- Média do tamanho dos contigs
- Valor de N50
- Quantidade de gaps

**Geneious:** Bioinformatics Software for Sequence Data Analysis  
<https://www.geneious.com/>



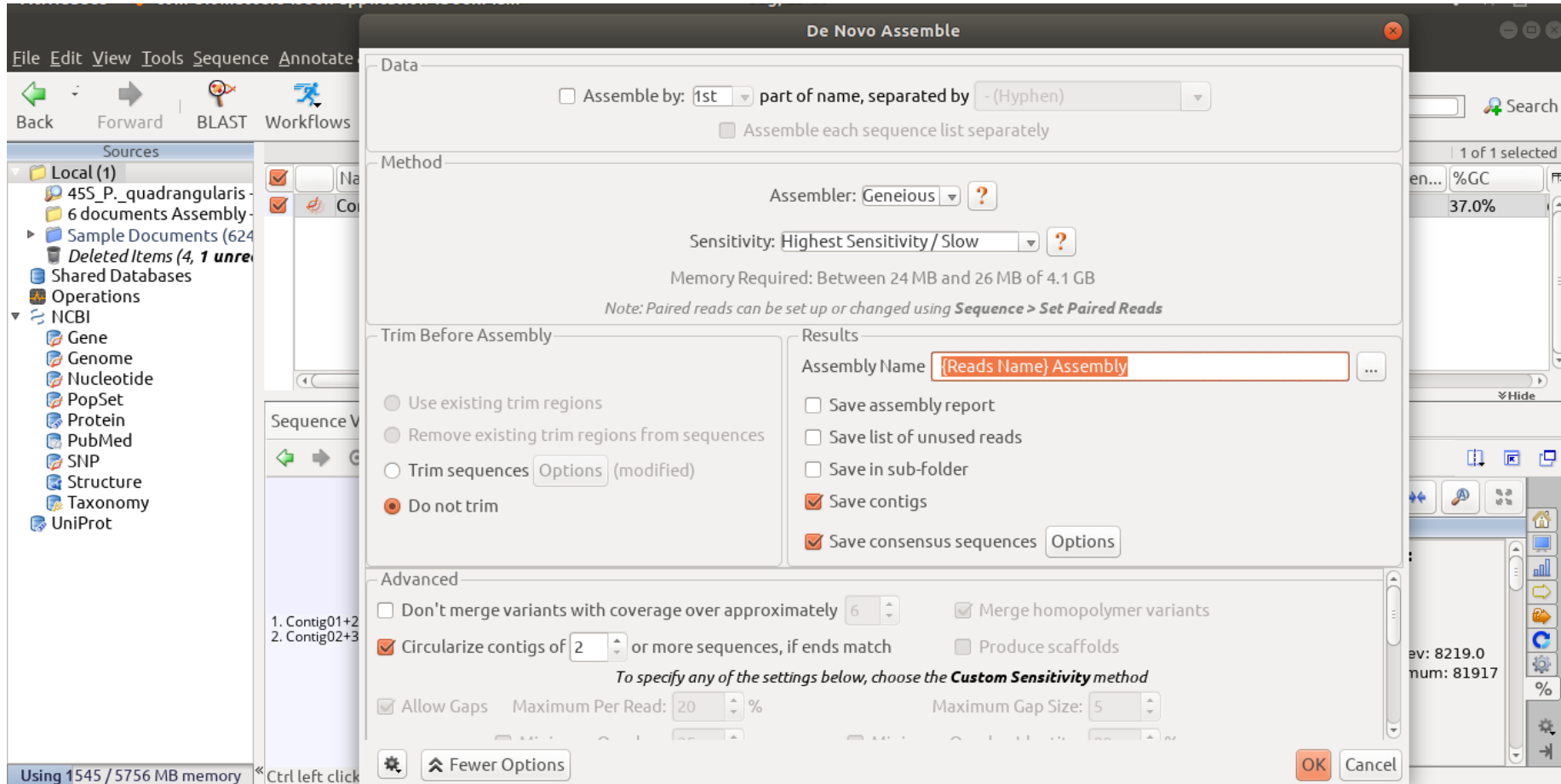
# GENEIOUS

Para fazer o *upload* de um arquivo: File / Import



# GENEIOUS

Para fazer montagem a partir de *contigs* ou de *reads*: Align/Assemble e depois De novo Assemble



# GENEIOUS

## Resultados da montagem: Observar se houve circularização do genoma

The screenshot displays the Geneious Prime interface. On the left, a 'Sources' tree shows the project structure. The main window features a table of sequence entries and a 'Sequence View' of a circular genome.

Name	Description	Modified	Sequenc...	# Sequen...
Consensus_1_Test Assembly	2 reads from Contigs_1_Test assembled usi...	15 Jun 2020 11:02 pm	146,536	-
Contigs_1_Test	-	15 Jun 2020 10:53 pm	-	2

The 'Sequence View' shows a circular genome for 'Consensus\_1\_Test Assembly' with a total length of 146,536 bp. The circular map is labeled with positions from 0 to 146,536 in increments of 10,000. The interface also includes a menu bar, a toolbar with various analysis tools, and a right-hand panel for 'Annotations and Tracks' which currently shows 'This sequence has no annotations.'

Using 1586 / 5756 MB memory

Ctrl left click on a sequence position or annotation, or select a region to zoom in. Ctrl-shift left click to zoom out.

# GENEIOUS

Mapear *reads* sobre a montagem: Fazer *upload* do arquivo Fastq com as *reads*.  
Abrir as opções em Align/Assemble e depois a função Map to Reference

The screenshot displays the Geneious software interface. In the background, the 'Align/Assemble' menu is open, and the 'Consensus\_1\_Test Assembly' is selected in the file browser. The 'Map to Reference' dialog box is the primary focus, showing the following settings:

- Data:** Reference Sequence: Consensus\_1\_Test Assembly - Local. A note states: *8\_R\_paired will be mapped to Consensus\_1\_Test Assembly*.
  - Assemble by: 1st part of name, separated by -(Hyphen)
  - Assemble each sequence list separately
- Method:** Mapper: Geneious. Sensitivity: Medium-Low Sensitivity / Fast.
  - Find structural variants, short insertions, and deletions of any size
  - Find short insertions and large deletions up to 1,000 bp
  - Fine Tuning: Iterate up to 5 times
  - Memory Required: Between 278 MB and 280 MB of 4.0 GB
  - Note: Paired reads can be set up or changed using **Sequence > Set Paired Reads**
- Trim Before Mapping:**
  - Use existing trim regions
  - Remove existing trim regions from sequences
  - Trim sequences (Options) (modified)
  - Do not trim
- Results:**
  - Assembly Name: {reads Name} assembled to {Reference I} ...
  - Save assembly report
  - Save list of unused reads
  - Save list of used reads  Include mates
  - Save in sub-folder
  - Save contigs
  - Save consensus sequences (Options)

At the bottom of the dialog, there is a 'More Options' button and 'OK' and 'Cancel' buttons. The status bar at the bottom left of the application shows 'Using 1614 / 5756 MB memory'.

# GENEIOUS

Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

*Esta função também permite obter a cobertura média do mapeamento*

The screenshot displays the Geneious Prime interface. On the left, a 'Sources' tree shows a project named 'Local (4)' containing '8\_Rpaired assembled to P\_loefgrenii'. The main window shows a table of sequence sources with the following data:

Name	Description	Modified	Sequenc...	# Sequen...
8_Rpaired	Paired reads created from 8_R1paired and...	15 Jun 2020 11:02 pm	-	34,144,258
8_Rpaired assembled to P_loefgrenii	5,879,055 reads from 8_Rpaired mapped t...	15 Jun 2020 11:03 pm	147,715	5,879,056
Consensus_1_Test Assembly	2 reads from Contigs_1_Test assembled usi...	15 Jun 2020 11:02 pm	146,536	-
Contigs_1_Test	-	15 Jun 2020 10:53 pm	-	2

The 'Contig View' for 'P\_loefgrenii' is shown, with a 'Coverage' graph. The graph shows a mean coverage of 2977.3 and a standard deviation of 586.3. The x-axis represents the contig position from 1 to 147,715 bp. A vertical line is positioned at 19,219 bp. The 'Statistics' panel on the right provides the following data:

- Stats include 1,178 hidden columns
- Nucleotide Statistics:**
- Length: 147,715 bp
- Sequences: 5,879,056
- Identical Sites: 29,054 (19.7%)
- Pairwise Identity: 99.8%
- Coverage of 147,791 bases:
- Mean: 2977.3 Std Dev: 586.3
- Minimum: 522 Maximum: 6488

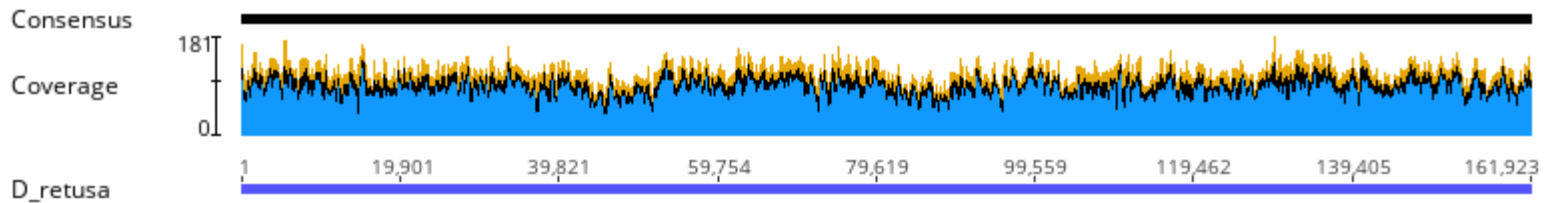
At the bottom, the status bar indicates 'Using 2088 / 5756 MB memory' and 'Cursor before column 19,219 (original base 19,074)'.



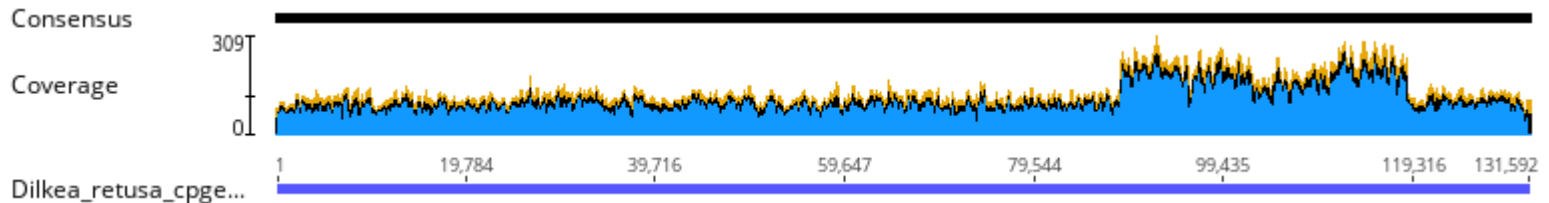
# GENEIOUS

Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

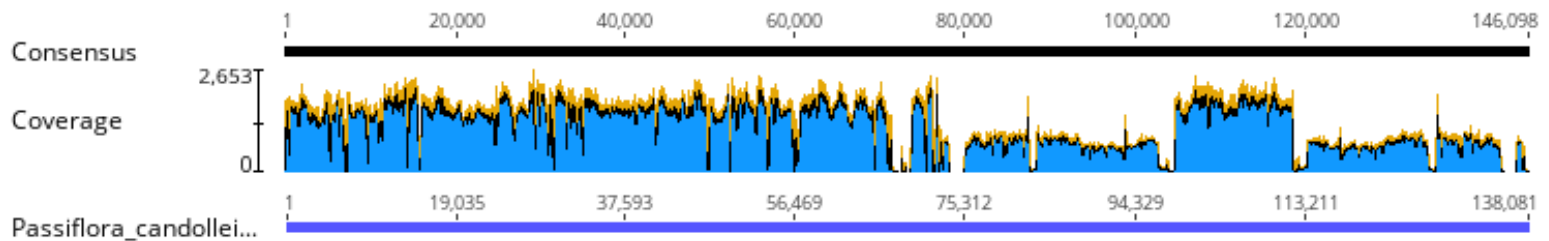
Exemplo: Montagem completa e mapeamento com cobertura contínua



Exemplo: Montagem com regiões erradas



Exemplo: Montagem com regiões erradas



# GENEIOUS

**Obter sequência Consensus:** Selecionar o arquivo de mapeamento e depois File / Export / Consensus sequence

The screenshot displays the Geneious Prime software interface. A dialog box titled "Consensus Sequence(s)" is open, showing settings for generating a consensus sequence. The "Threshold" is set to 65%. Other options include "Ignore Gaps" (unchecked), "Assign Quality" (set to "Highest"), "If no coverage call" (set to "?"), "Call" (set to "?") if Coverage < 2, "Trim to reference sequence" (checked), "Ignore reads mapped to multiple locations" (unchecked), and "Call Sanger Heterozygotes" (set to 50%). The "Append text to name of alignment" field is set to "consensus sequence".

In the background, a table shows sequence statistics for 4 selected items:

Name	Description	Modified	Sequenc...	# Sequen...
red and...		15 Jun 2020 11:02 pm	-	34,144,258
apped t...		15 Jun 2020 11:03 pm	147,715	5,879,056
bled usi...		15 Jun 2020 11:02 pm	146,536	-
		15 Jun 2020 10:53 pm	-	2

The "Nucleotide Statistics" panel shows the following data:

- Stats include 1,178 hidden columns
- Nucleotide Statistics:**
- Length: 147,715 bp
- Sequences: 5,879,056
- Identical Sites: 29,054 (19.7%)
- Pairwise Identity: 99.8%
- Coverage of 147,791 bases:
- Mean: 2977.3 Std Dev: 586.3
- Minimum: 522 Maximum: 6488

The status bar at the bottom indicates "Using 2117 / 5756 MB memory" and "Cursor before column 19,219 (original base 19,074). Mouse over column 120,973 (T) in Consensus".

# BWA

## Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

<http://bio-bwa.sourceforge.net/>

### Burrows-Wheeler Aligner

[Home](#)

#### Introduction

---

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

#### FAQ

---

##### How can I cite BWA?

The short read alignment component (bwa-short) has been published:

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60. [PMID: [19451168](#)]

If you use BWA-SW, please cite:

Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, Epub. [PMID: [20080505](#)]

(See also Errata below for a minor correction to the formulae in these papers.)

**There are three algorithms, which one should I choose?**

#### BWA:

[SF project page](#)  
[SF download page](#)  
[Mailing list](#)  
[BWA manual page](#)  
[Repository](#)

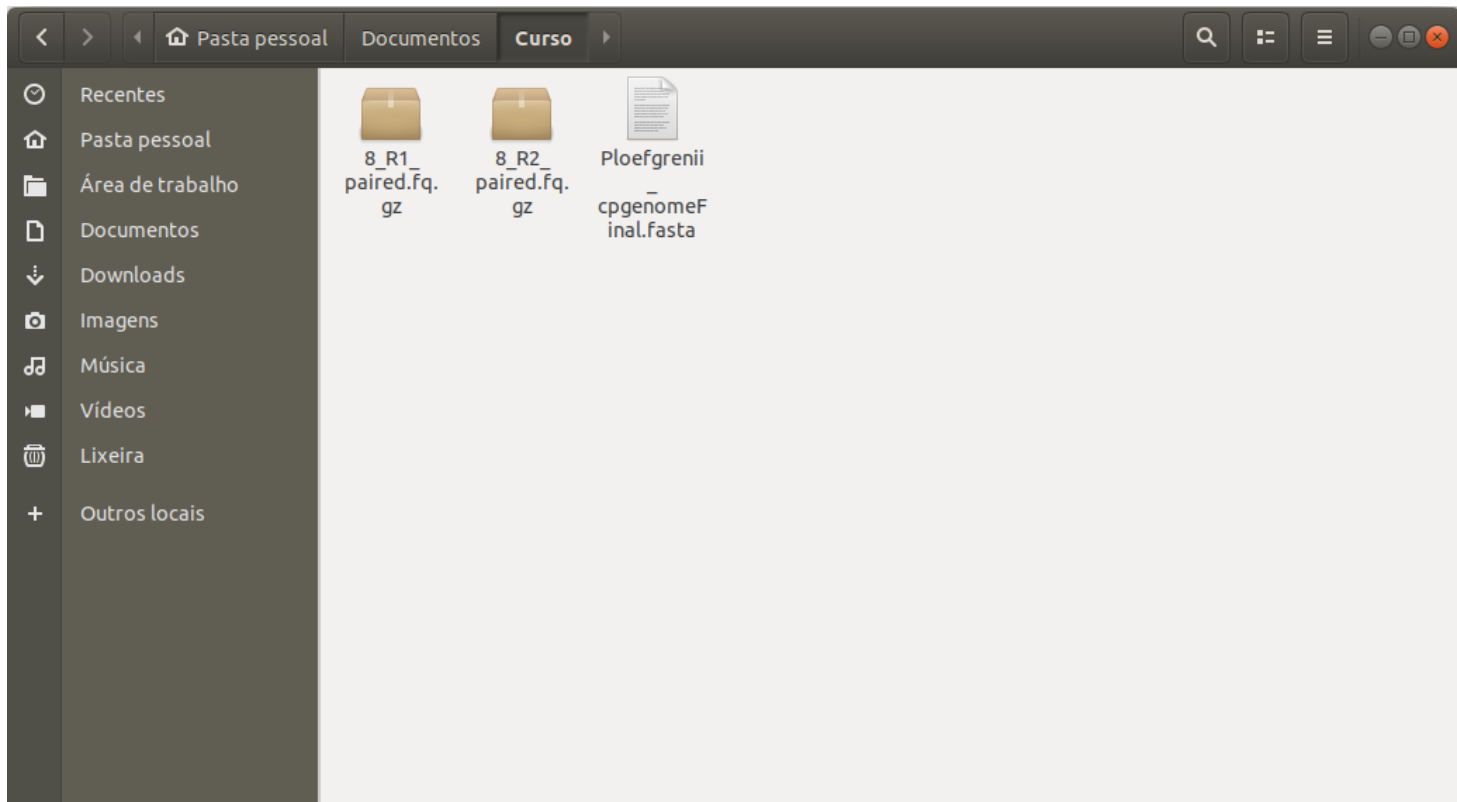
#### Links:

[SAMtools](#)  
[MAQ](#)

# BWA

## Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

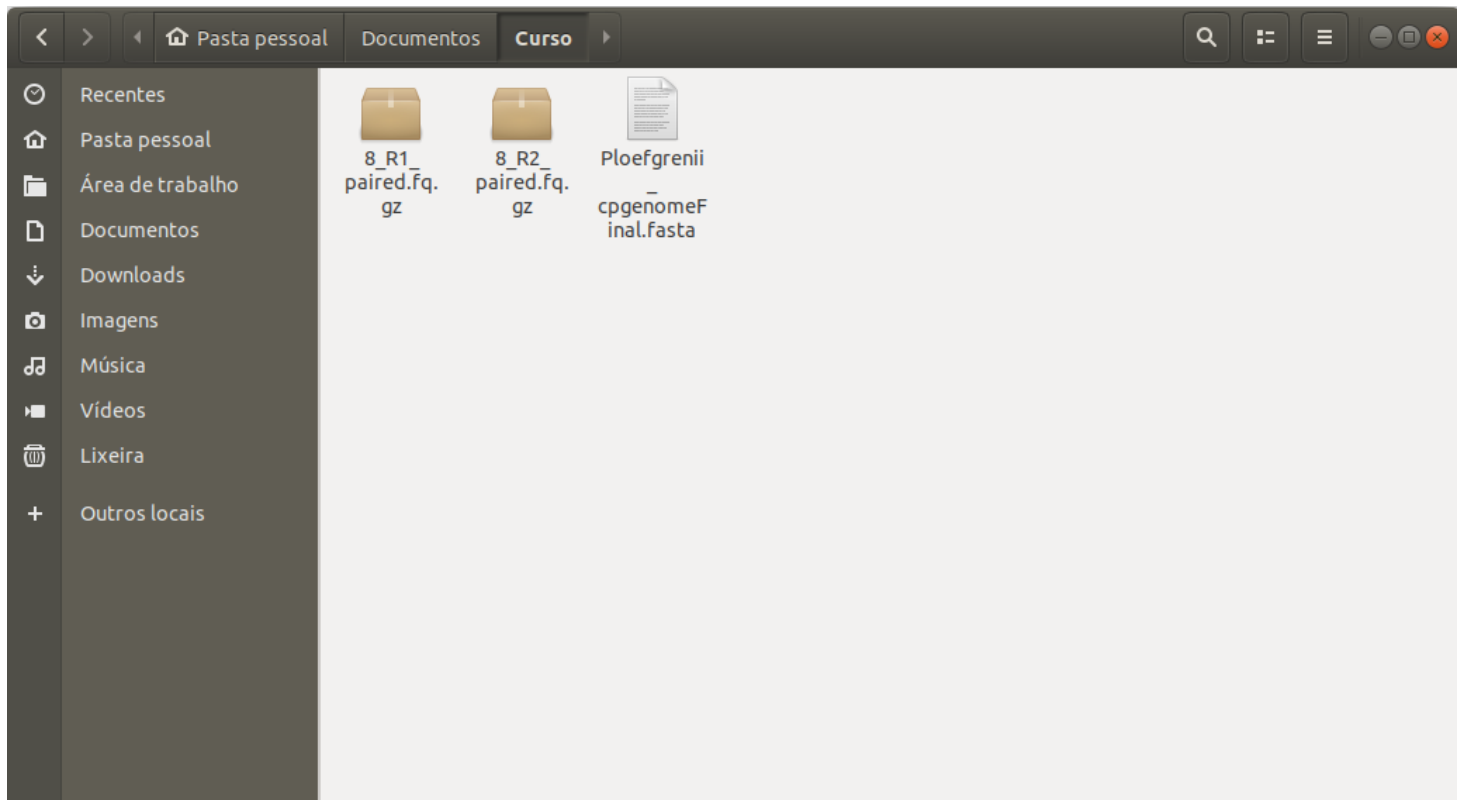
Após instalar o programa, criar uma pasta e incluir os arquivos .fastq contendo as *reads* e o arquivo .fasta com o genoma cp.



# BWA

## Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

Após instalar o programa, criar uma pasta e incluir os arquivos .fastq contendo as *reads* e o arquivo .fasta com o genoma cp.



# BWA

## Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

No terminal, entrar na pasta contendo os arquivos.

O primeiro passo será indexar a referência com o comando: `bwa index arquivoreferência.fasta`

```
luiz@luiz: ~/Documentos/Curso
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~$ cd Documentos/
luiz@luiz:~/Documentos$ cd Curso/
luiz@luiz:~/Documentos/Curso$ ls
8_R1_paired.fq.gz 8_R2_paired.fq.gz Ploefgrenii_cpgenomeFinal.fasta
luiz@luiz:~/Documentos/Curso$ bwa index Ploefgrenii_cpgenomeFinal.fasta
```



```
luiz@luiz: ~/Documentos/Curso
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~$ cd Documentos/
luiz@luiz:~/Documentos$ cd Curso/
luiz@luiz:~/Documentos/Curso$ ls
8_R1_paired.fq.gz 8_R2_paired.fq.gz Ploefgrenii_cpgenomeFinal.fasta
luiz@luiz:~/Documentos/Curso$ bwa index Ploefgrenii_cpgenomeFinal.fasta
[bwa_index] Pack FASTA... 0.01 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.10 seconds elapse.
[bwa_index] Update BWT... 0.01 sec
[bwa_index] Pack forward-only FASTA... 0.01 sec
[bwa_index] Construct SA from BWT and Occ... 0.09 sec
[main] Verston: 0.7.17-r1188
[main] CMD: bwa index Ploefgrenii_cpgenomeFinal.fasta
[main] Real time: 0.547 sec; CPU: 0.216 sec
luiz@luiz:~/Documentos/Curso$
```

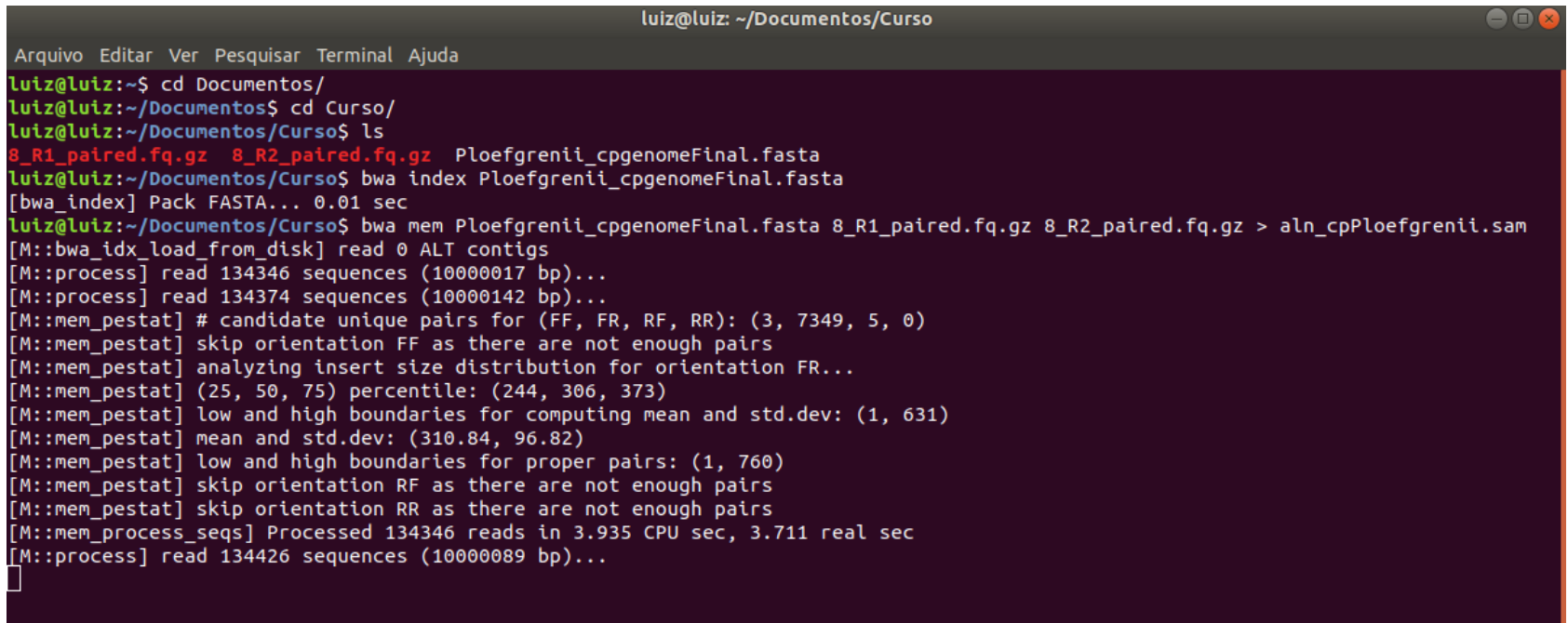
# BWA

## Mapear *reads* sobre a montagem: Avaliar a continuidade da cobertura no alinhamento

Após indexar a referência, o próximo passo é conduzir o mapeamento.

Para o mapeamento podemos utilizar o comando:

```
bwa mem arquivoreferência.fasta Arquivo_R1.fastq Arquivo_R2.fastq > Nomearquivodesaída.sam
```



```
luiz@luiz: ~/Documentos/Curso
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~$ cd Documentos/
luiz@luiz:~/Documentos$ cd Curso/
luiz@luiz:~/Documentos/Curso$ ls
8_R1_paired.fq.gz 8_R2_paired.fq.gz Ploefgrenii_cpgenomeFinal.fasta
luiz@luiz:~/Documentos/Curso$ bwa index Ploefgrenii_cpgenomeFinal.fasta
[bwa_index] Pack FASTA... 0.01 sec
luiz@luiz:~/Documentos/Curso$ bwa mem Ploefgrenii_cpgenomeFinal.fasta 8_R1_paired.fq.gz 8_R2_paired.fq.gz > aln_cpPloefgrenii.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 134346 sequences (10000017 bp)...
[M::process] read 134374 sequences (10000142 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (3, 7349, 5, 0)
[M::mem_pestat] skip orientation FF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation FR...
[M::mem_pestat] (25, 50, 75) percentile: (244, 306, 373)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 631)
[M::mem_pestat] mean and std.dev: (310.84, 96.82)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 760)
[M::mem_pestat] skip orientation RF as there are not enough pairs
[M::mem_pestat] skip orientation RR as there are not enough pairs
[M::mem_process_seqs] Processed 134346 reads in 3.935 CPU sec, 3.711 real sec
[M::process] read 134426 sequences (10000089 bp)...
```





# SAMTOOLS

## Manipular arquivos em formato .sam

<http://samtools.sourceforge.net/>

SAMtools



Home

### SAMTools in C

---

This is the C implementation of SAMTools. This implementation comes as a library in C and a command line tool that packages several utilities including:

- **import**: SAM-to-BAM conversion
- **view**: BAM-to-SAM conversion and subalignment retrieval
- **sort**: sorting alignment
- **merge**: merging multiple sorted alignments
- **index**: indexing sorted alignment
- **faidx**: FASTA indexing and subsequence retrieval
- **tview**: text alignment viewer
- **pileup**: generating position-based output and [consensus/indel calling](#)

API documentation is available [here](#). The full manual page of the command line tool is [here](#). Pileup format is further explained in [this page](#). **Most examples showed on this website are extracted from the example alignment that comes with the source code package.**

### General Information

[SAM Spec v1.4](#)  
[SF Project Page](#)  
[SF Download Page](#)  
[GitHub Project Page](#)  
[Mailing Lists](#)  
[Related Software](#)  
[FAQ](#)

### SAMtools in C

[General Introduction](#)  
[Manual Pages](#)  
[Variant Calling \(mpileup\)](#)  
[Text Alignment Viewer](#)  
[API Documentation](#)  
[Example C Program](#)  
[Working on a Stream](#)  
[Open Tasks](#)  
[Var Calling \(deprecated\)](#)

# SAMTOOLS

## Converter o arquivo .sam em um arquivo .bam

No terminal, entrar na pasta contendo os arquivos do mapeamento.

```
luiz@luiz:~/Documentos/Curso$ samtools
Program: samtools (Tools for alignments in the SAM format)
Version: 1.7 (using htslib 1.7-2)

Usage:  samtools <command> [options]

Commands:
-- Indexing
  dict      create a sequence dictionary file
  faidx     index/extract FASTA
  index     index alignment

-- Editing
  calmd     recalculate MD/NM tags and '=' bases
  fixmate   fix mate information
  reheader  replace BAM header
  targetcut cut fosmid regions (for fosmid pool only)
  addreplacerg adds or replaces RG tags
  markdup   mark duplicates

-- File operations
  collate   shuffle and group alignments by name
  cat       concatenate BAMs
  merge     merge sorted alignments
  mpileup   multi-way pileup
  sort      sort alignment file
  split     splits a file by read group
  quickcheck quickly check if SAM/BAM/CRAM file appears intact
  fastq     converts a BAM to a FASTQ
  fasta     converts a BAM to a FASTA

-- Statistics
  bedcov    read depth per BED region
  depth     compute the depth
  flagstat  simple stats
  idxstats  BAM index stats
  phase     phase heterozygotes
  stats     generate stats (former bamcheck)

-- Viewing
  flags     explain BAM flags
  tview     text alignment viewer
  view      SAM<->BAM<->CRAM conversion
  depad     convert padded BAM to unpadded BAM
```

# SAMTOOLS

## Converter o arquivo .sam em um arquivo .bam

Na pasta contendo os arquivos, utilizar a opção `samtools view` para fazer a conversão do arquivo.

```
luiz@luiz: ~/Documentos/Curso
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~/Documentos/Curso$ samtools view -o aln_cp.bam aln_cpPloefgrenii.sam
```

Após converter o arquivo, utilizar a opção `samtools sort` para ordenar as leituras mapeadas no arquivo .bam

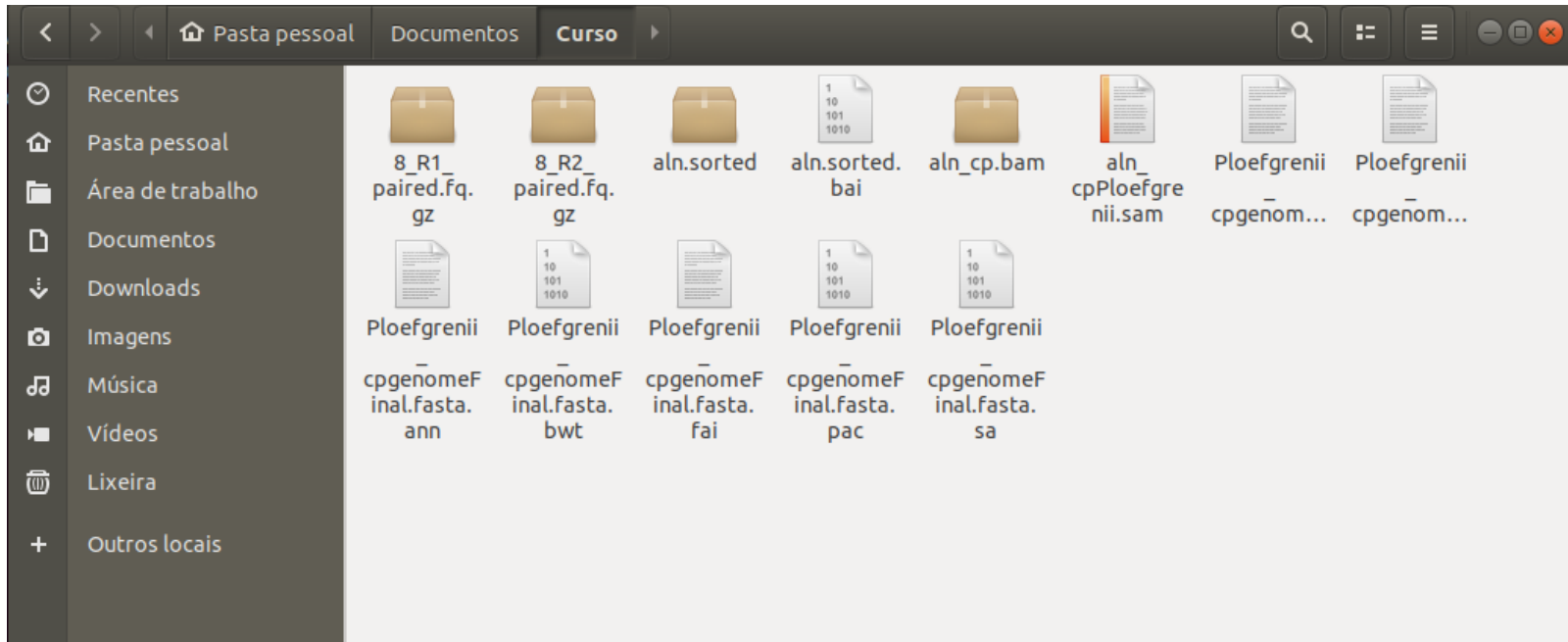
```
luiz@luiz: ~/Documentos/Curso
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~/Documentos/Curso$ samtools sort -T /tmp/aln.sorted -o aln.sorted.bam aln_cp.bam
```

Ao final, utilizar a opção `samtools index` para indexar o arquivo sorted.bam

```
luiz@luiz: ~/Documentos/Curso
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~/Documentos/Curso$ samtools sort -T /tmp/aln.sorted -o aln.sorted.bam aln_cp.bam
[bam_sort_core] merging from 10 files and 1 in-memory blocks...
luiz@luiz:~/Documentos/Curso$ samtools index aln.sorted.bam
```

# SAMTOOLS

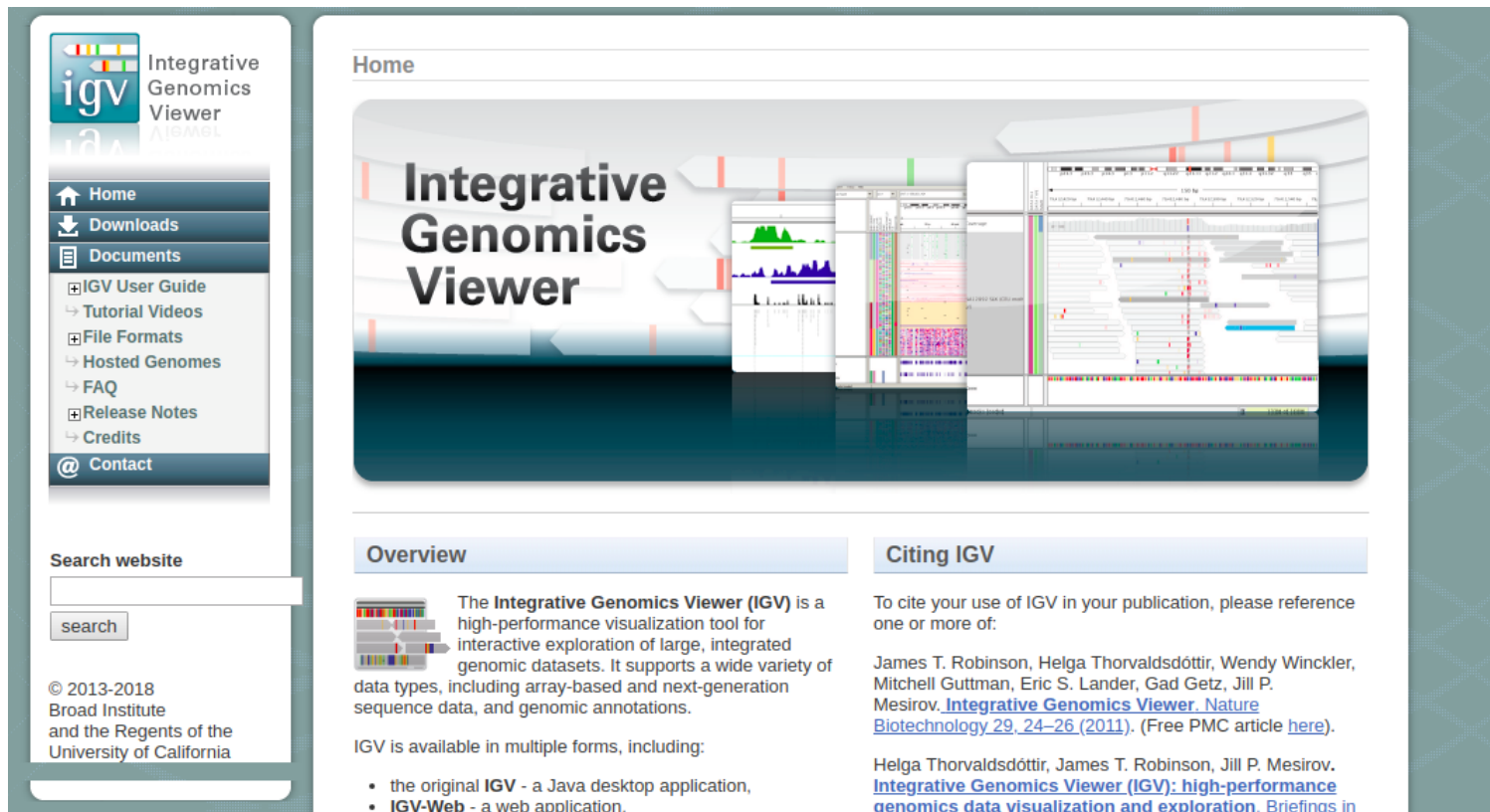
Ao final será gerado um arquivo .bai que será utilizado para visualizar o alinhamento.



# INTEGRATIVE GENOMICS VIEWER (IGV)

Para visualizar o resultado do mapeamento utilizar o IGV.

<http://software.broadinstitute.org/software/igv/>



**igv** Integrative Genomics Viewer

- Home
- Downloads
- Documents
  - IGV User Guide
  - Tutorial Videos
  - File Formats
  - Hosted Genomes
  - FAQ
  - Release Notes
  - Credits
- Contact

Search website

search

© 2013-2018  
Broad Institute  
and the Regents of the  
University of California

## Home

# Integrative Genomics Viewer

### Overview

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

IGV is available in multiple forms, including:

- the original **IGV** - a Java desktop application,
- IGV-Web** - a web application,

### Citing IGV

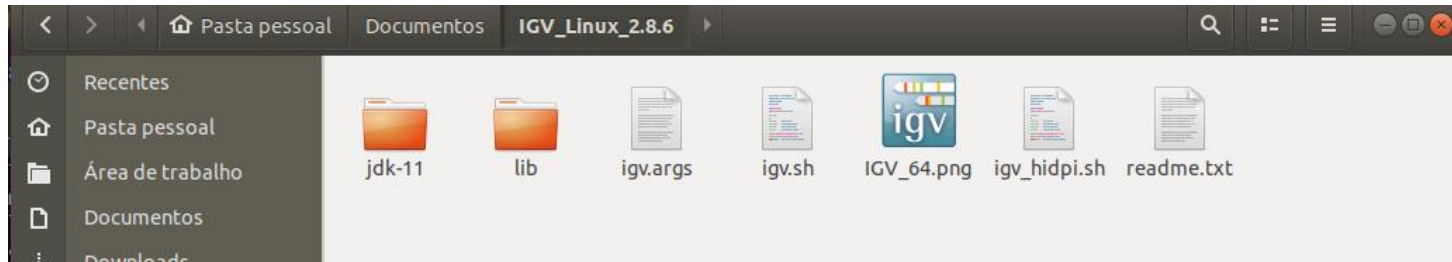
To cite your use of IGV in your publication, please reference one or more of:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011). (Free PMC article [here](#)).

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. [Integrative Genomics Viewer \(IGV\): high-performance genomics data visualization and exploration](#). *Briefings in*

# INTEGRATIVE GENOMICS VIEWER (IGV)

Após fazer o download do programa, utilizar o arquivo executável **igv.sh**

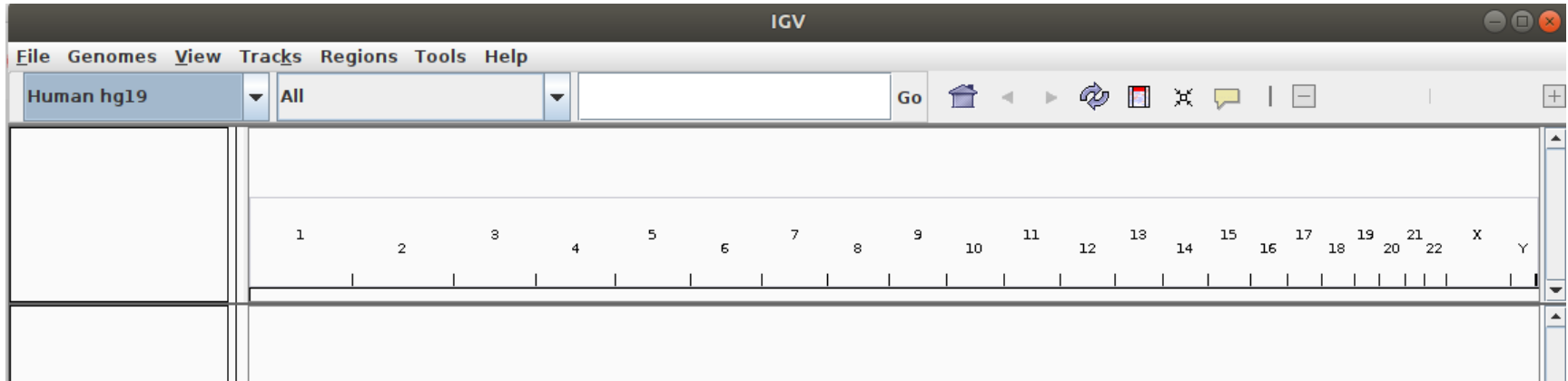


Para executar o programa, após localizar a pasta no terminal digitar: `bash igv.sh`

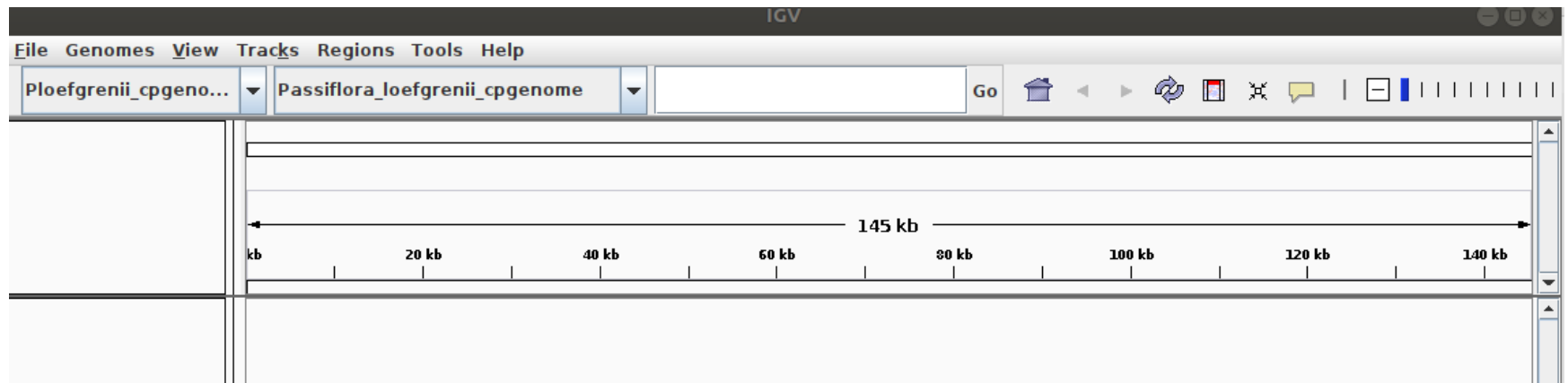
```
luiz@luiz: ~/Documentos/IGV_Linux_2.8.6
Arquivo Editar Ver Pesquisar Terminal Ajuda
luiz@luiz:~$ cd Documentos/
luiz@luiz:~/Documentos$ cd IGV_Linux_2.8.6/
luiz@luiz:~/Documentos/IGV_Linux_2.8.6$ ls
IGV_64.png  igv.args  igv_hidpi.sh  igv.sh  jdk-11  lib  readme.txt
luiz@luiz:~/Documentos/IGV_Linux_2.8.6$ bash igv.sh
```

# INTEGRATIVE GENOMICS VIEWER (IGV)

Após abrir o programa:

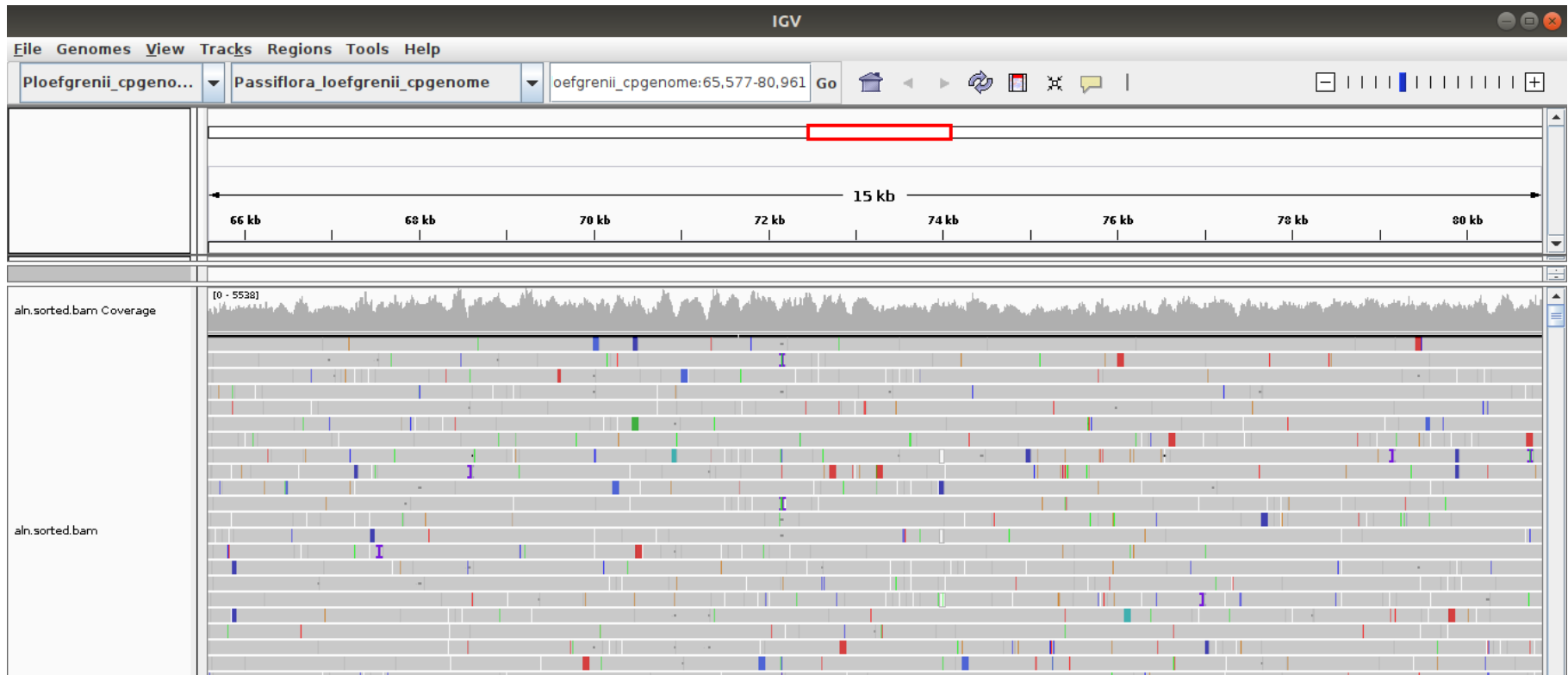


Para abrir a sequência referência clicar em Genomes e depois em Load Genome from File



# INTEGRATIVE GENOMICS VIEWER (IGV)

Abrir o arquivo de mapeamento clicando em File e depois em Load from File  
Abrir o arquivo .bam (**O arquivo .bai deve estar na mesma pasta**)





# GENEIOUS

Estimar o tamanho das regiões repetidas invertidas IRs: Fazer o *upload* da sequência final da montagem

The screenshot displays the Geneious Prime software interface. The top menu bar includes File, Edit, View, Tools, Sequence, Annotate & Predict, and Help. Below the menu is a toolbar with icons for Back, Forward, BLAST, Workflows, Align/Assemble, Tree, Primers, Cloning, Back Up, Contact Us, and Help. A search bar is located on the right side of the toolbar.

The left sidebar shows a tree view of sources, including Local (5), Shared Databases, and Operations. The main window displays a table of sequence assemblies:

Name	Description	Modified	Sequenc...	# Sequen...
8_R_paired	Paired reads created from 8_R1_paired and...	15 Jun 2020 11:02 pm	-	34,144,258
8_R_paired assembled to P_loefgrenii	5,879,055 reads from 8_R_paired mapped t...	15 Jun 2020 11:03 pm	147,715	5,879,056
Consensuss_1_Test Assembly	2 reads from Contigs_1_Test assembled usi...	15 Jun 2020 11:02 pm	146,536	-
Contigs_1_Test	-	15 Jun 2020 10:53 pm	-	2
P_loefgrenii_cpgenome_consensus_mapping	5,879,055 reads from 8_R_paired mapped t...	15 Jun 2020 11:26 pm	146,537	-

The bottom section of the interface shows the Sequence View for the selected assembly. It includes a sequence viewer with a scale from 0 to 146,537 bp. The right sidebar displays the Nucleotide Statistics:

**Nucleotide Statistics:**  
Length: 146,537 bp  
Rough Tm: 84.5°C

	Freq	%
A:	44,819	30.6%
C:	27,583	18.8%
G:	26,717	18.2%
T:	47,418	32.4%

At the bottom of the window, a status bar indicates "Using 1613 / 5756 MB memory" and provides instructions: "Ctrl left click on a sequence position or annotation, or select a region to zoom in. Ctrl-shift left click to zoom out."

# GENEIOUS

**Estimar o tamanho das regiões repetidas invertidas IRs: Abrir as opções em Annotate & Predict e depois a função Find Repeats**

Geneious Prime

File Edit View Tools Sequence Annotate & Predict Help

Back Forward BLAST Workflows Align/Assemble Tree Primers Cloning Back Up Contact Us Help

Sources

- Local (5)
- 45S\_P\_quadrangularis
- 6 documents Assembly
- Sample Documents (624)
- Deleted Items (3, 1 unreferenced)
- Shared Databases
- Operations
- NCBI
  - Gene
  - Genome
  - Nucleotide
  - PopSet
  - Protein
  - PubMed
  - SNP
  - Structure
  - Taxonomy
  - UniProt

Name	Description	Modified	Sequenc...	# Sequen...
8_R_paired	Paired reads created from 8_R1_paired and...	15 Jun 2020 11:02 pm	-	34,144,258
8_R_paired assembled to P_loefgrenii	5,879,055 reads from 8_R_paired mapped t...	15 Jun 2020 11:03 pm	147,715	5,879,056
Consensus	...	15 Jun 2020 11:02 pm	146,536	-
Contigs_1	...	15 Jun 2020 10:53 pm	-	2
P_loefgrenii	...	15 Jun 2020 11:26 pm	146,537	-

Find Repeats

Minimum repeat length: 1,000

Maximum mismatches: 0 %

Exclude repeats up to 10 bp longer than contained repeat

Exclude contained repeats when longer repeat has frequency at least 3

Only find repeats in selected region

Maximum repeats (approximate) to find: 10,000

OK Cancel

Sequence View

10,000 20,000 30,000 40,000 50,000 60,000 70,000 80,000 90,000 100,000 110,000 120,000 130,000 146,537

Annotations and Tracks

Filter

This sequence has no annotations.

Using 1256 / 5756 MB memory

Ctrl left click on a sequence position or annotation, or select a region to zoom in. Ctrl-shift left click to zoom out.

## Tamanho, localização e orientação das regiões repetidas invertidas:

The screenshot displays the Geneious Prime interface. The top menu bar includes File, Edit, View, Tools, Sequence, Annotate & Predict, and Help. The toolbar contains icons for Back, Forward, BLAST, Workflows, Align/Assemble, Tree, Primers, Cloning, Back Up, Contact Us, and Help. A search bar is located on the right side of the toolbar.

The left sidebar shows a tree view of sources, including Local (5), NCBI, and various databases like Gene, Genome, Nucleotide, etc.

The main window displays a table of sequence data:

Name	Description	Modified	Sequenc...	# Sequen...
8_R_paired	Paired reads created from 8_R1_paired and...	15 Jun 2020 11:02 pm	-	34,144,258
8_R_paired assembled to P_loefgrenii	5,879,055 reads from 8_R_paired mapped t...	15 Jun 2020 11:03 pm	147,715	5,879,056
Consensus_1_Test Assembly	2 reads from Contigs_1_Test assembled usi...	15 Jun 2020 11:02 pm	146,536	-
Contigs_1_Test	-	15 Jun 2020 10:53 pm	-	2
P_loefgrenii_cpgenome_consensus_mapping	5,879,055 reads from 8_R_paired mapped t...	15 Jun 2020 11:26 pm	146,537	-

Below the table, the 'Sequence View' tab is active, showing a genomic map with annotations. The map displays two 'Repeat 1' regions with start and end coordinates: 44,366 to 67,815 and 81,082 to 104,531. The total length of the sequence is 146,537 bases.

The bottom status bar indicates: Using 92 / 5756 MB memory. Selected 46,900 bases from 44,366 to 104,531. Mouse over base 142,444 (C).

**OBS:** Para a anotação e depósito da sequência, o ideal após identificar a localização das IRs é alterar o início da sequência do genoma cp. Considere como início da sequência a primeira base após a segunda IR

Desta forma o início da sequência será a primeira base da região LSC

# OBTER INFORMAÇÕES DAS REPETIÇÕES

<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>



## Navigation

- Tools
  - Alignment
  - Evolutionary Relationships
- Genome Comparison
  - AGenDA
  - AggloIndel
  - CEGeD
  - CG-CAT
  - DCJ
  - FFGC
  - Gecko
  - GEvolutionS
  - GraphTeams
  - MGA
  - newdist

## REPuter

[Welcome](#) [Submission](#) [WebService](#) [Download](#) [Manual](#) [References](#) [Reset Session](#)

Author: S. Kurtz

The repetitive structure of genomic DNA holds many secrets to be discovered. A systematic study of repetitive DNA on a genomic or inter-genomic scale requires extensive algorithmic support. The *REPuter* program was designed to serve as a fundamental tool in such studies. Efficient and complete detection of various types of repeats is provided together with an evaluation of significance and interactive visualization.

**REPuter**

The website provides a partly limited online version of *REPuter*. The uploaded data size is hard limited to 5 Mb and a maximum of 5000 repeats is calculated due our server capacity. There is a standalone version of *REPuter* available for Linux, OSX, Solaris, Irix and Alpha. If you are interested to obtain a standalone version, please download the [license agreement](#) and follow the instructions.

[privacy policy](#) [paRNAss](#)  
[Download](#) [Rose](#)  
[Download](#) [AltAVist](#)  
[CG-CAT](#) [References](#)  
[TCRProfiler](#) [Gecko](#) [acdc](#)  
[Welcome](#) **REPuter**  
[BiBiServ Team](#)  
[Statuscodes](#) [ClustalW](#)  
[References](#) [FFGC](#)  
[ROCOCO](#) [Intronserter](#)  
[Dialign](#) [libfid](#) [Welcome](#)  
[MoRAine](#) [license](#)  
[PoSSuMsearch2](#) [Metrans](#)  
[PoSSuMsearch](#) [mmfind](#)  
[Roci](#) [Manual](#) [AGT-SDP](#)  
[SplitsTree](#)  
[WebService](#) [Linklist](#)  
[genefisher2](#) [CEGeD](#)

- Tools
  - Alignment
  - Evolutionary Relationships
  - Genome Comparison
    - AGenDA
    - AggloIndel
    - CEGeD
    - CG-CAT
    - DCJ
    - FFGC
    - Gecko
    - GEvolutionS
    - GraphTeams
    - MGA
    - newdist
    - REPuter

Welcome Submission **WebService** Download Manual References Reset Session

**Nuclear Acid Repeat Calculation:** Compute repeats in nuclear acid sequences.

**DNA input sequence:** ? A nuclear acid sequence.

(1) Select data input method :

- File Upload
- Copy & Paste
- AWS support
- File on Server

Can be activated on local instances.

(2) File Upload:

+ Upload

File "Consensus\_Ploefgrenii.fasta" successfull uploaded!

Yes  No Skip validation and format-conversion step. If activated you can only use the following format: Fasta

◀ back example reset next ▶ 1/3

- Tools
  - Alignment
  - Evolutionary Relationships
  - Genome Comparison
    - AGenDA
    - AggloIndel
    - CEGeD
    - CG-CAT
    - DCJ
    - FFGC
    - Gecko
    - GEvolutionS

Welcome Submission **WebService** Download Manual References Reset Session

**Match Direction** (?)

- Forward (direct) [-f]
- Reverse [-r]
- Complement [-c]
- Palindromic [-p]

**Edit distance** (?)

**Hamming Distance** (?)

**Maximum Computed Repeats** (?)

**Minimal Repeat Size** (?)

◀ back example reset next ▶ 2/3

- Tools
  - Alignment
  - Evolutionary Relationships
  - Genome Comparison
    - AGenDA
    - AggloIndel
    - CEGeD
    - CG-CAT
    - DCJ
    - FFGC

Welcome Submission **WebService** Download Manual References Reset Session

(1) Select handling of the result :

- Download from BiBiServ2
- Upload to AWS Bucket

The result will be downloadable directly from BiBiServ2 after the tool finished.

◀ back No example available reset start calculation 3/3

# OBTER INFORMAÇÕES DAS REPETIÇÕES

## Navigation

- Tools
  - Alignment
  - Evolutionary Relationships
  - Genome Comparison
    - AGenDA
    - AggloIndel
    - CEGeD
    - CG-CAT
    - DCJ
    - FFGC
    - Gecko

## REPuter[reputer\_function\_nuclear\_acid\_repeat\_calculation] - Result

Welcome Submission WebService Download Manual References Reset Session

ID: bibiserv2\_2020-06-16\_044249\_g9snF

View/Filter Result using interactive viewer (recommend)

Open interactive viewer

Output represented in Format depending on Tool format

Download

PlainText-Visualizer

```
# /vol/bioapps/bin/repfind -c -f -p -r -l 30 -h 3 -best 50 /var/bibiserv2/anonymous/
# 146537_3_30 /var/bibiserv2/anonymous/reputer/16/04/42/bibiserv2_2020-06-16_044:
23450 44365 P 23450 81081 0 0.00e+00
103 145913 F 103 145904 -1 1.81e-30
97 48 F 97 156 -3 9.58e-43
96 222 F 96 146441 -3 3.71e-42
76 242 F 76 146461 0 1.06e-36
80 31934 F 80 32000 -3 9.17e-33
79 17034 F 79 19258 -3 3.53e-32
62 4048 P 62 4048 0 2.84e-28
71 104 F 71 212 -3 1.67e-27
64 31891 F 64 31957 -1 3.41e-27
67 17046 F 67 19270 -2 5.52e-27
63 109493 F 63 109524 -2 1.25e-24
63 145383 F 63 145650 -2 1.25e-24
54 74244 P 54 74244 0 1.86e-23
57 15316 F 57 15336 -1 4.97e-23
57 31957 F 57 32023 -1 4.97e-23
58 31890 F 58 32022 -2 1.08e-21
51 145461 F 51 145710 0 1.19e-21
54 234 F 54 288 -1 3.01e-21
50 292 F 50 146457 0 4.76e-21
52 145913 F 52 146015 -1 4.65e-20
56 95 F 56 203 -3 8.71e-19
46 242 F 46 296 0 1.22e-18
49 49250 P 49 49250 -1 2.80e-18
```



# ANOTAÇÃO

**CHLOROBOX:** Programas para análises com dados de genomas organelares

<https://chlorobox.mpimp-golm.mpg.de/>



Max Planck Institute of Molecular Plant Physiology

**CHLOROBOX** | [GeSeq](#) | [GB2sequin](#) | [OGDRAW](#) | [LOLA](#) | [GBSON](#) | [ISE-G 2015](#) | [Contact](#) | [Disclaimer](#)

The **CHLOROBOX** offers software tools for analyses of plant derived nucleic acid and protein sequences.

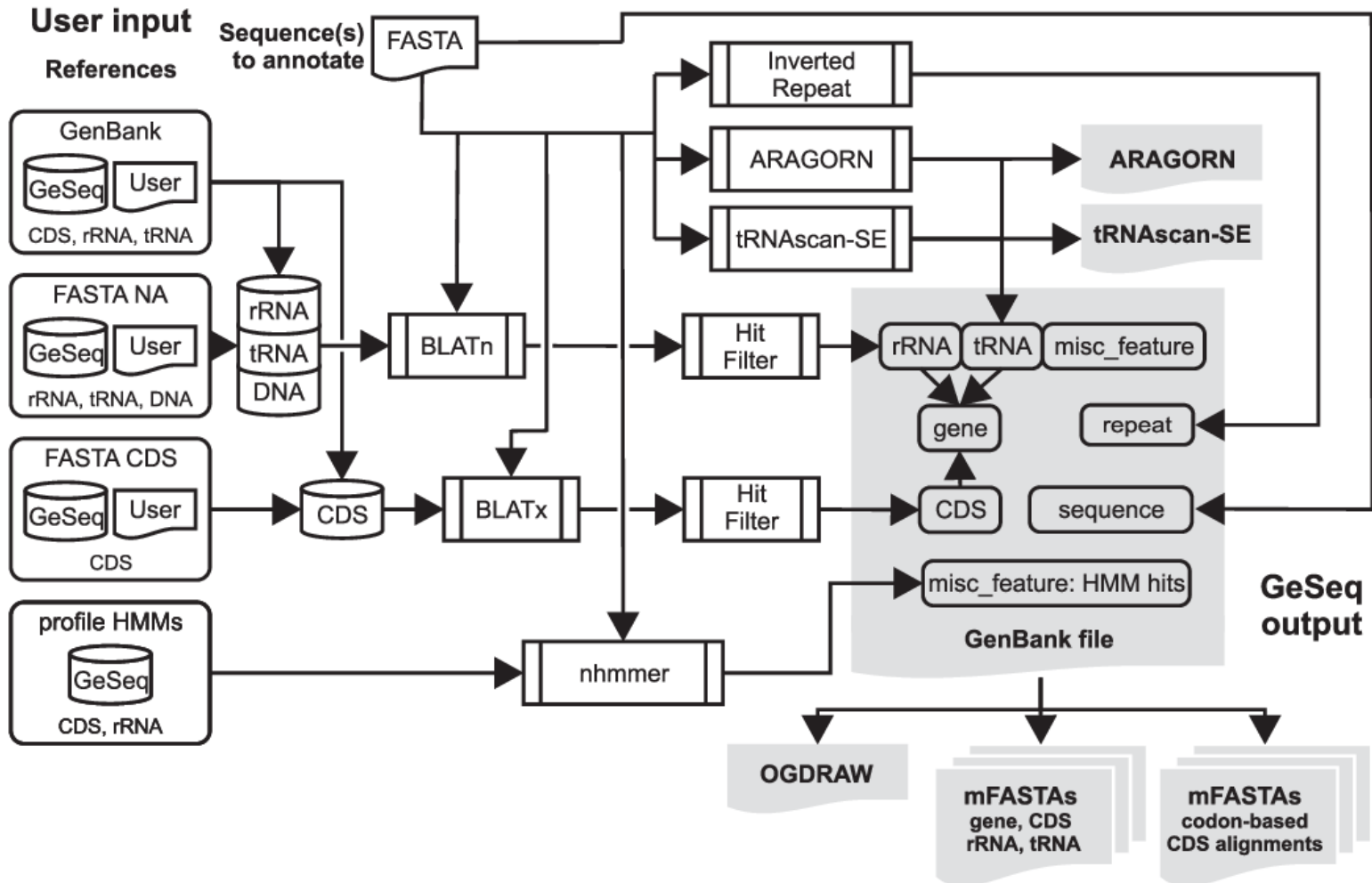
Open Chlorobox Programs

- GeSeq**  
Annotation of Organellar Genomes
- GB2sequin**  
Creation of NCBI Update or Submission Files from GenBank Files
- OGDRAW**  
Draw High-Quality Graphical Maps of Organellar Genomes
- LOLA**  
Extraction and Pairwise Comparison of DNA Sequence Segments
- Genbank-JSON Converter**  
Convert Genbank files to JSON and back using GBSON-Notation.

# GeSeq: Anotação de genomas organelares

(Tillich et al., 2017. Nucleic Acids Research)

<https://chlorobox.mpimp-golm.mpg.de/geseq.html>





# ANOTAÇÃO

**Geseq:** Anotação de genomas organelares  
(Tillich et al., 2017. Nucleic Acids Research)



<https://chlorobox.mpimp-golm.mpg.de/geseq.html>

The screenshot shows the GeSeq web application interface. At the top is a green navigation bar with the CHLOROBX logo and links for GeSeq, GB2sequin, OGDRAW, LOLA, GBSON, ISE-G 2015, Contact, and Disclaimer. A left sidebar contains a menu for CHLOROBX with options like Application, Documentation, Release Notes, Browser Compatibility, 3rd Party Software, Alternative Tools, Credits, and How to cite GeSeq? The main content area is titled "GeSeq - Annotation of Organellar Genomes" and contains introductory text, a link to documentation, and a citation for Tillich et al. (2017). At the bottom, there are three panels: "FASTA file(s) to annotate" (with an "Upload File(s)" button), "BLAT Reference Sequences" (with a "Server References" button), and "Actions" (with "Submit", "Reset", and "Example" buttons, and a checkbox for "I have read and accept the Disclaimer").

CHLOROBX | [GeSeq](#) | [GB2sequin](#) | [OGDRAW](#) | [LOLA](#) | [GBSON](#) | [ISE-G 2015](#) | [Contact](#) | [Disclaimer](#)

## GeSeq - Annotation of Organellar Genomes

GeSeq has been developed for a rapid and accurate annotation of organelle genomes, in particular chloroplast genomes.

Please take a look at our [documentation](#) which includes a quickstart section and do not hesitate to report bugs or suggestions for improvements by email.

*Citations keep this server running. If you use GeSeq please cite:*

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* 45: W6-W11

Hover over buttons for tips. To load an example job, press "Example" in the "Actions" field.

FASTA file(s) to annotate	BLAT Reference Sequences	Actions
0 files in list <a href="#">Upload File(s)</a>	Server References <a href="#">Add MGBL (FASTA)</a>	<a href="#">Submit</a> <a href="#">Reset</a> <a href="#">Example</a> <input type="checkbox"/> I have read and accept the Disclaimer

**FASTA file(s) to annotate**

1 file in list Upload File(s)

Ploefgrenii\_cpgenome.fasta ×

Circular  Linear

Sequence source

Plastid  Mitochondrial

**Options**

Generate multi-GenBank

Generate multi-GFF3

Generate multi-GBSON

Generate multi-FASTA

Generate codon-based alignments

Display raw BLAT output

**Annotation**

**BLAT search**

Annotate plastid IR

Annotate plastid trans-spliced rps12

Ignore genes annotated as locus tag

Ignore genes annotated as ORFs

Protein search identity

25

rRNA, tRNA, DNA search identity

60

**HMMER profile search**

Embryophyta chloroplast (CDS + rRNA)

**3rd Party tRNA annotators**

ARAGORN v1.2.38

Genetic code

Bacterial/Plant Chloroplast

Max intron length

3000 bp

Options

Allow overlaps

Fix intron

Report low scoring tRNAs

**BLAT Reference Sequences**

**Server References**

Add NCBI RefSeq(s)

no RefSeq selected

MPI-MP chloroplast references (Embryophyta CDS + rRNA)

**Custom References**

GenBank/ENA

0 files in list Upload File(s)

FASTA Nucleotide (CDS)

0 files in list Upload File(s)

FASTA Nucleotide (tRNA, rRNA, primer, other DNA or RNA)

0 files in list Upload File(s)

tRNAscan-SE v2.0.5

Sequence source

Organelar tRNAs

Search mode

Default

Genetic Code

Universal

Cut-off score for reporting tRNAs

15

Score and report output options

Disable pseudogene checking

Display detailed prediction output

Show origin of first-pass hits

Show primary and secondary structure components to scores

**Actions**

Submit Reset Example

I have read and accept the Disclaimer

**Results**

### Passos:

- Abrir o arquivo .fasta em Upload file;
- Escolher as opções de arquivo de saída (GenBank, multi-GFF3);
- Selecionar Annotate plastid IR e Annotate plastid trans-spliced rps12;
- Alterar os valores para Protein search identity de acordo com o seu conjunto de dados e das referências da anotação;
- Selecionar as referências para a anotação em BLAT Reference Sequences (Dica: Usar o MPI-MP chloroplast references);
- Submeter a análise clicando em Submit.

- Application
- Documentation
- Release Notes
- Browser Compatibility
- 3rd Party Software
- Alternative Tools
- Credits
- How to cite GeSeq?

### FASTA file(s) to annotate

1 file in list [Upload File\(s\)](#)

Ploefgrenii\_cpgenome.fasta ✕

Circular  Linear

Sequence source  
 Plastid  Mitochondrial

### Options

- Generate multi-GenBank
- Generate multi-GFF3
- Generate multi-GBSOM
- Generate multi-FASTA
- Generate codon-based alignments
- Display raw BLAT output

### Annotation

#### BLAT search

- Annotate plastid IR
- Annotate plastid trans-spliced rps12
- Ignore genes annotated as locus tag
- Ignore genes annotated as ORFs

### BLAT Reference Sequences

#### Server References

[Add NCBI RefSeq\(s\)](#)

no RefSeq selected

MPI-MP chloroplast references (Embryophyta CDS + rRNA)

#### Custom References

GenBank/ENA

0 files in list [Upload File\(s\)](#)

FASTA Nucleotide (CDS)

0 files in list [Upload File\(s\)](#)

FASTA Nucleotide (tRNA, rRNA, primer, other DNA or RNA)

0 files in list [Upload File\(s\)](#)

### Actions

[Submit](#) [Reset](#) [Example](#)

I have read and accept the Disclaimer

### Results

Organelle Genome Resources of the NCBI Reference Sequence Database (RefSeq)

Search  Type  All  Chloroplast  Mitochondrion [Clear selection](#)

*Gunneridae* [ADD](#)

*Pentapetalae* [ADD](#)

*Rosids* [ADD](#)

*Fabids* [ADD](#)

*Malpighiales* [ADD](#)

*Passifloraceae* [ADD](#)

*Passiflora* [ADD](#)

- Passiflora actinia* [ADD](#)
- Passiflora auriculata* [ADD](#)
- Passiflora biflora* [ADD](#)
- Passiflora cincinnata* [ADD](#)
- Passiflora edulis* [ADD](#)
- Passiflora laurifolia* [ADD](#)
- Passiflora ligularis* [ADD](#)
- Passiflora nitida* [ADD](#)
- Passiflora oerstedii* [ADD](#)
- Passiflora pittieri* [ADD](#)
- Passiflora quadrangularis* [ADD](#)
- Passiflora retipetala* [ADD](#)
- Passiflora serratifolia* [ADD](#)
- Passiflora serratodigitata* [ADD](#)
- Passiflora vitifolia* [ADD](#)

- Passiflora edulis* NC\_034285.1 ✕
- Passiflora actinia* NC\_038118.1 ✕
- Passiflora cincinnata* NC\_037690.1 ✕

Chloroplast ■ Mitochondrion ■ Other ■

[Ok](#) [Cancel](#)



# Resultados da anotação no Geseq:

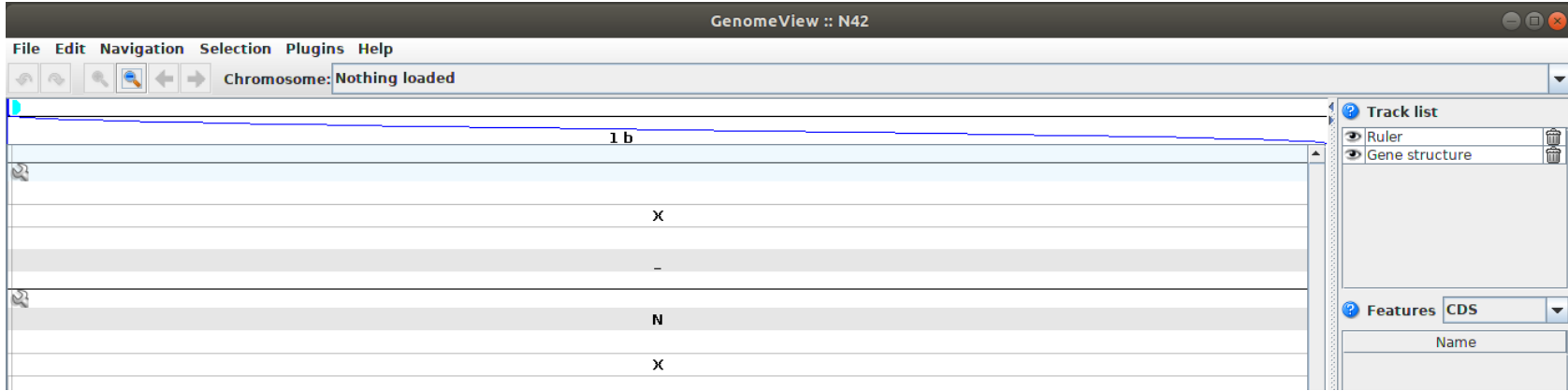
```
File GeSeqJob-20200602-13409_Passiflora_loefgrenii_cpgenome_GenBank.gb
misc_feature 1..86371
/organism="Passiflora"
/organelle="plastid:chloroplast"
/mol_type="genomic DNA"
misc_feature 1..86371
/note="large single copy (LSC), annotated by OGDRAW v1.4"
gene 229..1290
/gene="psbA"
/note="blatX_hit psbA_NC_038118.1, position 1 - 1062, psi
score 99.3, coverage 100.00%, match 99.25%"
CDS 229..1290
/gene="psbA"
/translation="MTAILERRESESLWGRFCNWTSTENRLYIGWGVLMIPTLLTA
TSVFIIAFIAAPPVDIDGIREPVSGLLYGNNIISGAIPTSAAGLHFYPIWEAASV
DEWLYNGGPYELIVLHFLLVACVMYMGREWELSFRLGMRPWIAVAYSAPVAAAAVFLI
YPIGQGSFSDGMPGLGISGTFNFMIVFQAEHNLMPFHMLGVAGVFGSLSFAMHGSL
VTSSLIRETTEENESANEGRYRFGQEEETYNIVAAGYFGRLIQYASFNNSRSLHFLA
AWPVGIGWFTALGISTMAFNLNGFNFNQSVVDSOGRVINTWADIINRANLGMVMHER
NAHNFPLDLAAVEAPSTNG"
1290..1290
```

Arquivo .gff3 com os dados da anotação:

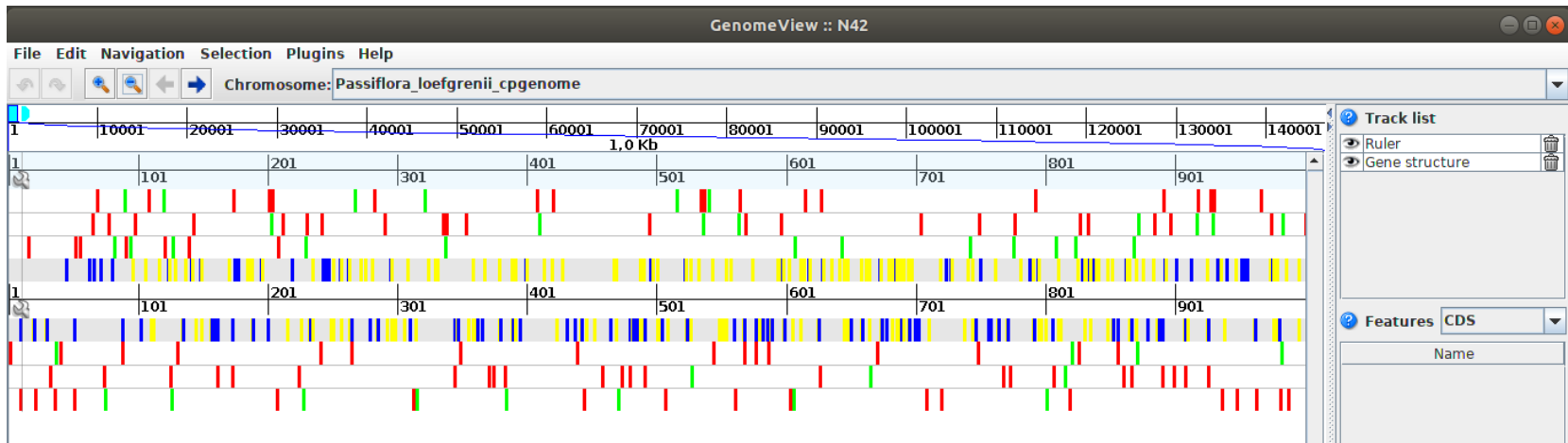
```
File GeSeqJob-20200602-13409_Passiflora_loefgrenii_cpgenome_GFF3.gff3
##gff-version 3
##source-version GeSeq 1.82
##sequence-region Passiflora_loefgrenii_cpgenome 1 146537
Passiflora_loefgrenii_cpgenome GeSeq region 1 146537 . + 0 Is_circular=true
Passiflora_loefgrenii_cpgenome GeSeq source 1 146537 . + 1 ID=source-null;gbkey=source
Passiflora_loefgrenii_cpgenome blatN gene 143980 144053 100.0 + 1 ID=gene-blatn_trnI-CAU_1;gbkey=gene;gene=trnI-CAU;gene_biotype=tRNA;Note=blatN_hit_trnI-CA
Passiflora_loefgrenii_cpgenome blatN tRNA 143980 144053 . + 1 ID=trna-blatn_trnI-CAU_1;gbkey=tRNA;gene=trnI-CAU
Passiflora_loefgrenii_cpgenome blatN gene 142096 142176 100.0 + 1 ID=gene-blatn_trnL-CAA_1;gbkey=gene;gene=trnL-CAA;gene_biotype=tRNA;Note=blatN_hit_trnL-CA
Passiflora_loefgrenii_cpgenome blatN tRNA 142096 142176 . + 1 ID=trna-blatn_trnL-CAA_1;gbkey=tRNA;gene=trnL-CAA
Passiflora_loefgrenii_cpgenome blatN gene 135720 135791 100.0 - 1 ID=gene-blatn_trnV-GAC_1;gbkey=gene;gene=trnV-GAC;gene_biotype=tRNA;Note=blatN_hit_trnV-GA
Passiflora_loefgrenii_cpgenome blatN tRNA 135720 135791 . - 1 ID=trna-blatn_trnV-GAC_1;gbkey=tRNA;gene=trnV-GAC
Passiflora_loefgrenii_cpgenome blatN gene 128579 128681 100.0 - 1 ID=gene-blatn_rrn4.5_1;gbkey=gene;gene=rrn4.5;gene_biotype=rRNA;Note=blatN_hit_rrn4.5_Vvin
Passiflora_loefgrenii_cpgenome blatN rRNA 128579 128681 . - 1 ID=rrna-blatn_rrn4.5_1;gbkey=rRNA;gene=rrn4.5
Passiflora_loefgrenii_cpgenome blatN gene 128579 128681 100.0 - 1 ID=gene-blatn_rrn4.5S_1;gbkey=gene;gene=rrn4.5S;gene_biotype=rRNA;Note=blatN_hit_rrn4.5S_N
Passiflora_loefgrenii_cpgenome blatN rRNA 128579 128681 . - 1 ID=rrna-blatn_rrn4.5S_1;gbkey=rRNA;gene=rrn4.5S
```

# CURADORIA MANUAL DA ANOTAÇÃO

GenomeView: <https://genomeview.org/>



**Primeiro passo:** Abrir o arquivo fasta com a sequência do genoma cloroplastidial em File/Load data



# CURADORIA MANUAL DA ANOTAÇÃO

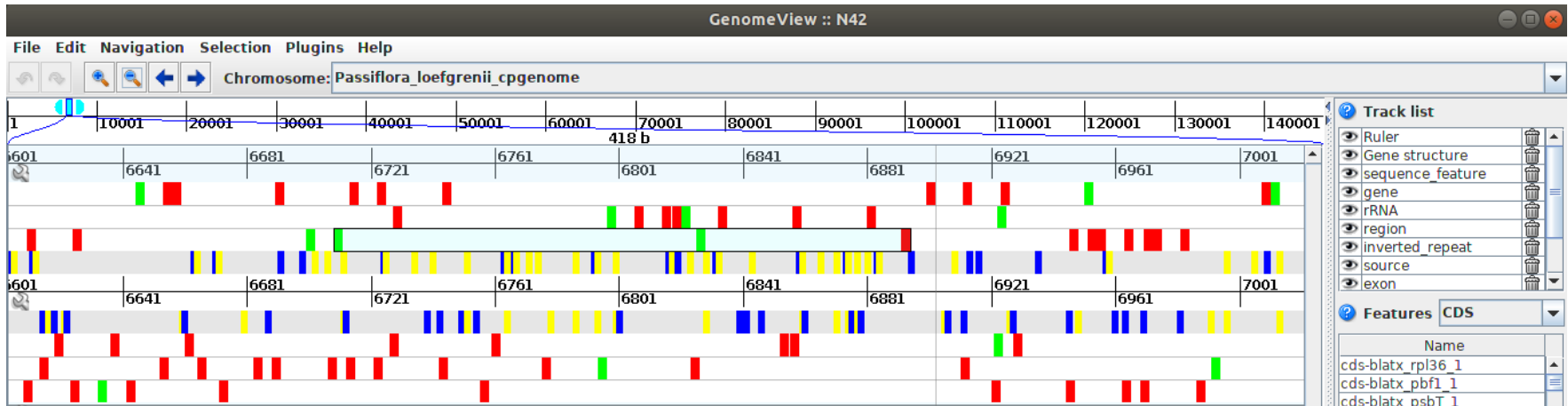
Abrir o arquivo .gff3 da anotação do genoma cloroplastidial em File / Load data

The screenshot displays the GenomeView software interface for a chloroplast genome annotation. The main window shows a genomic track for *Passiflora loefgrenii\_cpgenome* with a scale from 10001 to 140001. The track list on the right includes Ruler, Gene structure, sequence\_feature, gene, rRNA, region, inverted\_repeat, source, and exon. The Features panel shows a list of CDS features, including cds-blatax\_rpl36\_1, cds-blatax\_pbf1\_1, cds-blatax\_psbT\_1, cds-blatax\_petN\_1, cds-blatax\_psbM\_1, cds-blatax\_psbZ\_1, cds-blatax\_ndhJ\_1, and cds-blatax\_atnE\_1. The main visualization area shows a gene track with blue arrows indicating the direction of transcription, a rRNA track with green boxes, and a region track with repeat regions. The status bar at the bottom indicates the current position: 32 / 1920 (Mb).

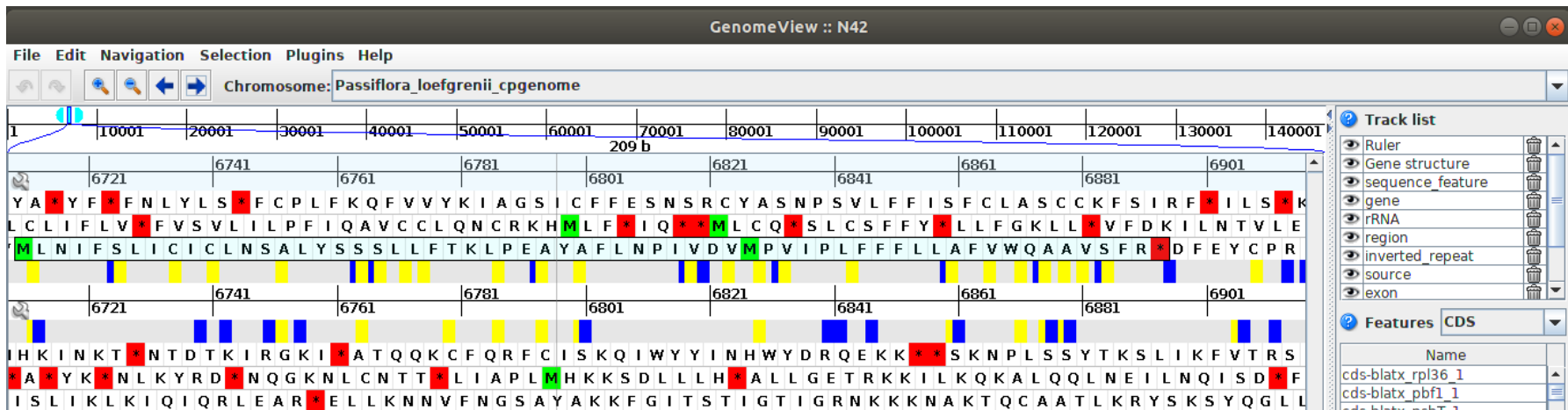
# CURADORIA MANUAL DA ANOTAÇÃO

**GenomeView:** Permite a visualização e edição das estruturas dos genes que foram preditos

Exemplo: Estrutura do gene *psbK*:



Exemplo: Sequência peptídica codificada pelo gene *psbK*:





# Corrigir a anotação de seqüências faltando *stop* códons:

GenomeView :: N42

File Edit Navigation Selection Plugins Help

Chromosome: Passiflora\_loefgrenii\_cpgenome

1 10001 20001 30001 40001 50001 60001 70001 80001 90001 100001 110001 120001 130001 140001

97 b

80272 80281 80290 80299 80308 80317 80326 80335 80344 80353 80362

G G \* R \* K \* I \* I Q R L \* T G F L S K S I A K C F S I S R M L N

G M K M K M N I N P A A L N W I F E \* I N C E M L F H F \* N A \*

G D E D E N E Y K S S G F K L D F \* V N Q L R N A F P F L E C L

GGGGATGAGATGAAAATGAATATAAATCCAGCGGCTTTAAACTGGATTTTGTAAATCAATTGCGAAATGCTTTTCCATTTCTAGAATGCTTAA

80272 80281 80290 80299 80308 80317 80326 80335 80344 80353 80362

CCCCTACTTCTACTTTTACTTATATTTAGGTGCGCGAAATTTGACCTAAAAACTCATTTAGTTAAAGCTTTTACGA AAAGGTAAGGATCTTACGAATT

P I F I F I F I F G A A K F Q I K S Y I L Q S I S K W K \* F A \*

P H L H F H I Y I W R S \* V P N K L L D I A F H K E M E L I S L

P S S S S F S Y L D L P K L S S K Q T F \* N R F A K G N R S H K I

sequence\_feature

gene

gene-blax\_rpsA\_1

Track list

- Ruler
- Gene structure
- sequence\_feature
- gene
- rRNA
- region
- inverted\_repeat
- source
- exon

Features CDS

Name
cds-blax_rps3_1
cds-blax_ndh1_1
cds-blax_petD_1
cds-blax_rps7_1
cds-blax_rps7_2
cds-blax_matK_1
cds-blax_atrA_1

Exemplo: Completar a seqüência adicionando na predição as bases que faltavam para o stop códon:

Atividades net-sf-genomeview-gui-GenomeView

ter, 02:21

GenomeView :: N42

File Edit Navigation Selection Plugins Help

Chromosome: Passiflora\_loefgrenii\_cpgenome

1 10001 20001 30001 40001 50001 60001 70001 80001 90001 100001 110001 120001 130001 140001

97 b

80272 80281 80290 80299 80308 80317 80326 80335 80344 80353 80362

G G \* R \* K \* I \* I Q R L \* T G F L S K S I A K C F S I S R M L N

G M K M K M N I N P A A L N W I F E \* I N C E M L F H F \* N A \*

G D E D E N E Y K S S G F K L D F \* V N Q L R N A F P F L E C L

GGGGATGAGATGAAAATGAATATAAATCCAGCGGCTTTAAACTGGATTTTGTAAATCAATTGCGAAATGCTTTTCCATTTCTAGAATGCTTAA

80272 80281 80290 80299 80308 80317 80326 80335 80344 80353 80362

CCCCTACTTCTACTTTTACTTATATTTAGGTGCGCGAAATTTGACCTAAAAACTCATTTAGTTAAAGCTTTTACGA AAAGGTAAGGATCTTACGAATT

P I F I F I F I F G A A K F Q I K S Y I L Q S I S K W K \* F A \*

P H L H F H I Y I W R S \* V P N K L L D I A F H K E M E L I S L

P S S S S F S Y L D L P K L S S K Q T F \* N R F A K G N R S H K I

sequence\_feature

Track list

- Ruler
- Gene structure
- sequence\_feature
- gene
- rRNA
- region
- inverted\_repeat
- source
- exon

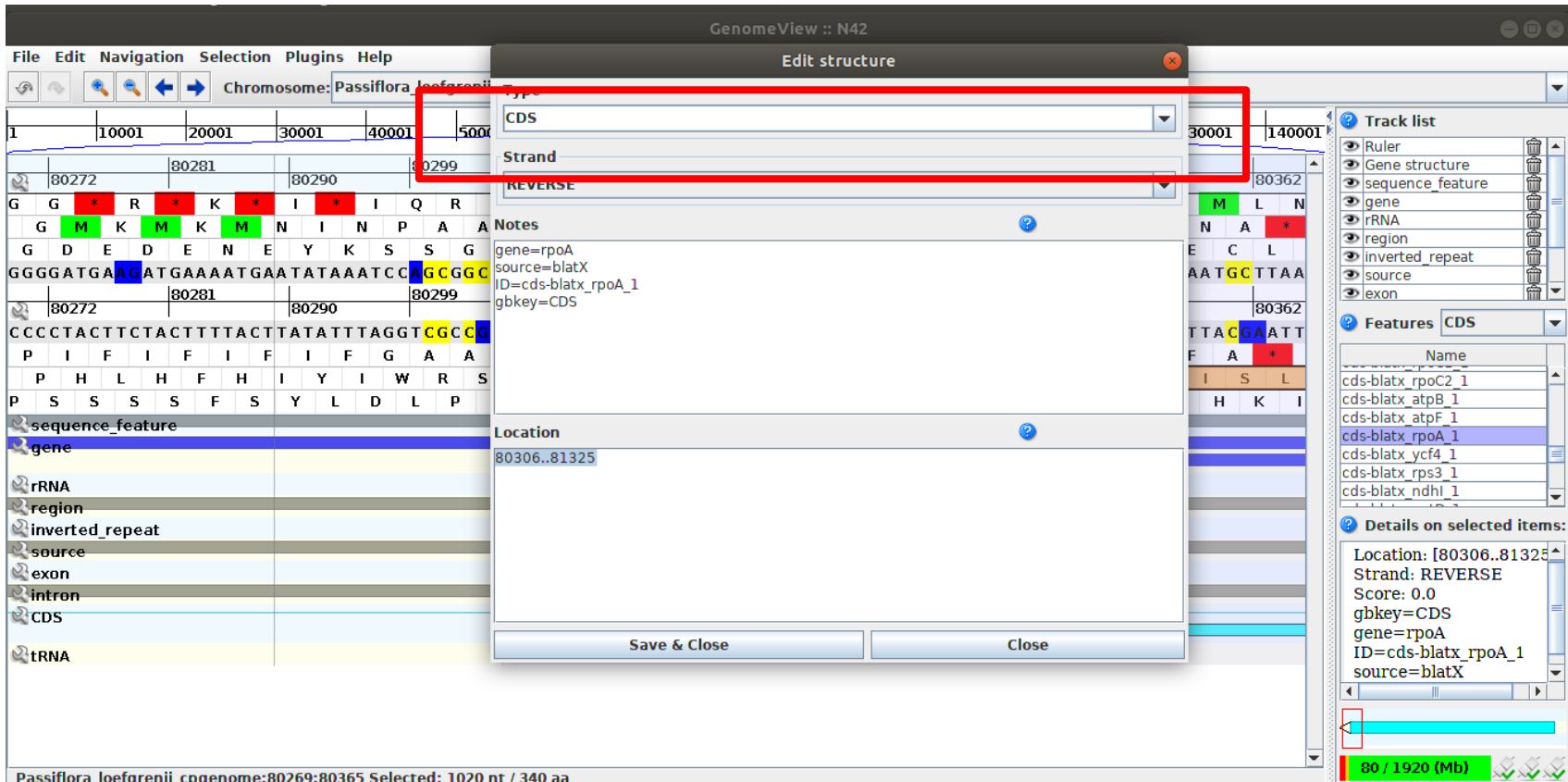
Features CDS

Name
cds-blax_rps3_1
cds-blax_ndh1_1
cds-blax_petD_1
cds-blax_rps7_1

Para corrigir a anotação da estrutura de um gene e da CDS: Clique na estrutura do gene, e com a estrutura selecionada clique em Edit na barra de ferramentas. Agora procure pela opção Edit structure

Na caixa Edit structure alterar as coordenadas das posições do gene na opção Location

Adicionar as novas coordenadas que contempla a estrutura completa com o stop códon:



**Obs: A primeira correção alterou apenas as posições da CDS**

**Para corrigir a anotação da estrutura do gene: Clique na linha azul marinho correspondente ao gene, depois clique em Edit na barra de ferramentas e procure por Edit structure**

GenomeView :: N42

File Edit Navigation Selection Plugins Help

Chromosome: Passiflora\_loefgrenii\_cpgenome

1 10001 20001 30001 40001 50001 60001 70001 80001 90001 100001 110001 120001 130001 140001

97 b

80272 80281 80290 80299 80308 80317 80326 80335 80344 80353 80362

G G \* R \* K \* I \* I Q R L \* T G F L S K S I A K C F S I S R M L N

G M K M K M N I N P A A L N W I F E \* I N C E M L F H F \* N A \*

G D E D E N E Y K S S G F K L D F \* V N Q L R N A F P F L E C L

GGGGATGAAGATGAAAATGAATATAAATCCAGCGGC TTTAAACTGGATTTTTGAGTAAATCAATTGCCGAAATGCTTTTCATTTCTAGAATGCTTAA

80272 80281 80290 80299 80308 80317 80326 80335 80344 80353 80362

CCCTACTTCTACTTTTACTTATATTTAGGTGCGCCGAAATTTGACCTAAAAACTCATTTAGTTAAACGCTTTACGAAAAGGTAAAAGATCTTACGAAATT

P I F I F I F I F G A A K F Q I K S Y I L Q S I S K W K \* F A \*

P H L H F H I Y I W R S \* V P N K L L D I A F H K E M E L I S L

P S S S S F S Y L D L P K L S S K Q T F \* N R F A K G N R S H K I

sequence\_feature

gene

rRNA

region

inverted\_repeat

source

exon

intron

CDS

tRNA

Name : gene-blatx\_rpoA\_1  
Start : 80308  
End : 81325

Track list

- Ruler
- Gene structure
- sequence\_feature
- gene
- rRNA
- region
- inverted\_repeat
- source
- exon

Features CDS

Name

- cds-blatx\_rpoC2\_1
- cds-blatx\_atpB\_1
- cds-blatx\_atpF\_1
- cds-blatx\_rpoA\_1
- cds-blatx\_ycf4\_1
- cds-blatx\_rps3\_1
- cds-blatx\_ndhI\_1

Details on selected items:

gbkey= gene  
gene=rpoA  
gene\_biotype=protein\_coding  
ID=gene-blatx\_rpoA\_1  
Note=blatX\_hit\_rpoA\_N  
score=99.0  
source=blatX

Passiflora\_loefgrenii\_cpgenome:80269:80365 (80333) Selected: 1018 nt / 339 aa

58 / 1920 (Mb)

Na caixa Edit structure altere as coordenadas das posições do gene na opção Location

Adicionar as novas coordenadas que contempla a estrutura completa com o stop códon:

The screenshot displays the GenomeView application interface. The main window shows a genomic track for *Passiflora loefgrenii* with a selected region from 80269 to 80365. The 'Edit structure' dialog box is open, showing the following details:

- Type:** gene
- Strand:** REVERSE
- Location:** 80306..81325

The dialog also includes a 'Notes' section with the following text:

```
score=99.0
gene=rpoA
Note=blatX_hit_rpoA_NC_038118.1%2C_position_1_-_1018%2C_psl_score_99.0%2C_coverage_99.0
source=blatX
ID=gene-blatx_rpoA_1
gene_biotype=protein_coding
gbkey=gene
```

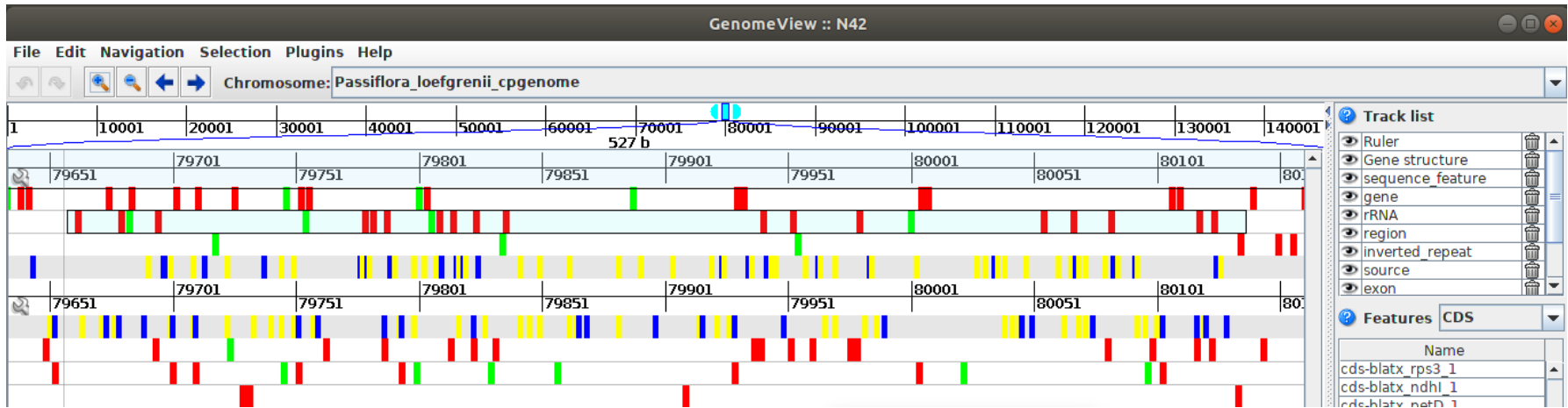
The background interface shows a sequence alignment with various features like exons and introns, and a track list on the right side. The track list includes items like Ruler, Gene structure, sequence feature, gene, rRNA, region, inverted\_repeat, source, and exon. The 'Features' section is set to 'CDS' and shows a list of CDS features such as cds-blatx\_rpoC2\_1, cds-blatx\_atpB\_1, cds-blatx\_atpF\_1, cds-blatx\_rpoA\_1, cds-blatx\_ycf4\_1, cds-blatx\_rps3\_1, and cds-blatx\_ndh1\_1. The 'Details on selected items' section shows the following information:

```
gbkey=gene
gene=rpoA
gene_biotype=protein_coding
ID=gene-blatx_rpoA_1
Note=blatX_hit_rpoA_NC_038118.1%2C_position_1_-_1018%2C_psl_score_99.0%2C_coverage_99.0
source=blatX
```

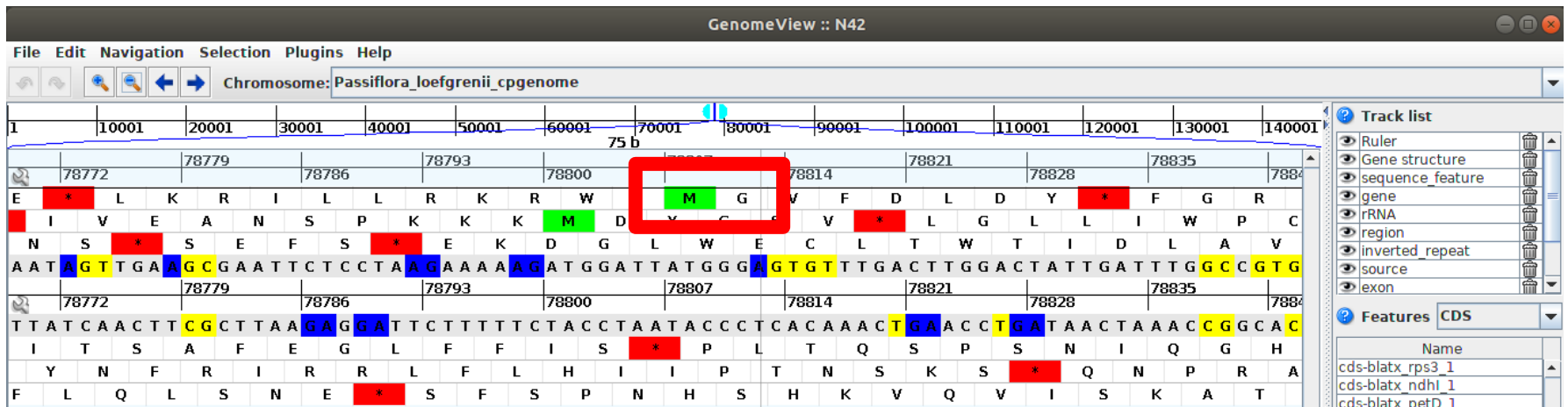
The status bar at the bottom indicates the selected region: Passiflora\_loefgrenii\_cpgenome:80269:80365 Selected: 1018 nt / 339 aa. The bottom right corner shows a scale bar for 65 / 1920 (Mb).

**Corrigir a anotação de sequências contendo *stop códons* internos:** Observar os outros quadros leitura para esta sequência, e procurar por aquele que não apresenta stop códons

Exemplo: Estrutura predita do gene *petD* contendo stop códons:



Exemplo: Correção da anotação do gene *petD*:



**Para corrigir a anotação da estrutura de um gene:** Clicar sobre a estrutura do gene, clicar na opção Edit na barra de ferramentas, e escolher a função Edit structure

The screenshot displays the GenomeView application window titled "GenomeView :: N42". The main interface shows a genomic track for *Passiflora\_loefgrenii* with a selected region from 79780 to 79801. The "Edit structure" dialog box is open, showing the following details:

- Type:** CDS
- Strand:** FORWARD
- Notes:** gene=petD, source=blatX, ID=cds-blatx\_petD\_1, gbkey=CDS
- Location:** 79658..80136

The dialog box has "Save & Close" and "Close" buttons. The background interface includes a menu bar (File, Edit, Navigation, Selection, Plugins, Help), a track list on the right, and a details panel for the selected CDS feature. The details panel shows the location [79658..80136], strand (FORWARD), score (0.0), and gene (petD). The status bar at the bottom indicates "Passiflora\_loefgrenii\_cpgenome:79777:79851 Selected: 479 nt / 159 aa".

# Para corrigir a anotação da estrutura de um gene: Na caixa Edit structure altere as coordenadas das posições do gene na opção Location

Exemplo: A localização do novo éxon para o gene *petD* foi incluída na opção Location:

The screenshot displays the GenomeView software interface. The main window shows a genomic track for *Passiflora\_loefgrenii* with a sequence viewer. The sequence viewer shows the following DNA sequence: `T Y P N N K K T * P` (top line) and `L S Q * Q K N L T` (second line). Below this, the protein sequence is shown: `P I P I T K K P D` (top line) and `R D W Y C F F R V` (second line). A red asterisk is present in the protein sequence at the position corresponding to the asterisk in the DNA sequence. The 'Edit structure' dialog box is open, showing the following fields:

- Type: CDS
- Strand: FORWARD
- Notes: gene=petD, source=blatX, ID=cds-blatx\_petD\_1, gbkey=CDS
- Location: 78807..78812,79658..80136

The dialog box has 'Save & Close' and 'Close' buttons. The background track list on the right shows the following features:

- Ruler
- Gene structure
- sequence feature
- gene
- rRNA
- region
- inverted\_repeat
- source
- exon

The 'Features' section shows the following items:

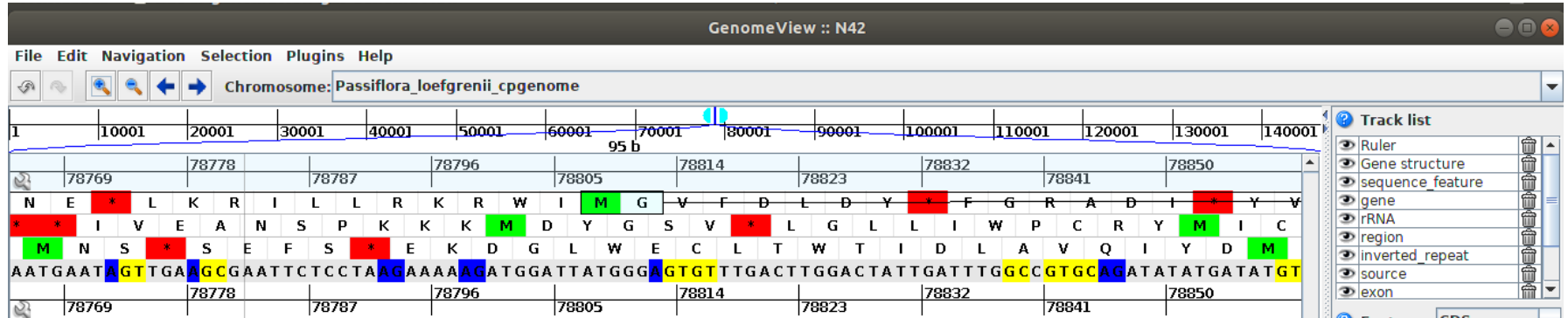
- cds-blatx\_rps3\_1
- cds-blatx\_ndh1\_1
- cds-blatx\_petD\_1
- cds-blatx\_rps7\_1
- cds-blatx\_rps7\_2
- cds-blatx\_matK\_1
- cds-blatx\_atpA\_1
- cds-blatx\_rns11\_1

The 'Details on selected items:' section shows the following information:

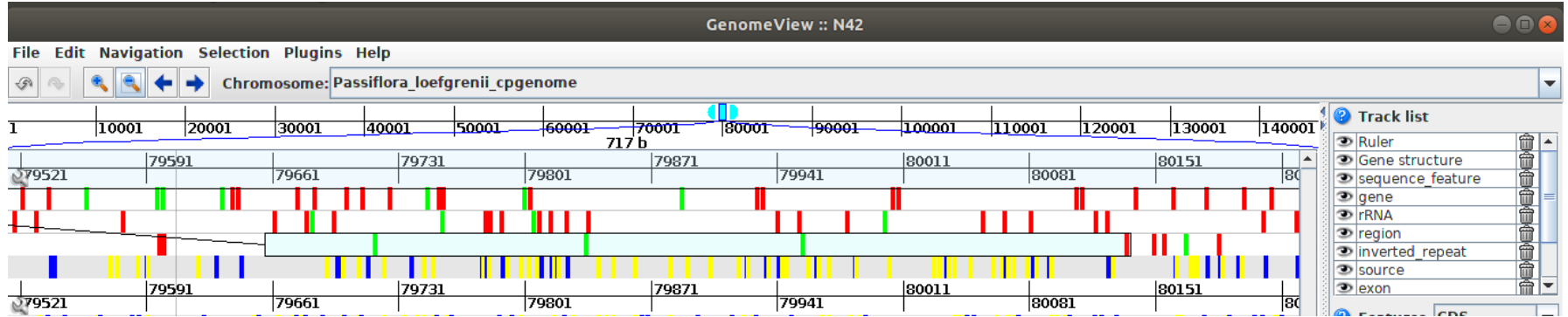
- Location: [78807..78812
- Strand: FORWARD
- Score: 0.0
- gbkey=CDS
- gene=petD
- ID=cds-blatx\_petD\_1
- source=blatX

The status bar at the bottom shows: `Passiflora_loefgrenii_cpgenome:79651:79725 Selected: 6 nt / 2 aa`

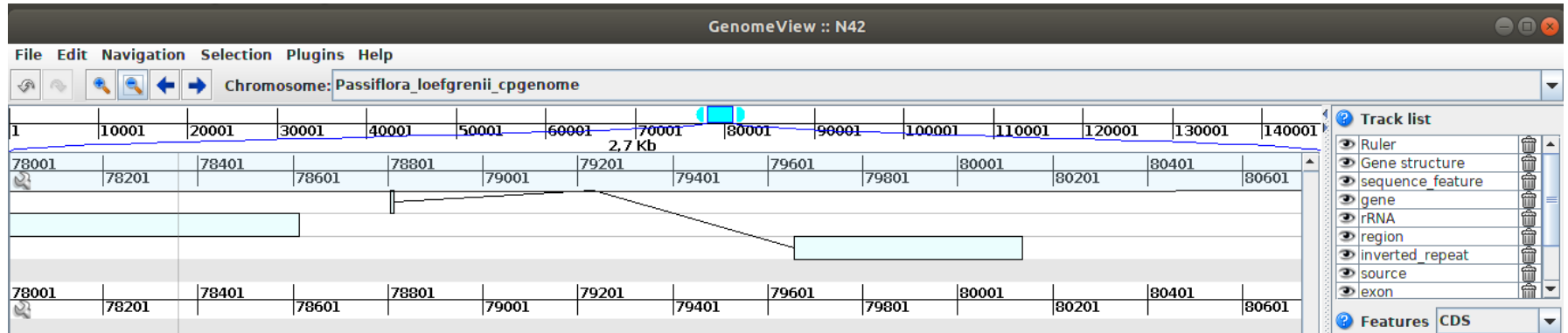
## Exemplo: Primeiro éxon do gene *petD*:



## Exemplo: Segundo éxon do gene *petD*:



## Exemplo: Estrutura completa do gene *petD*:





# Conferiu a anotação de todos os genes?

Para salvar o novo arquivo de anotação: Clique em File, e depois em Save annotation\_

GenomeView :: N42

File Edit Navigation Selection Plugins Help

Chromosome: Passiflora\_loefgrenii\_cpgenome

1 10001 20001 30001 40001

80272 80281 80290

G G \* R \* K \* I \* I

G M K M K M N I N

G D E D E N E Y K

GGGGATGAAATGAAATGAAATATAAAT

80272 80281 80290

CCCCTACTTCTACTTTTACTTATATTTA

P I F I F I F I F

P H L H F H I Y I

P S S S S F S Y L D

sequence\_feature

gene

rRNA

region

inverted\_repeat

source

exon

intron

CDS

tRNA

Save Dialog

Location to save to  
ome/luiz/Documents/Dados\_DocLuiz/Assembly/P\_loefgrenii/Ploefgrenii\_Annot\_manual Browse...

File format options  
GFF3Parser Include sequence

Select entries to save  
 Passiflora\_loefgrenii\_cpgenome  
Select all entries  
Deselect all entries

Annotation types to save  
 CDS  
 GeSeqJob-20200602-13409\_Passiflora\_loefgrenii\_cpgenome\_OGDRAW.jpg  
 region  
 source  
 gene  
 tRNA  
 rRNA  
 exon  
 intron  
 inverted\_repeat  
 sequence\_feature  
Select all types  
Deselect all types

Save Cancel

Track list

- Ruler
- Gene structure
- sequence\_feature
- gene
- rRNA
- region
- inverted\_repeat
- source
- exon

Features CDS

Name
cds-blatx_rpoC2_1
cds-blatx_atpB_1
cds-blatx_atpF_1
cds-blatx_rpoA_1
cds-blatx_ycf4_1
cds-blatx_rps3_1
cds-blatx_ndhI_1

Details on selected items:

Location: 10000..01000  
Strand: REVERSE  
Score: 0.0  
gbkey=CDS  
gene=rpoA  
ID=cds-blatx\_rpoA\_1  
source=blatX

Passiflora\_loefgrenii\_cpgenome:80269:80365 Selected: 1020 nt / 340 aa

62 / 1920 (Mb)

# GENÔMICA COMPARATIVA

## MAUVE

<http://darlinglab.org/mauve/mauve.html>

the Darling lab | computational (meta)genomics

[Blog](#) [Mauve](#) [Openings](#) [People](#) [Projects](#) [Publications](#) [Reviews](#) [Software](#) [Tutorials](#)

### [Download Mauve](#)

#### User Guide:

- [Introduction](#)
- [Installing](#)
- [Aligning genomes](#)
- [The viewer](#)
- [File formats](#)
- [Reordering contigs](#)
- [mauveAligner](#)
- [progressiveMauve](#)
- [Version history](#)
- [Screenshots](#)

#### Developer Guide:

- [Overview](#)
- [Building from source](#)
- [Developing the GUI](#)
- [Windows builds](#)
- [Mac builds](#)
- [Deploying Mauve](#)
- [Benchmarking](#)



# mauve

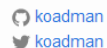
## Multiple Genome Alignment

Mauve is a system for constructing multiple genome alignments in the presence of large-scale evolutionary events such as rearrangement and inversion. Multiple genome alignments provide a basis for research into comparative genomics and the study of genome-wide evolutionary dynamics.

Mauve has been developed with the idea that a multiple genome aligner should require only modest computational resources. It employs algorithmic techniques that scale well in the lengths of sequences being aligned. For example, a pair of *Y. pestis* genomes can be aligned in under a minute, while a group of 9 divergent Enterobacterial genomes can be aligned in a few hours. However, the current algorithm's compute time (progressiveMauve) scales cubically in the number of genomes to align, making it unsuitable for datasets containing more than 50-100 bacterial genomes.

Mauve development began at the University of Wisconsin-Madison with a team including Aaron Darling, Bob Mau, and Nicole Perna. Several others have contributed development to aspects of the Mauve software in the time since.

the Darling lab | computational  
(meta)genomics  
[aaron.darling@uts.edu.au](mailto:aaron.darling@uts.edu.au)

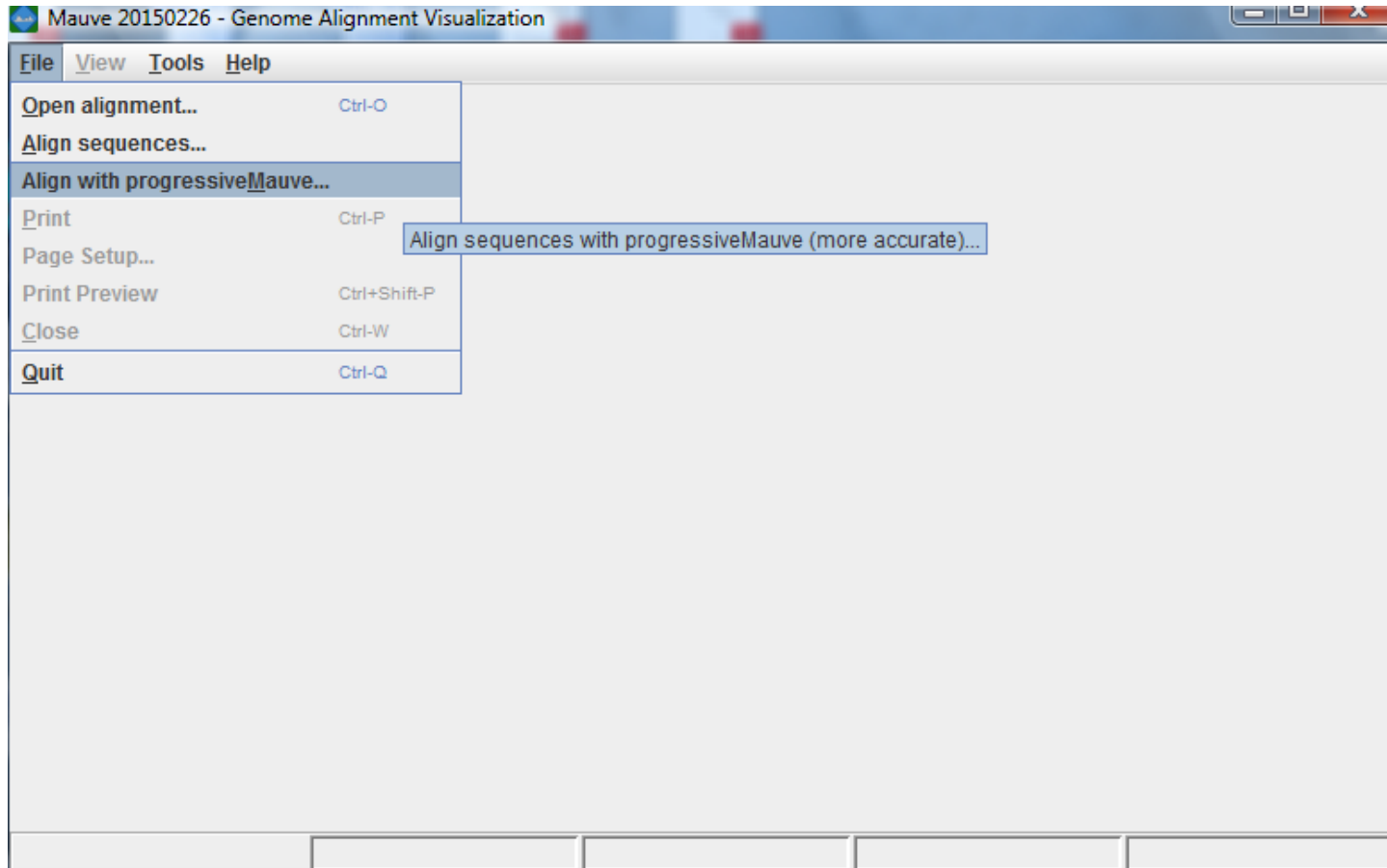


The Darling lab at the University of Technology  
Sydney. We develop computational and molecular  
techniques to characterize the hidden world of

# MAUVE

## Alinhamento das seqüências com a função progressiveMauve

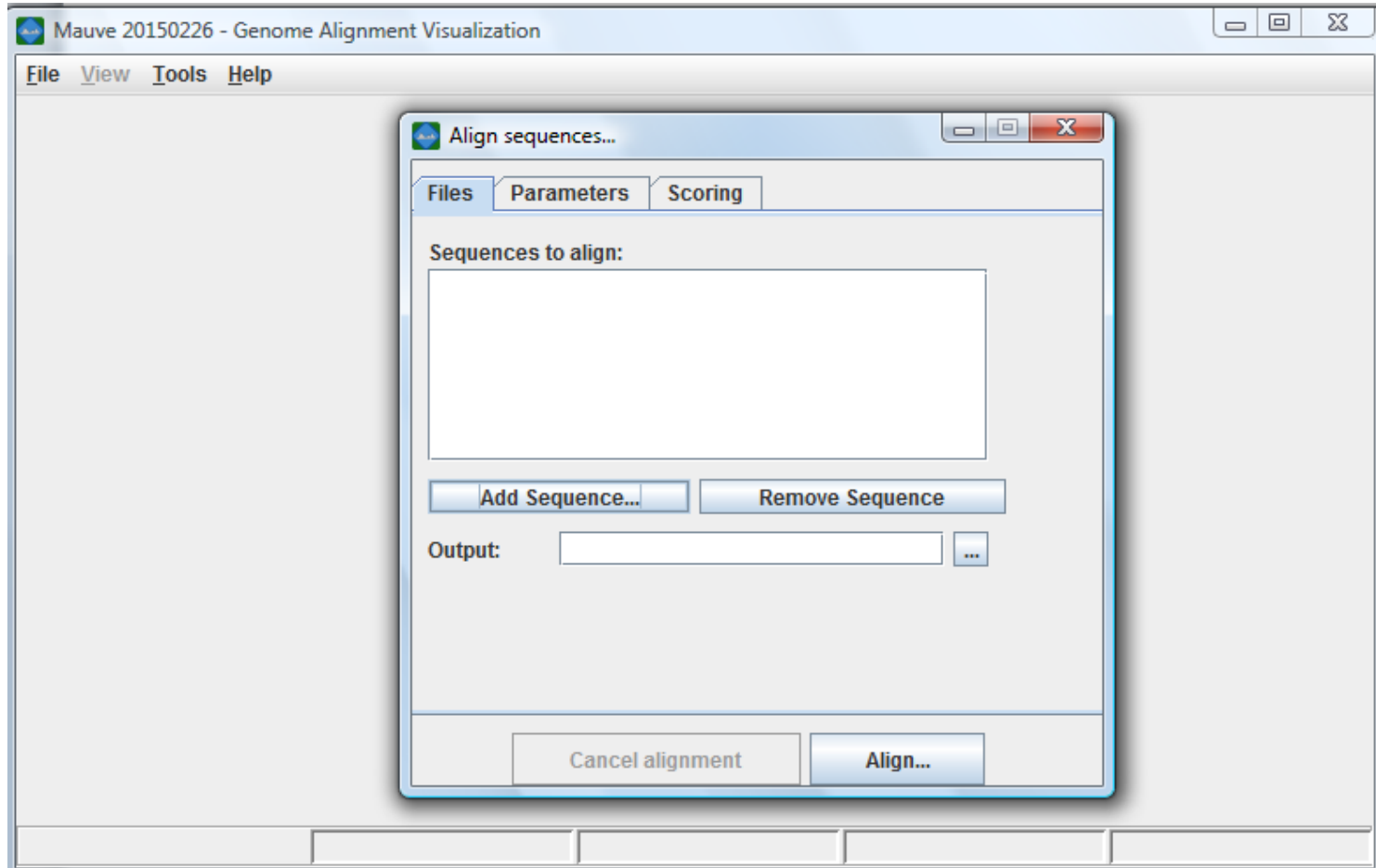
- Abrir o programa
- Clicar em File e escolher Align with progressiveMauve



# MAUVE

## Alinhamento das seqüências com a função progressiveMauve

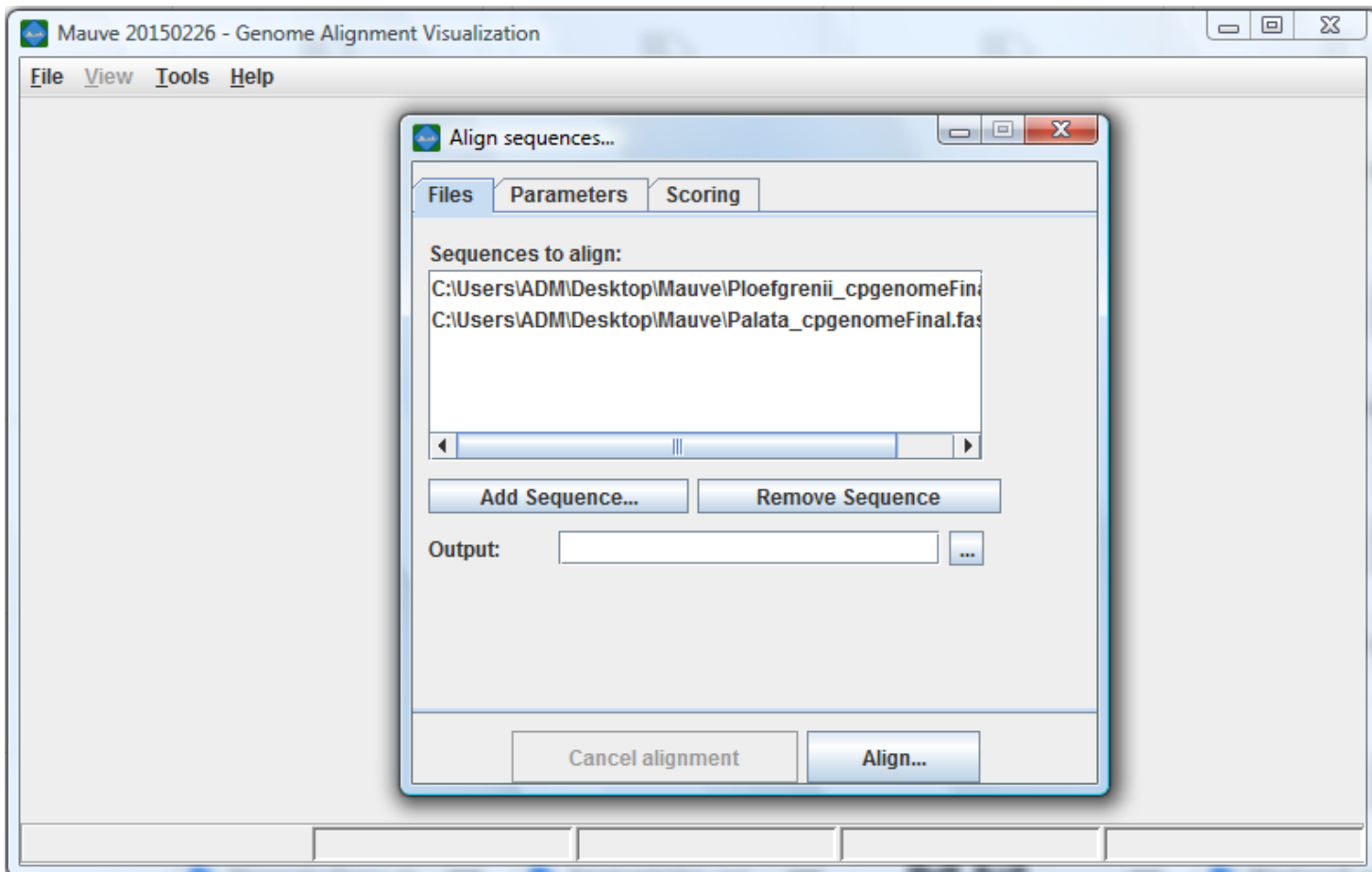
- Adicionar as seqüências .fasta em Add Sequences



# MAUVE

## Alinhamento das sequências com a função progressiveMauve

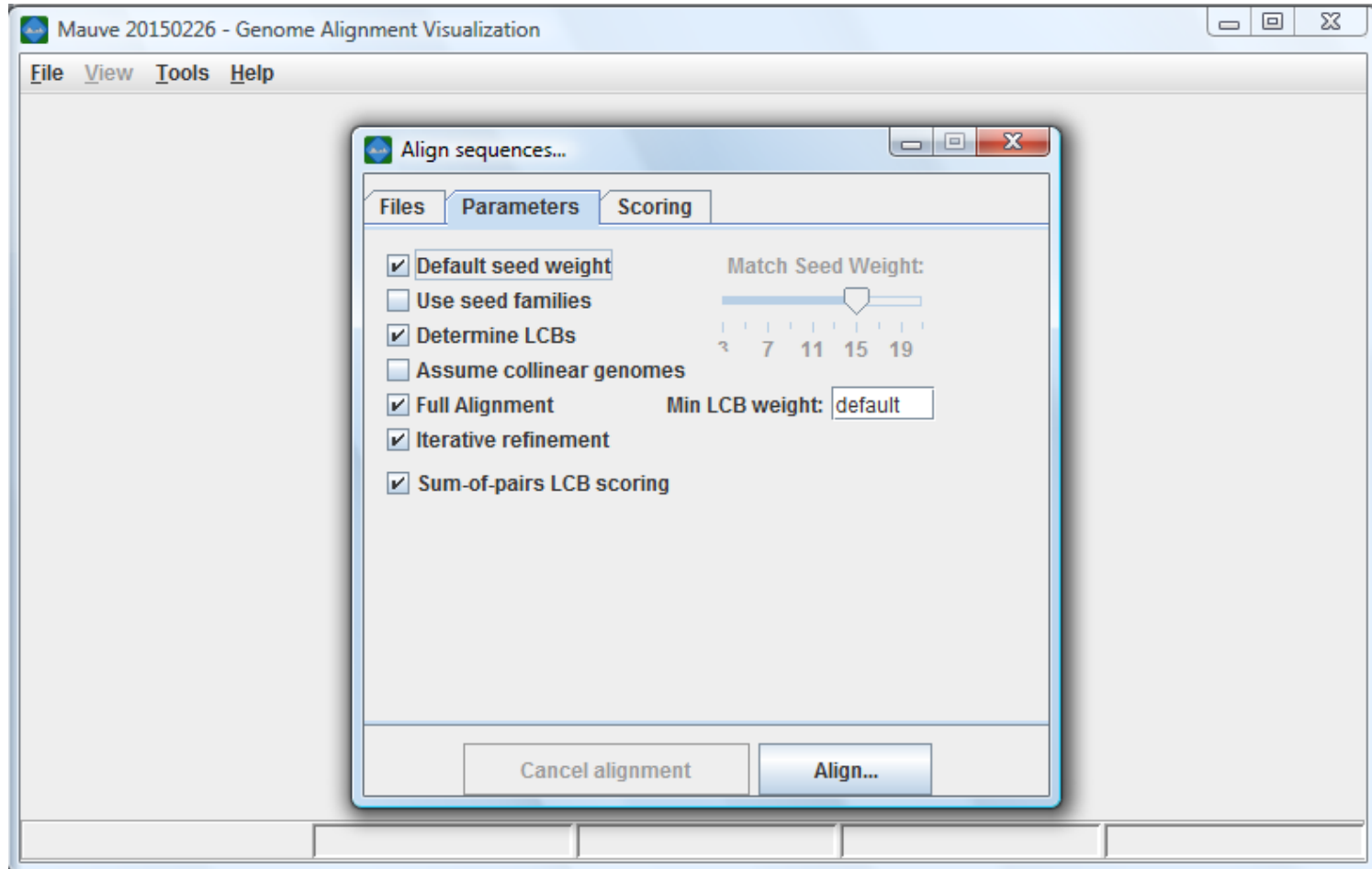
- Adicionar as sequências .fasta em Add Sequences



# MAUVE

## Alinhamento das seqüências com a função progressiveMauve

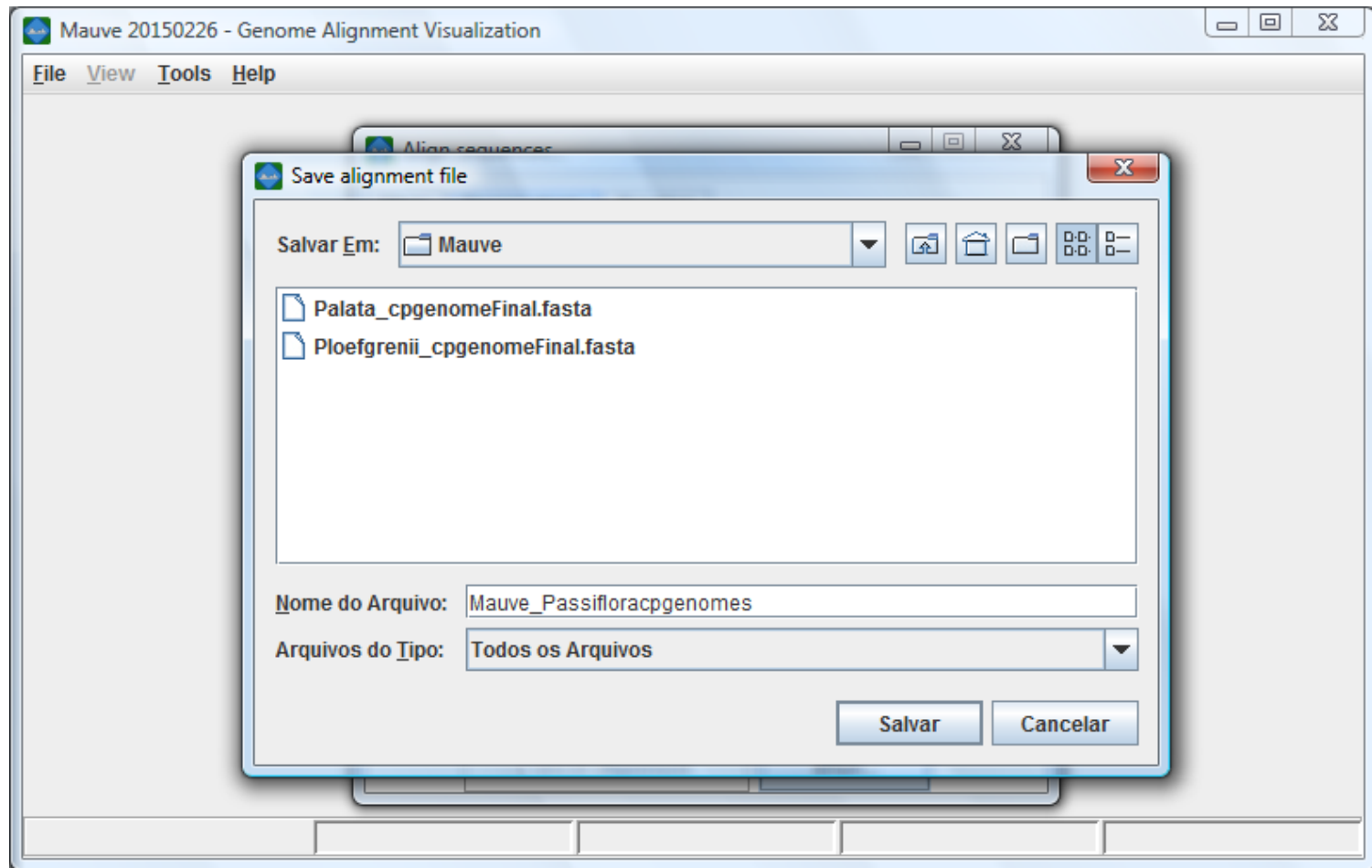
- Definir os parâmetros de alinhamento em Parameters



# MAUVE

## Alinhamento das seqüências com a função progressiveMauve

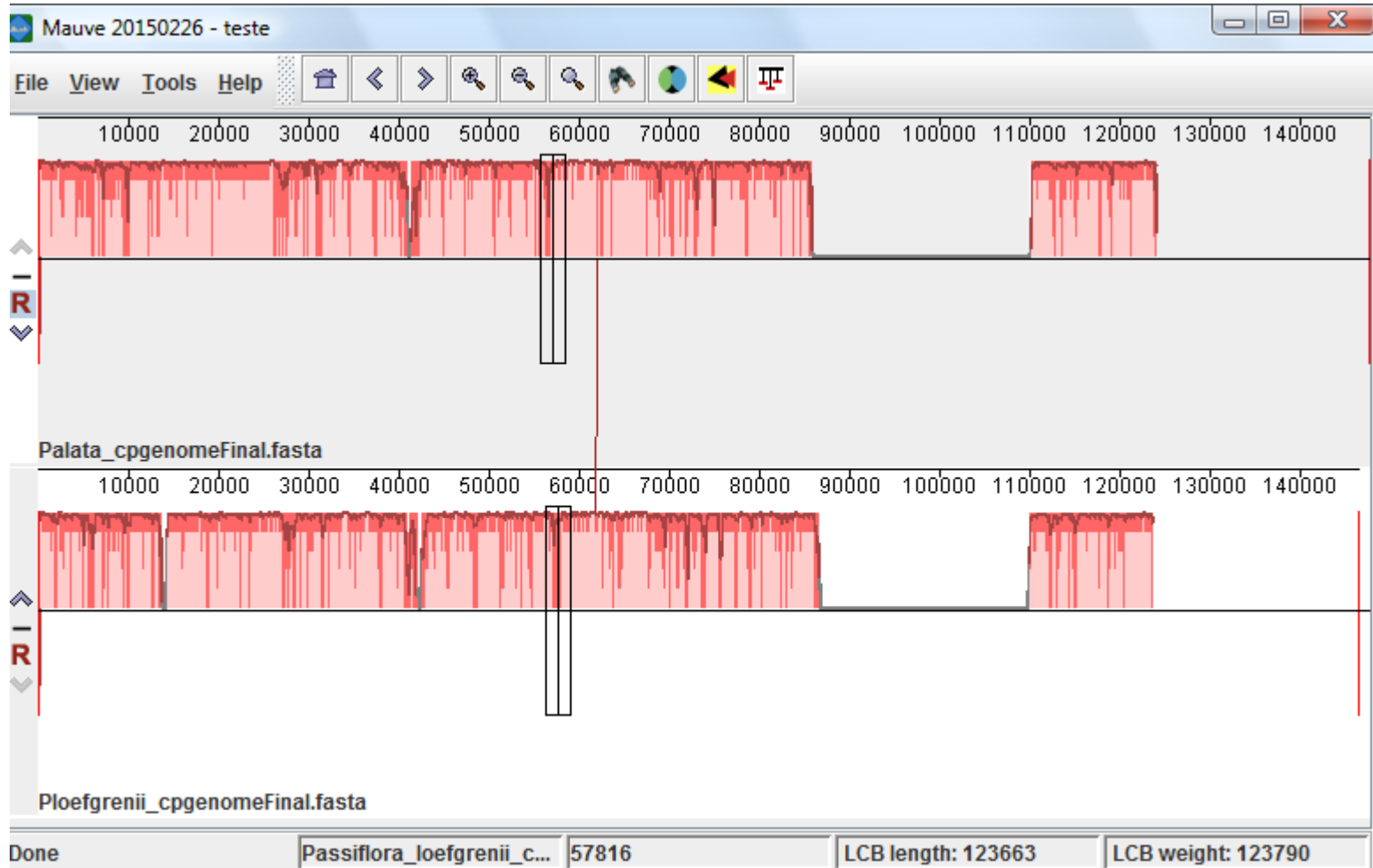
- Escolher uma pasta para salvar os arquivos de saída do programa.



# MAUVE

## Resultados

- Sintenia entre as seqüências analisadas.

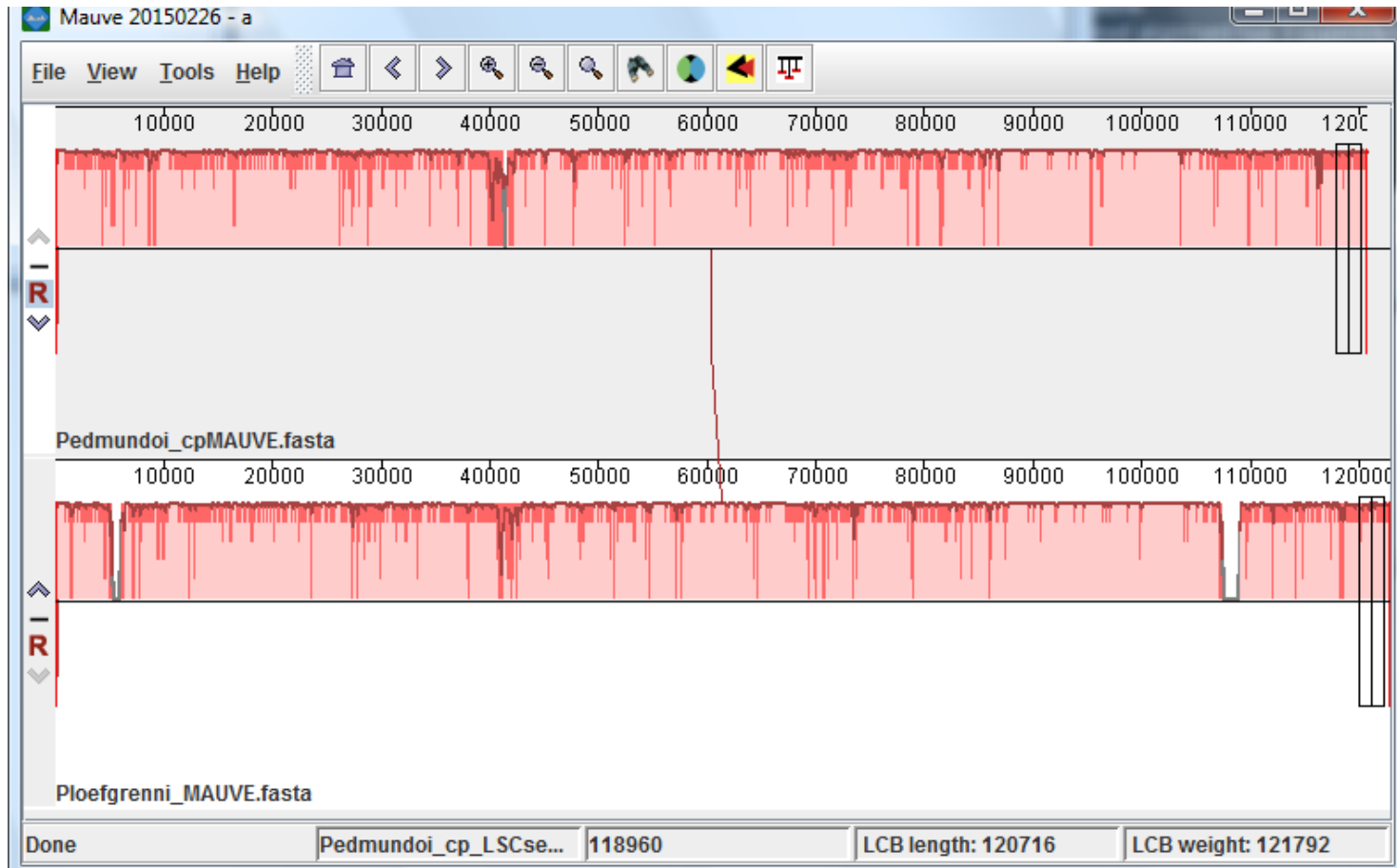




# MAUVE

## Resultados

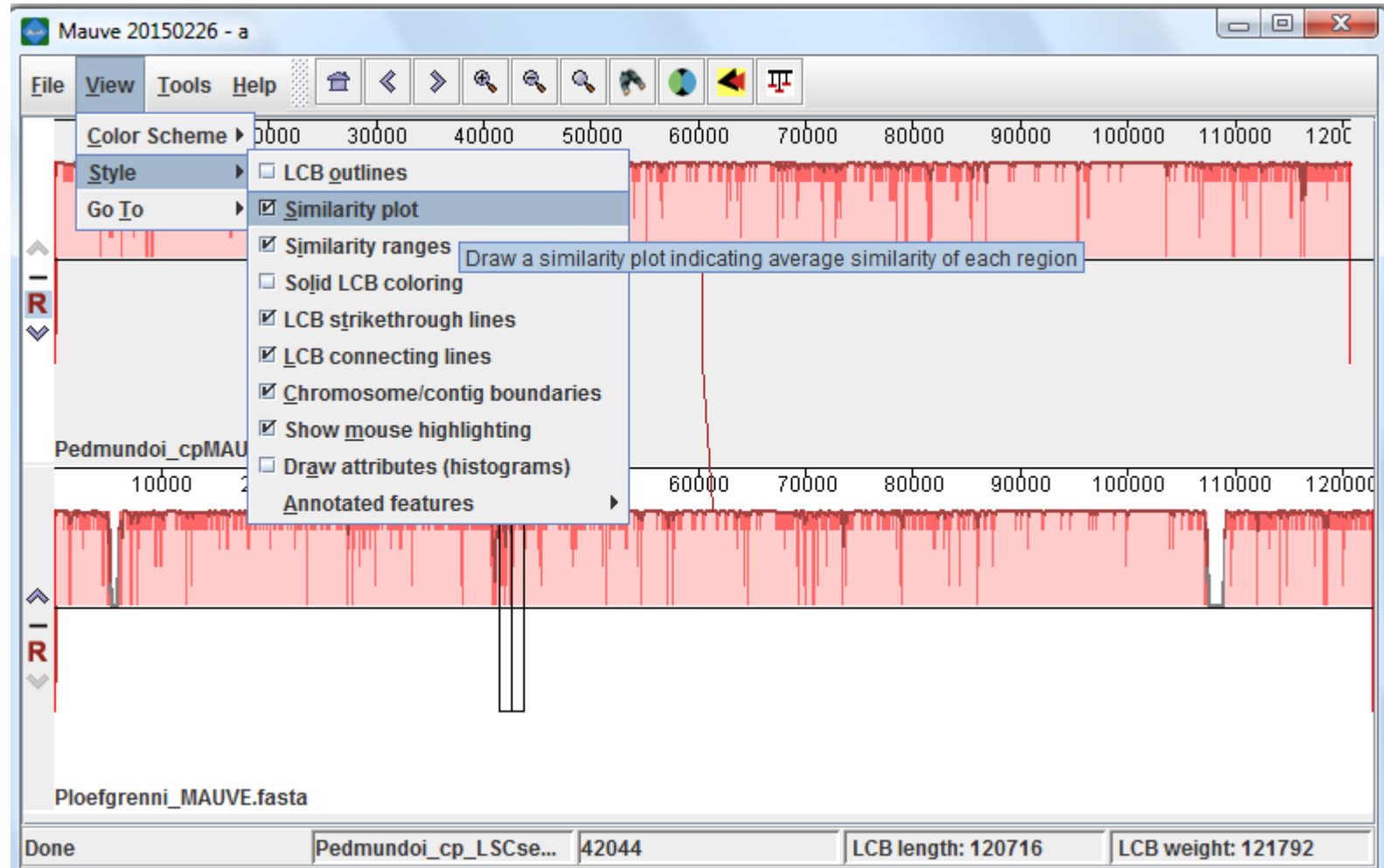
- Utilizando apenas uma das IRs.



# MAUVE

## Resultados

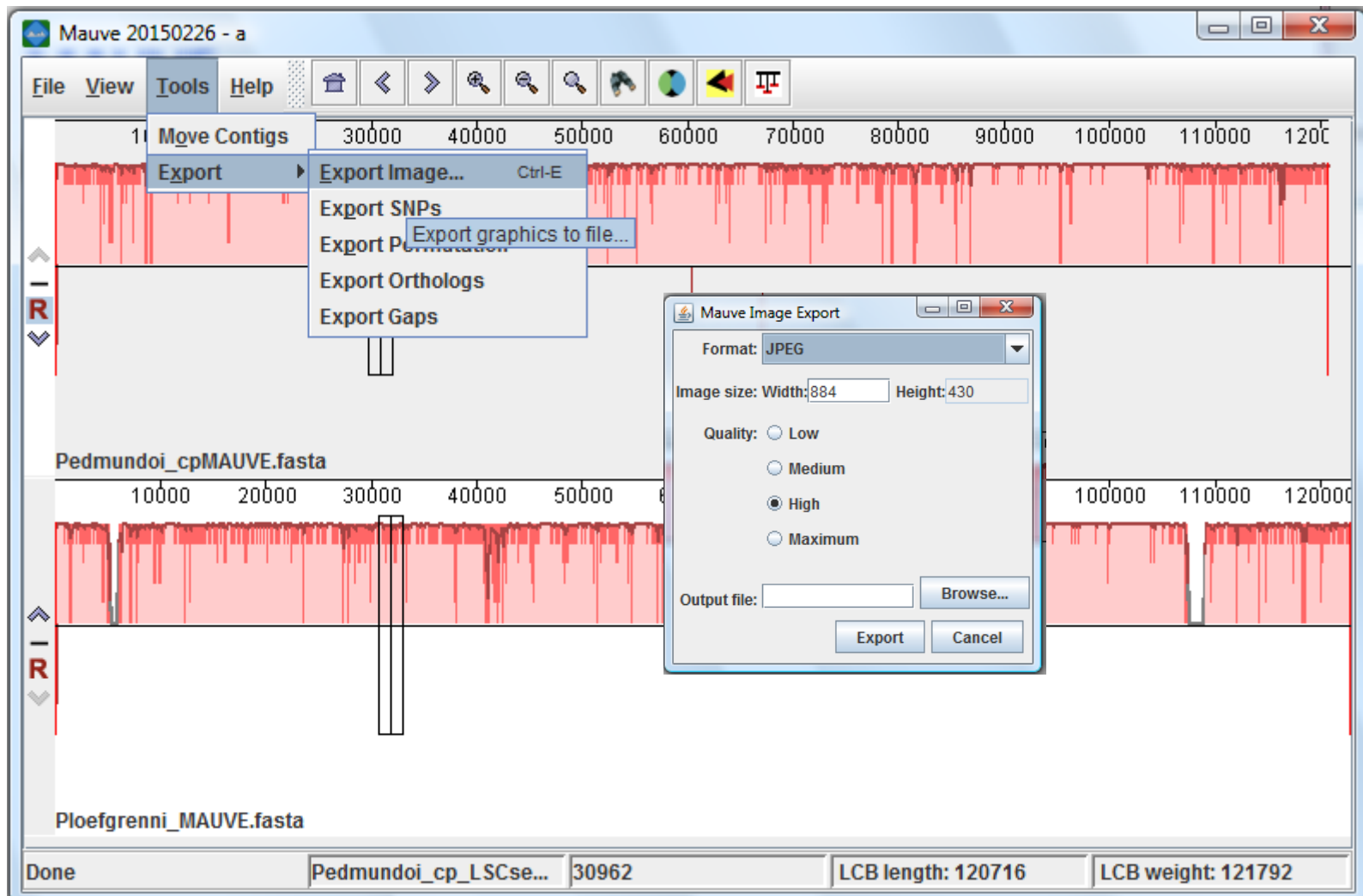
- Fazer alterações na imagem do alinhamento clicando em View e depois em Style.



# MAUVE

## Resultados

- Para salvar os arquivos de saída clicar em Tools e depois em Export.



# MAUVE

## Resultados

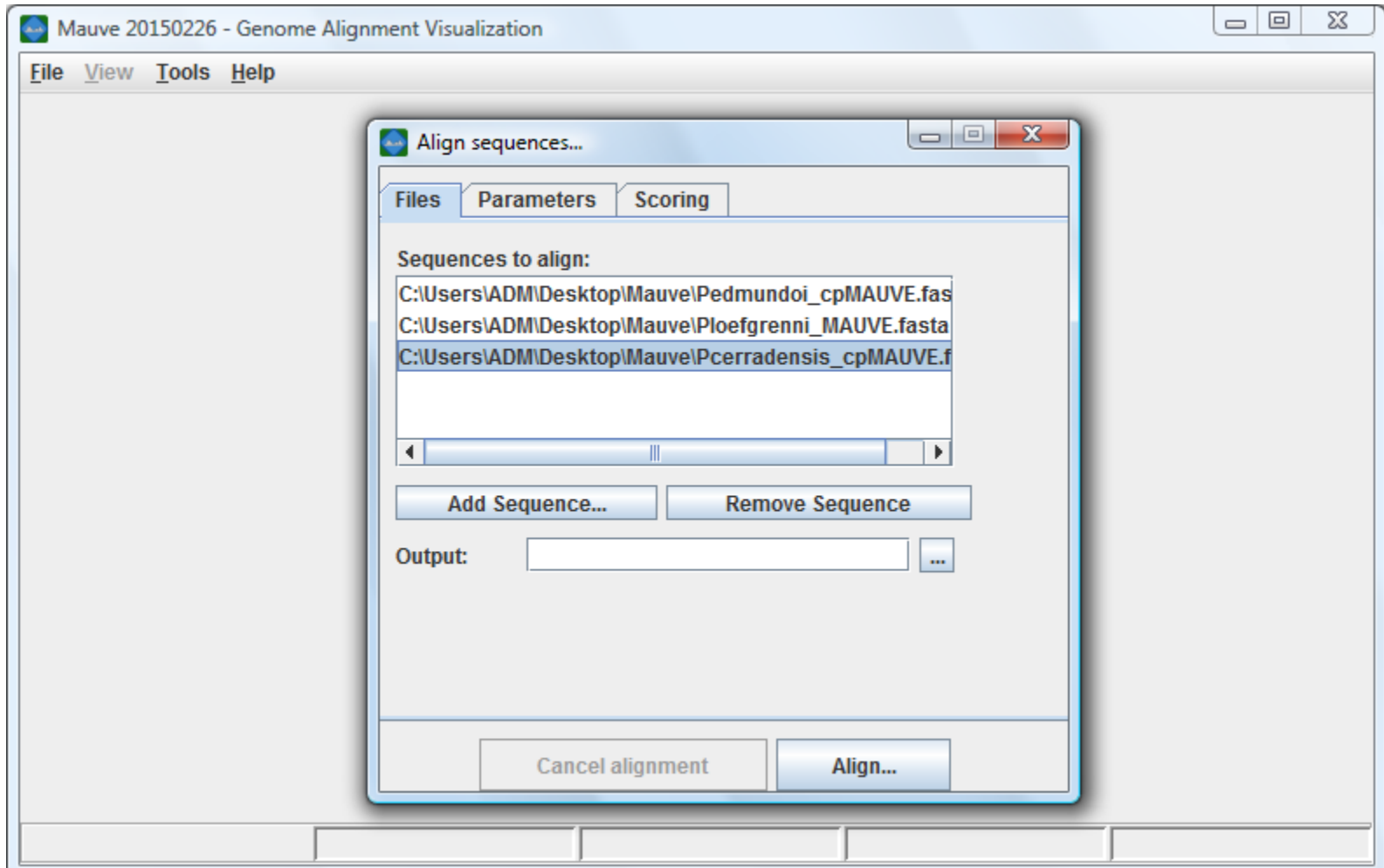
Nome	Modificado em	Tipo	Tamanho
Mauve_Passifloracpgenomes	11/07/2020 17:24	Arquivo	307 KB
Mauve_Passifloracpgenomes.backbone	11/07/2020 17:24	Arquivo BACKBONE	3 KB
Mauve_Passifloracpgenomes.bbcols	11/07/2020 17:24	Arquivo BBCOLS	2 KB
Mauve_Passifloracpgenomes.guide_tree	11/07/2020 17:24	Arquivo GUIDE_TR...	1 KB
Palata_cpgenomeFinal	11/07/2020 17:15	Arquivo FASTA	145 KB
Palata_cpgenomeFinal.fasta.sslis	11/07/2020 17:24	Arquivo SSLIST	616 KB
Pcerradensis_cpgenomeFinal	11/07/2020 17:45	Arquivo FASTA	161 KB
Ploefgrenii_cpgenomeFinal	11/07/2020 17:15	Arquivo FASTA	144 KB
Ploefgrenii_cpgenomeFinal.fasta.sslis	11/07/2020 17:24	Arquivo SSLIST	611 KB

```
Mauve_Passifloracpgenomes - Bloco de notas
Arquivo  Editar  Formatar  Exibir  Ajuda
#FormatVersion Mauve1
#Sequence1File C:\Users\ADM\Desktop\Mauve\Ploefgrenii_cpgenomeFinal.fasta
#Sequence1Format FastA
#Sequence2File C:\Users\ADM\Desktop\Mauve\Palata_cpgenomeFinal.fasta
#Sequence2Format FastA
#BackboneFile C:\Users\ADM\Desktop\Mauve\Mauve_Passifloracpgenomes.bbcols
> 1:1-146537 + C:\Users\ADM\Desktop\Mauve\Ploefgrenii_cpgenomeFinal.fasta
TCAAAATCCTATTTTGAACCTATTTCCACTTTATTTCATTCAAAGAATTCTAATAAATCGAGATAGATAAGATCATAAA
GATGGATCTATGAATGTTGATCTTGGTTGACACGGGTATATGAGTCATGTTATAC TGTTGAGTAACAGCCCCAACTC-
----TCTATTTTTTGTCTAGAGAATTGGTGTGCTTGGGAGTCCCTGATGATTAATAAACCAAGATTTACCATGACTG
CAATTTTAGAGAGACCGAAAGCGAAAGCCTATGGGGTCTTTCTGTAAC TGGATAACAGCACCGAAACCGTCTTTAC
ATTGGATGGTTGGTGTTTGATGATCCCTACTTTATTGACCGCAACTTCTGATTTTATTATCGCTTTCATTGCTGCCCC
TCCAGTGGATATTGATGGTATTCGTGAACCGGTTCTGGATCTTTACTTTACGGAAACAATAATTATTTCTGGTGCCATTA
TTCCCTACTTCTGCCGCTATAGGTTTGCACCTTTACCCAATATGGGAAGCTGCATCCGTTGATGAATGGTTATACAACGGC
GGTCTTATGAACAAATGTTCTACACTTCTTACTTGGTGTAGCTTGTACATGGGCCGTGAGTGGGAAC TTAGTTTCCG
TTTGGGTATGCGCCCTGGATTGCTGTTGCATATTCAGCTCCGTTGTCAGCGGCTGCCGCTGCTTCTTGATTTACCCAA
TTGGTCAAGGAAGTTTTTCGGACGGTATGCC TTTAGGAATCTCTGGTACTTTCACTTTATGATTTGATTCCAGGCTGAG
CACAAACATCCTTTGACACCACTTTTCATATGTTAGGCGTAGCAGGTGATTCGGGGCTCCCTATTTCAGTGCATGATG
TTCC TTGGTAACCTCTAGTTT GATCAGGGAACACACAGAAATGAAATCCGCTAATGAAGGTTACAGATTTGGTCAAGAAG
AGGAAC TTATAATATCGTAGCCGCTCATGGTTATTTGGCCGATTGATCTTCCAATATGCTAGTTTCAACAACCTCTCGT
TCATTACACTTCTTCTAGCTGCTTGGCTGTAATAGGTATCTGGTTCACCCTTATAGGTATTAGCACTATGGCTTTCAA
CTTAAATGGTTCAATTTCAACCAATCTGTAGTTGATAGTCAAGGCTGTGTAATTAACACTTGGGCTGATATTATCAACC
GTGCTAACCTTGGTATGGAAAGTTATGCATGAACGTAATGCTCACAAC TTCCCTCTAGACCTAGCTGCTGTTGAAGCTCCA
TCTACAATGGGTAAAGACTTTGGTCTTAGTGTGACAAGTTTCAATGAAATAAATAAAGGAGCAATAACAATCTTCTTGATA
TAACAAGAAATGGCTATTGCTCCCTTCTCATATTTTTTTTATTTAGTACTTTTTTTAGTCTTTGAGTTTCAATAAATT
TTTTCTTTTATTTCTTTCTATTTTTTAAGAAATAAATAATAGAA--AGAAATAAAGAAATGAAATTC TGGTAATTT
TTAGTGGTAAATTTTACGTAGTTTTAAATAGAGTTTTGGGGCGGATGTAGCCAAGTGGATAAAGCGGTTGGATTGTGAA
TCCAGCCGCGGCTTATTCGGCTATTCGGCCATAGCCATAAATAAGATAGATAGATTTTCAGATTCAG
```

# MAUVE

## Alinhamento das seqüências com a função progressiveMauve

- Adicionar as seqüências .fasta em Add Sequences



# MAUVE

## Resultados

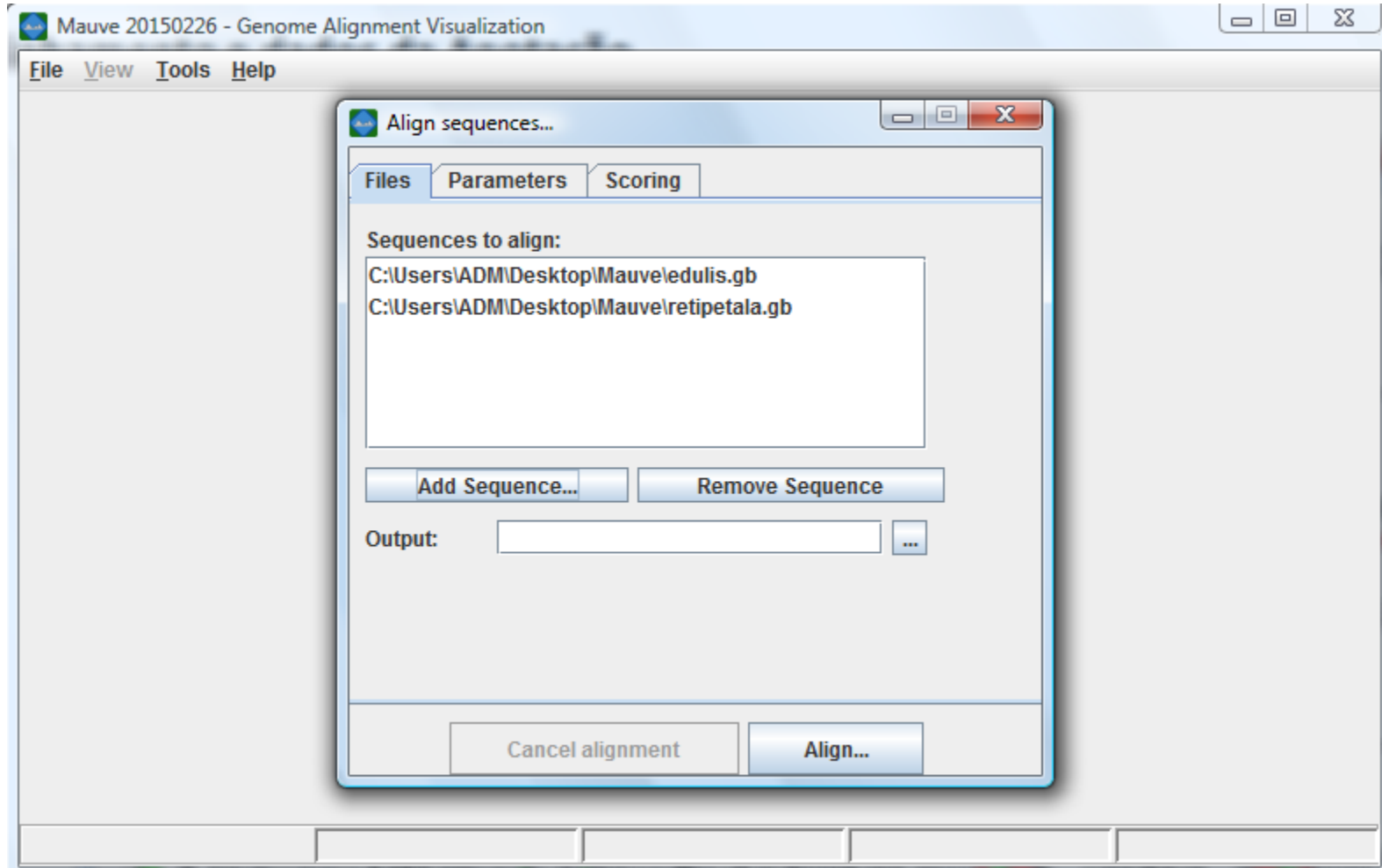
- Rearranjos identificados entre os genomas cloroplastidiais.



# MAUVE

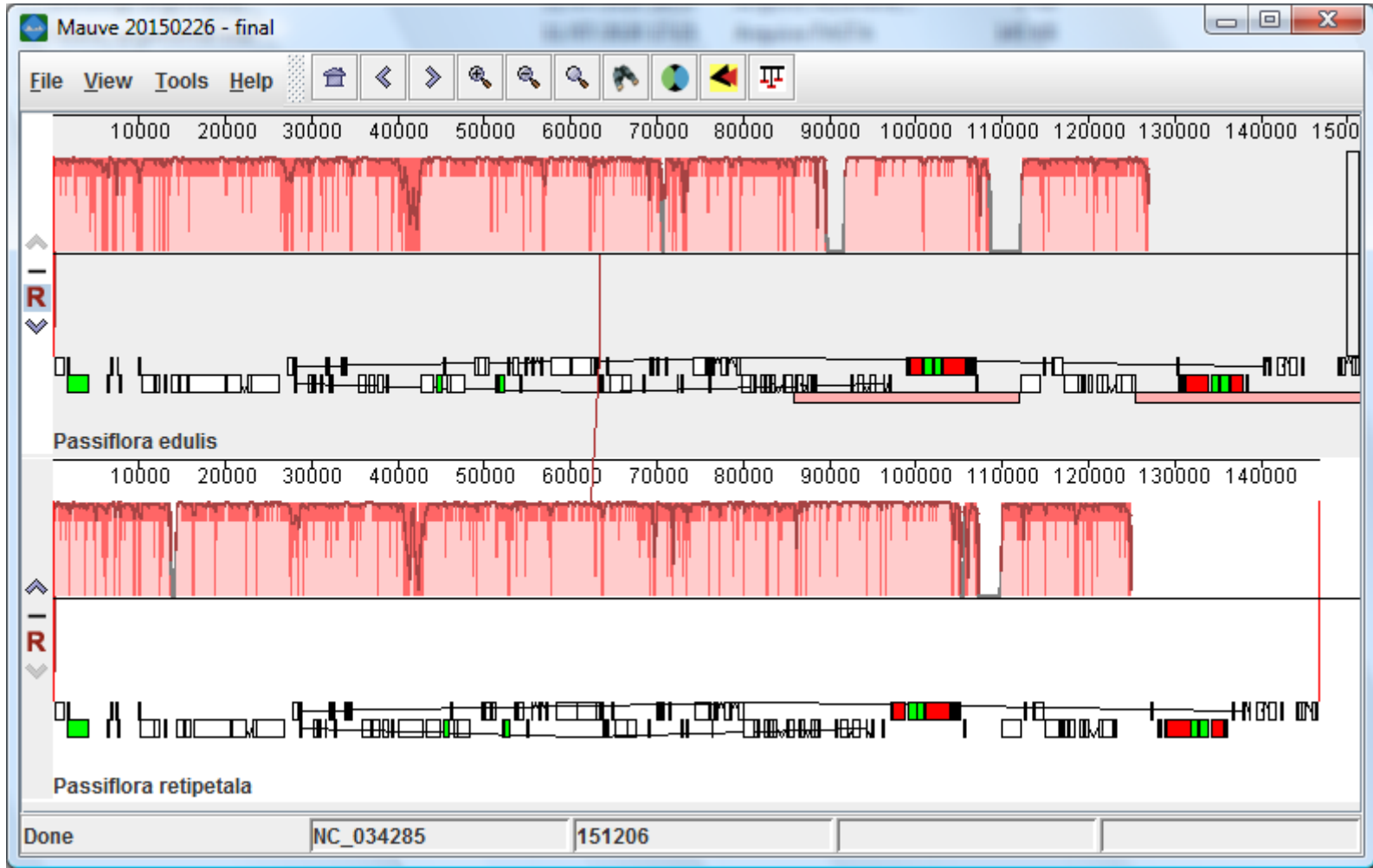
## Alinhamento e dados da Anotação

- Adicionar os arquivos Genbank contendo as sequências fasta e anotação em Add Sequences



# MAUVE

## Alinhamento e dados da Anotação





# OBRIGADO!

luizcauz@usp.br

