UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA 'LUIZ DE QUEIROZ'
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS
Laboratório de Genética Molecular de Plantas Cultivadas

Minicurso

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

Dr. Luiz Augusto Cauz dos Santos – luizcauz@usp.br
Dra. Zirlane Portugal da Costa – 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

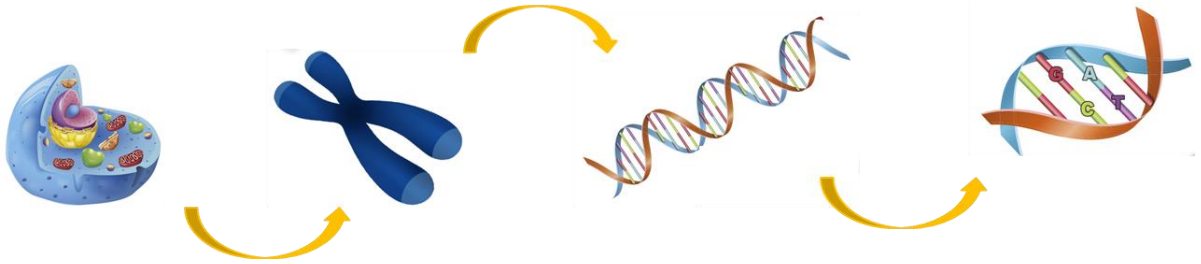
INTRODUÇÃO À GENÔMICA

- **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

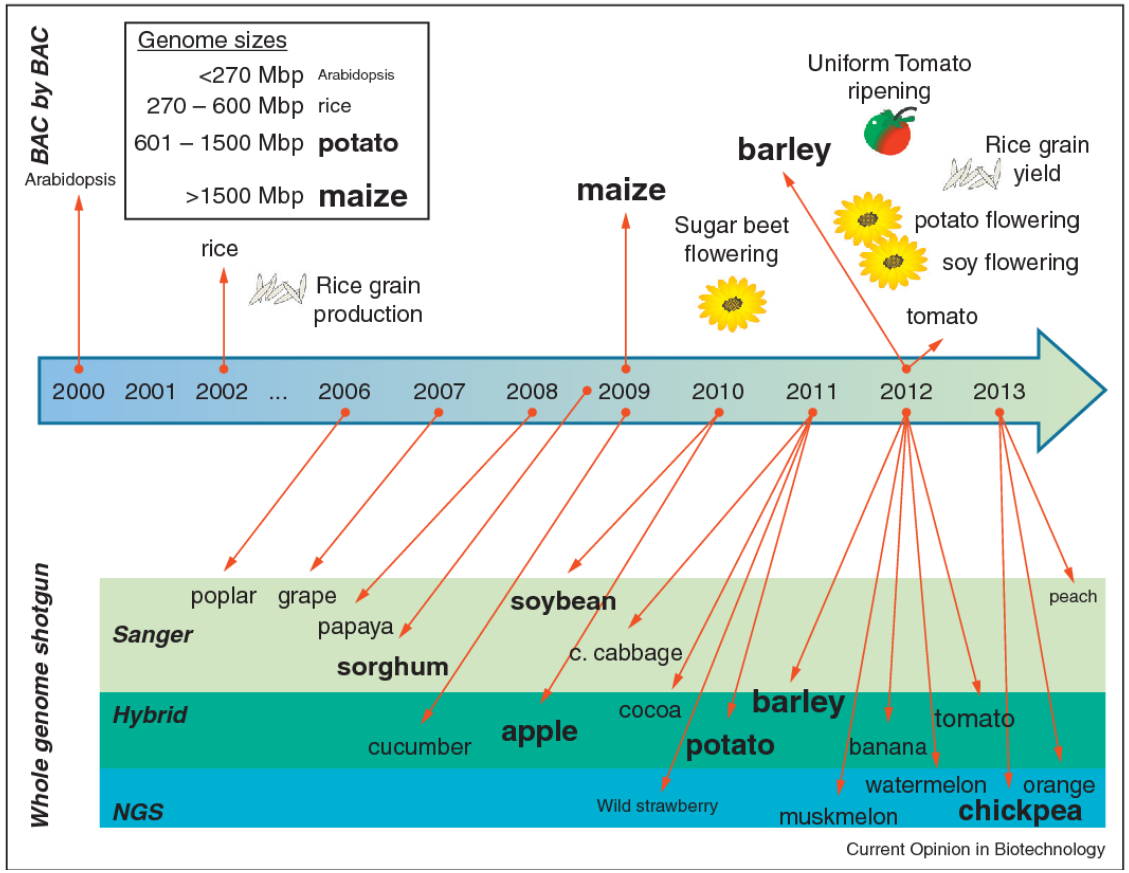
FGS (*First Generation Sequencing*)
- Sanger

NGS (*Next Generation Sequencing*)
- 454
- Illumina
- SOLiD

TGS (*Third Generation Sequencing*)
- Plataforma PacBio
- Oxford nanopore

Evolução dos métodos de sequenciamento de DNA

Genomas de plantas sequenciados



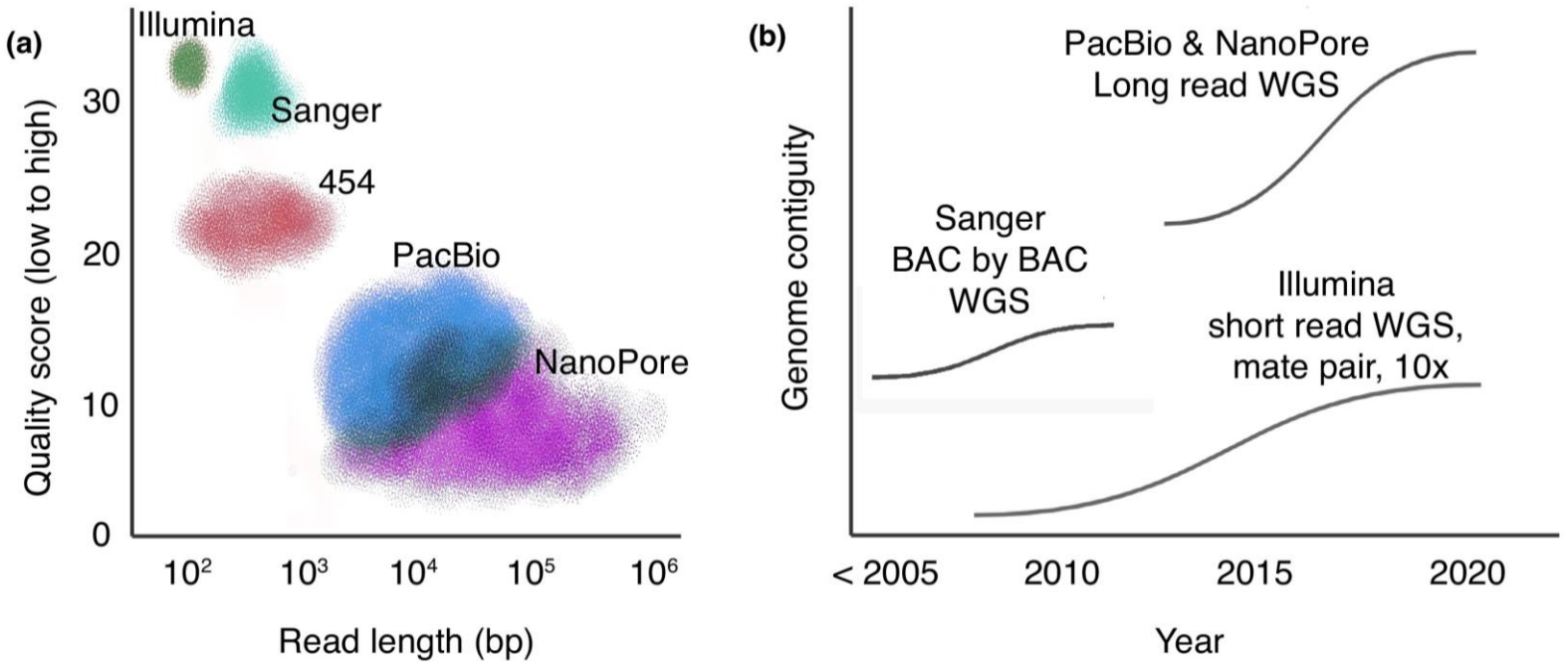
INTRODUÇÃO À GENÔMICA

illumina®

MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series	NovaSeq Series
Power and simplicity for targeted sequencing.	Small genome and targeted sequencing.	Everyday genome, exome transcriptome sequencing, and more.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale human whole-genome sequencing.	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[1]
MAX READ LENGTH 2×150 bp	MAX READ LENGTH 2×300 bp	MAX READ LENGTH 2×150 bp	MAX READ LENGTH 2×150 bp	MAX READ LENGTH 2×150 bp	MAX READ LENGTH 2×150 bp

INTRODUÇÃO À GENÔMICA

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



Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER

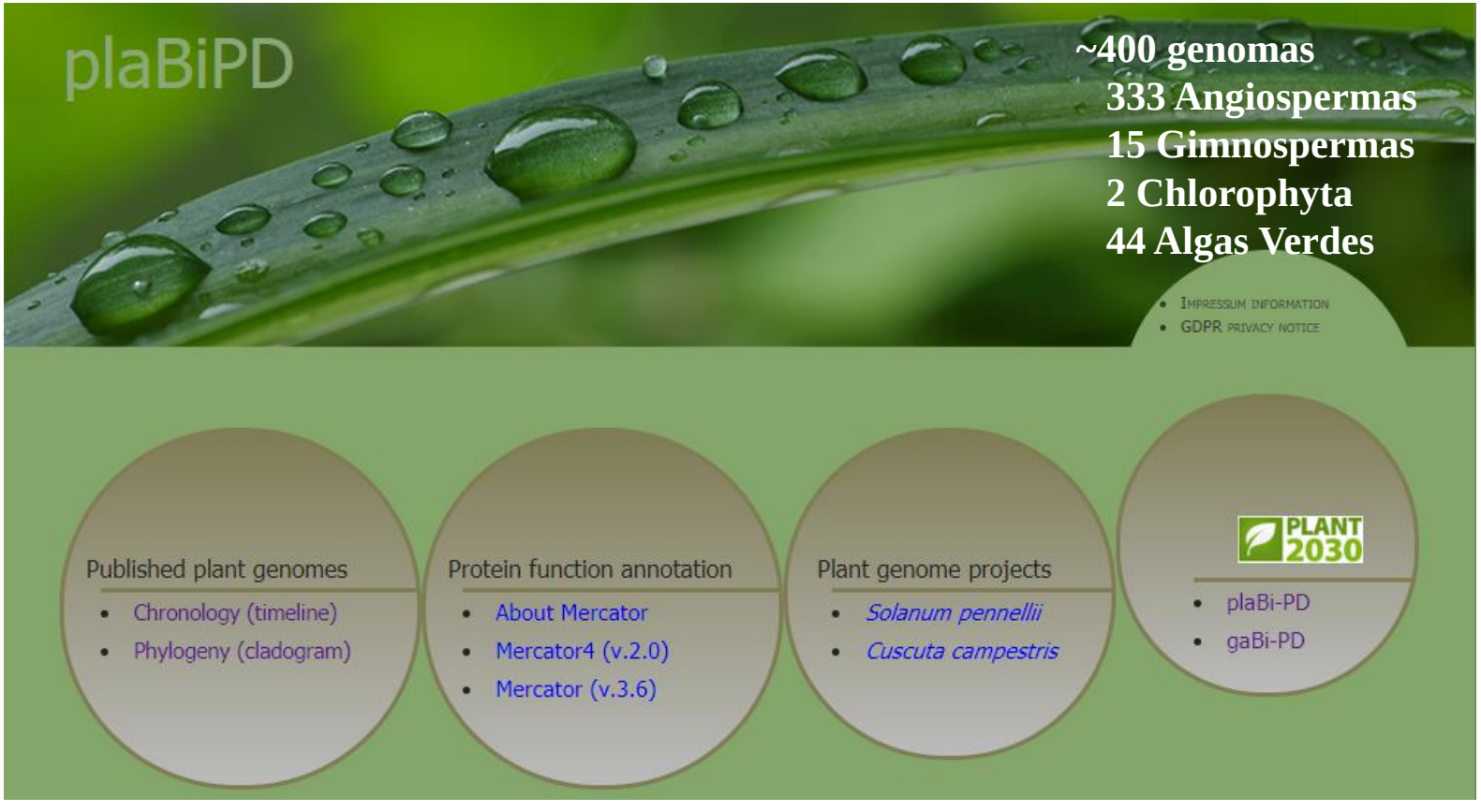
Building near-complete plant genomes

Todd P Michael[1] and Robert VanBuren[2,3]

INTRODUÇÃO À GENÔMICA

Genomas de plantas sequenciados

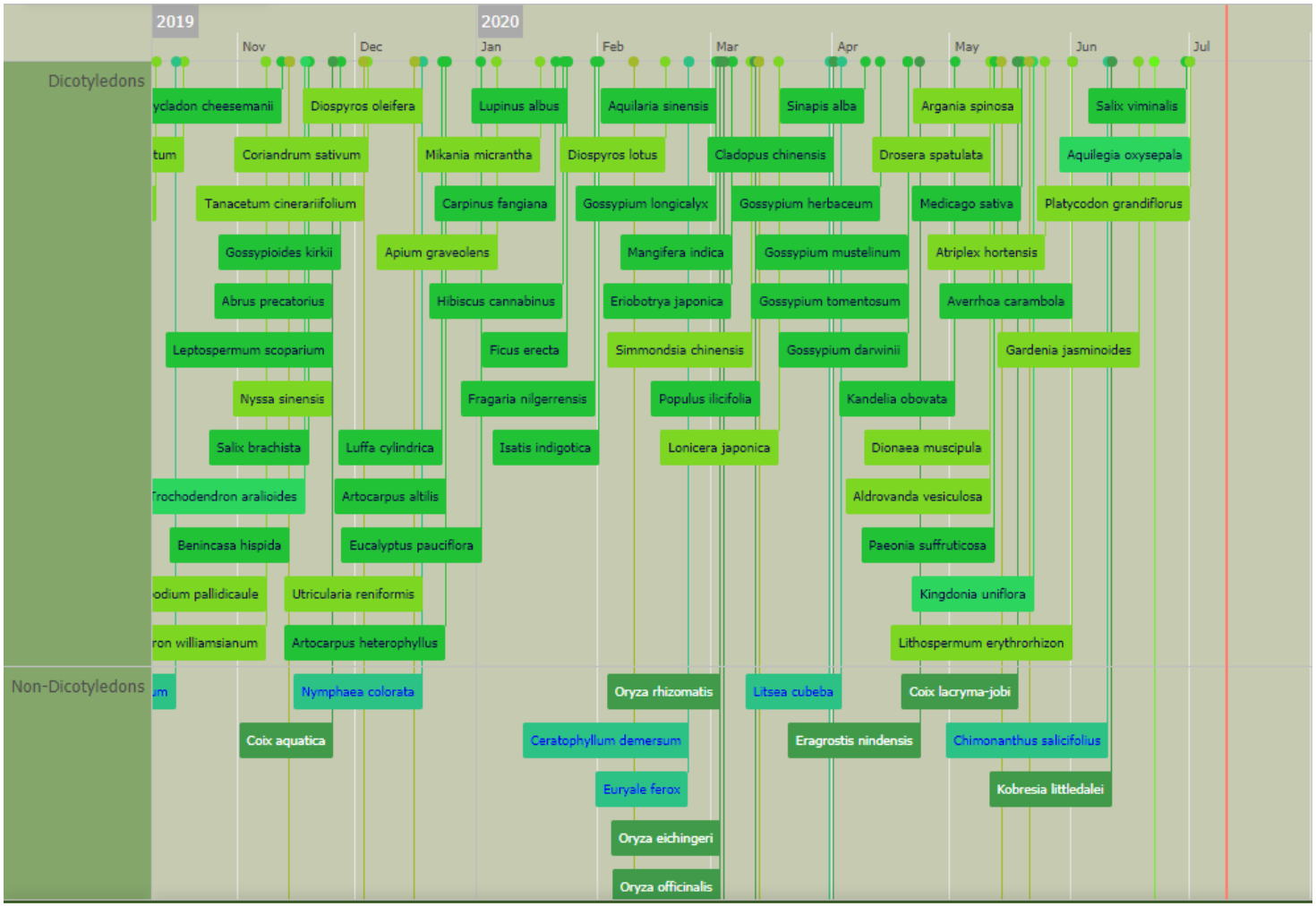
https://www.plabipd.de/index.ep



INTRODUÇÃO À GENÔMICA

Genomas de plantas sequenciados

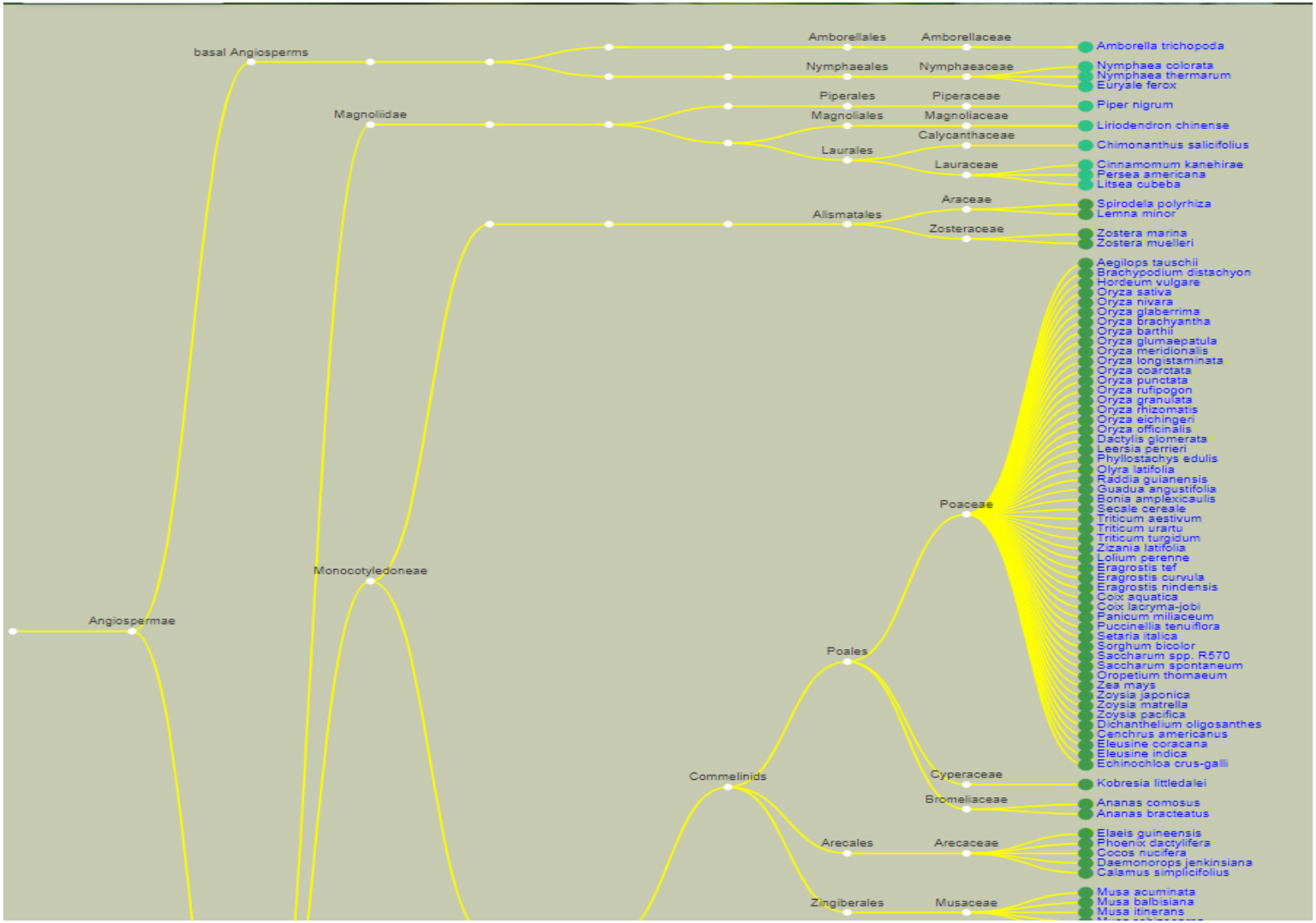
https://www.plabipd.de/index.ep - apresentação cronológica das espécies com genomas publicados



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

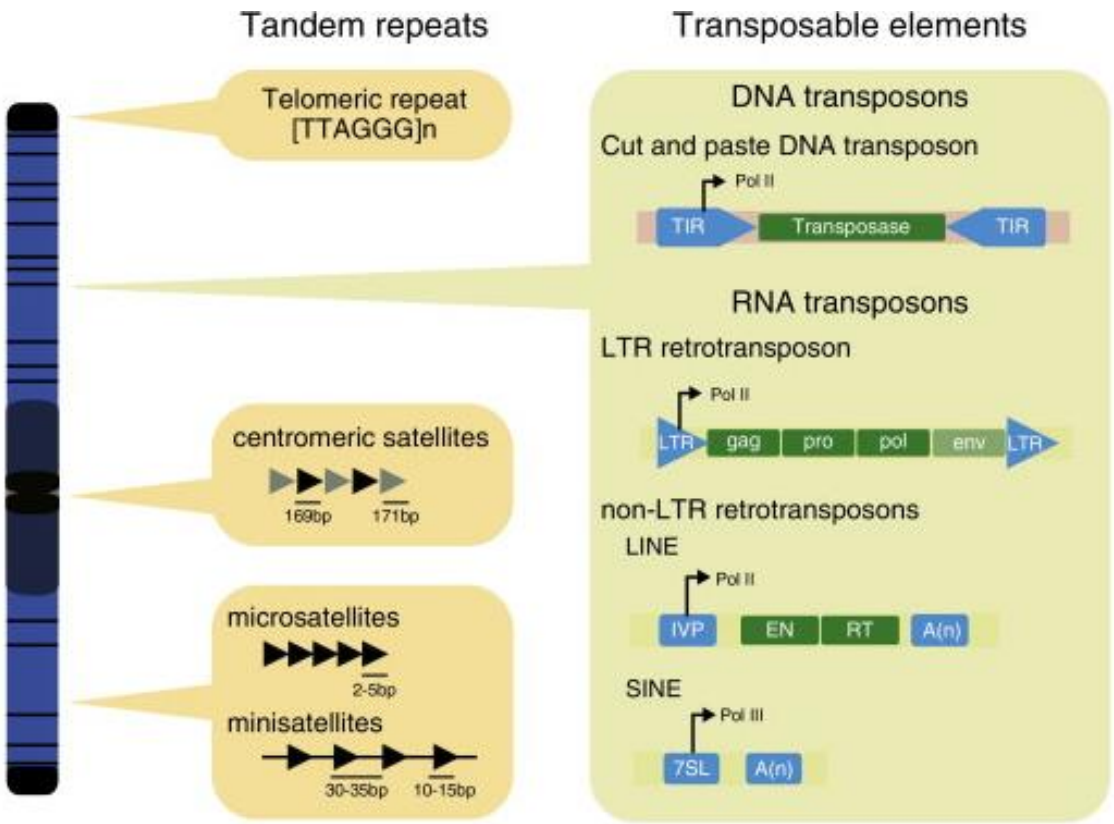
Variação no tamanho de genomas

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

INTRODUÇÃO À GENÔMICA

- Complexidade dos genomas eucariotos

Porção repetitiva do genoma



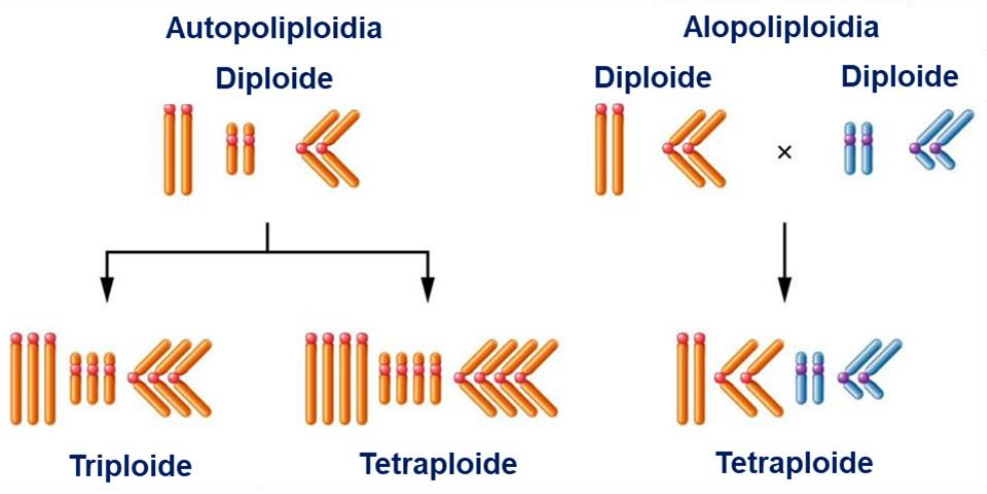
Representação esquemática das classes de repetição, sua distribuição e suas características estruturais

INTRODUÇÃO À GENÔMICA

- Complexidade dos genomas eucariotos

➢ **Poliploidia: condição na qual uma célula ou organismo adquire um ou mais conjuntos adicionais de cromossomos**

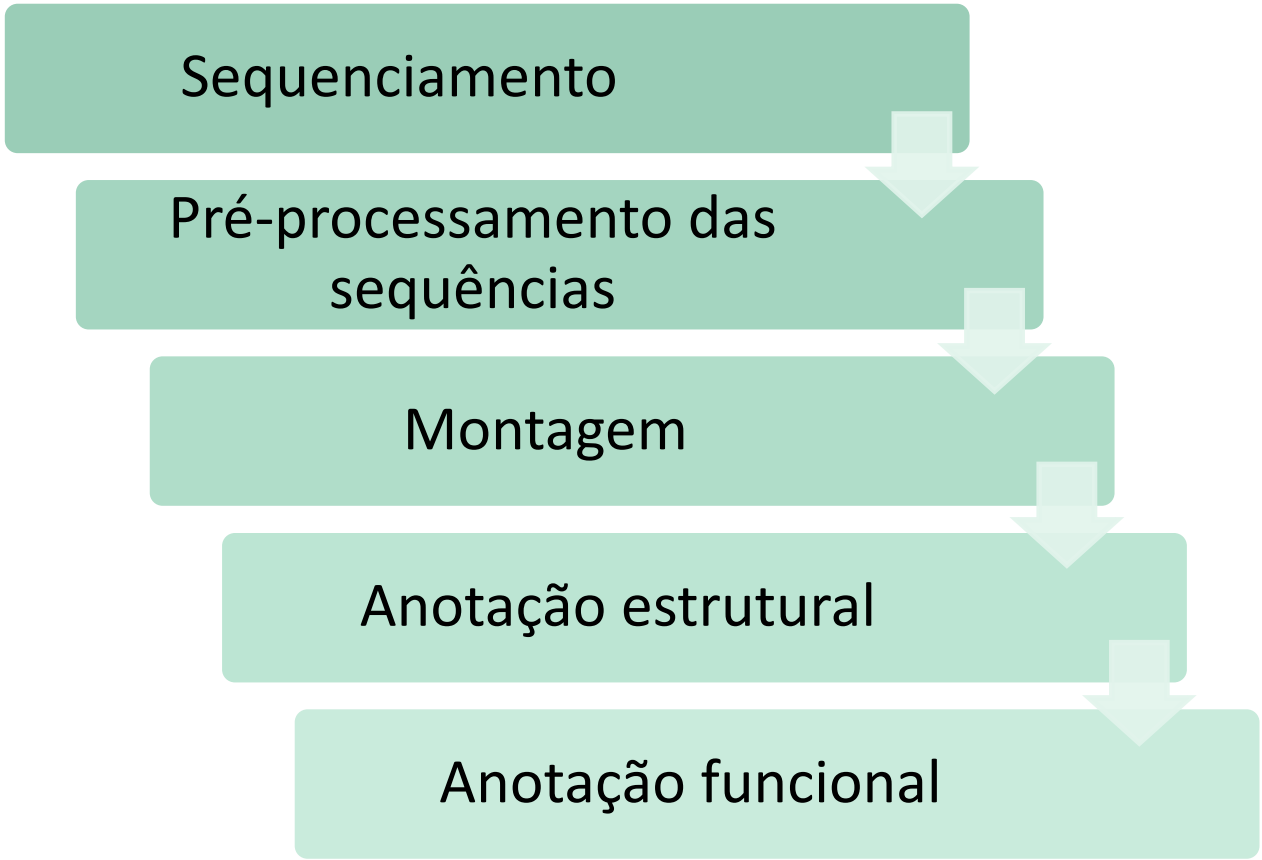
- Comum entre plantas, bem como entre certos grupos de animais
- Autopoliploide: conjunto cromossômico derivado de uma única espécie
- Alopoliploide: conjunto cromossômico derivado de diferentes espécies



INTRODUÇÃO À GENÔMICA

- Obtenção de genomas

OVERVIEW



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

FastQC Report

Qua 8 jul 2020
phaemal1r1.fastq

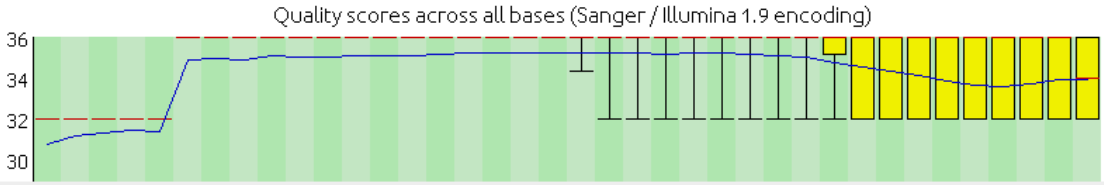
Summary

- 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	phaemal1r1.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



Produced by FastQC (version 0.11.8)

Pré-processamento das sequências

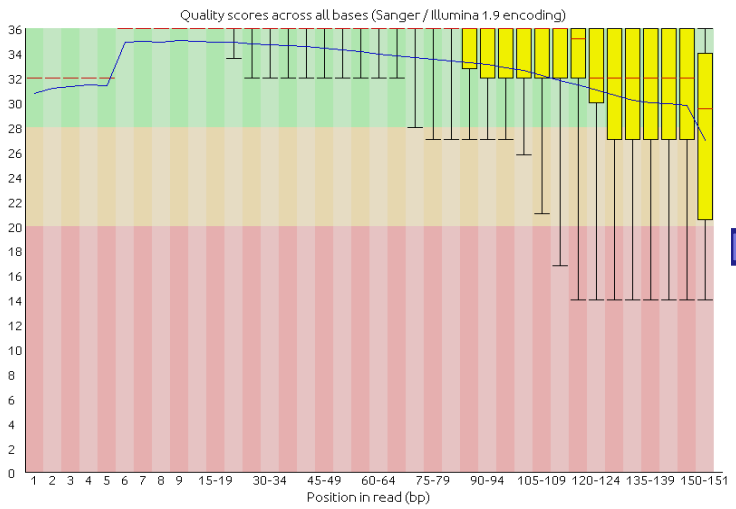
- Controle de qualidade dos dados brutos

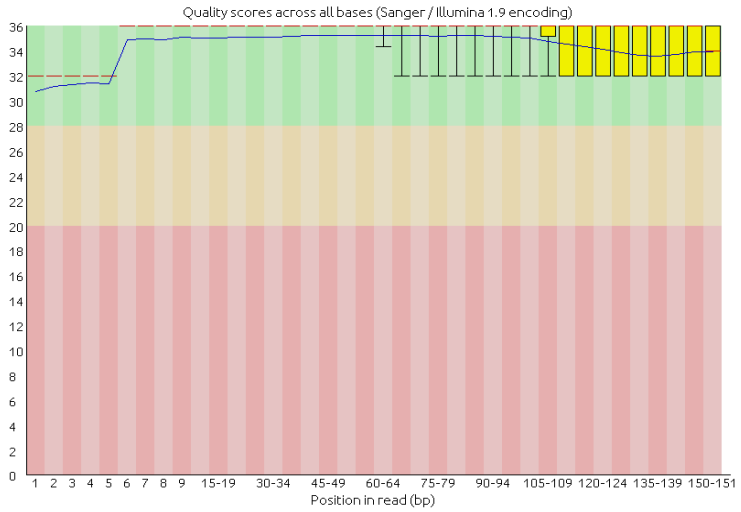
1. Qualidade das bases

- *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





Bom

Razoável

Ruim

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

L1R1_1.fastq	L1R1_2.fastq	L1R1.fasta
815,6 Mb	546,8 Mb	216,7 Mb

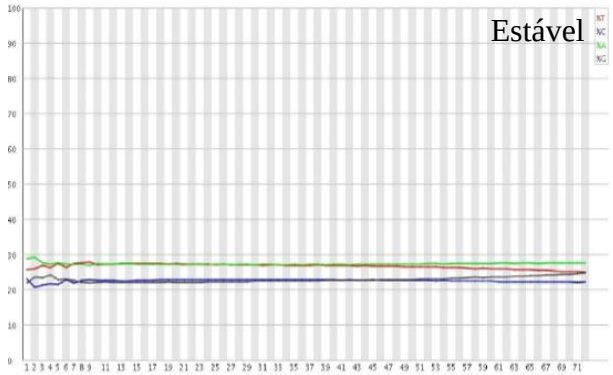
Pré-processamento das sequências

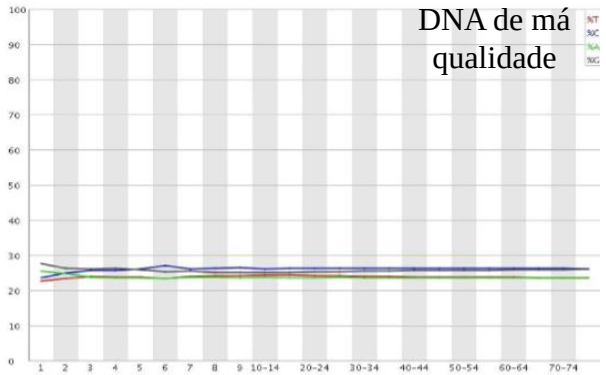
- Controle de qualidade dos dados brutos

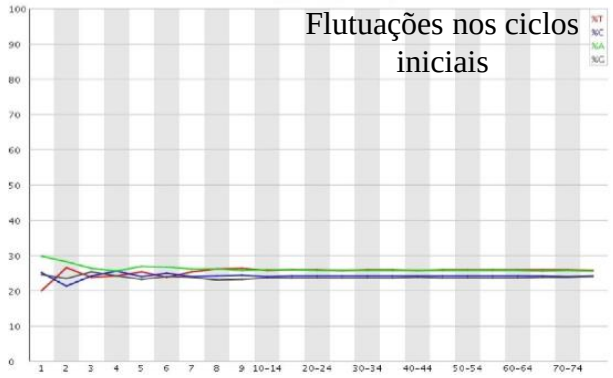
2. Distribuição nucleotídica

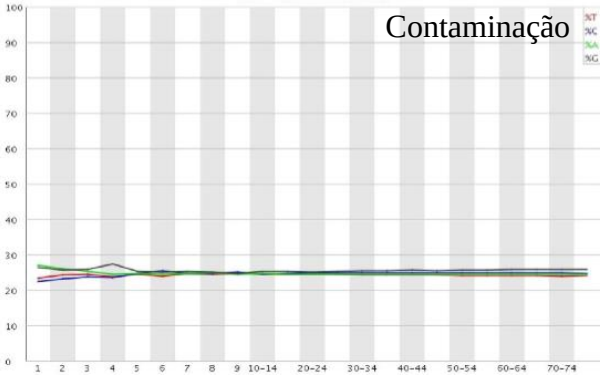
- Deve permanecer estável ao longo das reads;
- Associação com a qualidade das bases;

%T
%C
%A
%G







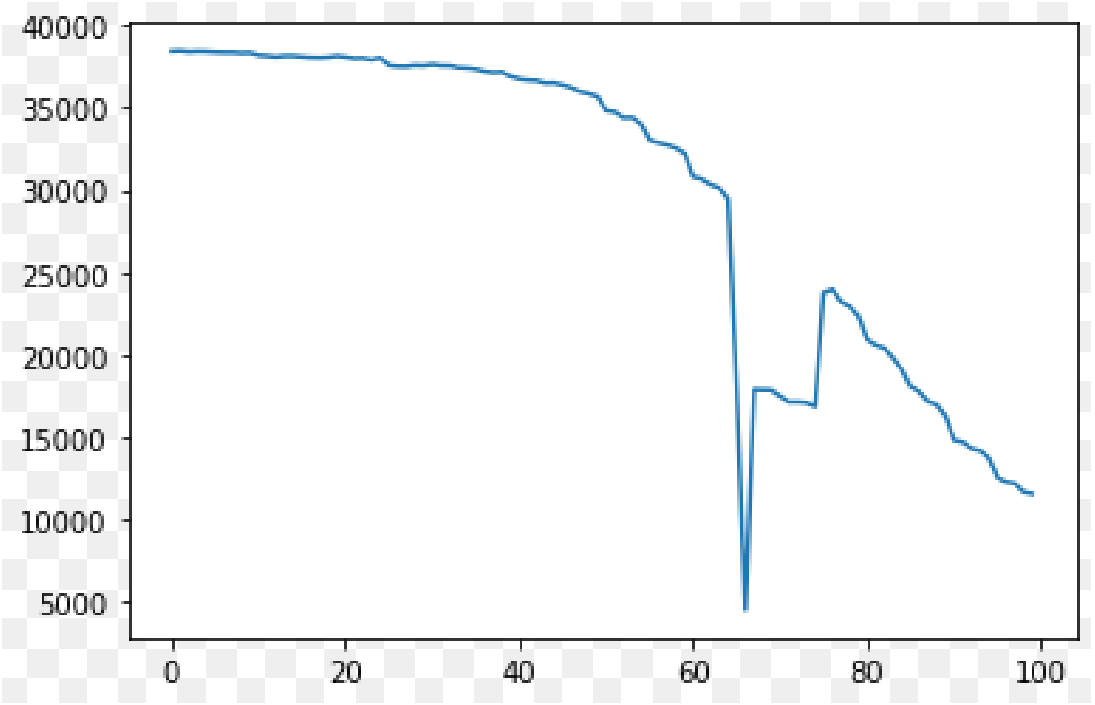


Pré-processamento das sequências

- 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 cerevisae* e *Mycobacterium tuberculosis*;
- >10% de desvio pode indicar contaminação;



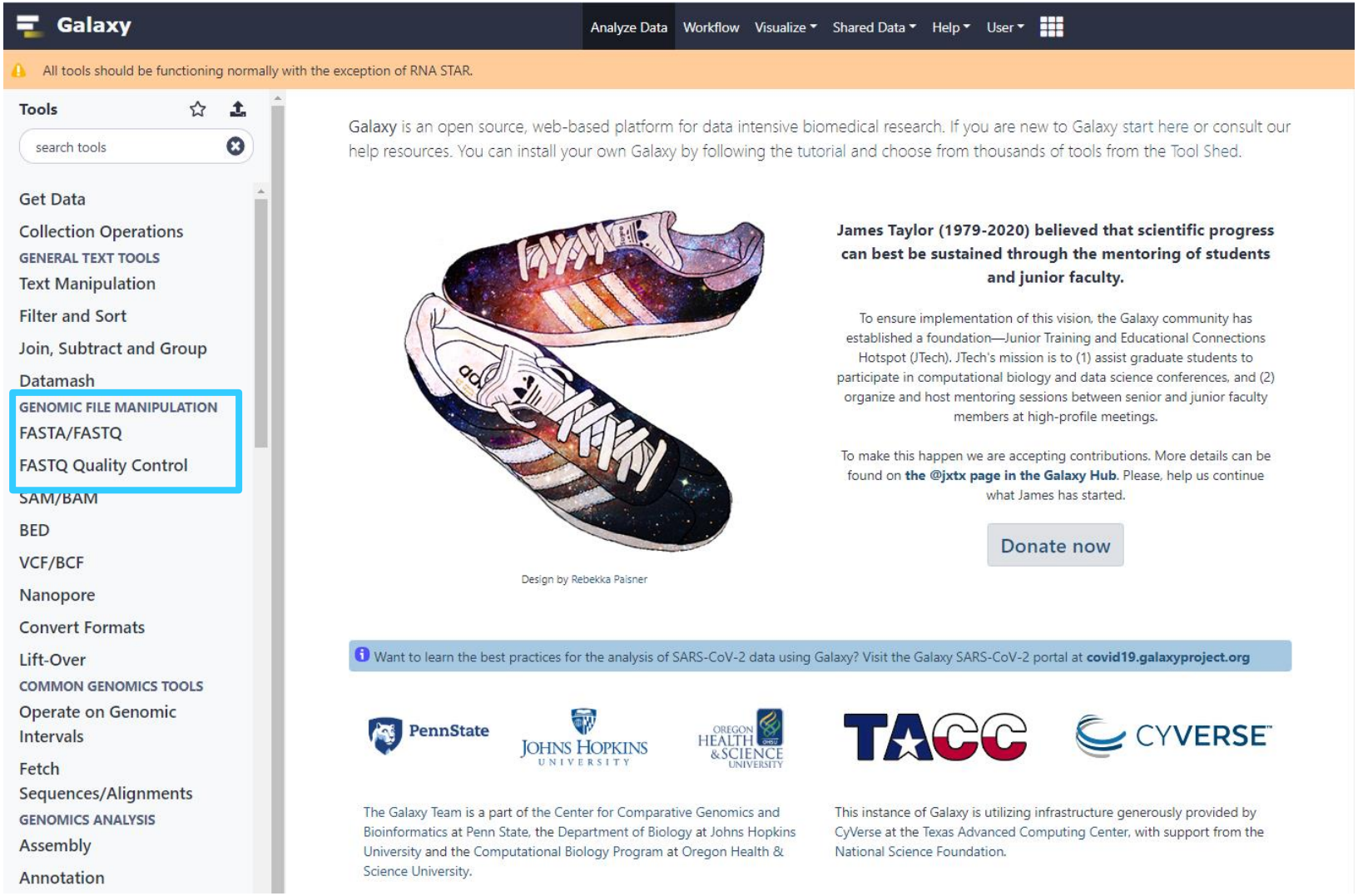
Pré-processamento das sequências

- Controle de qualidade dos dados brutos

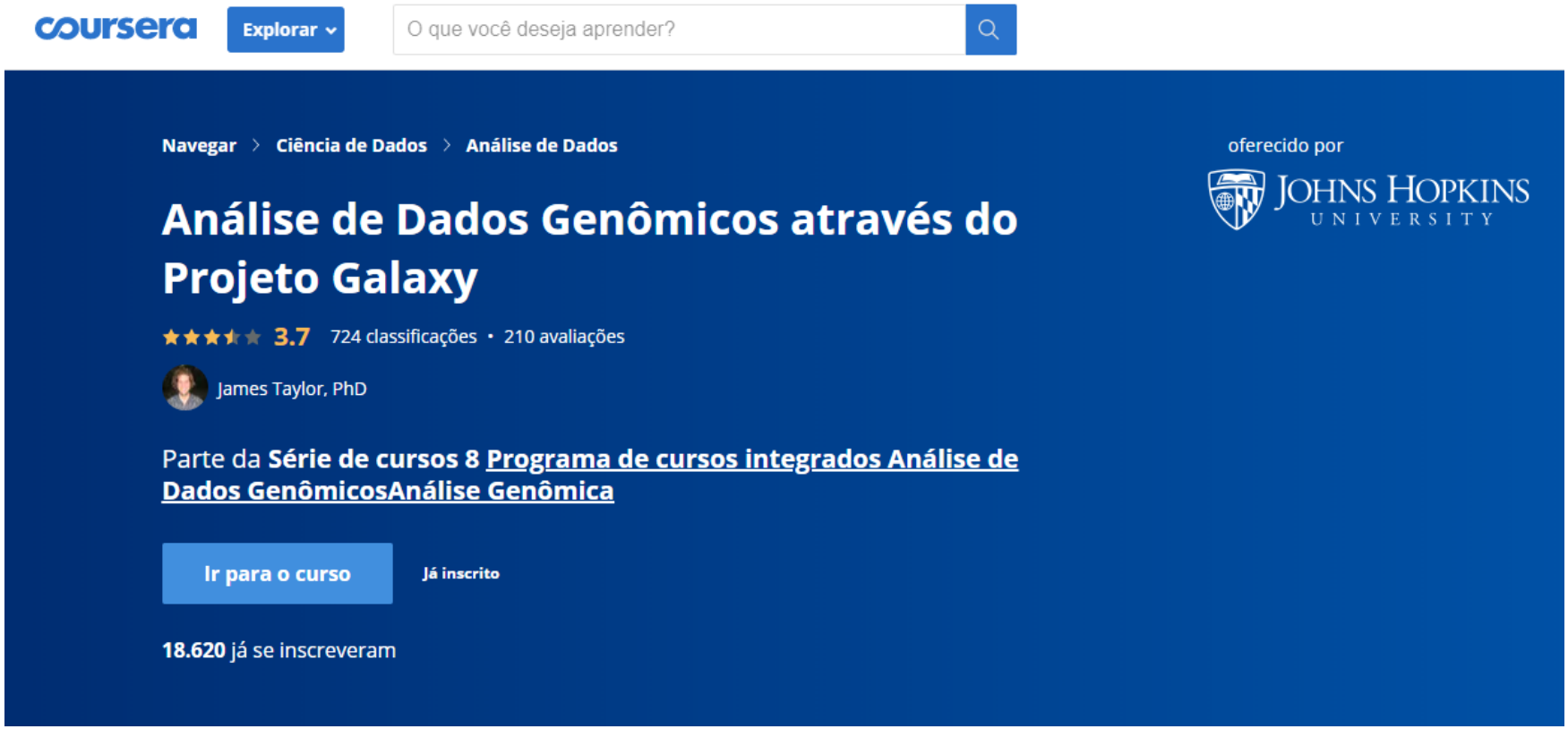
SOFTWARES

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

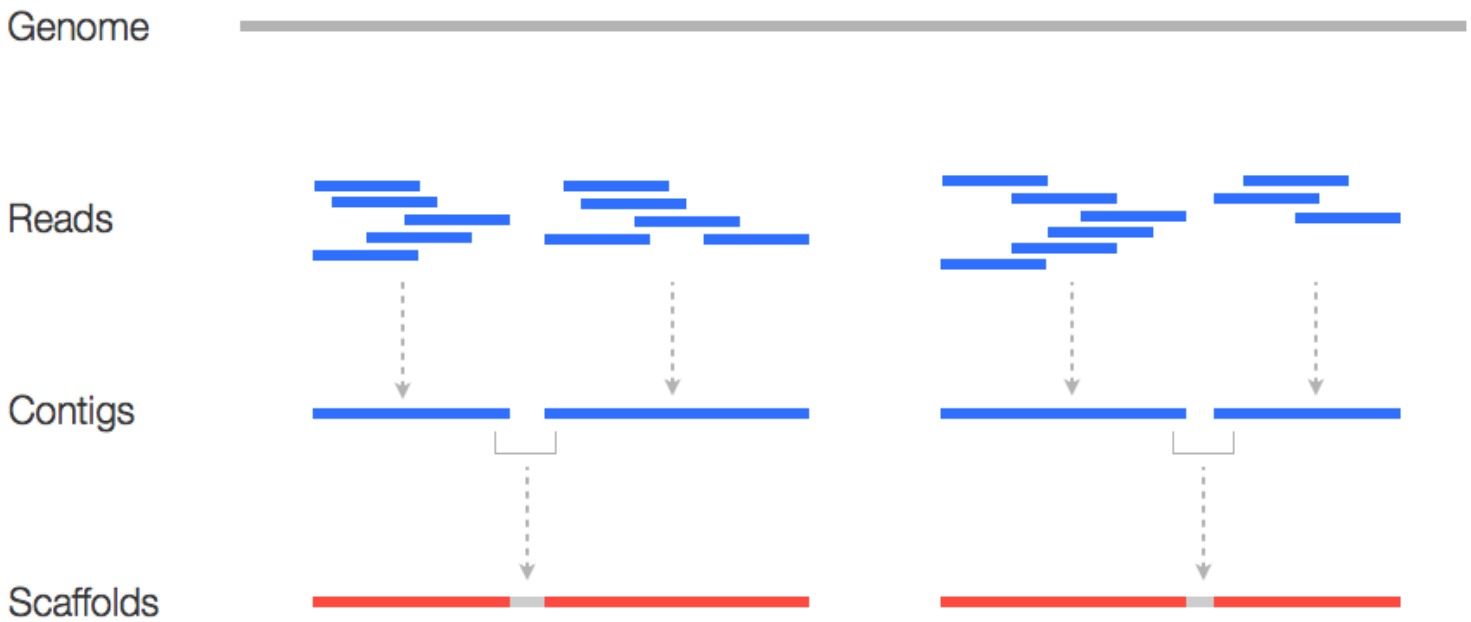


INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Montagem

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

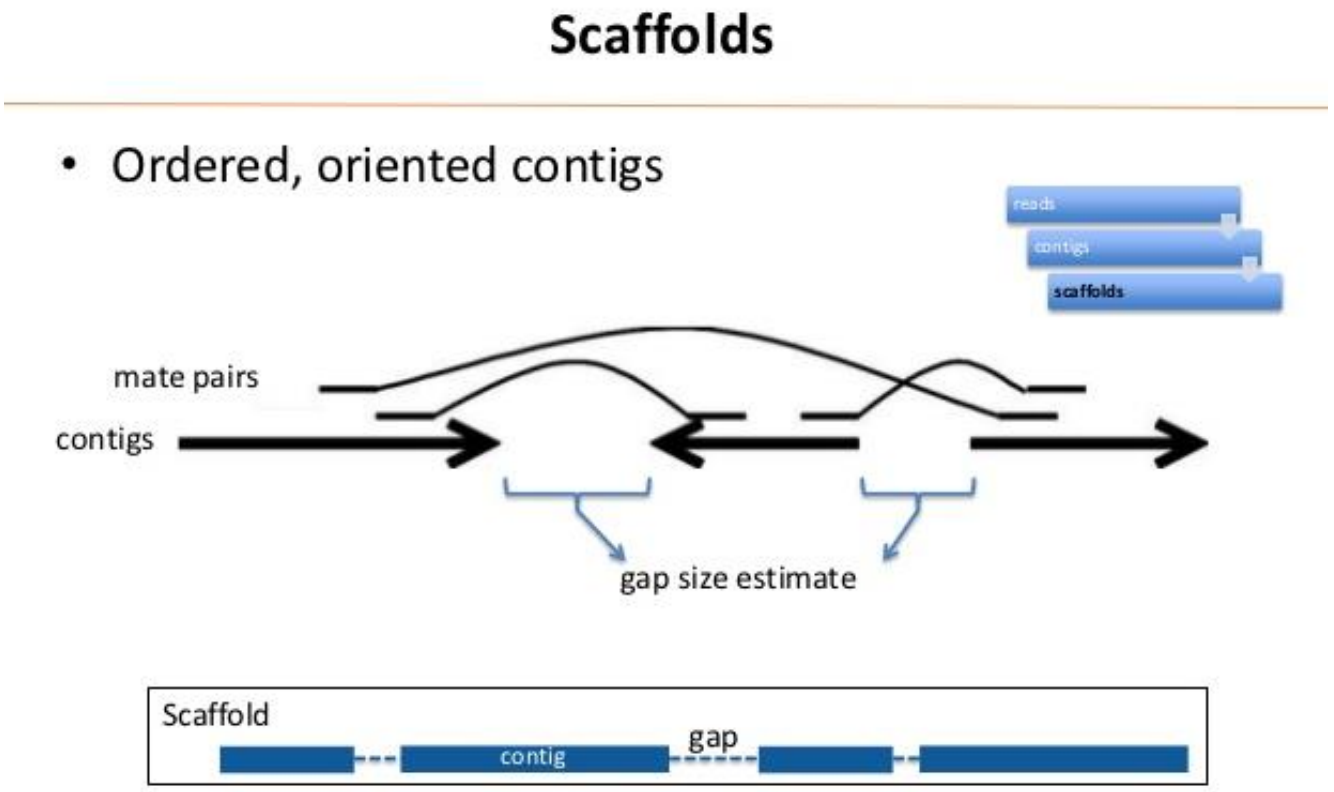
***Read*:** Fragmento que foi sequenciado
***Contigs*:** Pedaço contíguo de sequência formado a partir da sobreposição de *reads*
***Scaffolds*:** Resultado da conexão entre *contigs*



INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Montagem

Obtenção dos *scaffolds*



INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Montagem

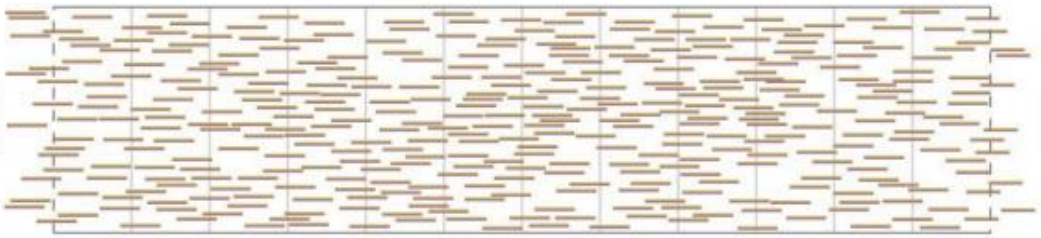
- De-novo

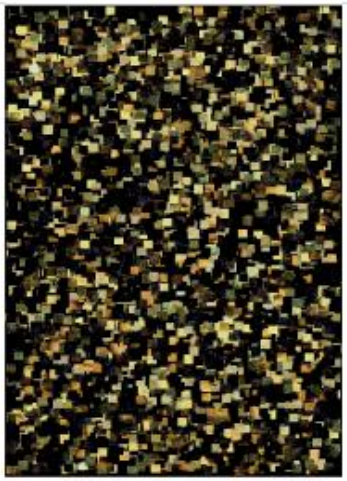
- Reference

INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

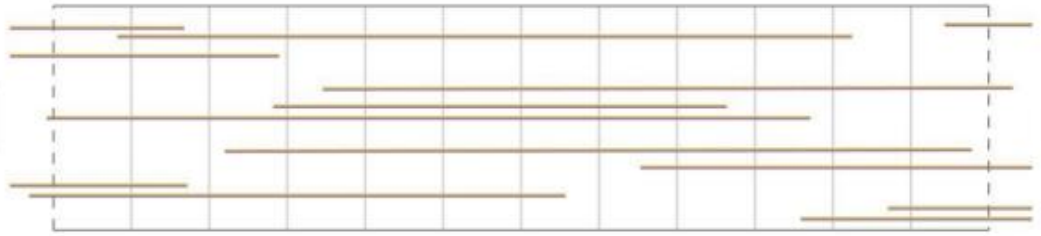
➢ Montagem

Short Reads



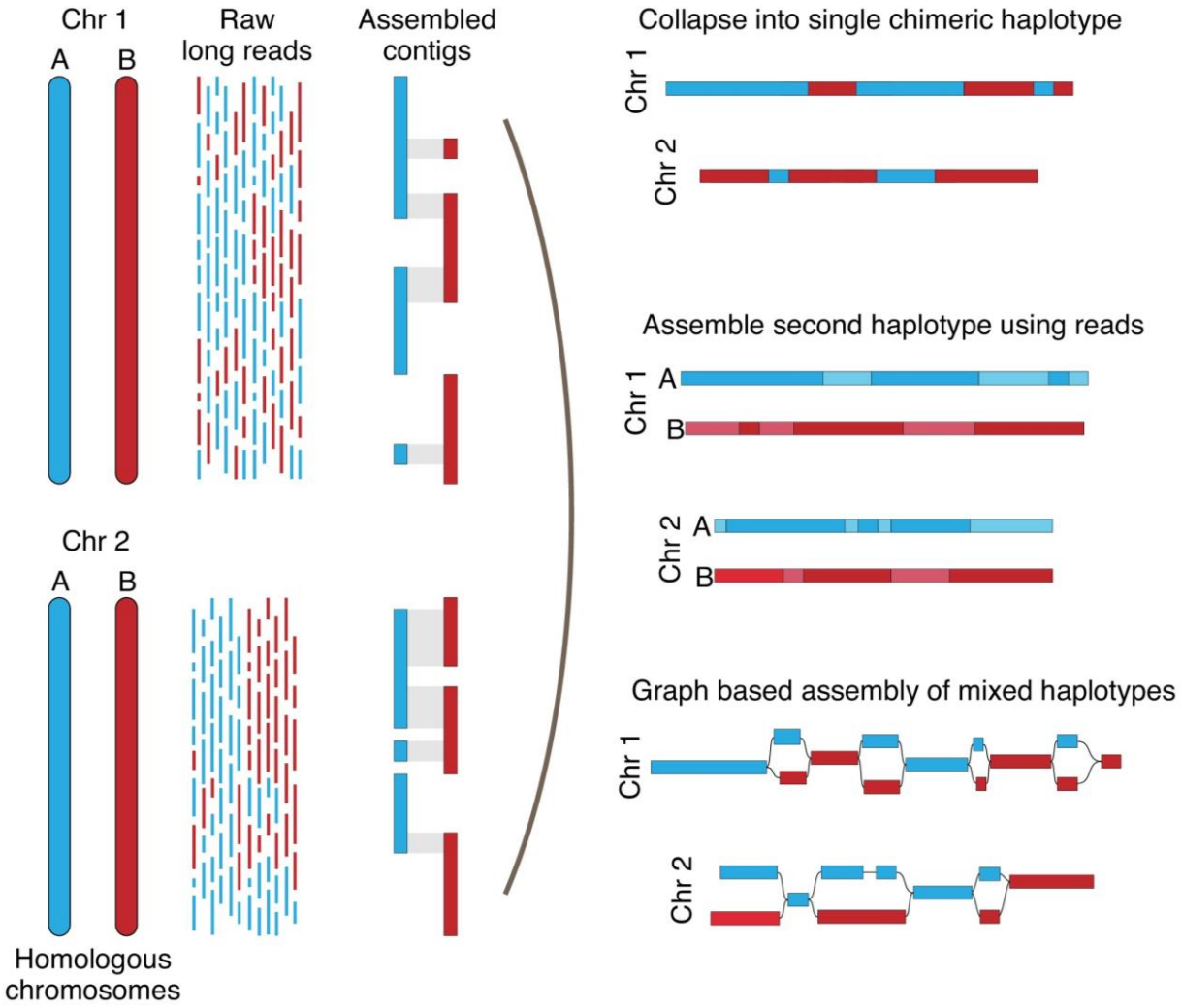


Long Reads



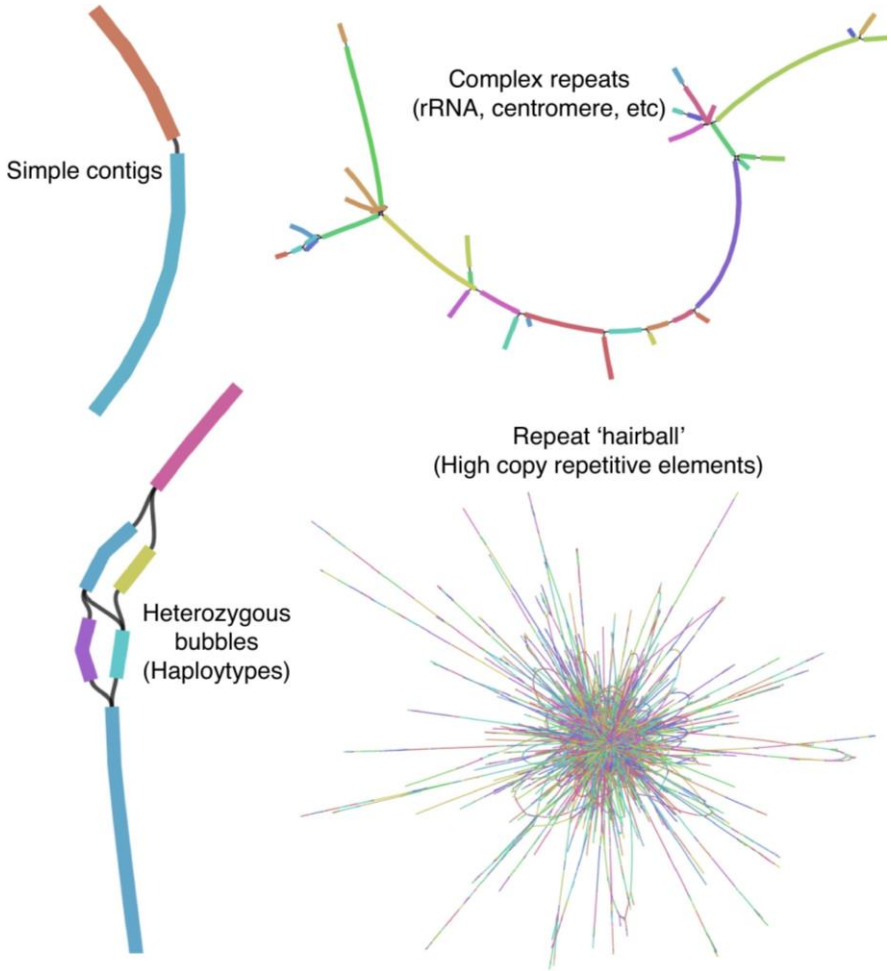
INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

- **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

ScienceDirect

ELSEVIER

Building near-complete plant genomes

Todd P Michael[1] and Robert VanBuren[2,3]

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

Genome analysis

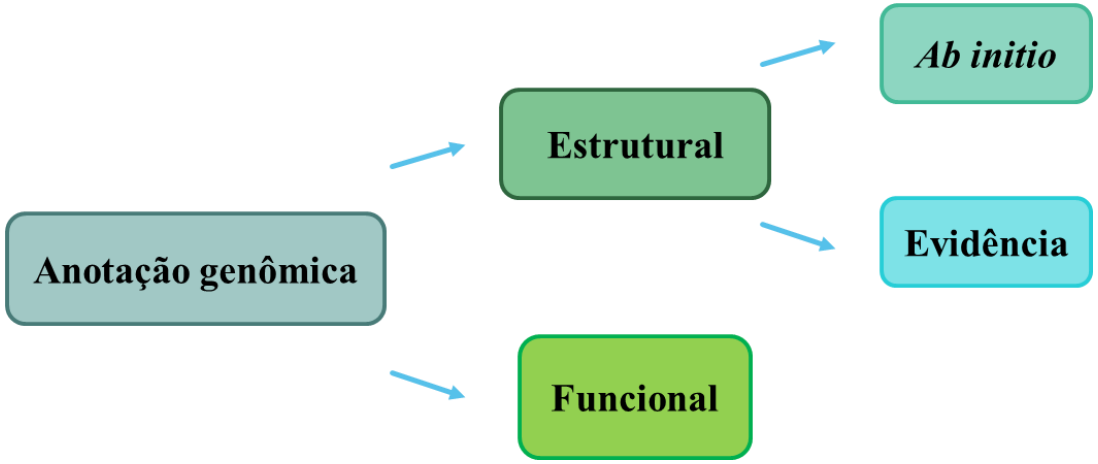
Bandage: interactive visualization of *de novo* genome assemblies

Ryan R. Wick[1,*], Mark B. Schultz[1], Justin Zobel[2] and Kathryn E. Holt[1]

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

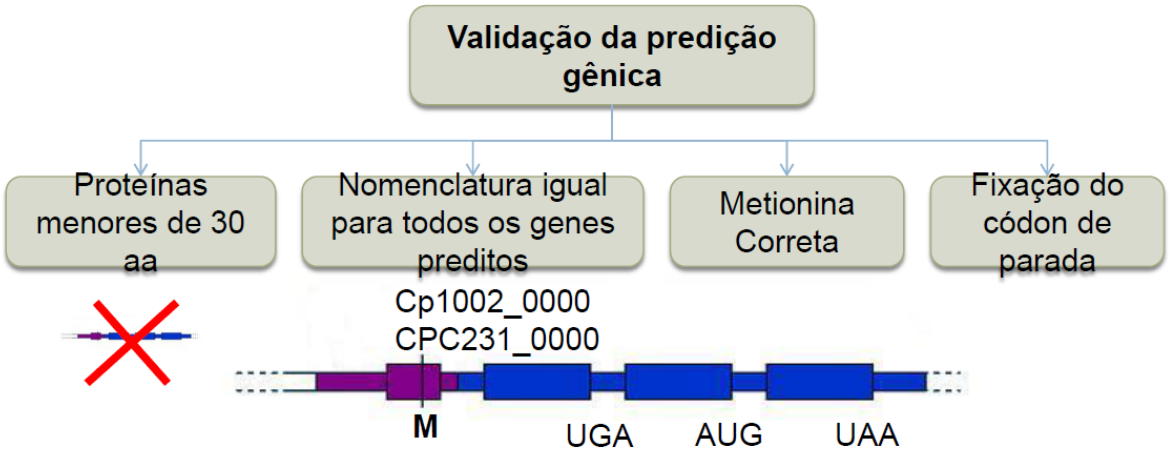
- ✓ **Anotação funcional**: processo de atribuir informações, como termos de ontologia gênica, à anotação estrutural

INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

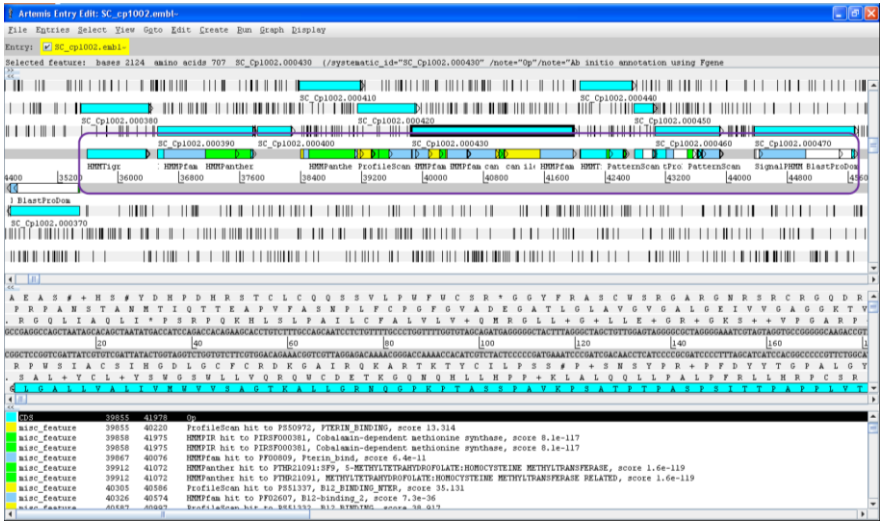
➢ Anotação estrutural

INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

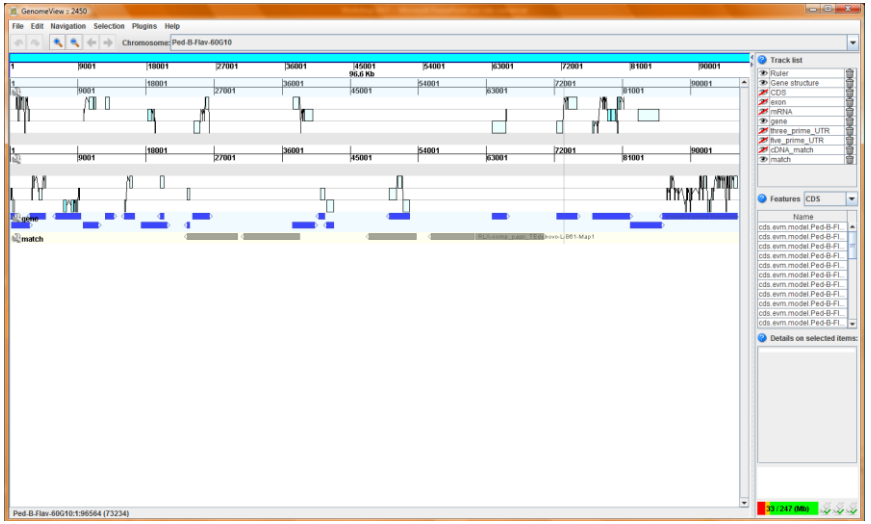
➢ Curadoria Manual



✓ **Artemis**



✓ **GenomeView**



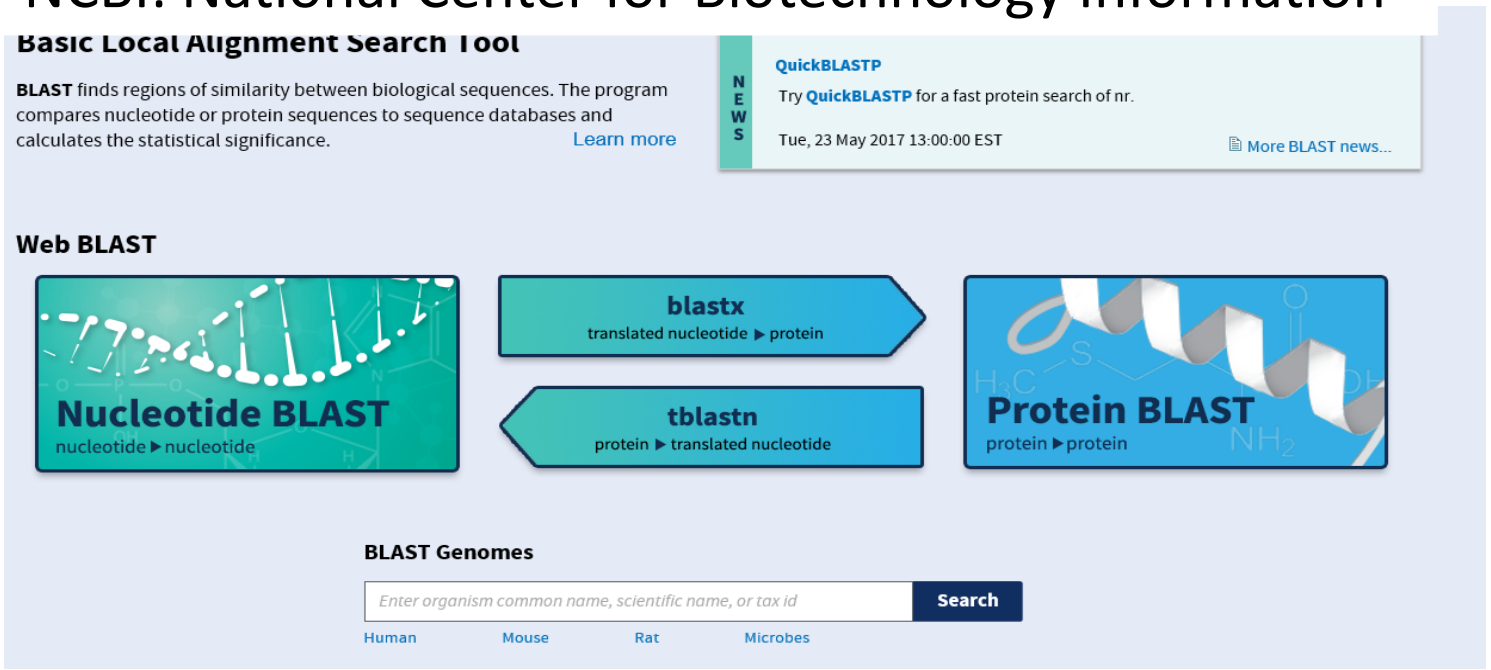
INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Anotação Funcional

✓ BLAST: Basic Local Alignment Search Tool

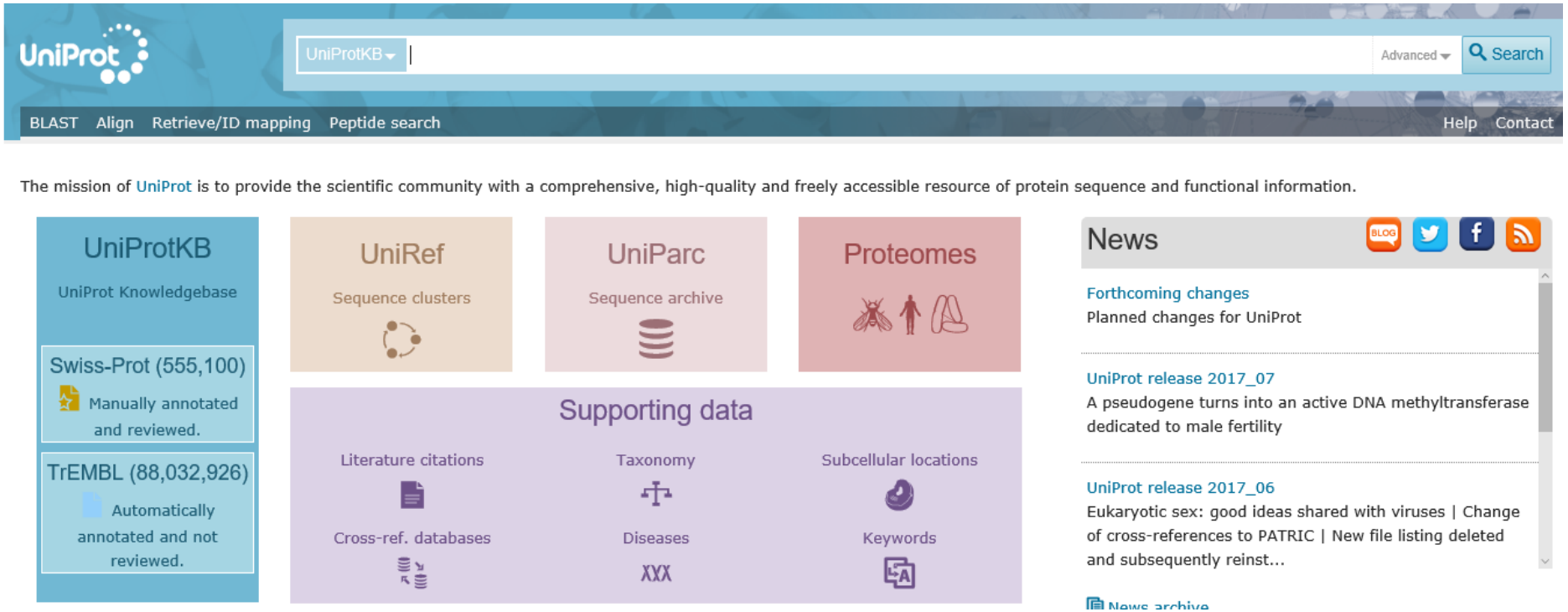
- BLASTX : verificar similaridade com proteínas nos bancos de dados
- BLASTN: verificar similaridade com ESTs e RNA ribossomal

NCBI: National Center for Biotechnology Information



INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢Anotação Funcional



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
```

www.nature.com/scientificreports

SCIENTIFIC REPORTS

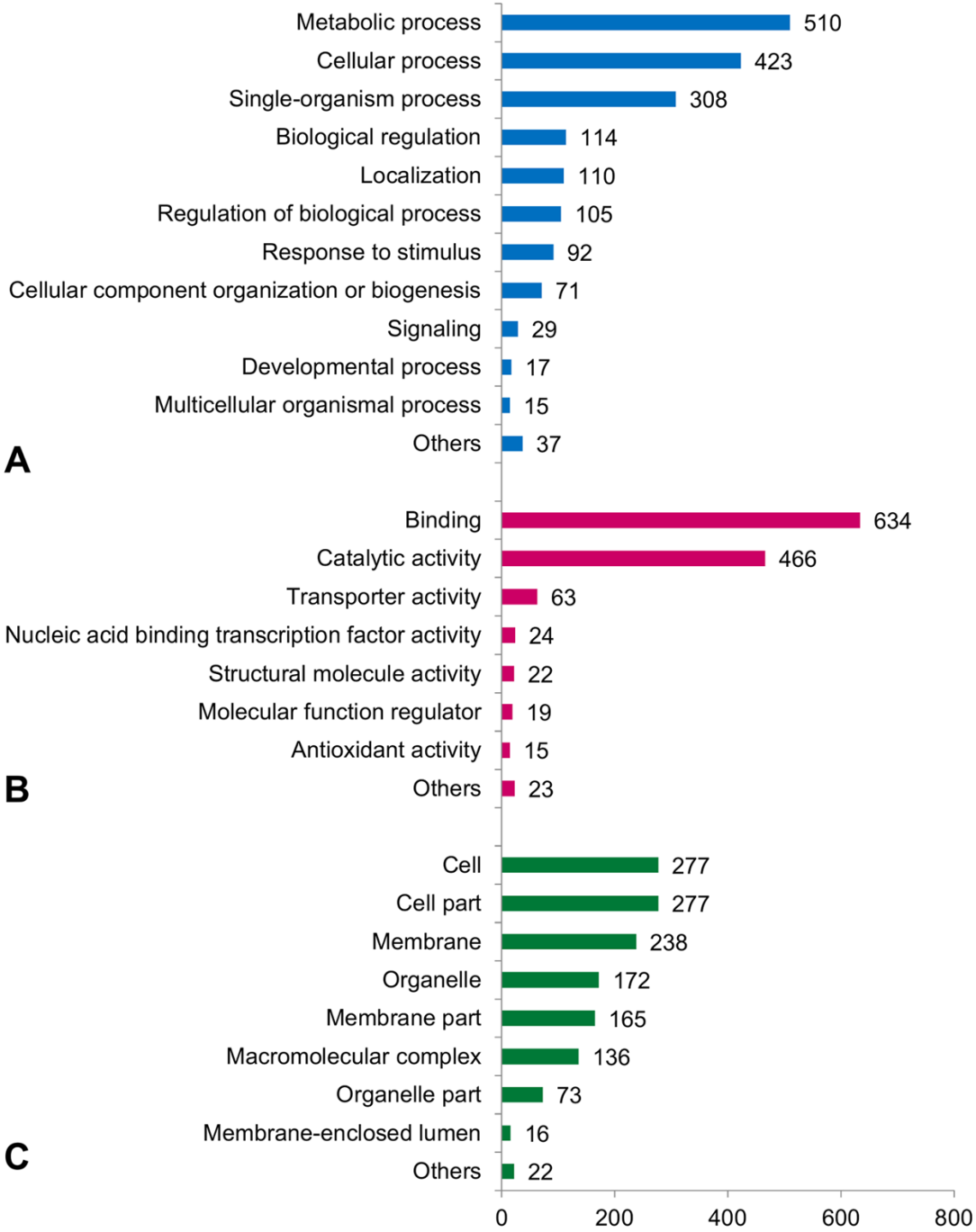
OPEN

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[1], Zirlane Portugal Costa[1], Luiz Augusto Cauz-Santos[1], Alina Carmen Egoávil Reátegui[1], Nathalie Rodde[2], Stéphane Cauet[2], Marcelo Carnier Dornelas[3], Philippe Leroy[4], Alessandro de Mello Varani[5], Hélène Bergès[2] & Maria Lucia Carneiro Vieira[1]

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 *Populus 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.

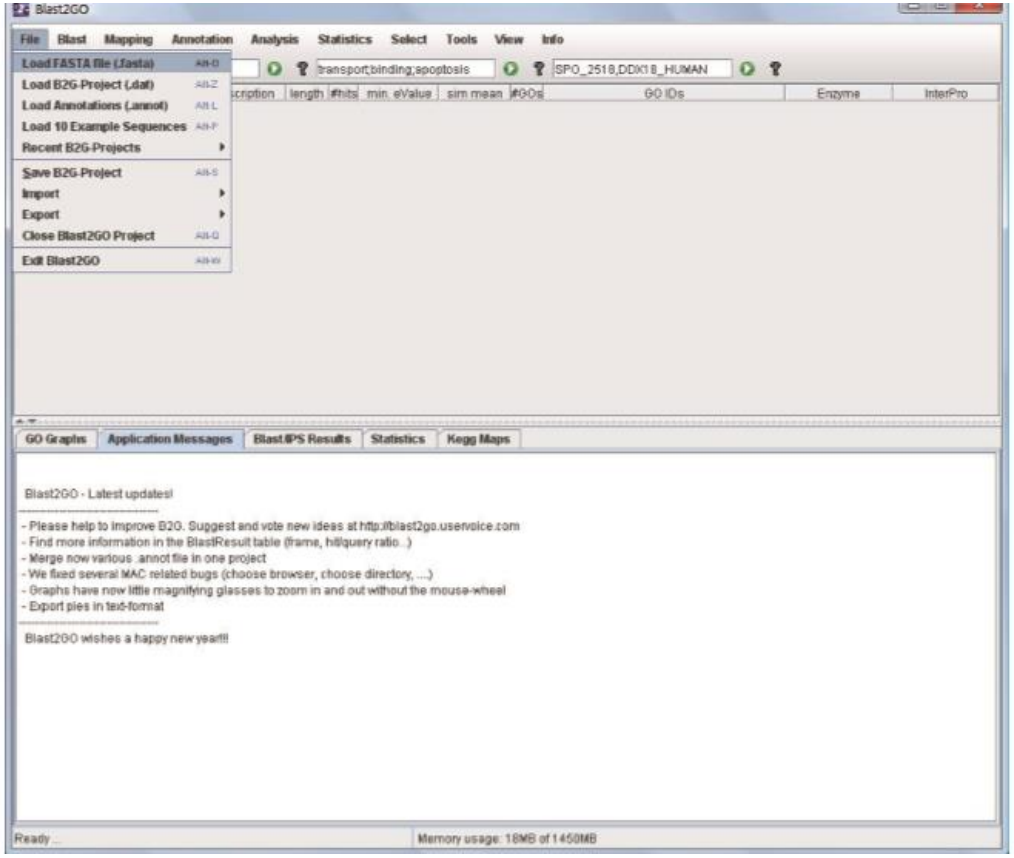


INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Anotação Funcional

blast2go

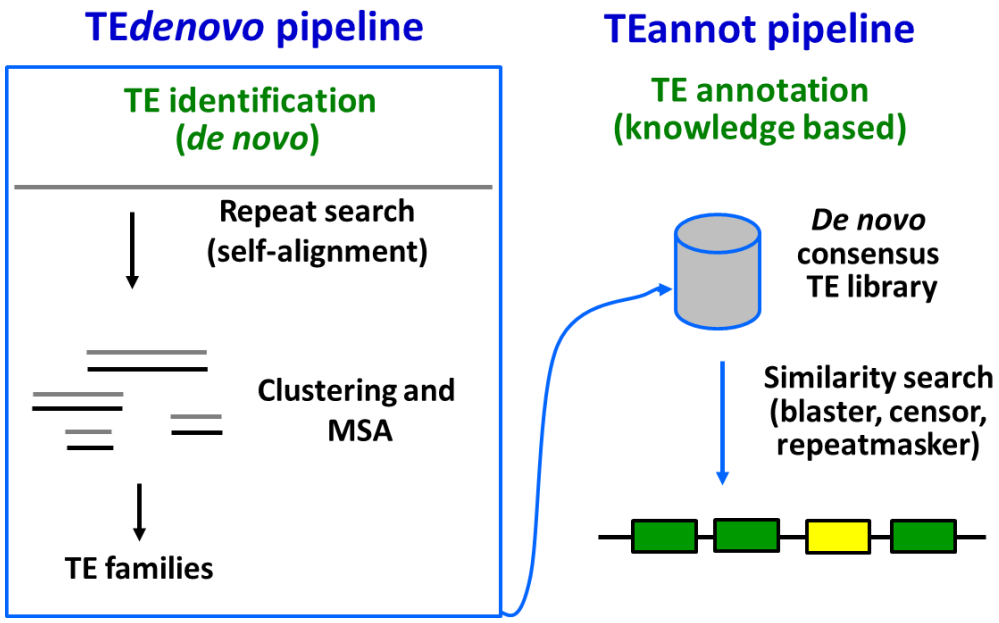
✓ Anotação automática



INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Anotação de elementos de transposição

- Identificação e Classificação: REPET/PASTEC

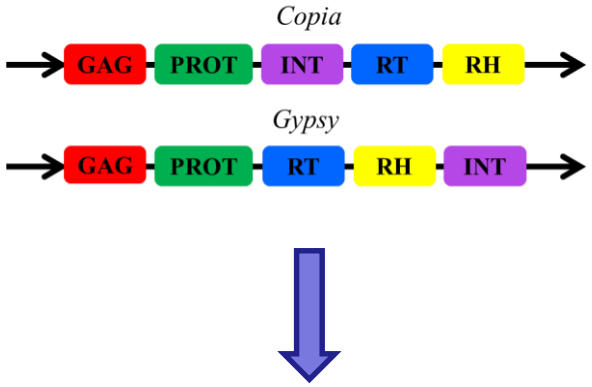


Métodos utilizados para a identificação de elementos de transposição

INTRODUÇÃO A MONTAGEM E ANOTAÇÃO DE GENOMAS

➢ Anotação de elementos de transposição

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

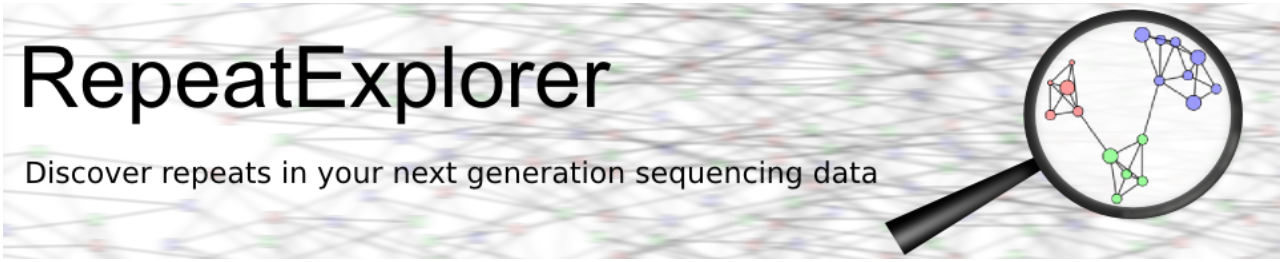


Linhagens evolutivas

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

giri REPBASE

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

GyDB
Gypsy Database 2.0

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*, Petr Novák, Nina Hoštáková and Jiří Macas

Abstract

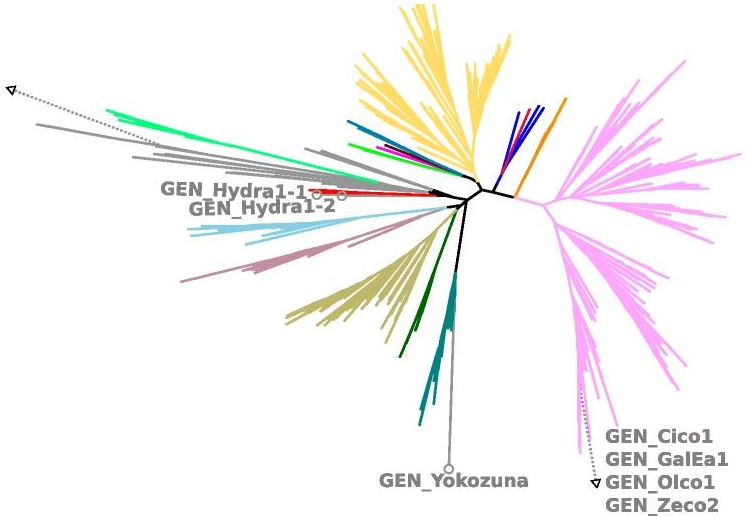
Background: Plant LTR-retrotransposons are classified into two superfamilies, Ty1/copia and Ty3/gypsy. They are further divided into an enormous number of families which are, due to the high diversity of their nucleotide sequences, usually specific to a single or a group of closely related species. Previous 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 numbers of elements. Furthermore, there is no reference database that allows for similarity based classification of LTR-retrotransposons.

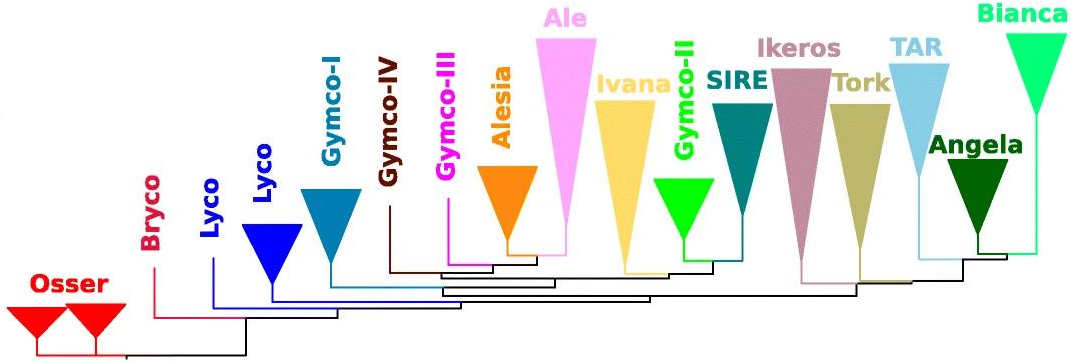
Results: We have assembled a database of retrotransposon encoded polyprotein domains sequences extracted from 5410 Ty1/copia elements and 8453 Ty3/gypsy elements sampled from 80 species representing major groups of green plants (Viridiplantae). Phylogenetic analysis of the three most conserved polyprotein 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-elixir.cerit-sc.cz/) or using a standalone version of REXdb that can be downloaded seaparately 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

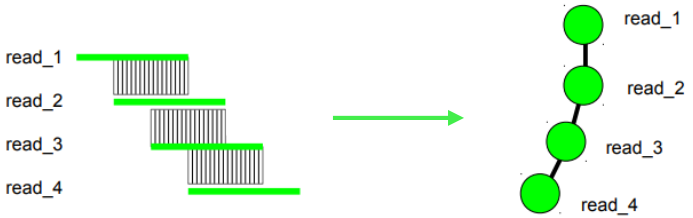




ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢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 GATTTGACGGACACACTAACTAGCTA
>0002_f ACTCATTTGGACTTAACTTTGATAAT
>0002_r TATGTTGAAAAATTGAATTTCGGGAC
>0003_f TGACATTTGTGAACGTTAATGTTCAA
>0003_r TATTGAAATACTGGACACAAATTGGA
```

2. Cálculo da quantidade de sequências para a cobertura desejada

$$\textbf{Genome Coverage} = \frac{(N° \; de \; reads \; x \; Tamanho \; das \; reads)}{Tamanho \; do \; genoma}$$

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

$$0{,}25 = \frac{(N° \; de \; reads \; x \; 100)}{1.000 \; x \; 10^6}$$

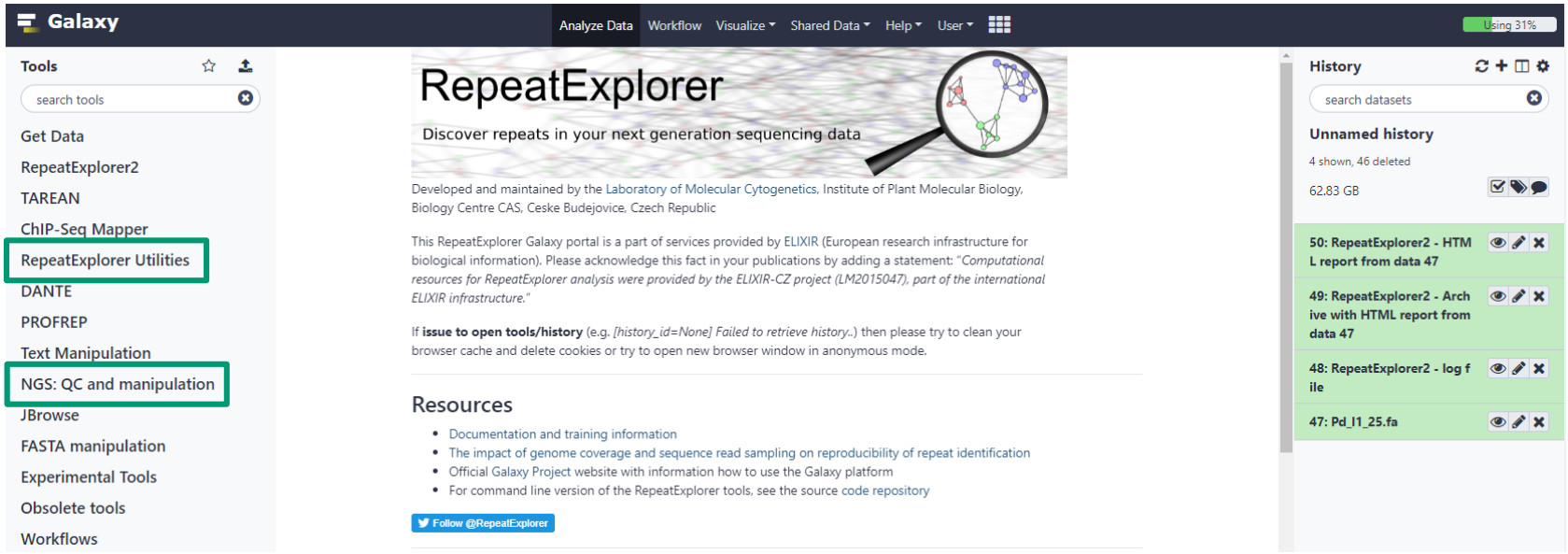
$$N° \; de \; reads = 2.500.000$$

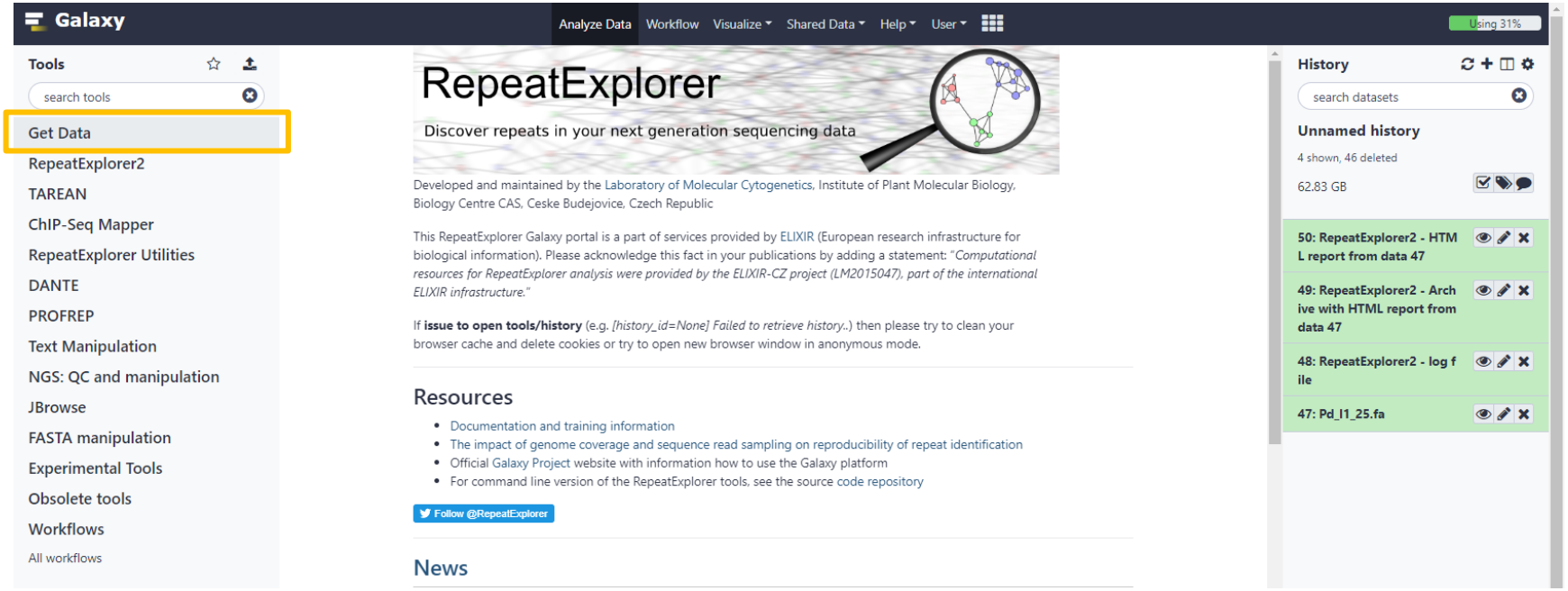
$$0{,}25 \; x \; 1.000 \; x \; 10^6 = 250.000.000 \text{ pb}$$

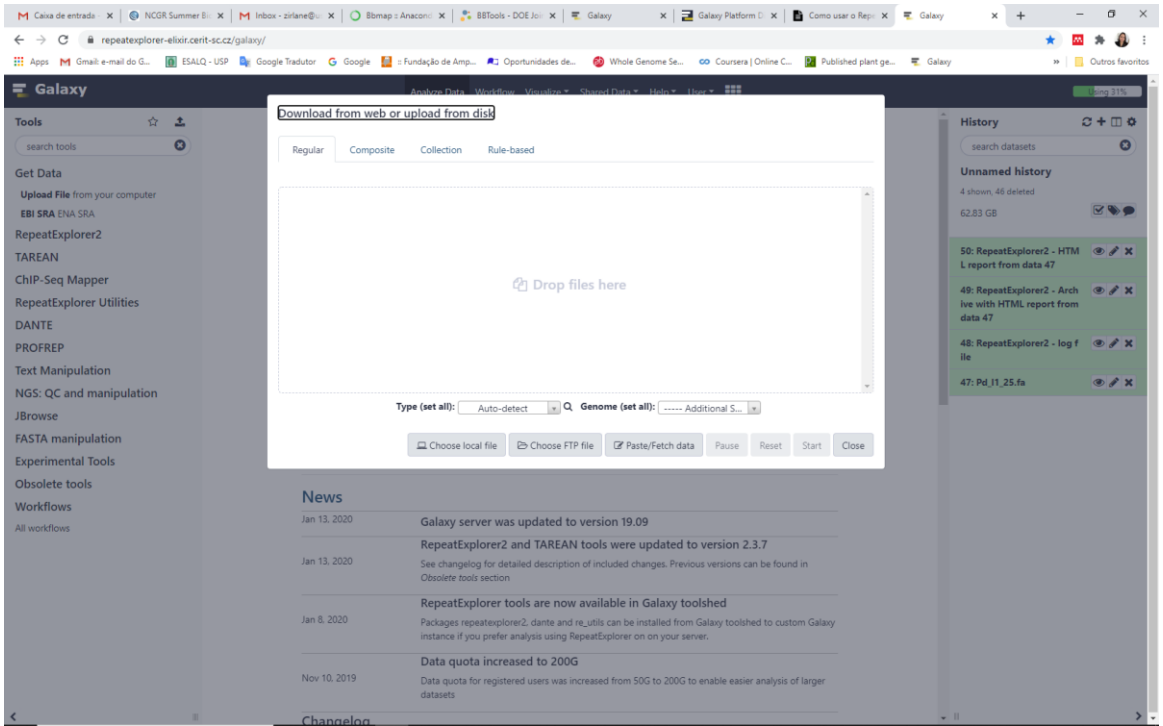
ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

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

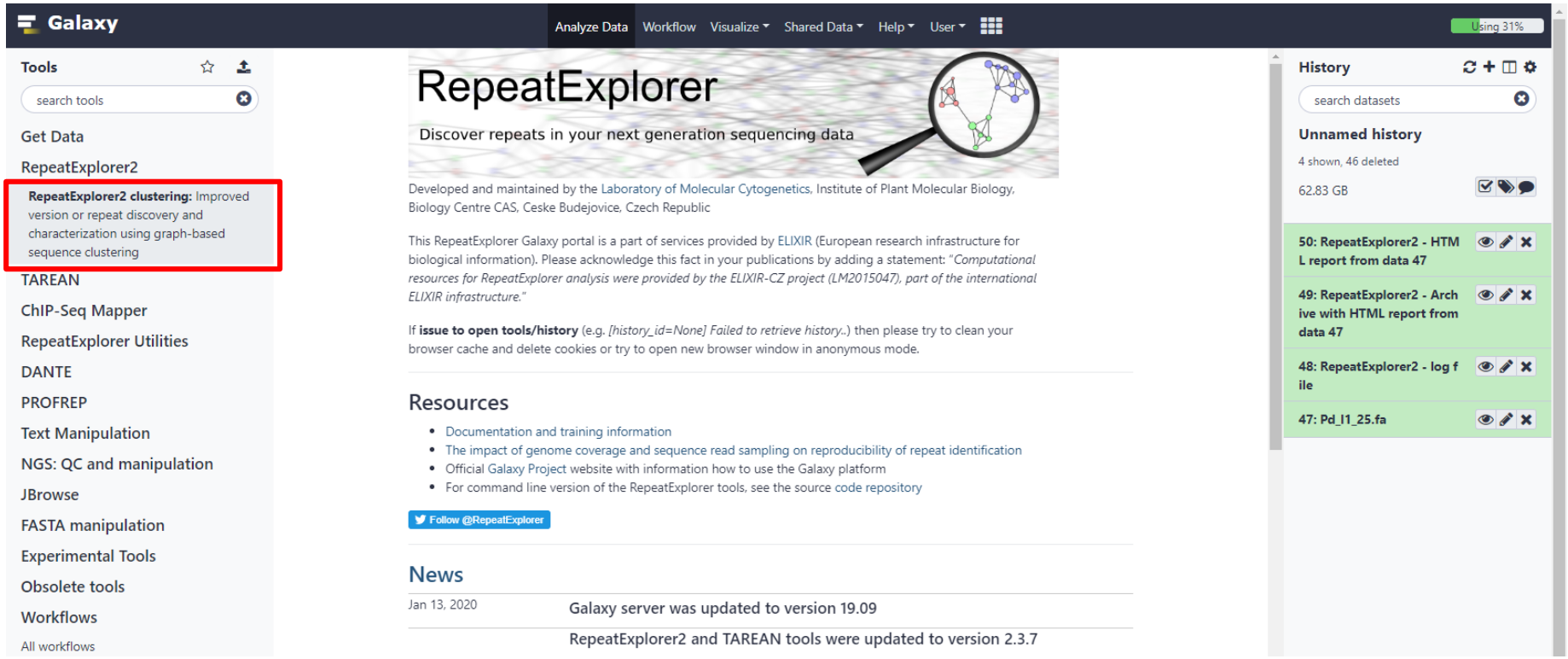



Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 31%
Tools
search tools
Get Data
RepeatExplorer2
TAREAN
ChIP-Seq Mapper
RepeatExplorer Utilities
DANTE
PROFREP
Text Manipulation
NGS: QC and manipulation
JBrowse
FASTA manipulation
Experimental Tools
Obsolete tools
Workflows
All workflows
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 - log file
47: Pd_l1_25.fa


Galaxy
Download from web or upload from disk
Regular
Composite
Collection
Rule-based
Drop files here
Type (set all):
Auto-detect
Genome (set all):
Choose local file
Choose FTP file
Paste/Fetch data
Pause
Reset
Start
Close
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
Jan 8, 2020
RepeatExplorer tools are now available in Galaxy toolshed
Nov 10, 2019
Data quota increased to 200G
Changelog

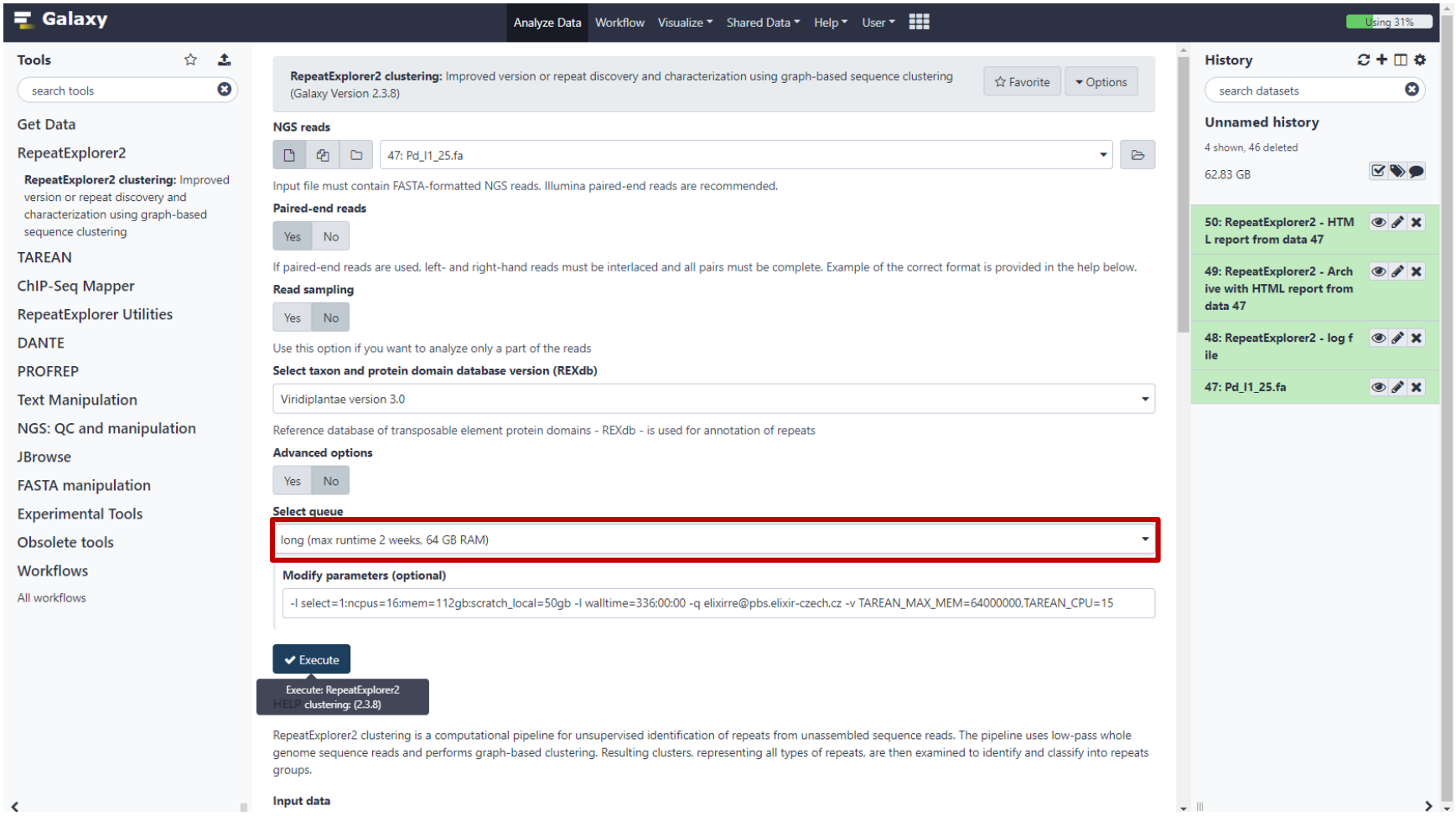
ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

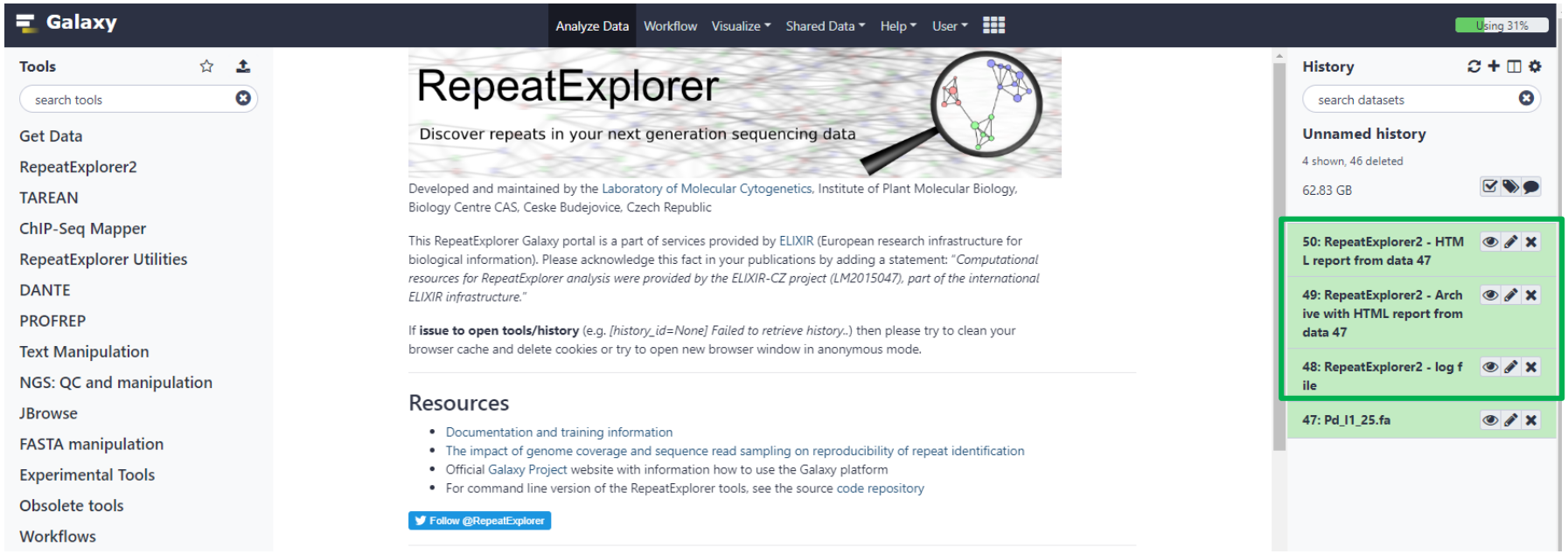
➢RepeatExplorer



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

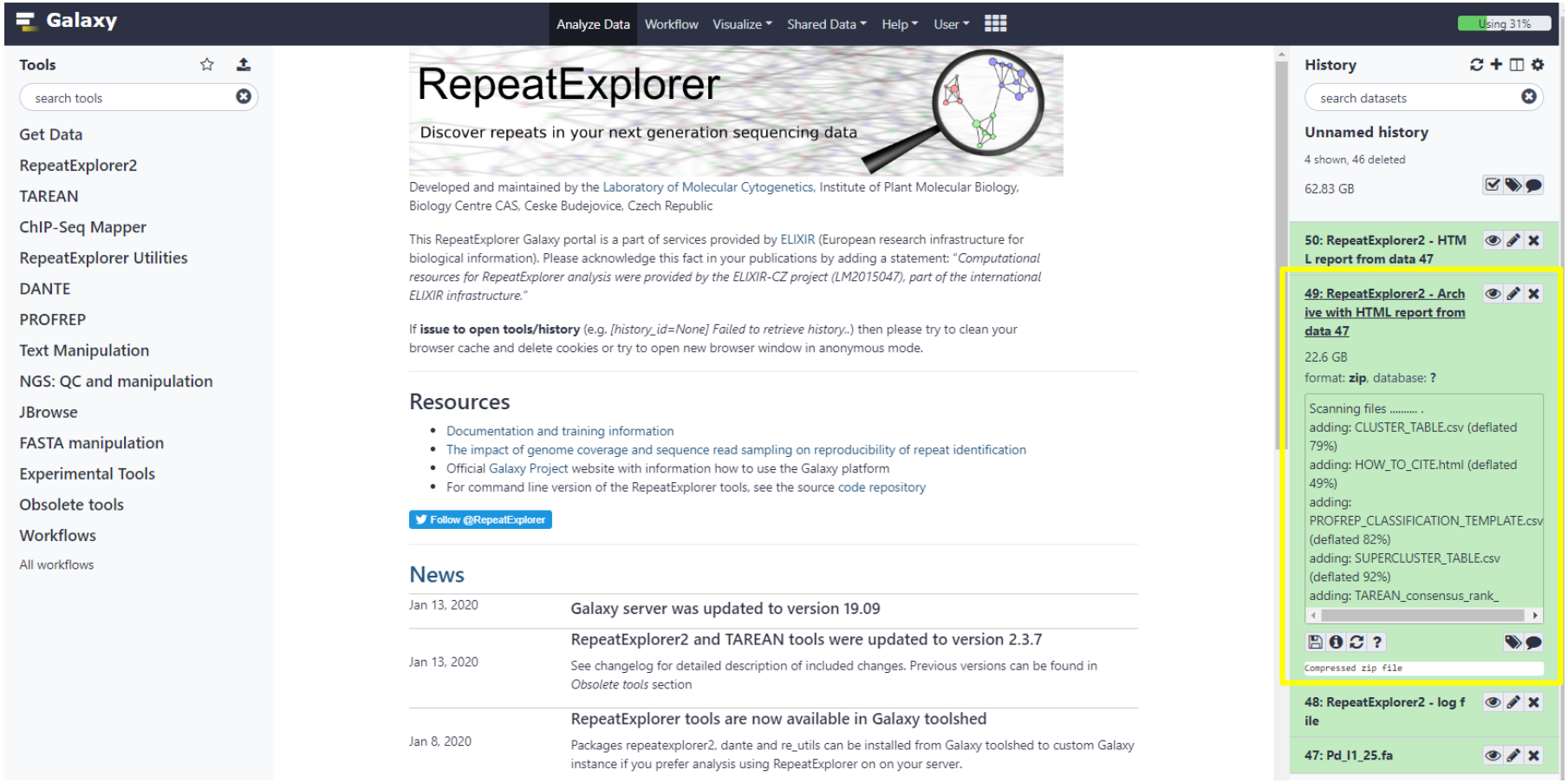
✓ **Resultados**



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

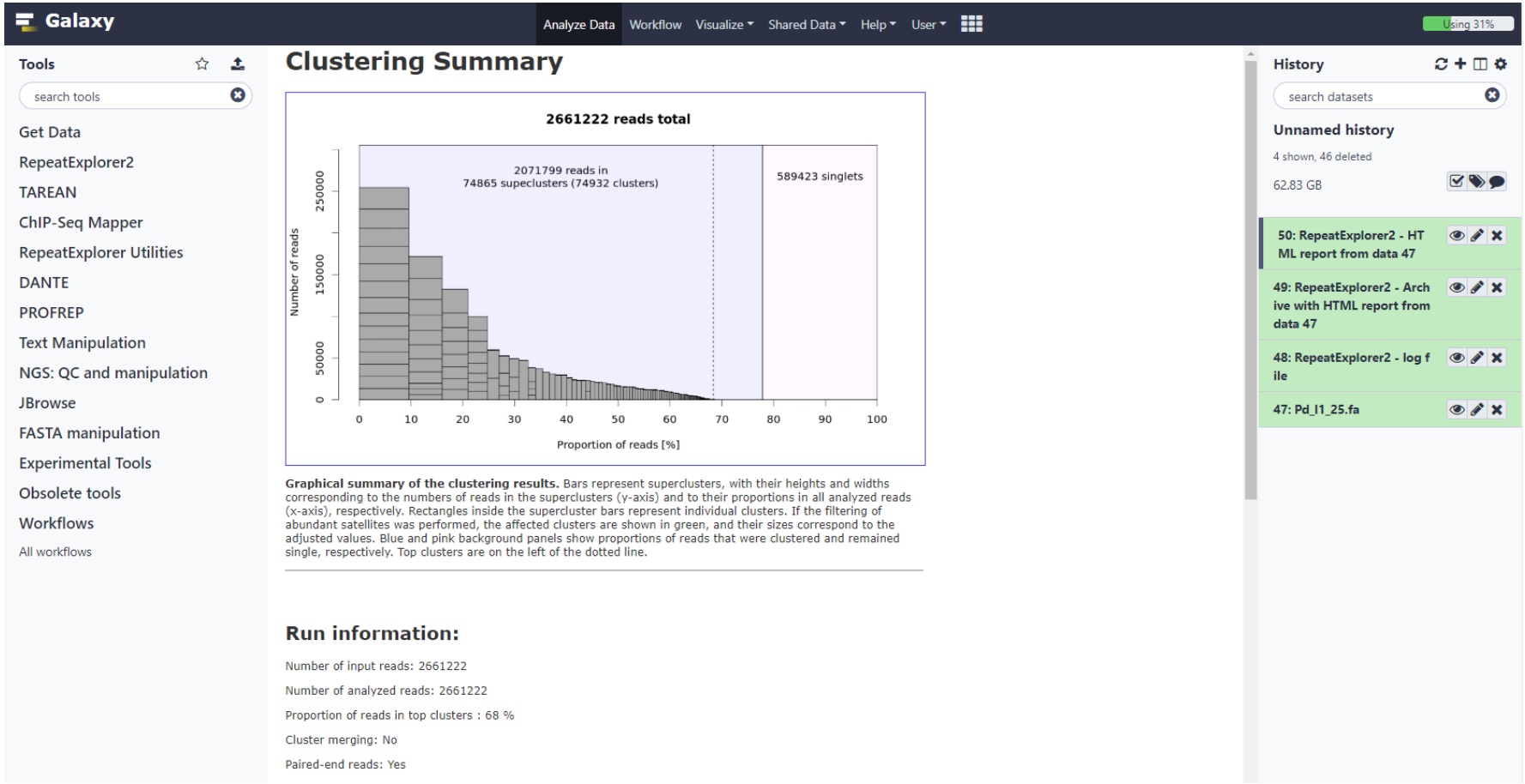
✓ **Resultados**



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

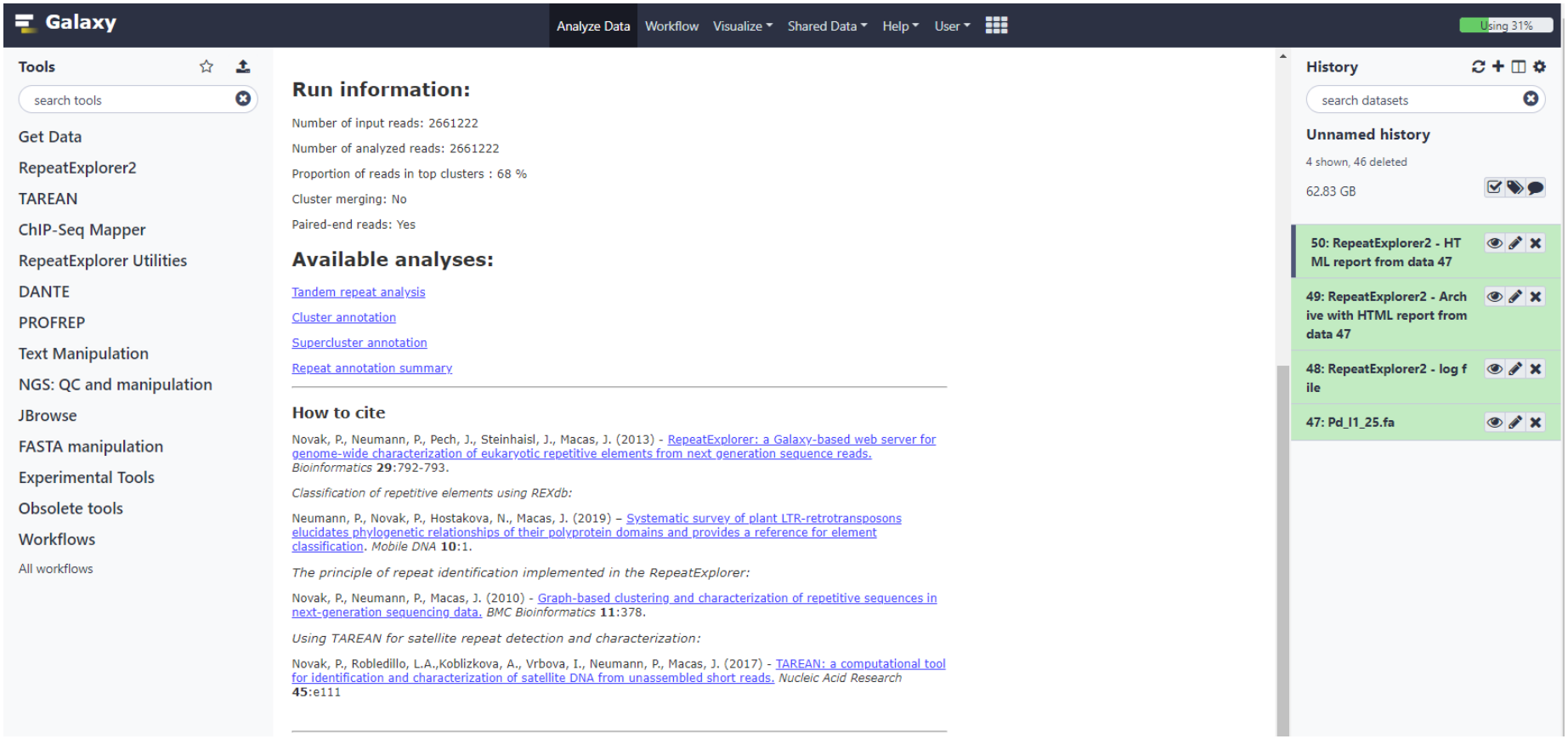
✓ **Resultados**



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

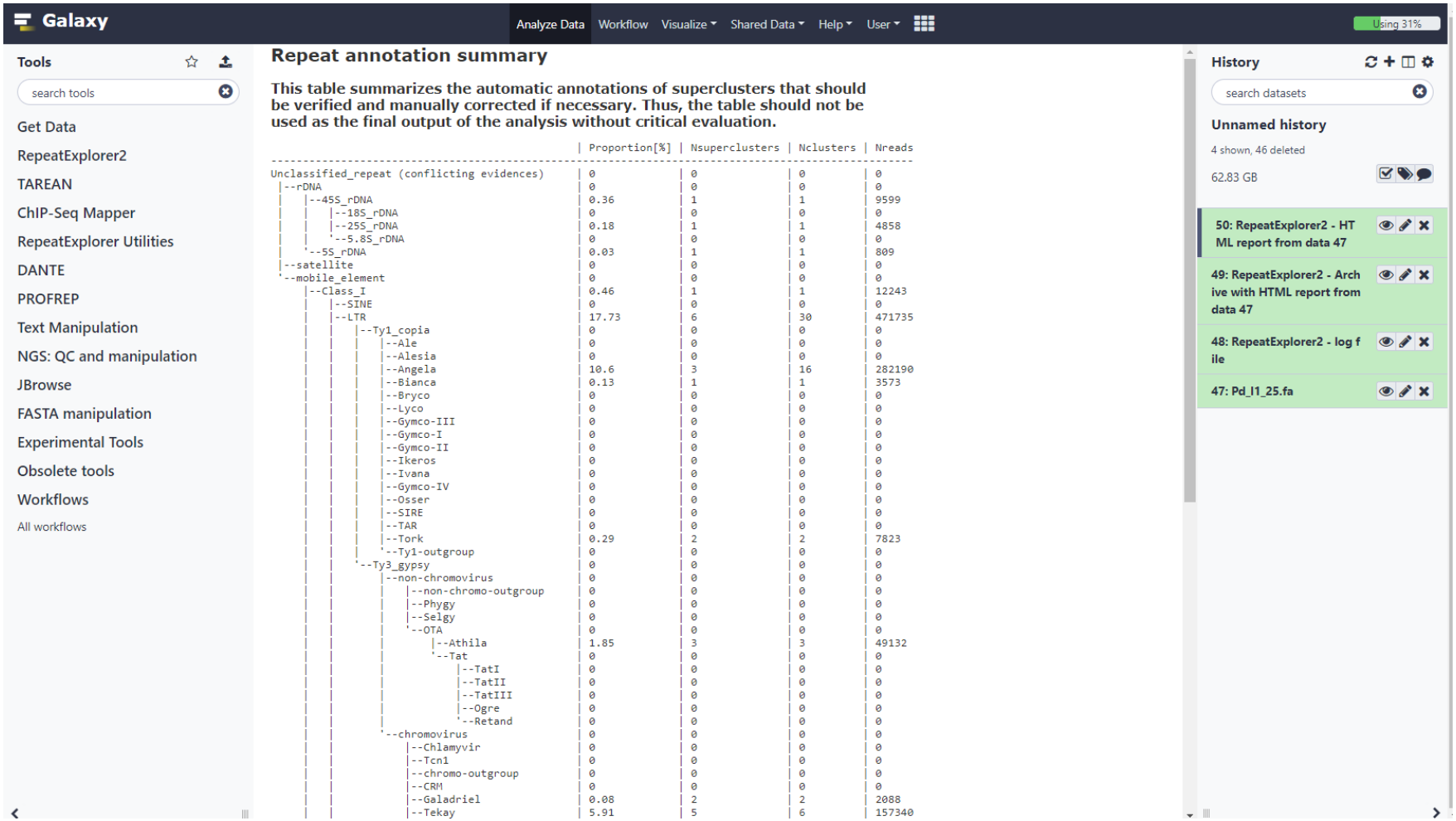
✓ Resultados



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

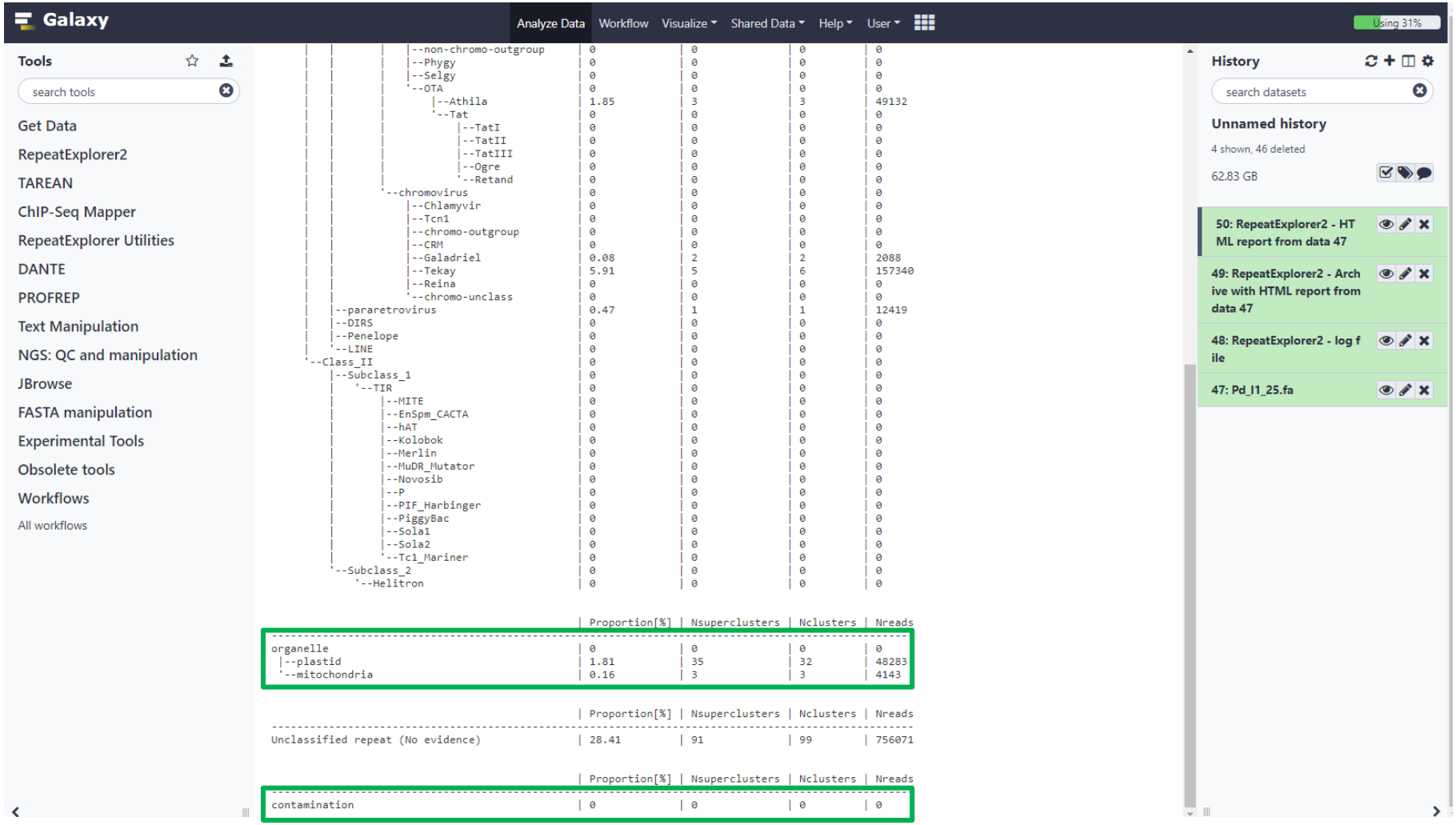
✓ **Resultados: Sumário dos repeats anotados**



ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

✓ **Resultados: Sumário dos repeats anotados**



|--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
|--Reina	0	0	0	0
'--chromo-unclass	0	0	0	0
|--pararetrovirus	0.47	1	1	12419
|--DIRS	0	0	0	0
|--Penelope	0	0	0	0
'--LINE	0	0	0	0
'--Class_II	0	0	0	0
|--Subclass_1	0	0	0	0
'--TIR	0	0	0	0
|--MITE	0	0	0	0
|--EnSpm_CACTA	0	0	0	0
|--hAT	0	0	0	0
|--Kolobok	0	0	0	0
|--Merlin	0	0	0	0
|--MuDR_Mutator	0	0	0	0
|--Novosib	0	0	0	0
|--P	0	0	0	0
|--PIF_Harbinger	0	0	0	0
|--PiggyBac	0	0	0	0
|--Sola1	0	0	0	0
|--Sola2	0	0	0	0
'--Tc1_Mariner	0	0	0	0
'--Subclass_2	0	0	0	0
'--Helitron	0	0	0	0

	Proportion[%]	Nsuperclusters	Nclusters	Nreads
organelle	0	0	0	0
|--plastid	1.81	35	32	48283
'--mitochondria	0.16	3	3	4143

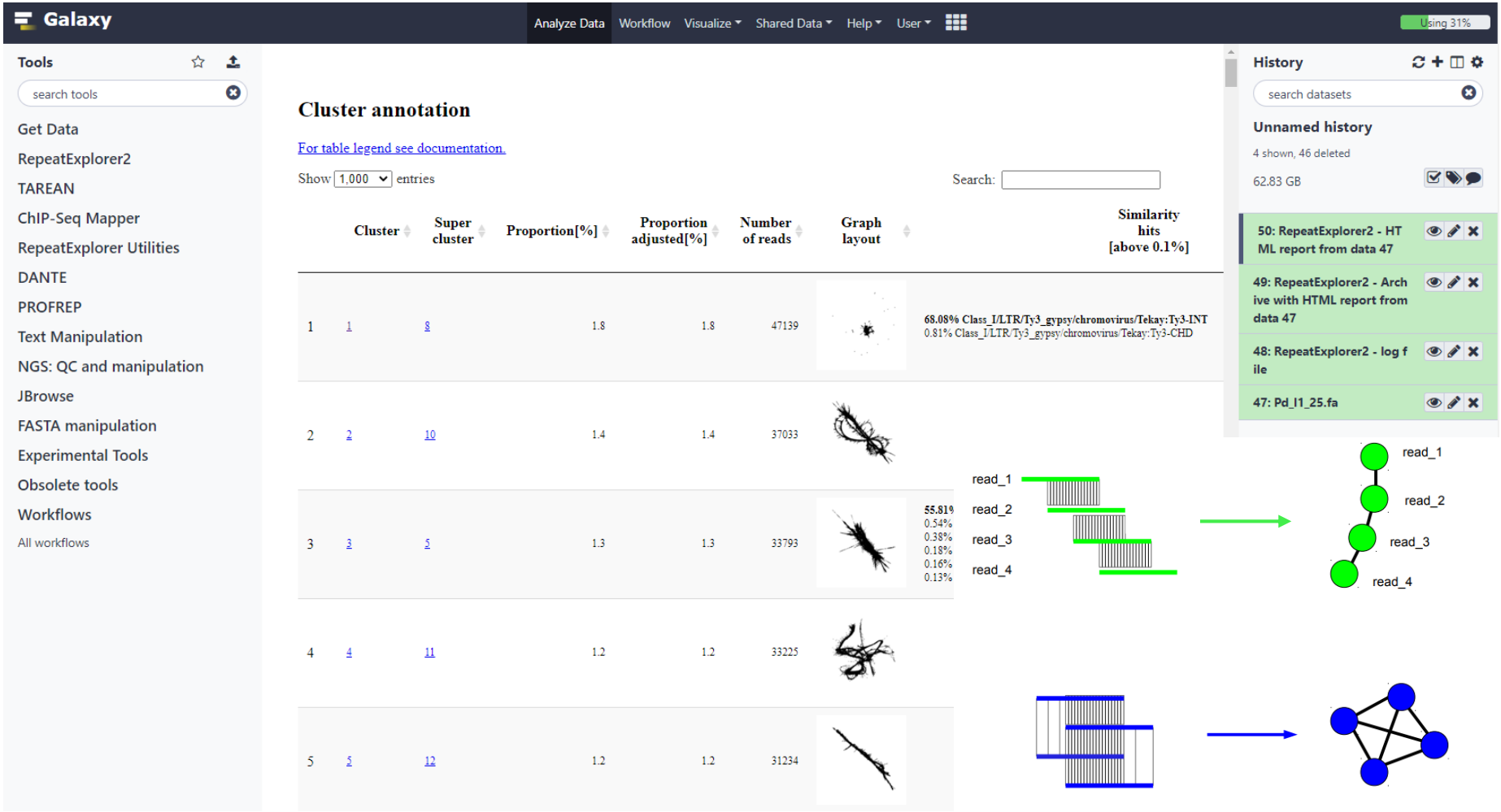
	Proportion[%]	Nsuperclusters	Nclusters	Nreads
Unclassified repeat (No evidence)	28.41	91	99	756071

	Proportion[%]	Nsuperclusters	Nclusters	Nreads
contamination	0	0	0	0

ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢RepeatExplorer

✓ **Resultados: anotação dos clusters**

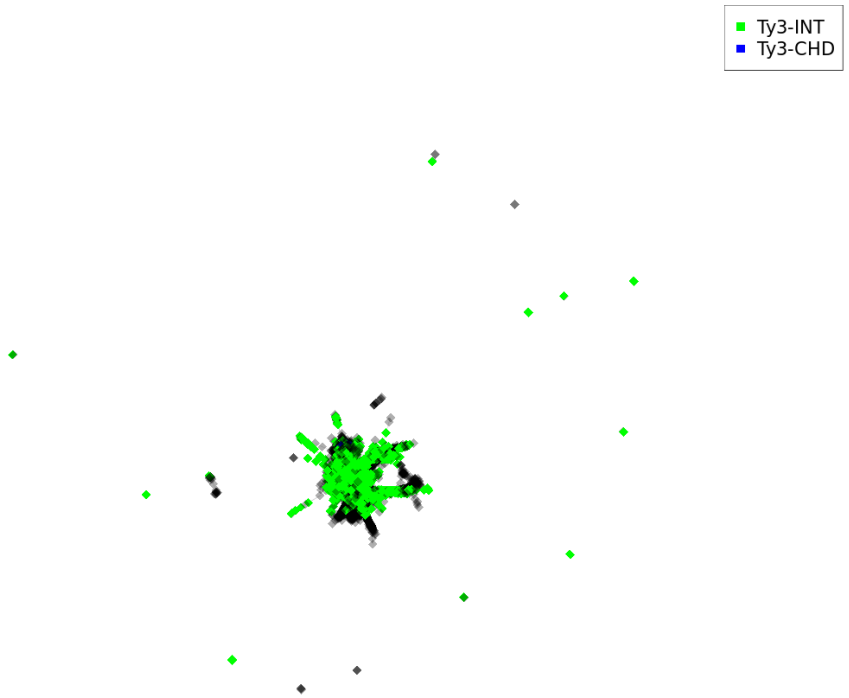


ANÁLISE DOS DADOS ILLUMINA PARA IDENTIFICAR REPEATS NO GENOMA NUCLEAR

➢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

G42

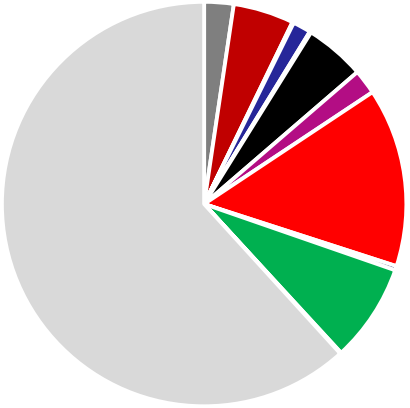
	A	B	C	D	E	F	G	H	I
7	Cluster	Supercluster	Size	Size_adjusted	Automatic_annotation	TAREAN_annotation	Final_annotation		
8	1	8	47139	47139	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	Other			
9	2	10	37033	37033	All	Other			
10	3	5	33793	33793	All/repeat/mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	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

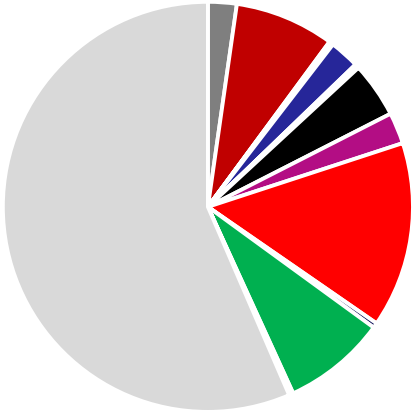
➢RepeatExplorer

✓ **Testes com diferentes coberturas**

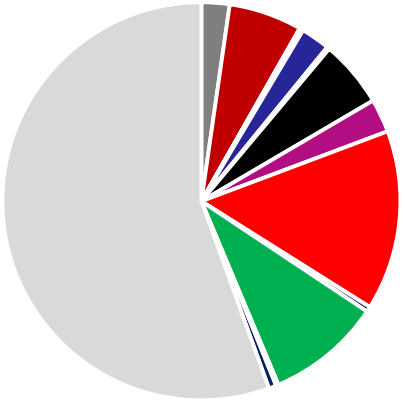
Cobertura de 0,1: 35,30% de TEs

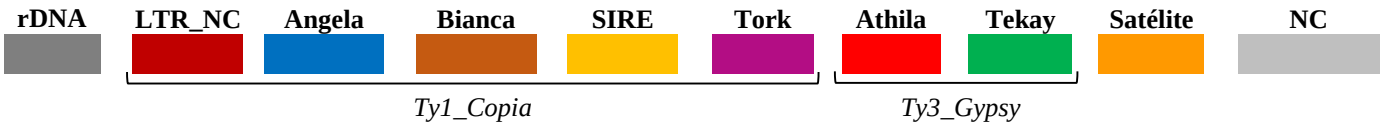


Cobertura de 0,25: 38,01% de TEs



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		
				Angela	**11,64**
				Bianca	0,19
				Ivana	0,02
				SIRE	1,62
				Tork	0,16
			Ty3_gypsy		
				Athila	**9,92**
				Tekay	**5,96**
Unclassified repeat					42,23
					74,78

✓ **2n = 18**

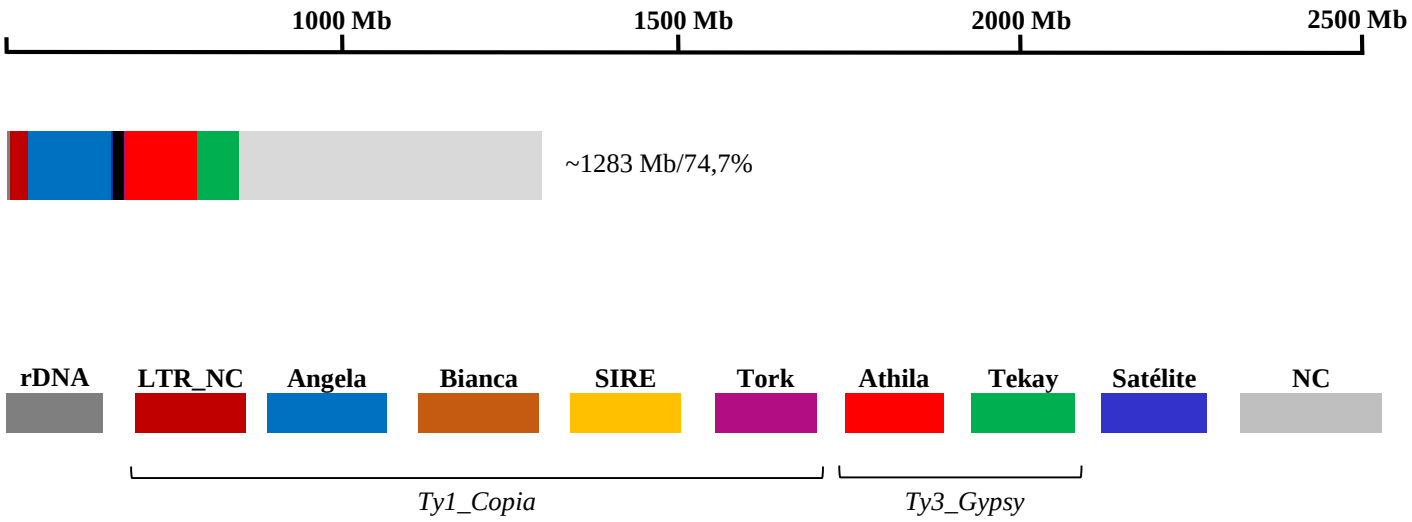
✓ **Genoma: 1283,80 Mb**

✓ **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

PARTE II

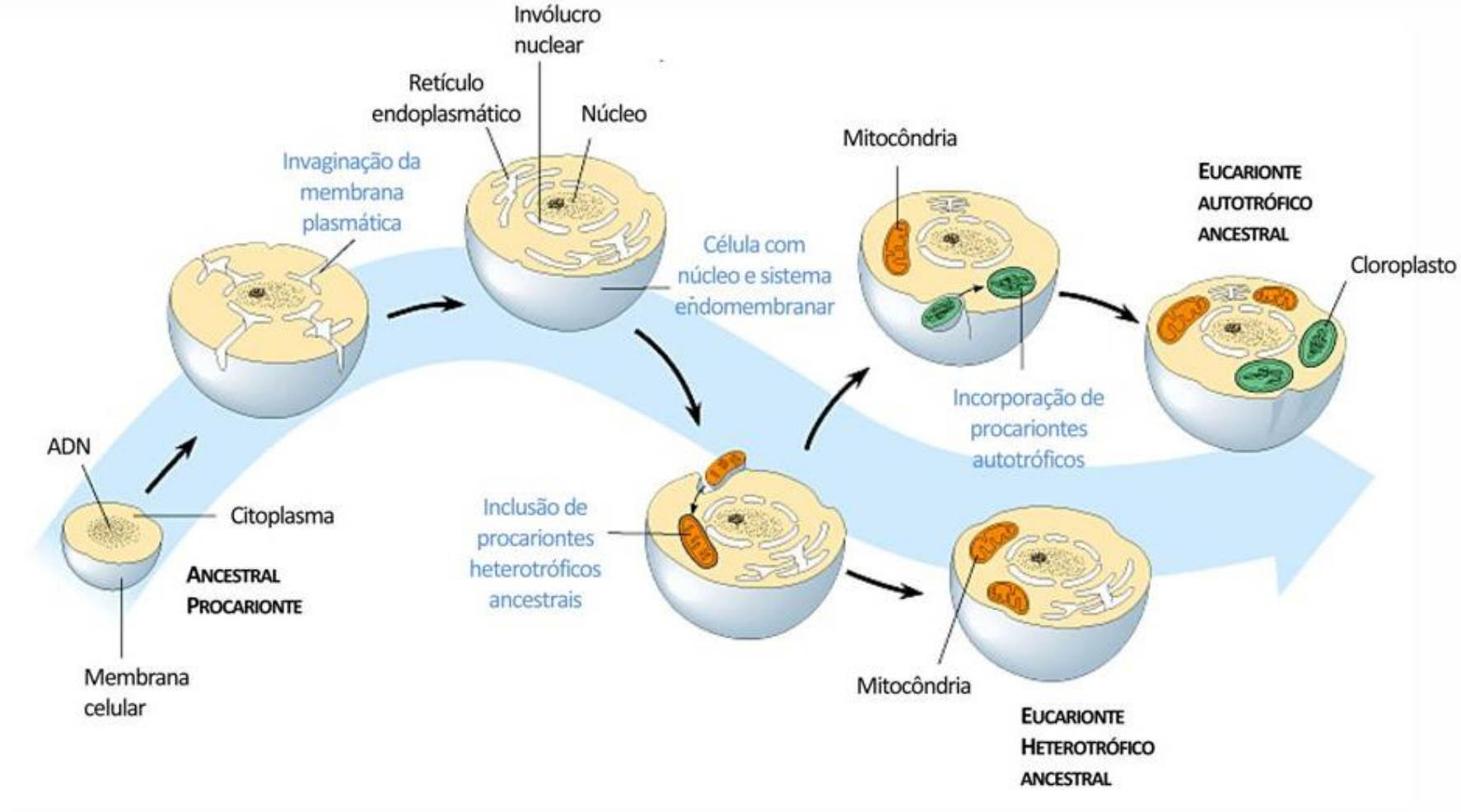
UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA 'LUIZ DE QUEIROZ'
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS
Laboratório de Genética Molecular de Plantas Cultivadas

Análise de genomas cloroplastidiais

Luiz Augusto Cauz dos Santos

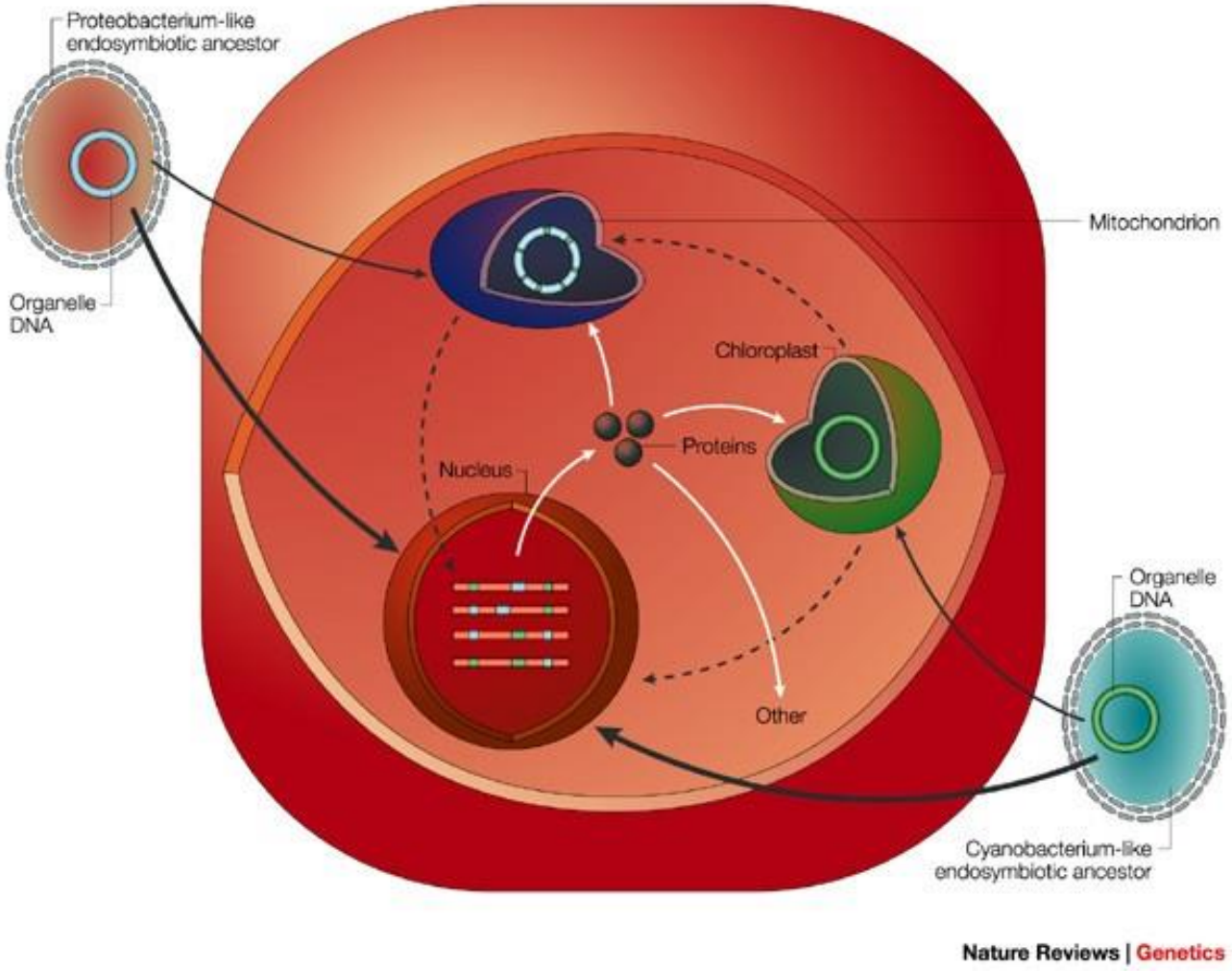
➢ Origem dos genomas extranucleares

A teoria da endossimbiose



➢ Origem dos genomas extranucleares

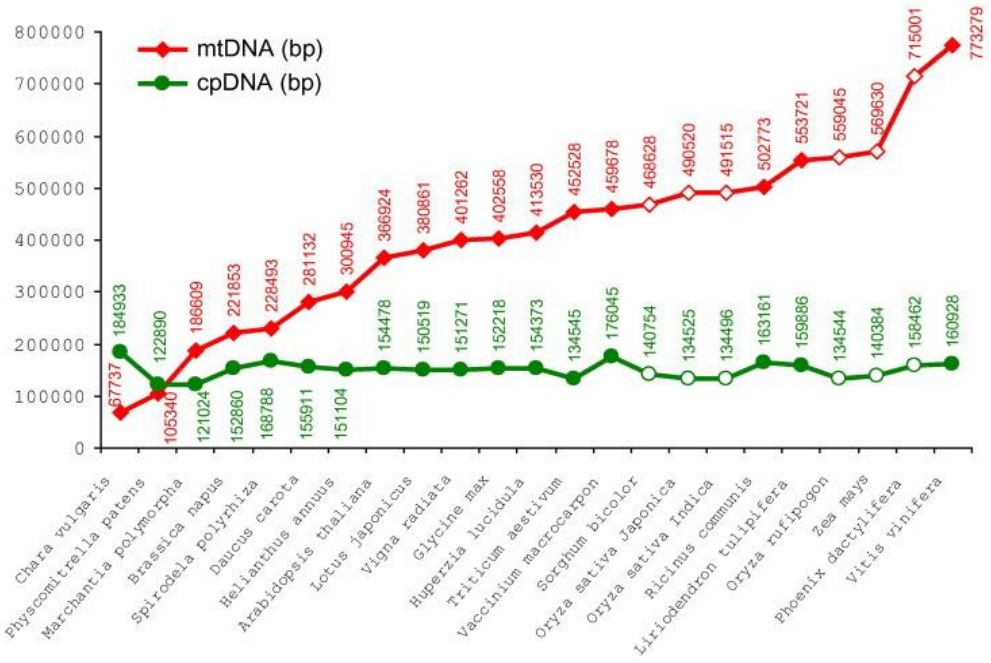
A transferência de material genético entre os genomas

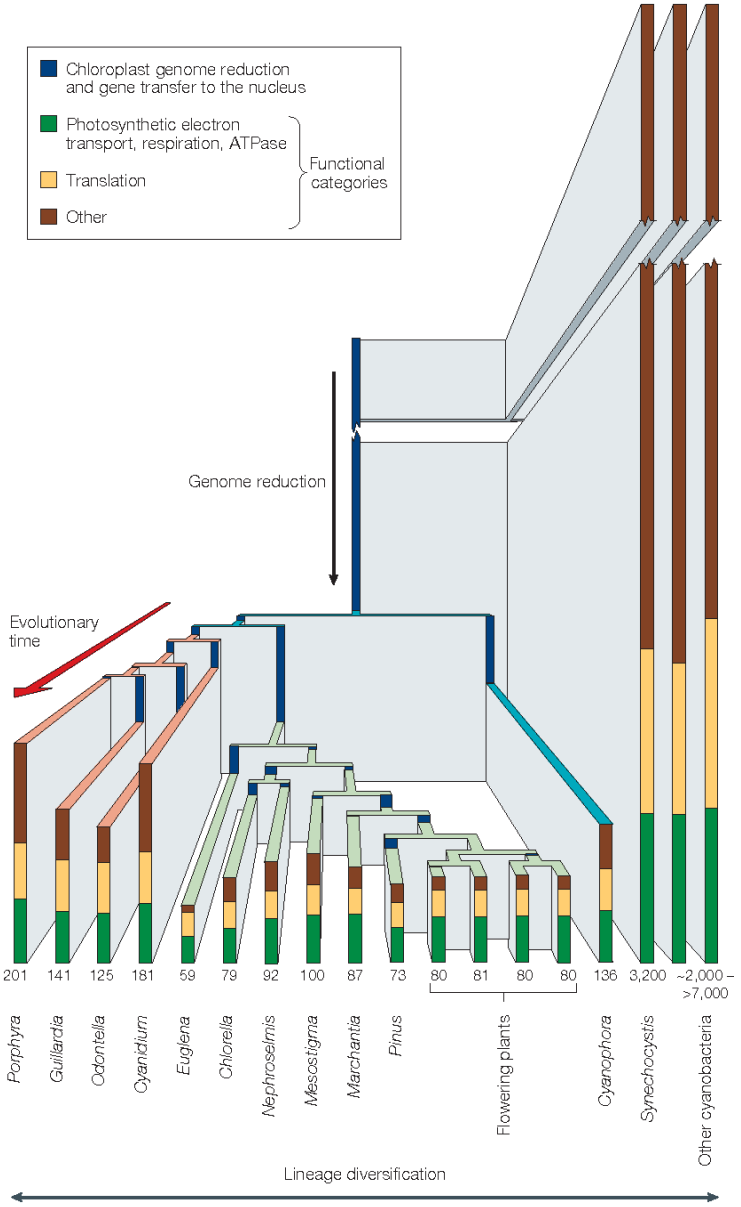


➢ Origem dos genomas extranucleares

Redução do genoma cloroplastidial

Variação de tamanho dos genomas mitocondriais





GENOMA MITOCONDRIAL

Moléculas circulares

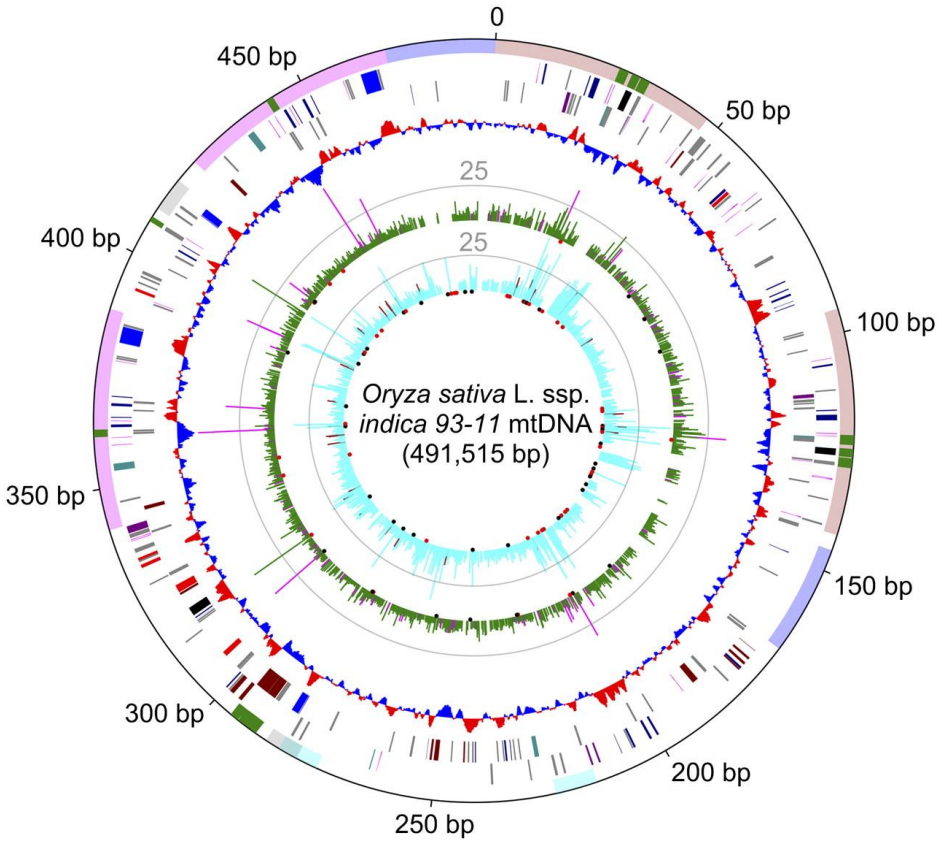
Tamanho entre 69.2kb a 521kb (Algas)
200 a 750kb (Angiospermas)

Múltiplas copias 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

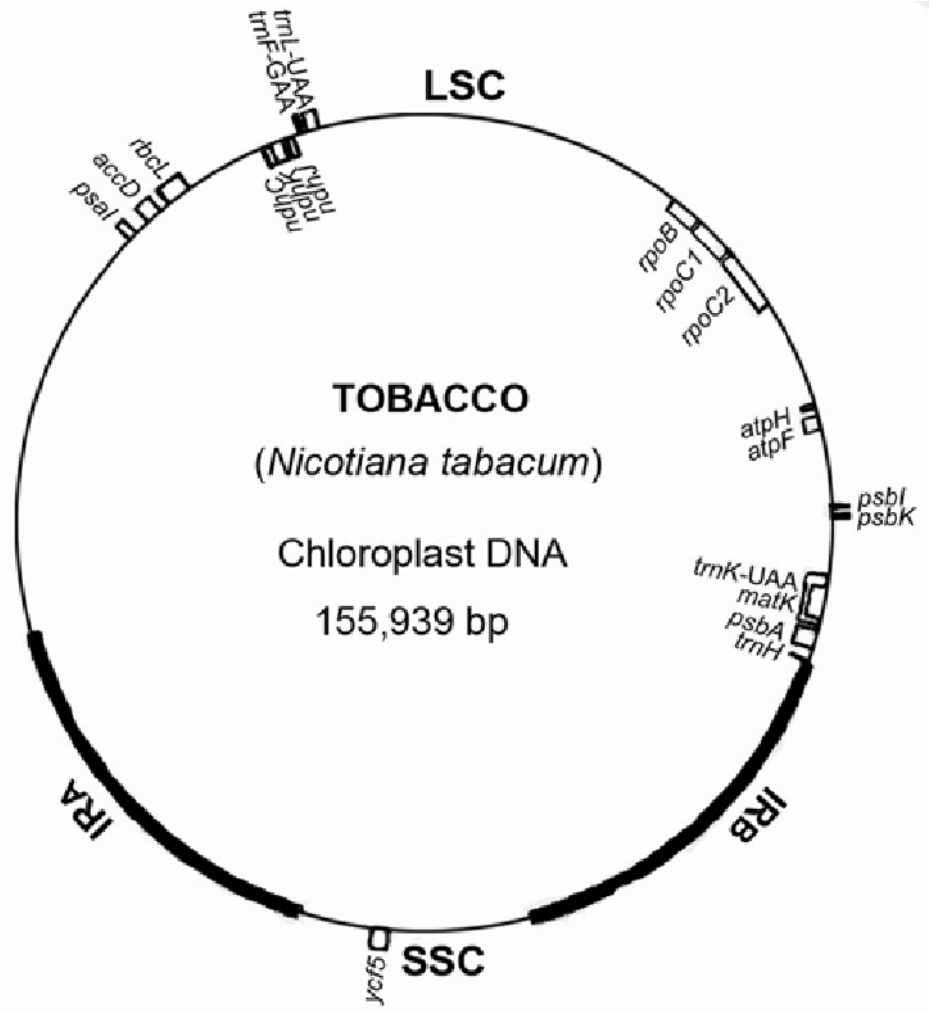
Estrutura Quadripartida

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

Cauz-Santos *et al.,* 2020. Unpublished

- **Os genomas cp são altamente conservados em Angiospermas**

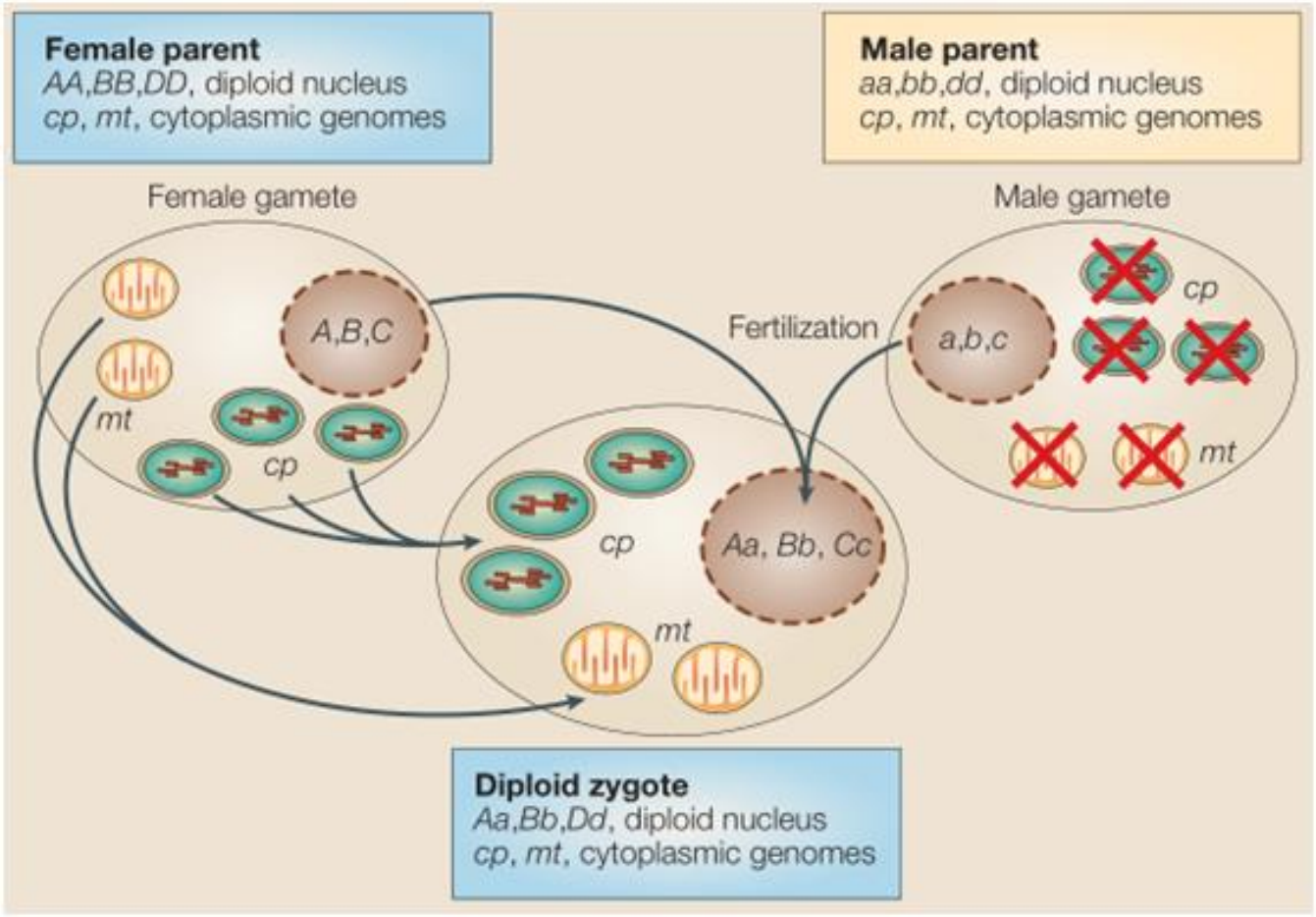
- **Exemplo na ordem Malpighiales:** Oito regiões adjacentes de sintenia nas sequências dos genomas cp: as regiões incluem genes, tRNAs e rRNAs

Alinhamento de 12 genomas cp completos de Malpighiales

Cauz-Santos *et al.,* 2017. Frontiers in Plant Science

➢ Herança citoplasmática

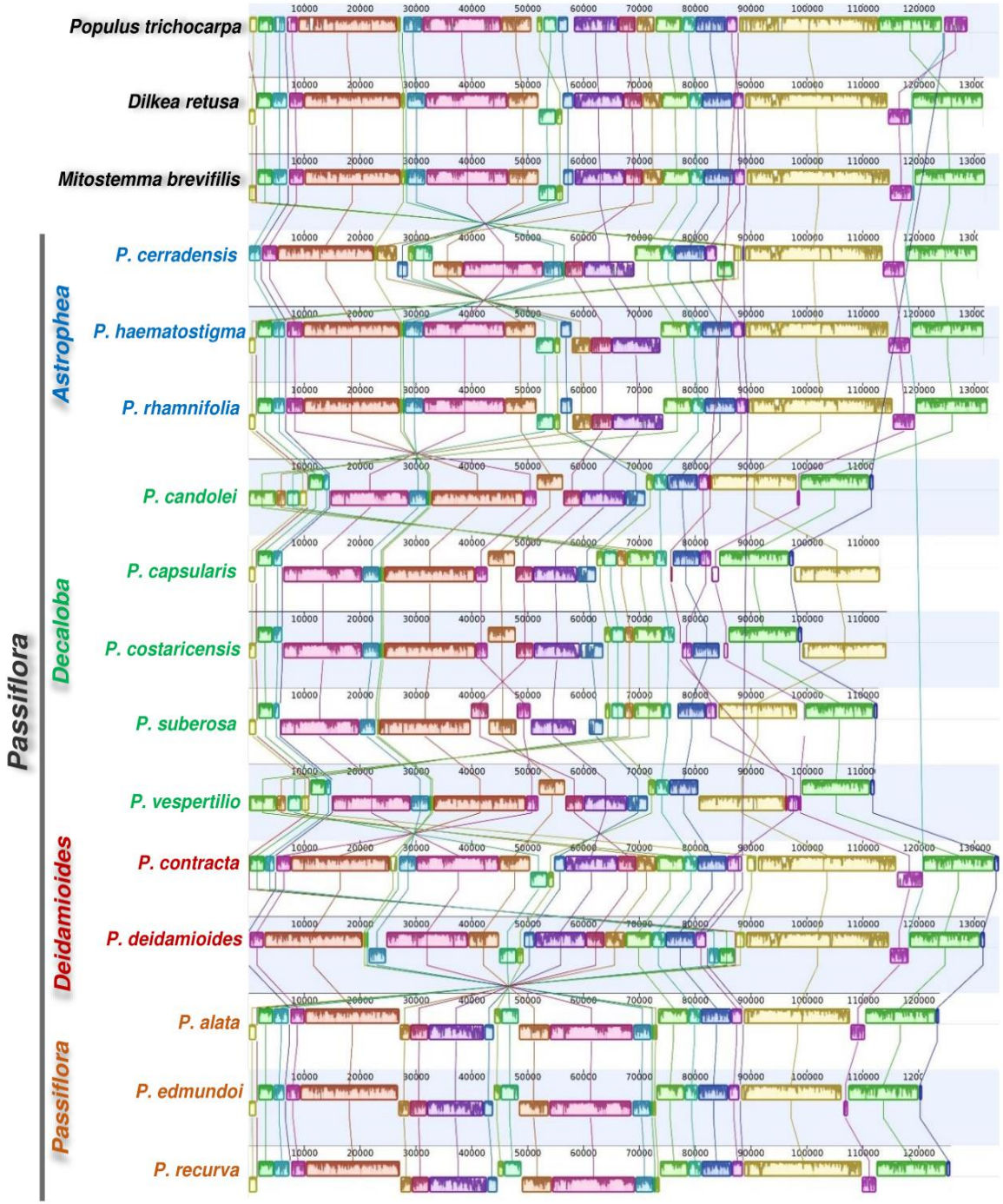
Herança uniparental de cloroplastos e mitocôndrias



Cerca de 20% das angiospermas poderiam apresentar herança biparental

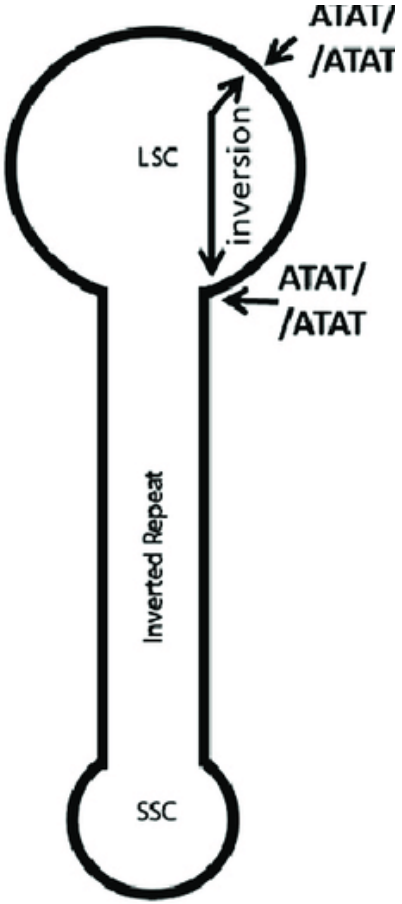
➢ Rearranjos

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

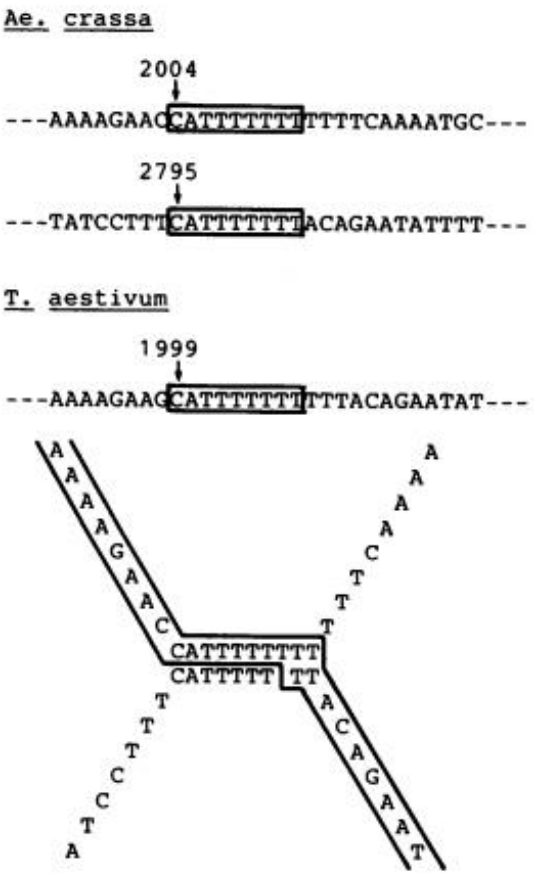


➢ Rearranjos

Recombinação homologa intramolecular



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



Ogihara *et al.,* 1988. PNAS

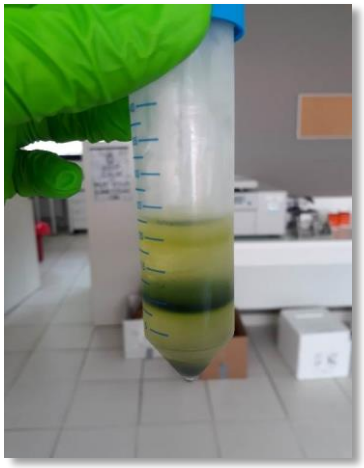
AMOSTRAGEM

Material vegetal

Enriquecimento de cpDNA
Isolamento de Cloroplastos
(Gradientes Percoll ou sacarose)

Extraçao de DNA total
Cerca de 1-3% de cpDNA

Sequenciamento
Ao menos 2GB de dados por amostra



Isolamento de cloroplastos em gradiente de sacarose

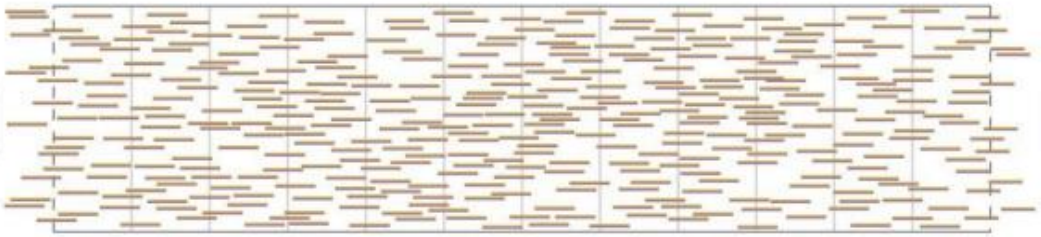
NextSeq 500 System Performance Parameters[§]

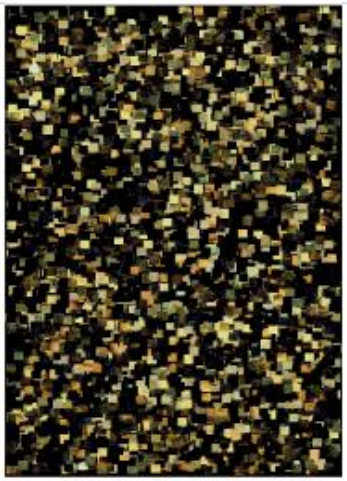
Flow Cell Configuration	Read Length (bp)	Output (Gb)	Run Time
High Output Flow Cell	2 × 150	100–120	29 hours
Up to 400 M single reads	2 × 75	50–60	18 hours
Up to 800 M paired-end reads	1 × 75	25–30	11 hours
Mid Output Flow Cell	2 × 150	32–39	26 hours
Up to 130 M single reads			
Up to 260 M paired-end reads	2 × 75	16–19	15 hours

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

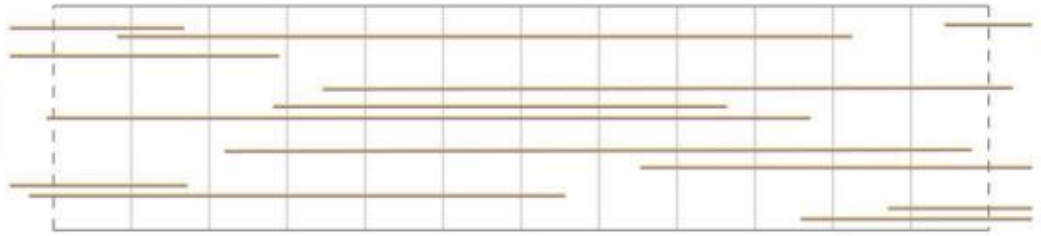
➢ Montagem de Genomas

Short Reads





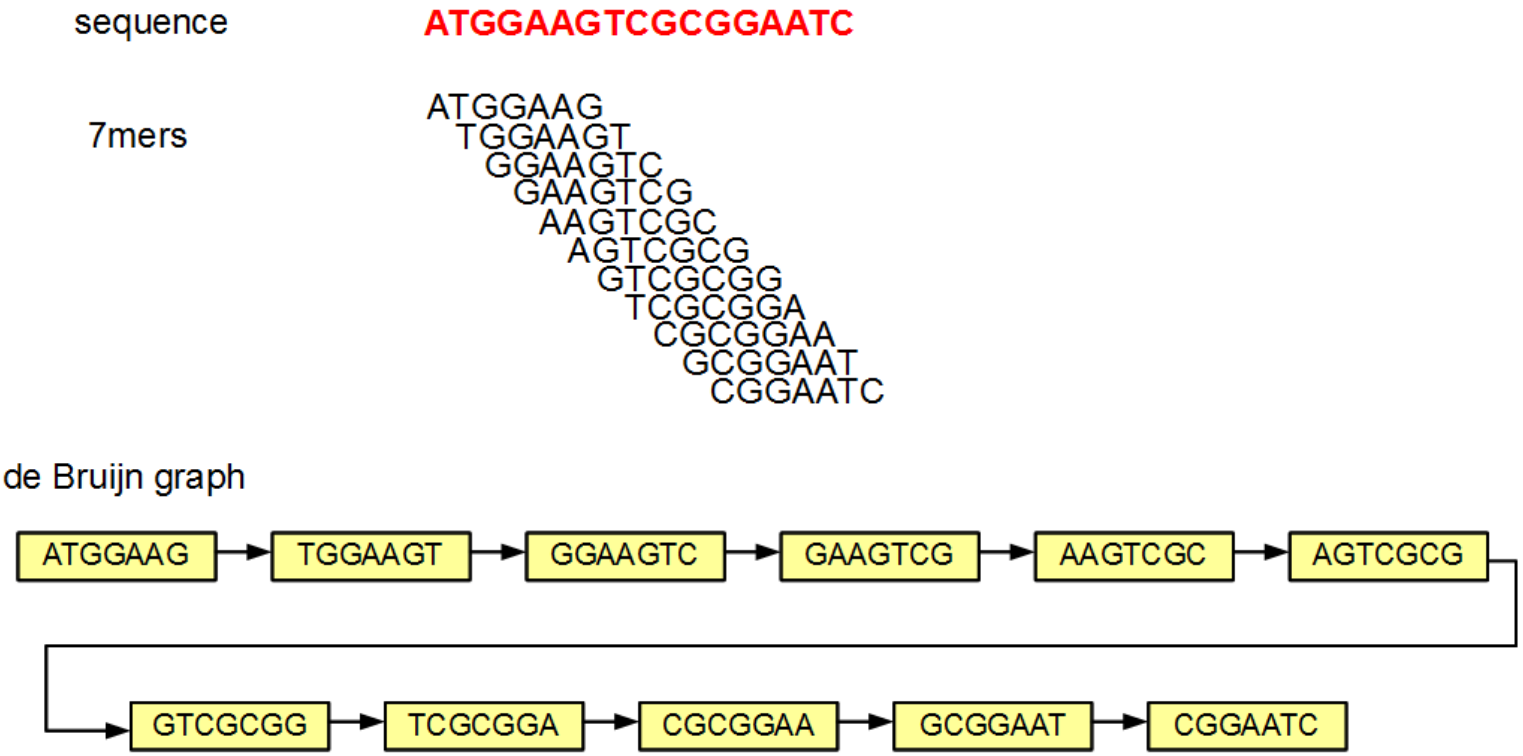
Long Reads



➢ 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

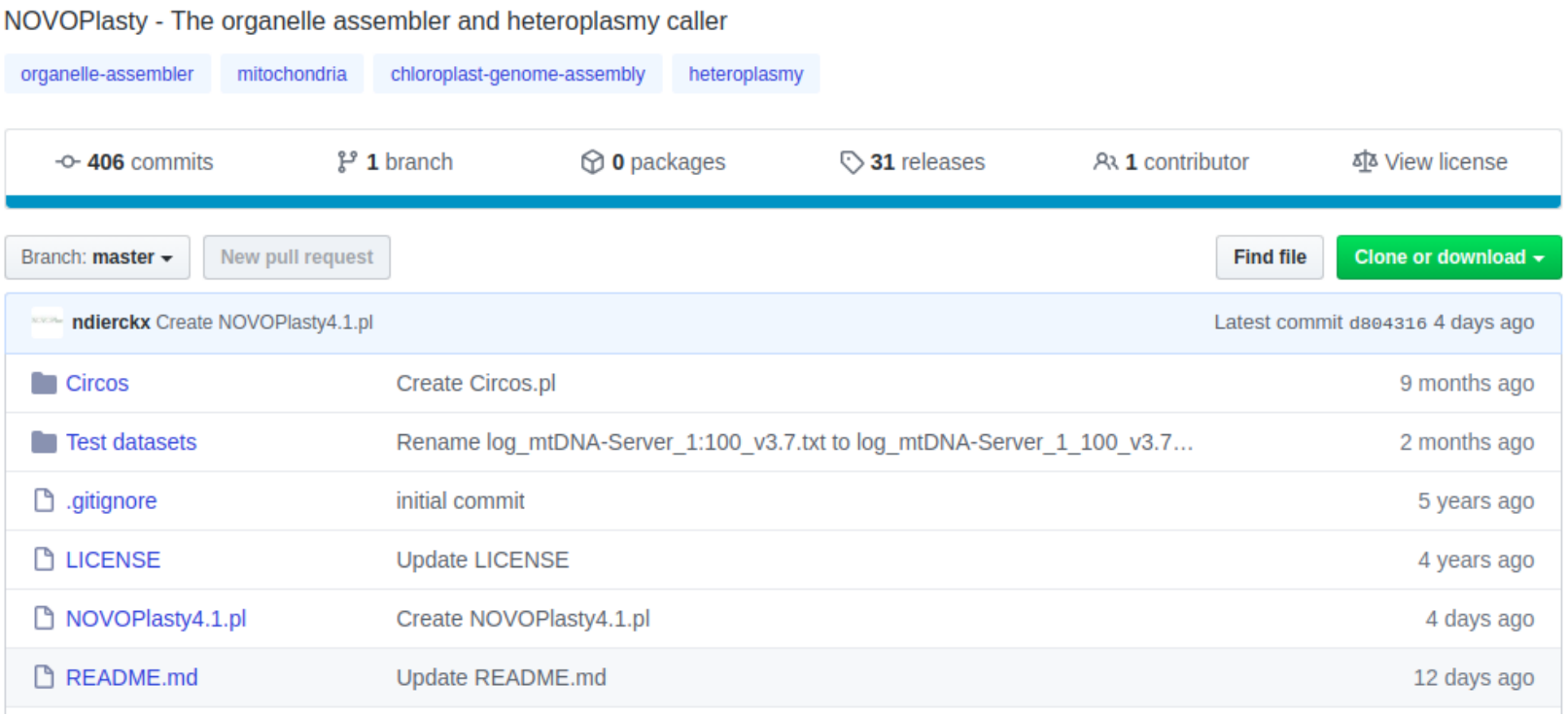
N50 = 54Mb

NOVOPLASTY

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

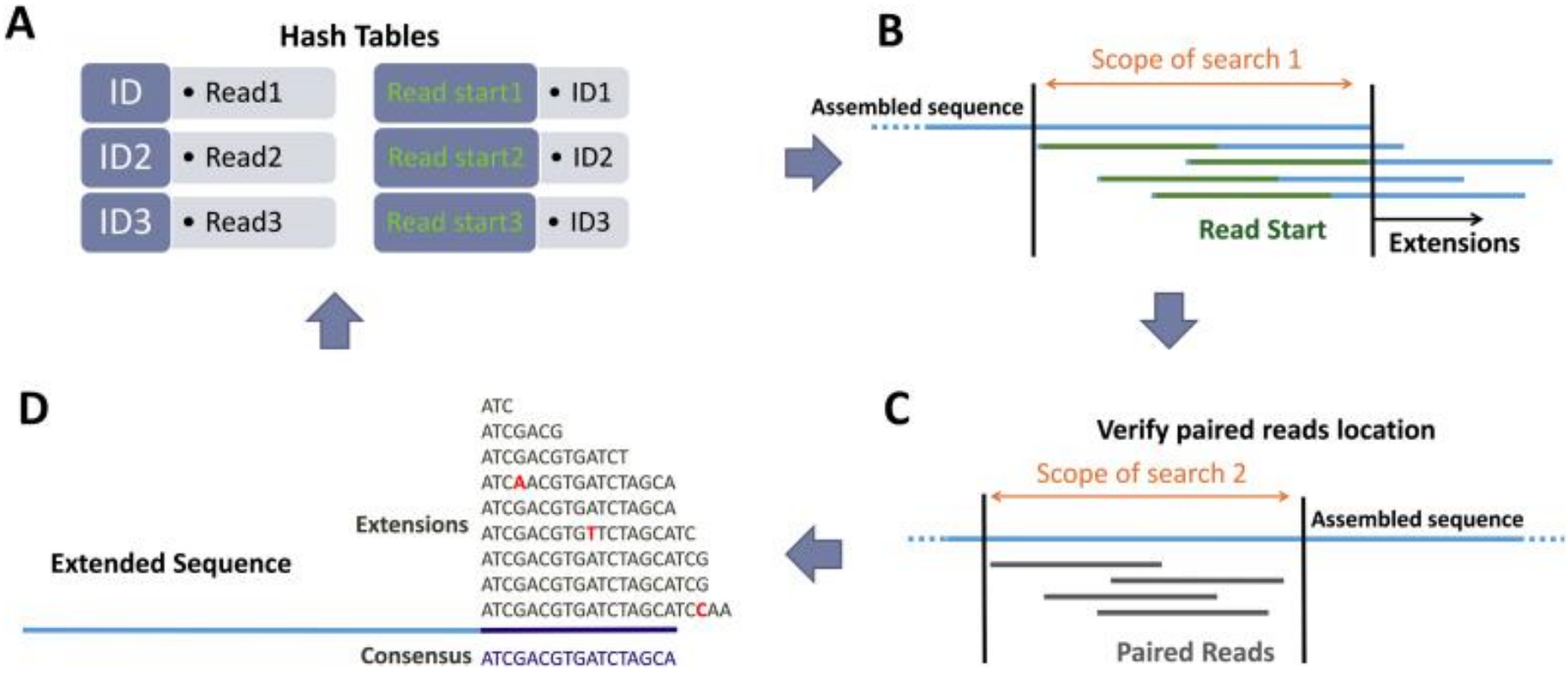
A montagem se baseia na extensão a partir de uma sequência semente

https://github.com/ndierckx/NOVOPlasty



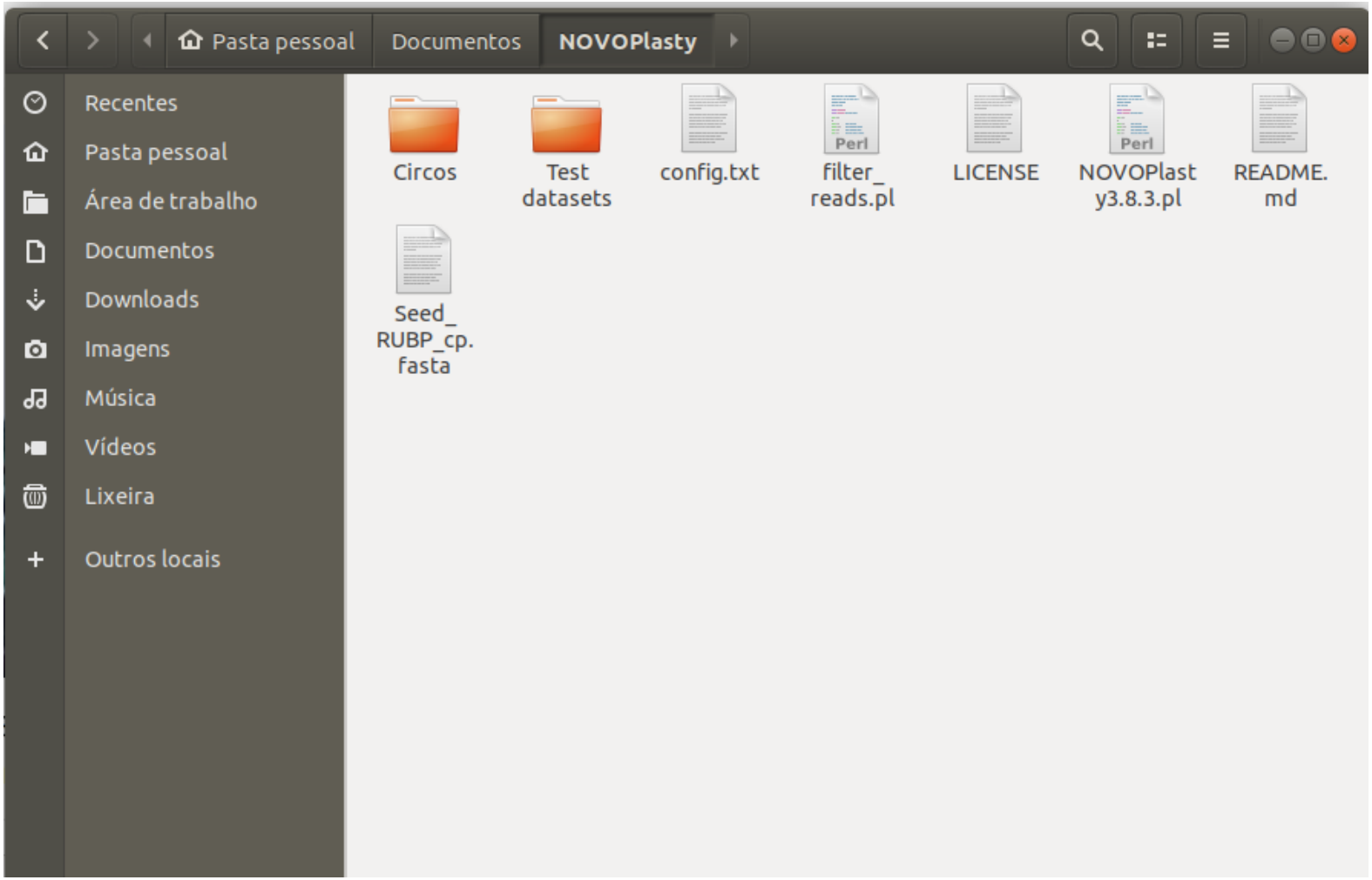
NOVOPLASTY

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 maximo a ser utilizado caso a capacidade de memoria 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: GATTTTACCATGACTGCAATTTTAGAGAGACGCGAAAGCGAAAGCCTATGGGGTCGTTTCTGTAACTGGATAACC

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: GATTTTACCATGACTGCAATTTTAGAGAGACGCGAAAGCGAAAGCCTATGGGGTCGTTTCTGTAACTGGATAACC

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:
I**nsert 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

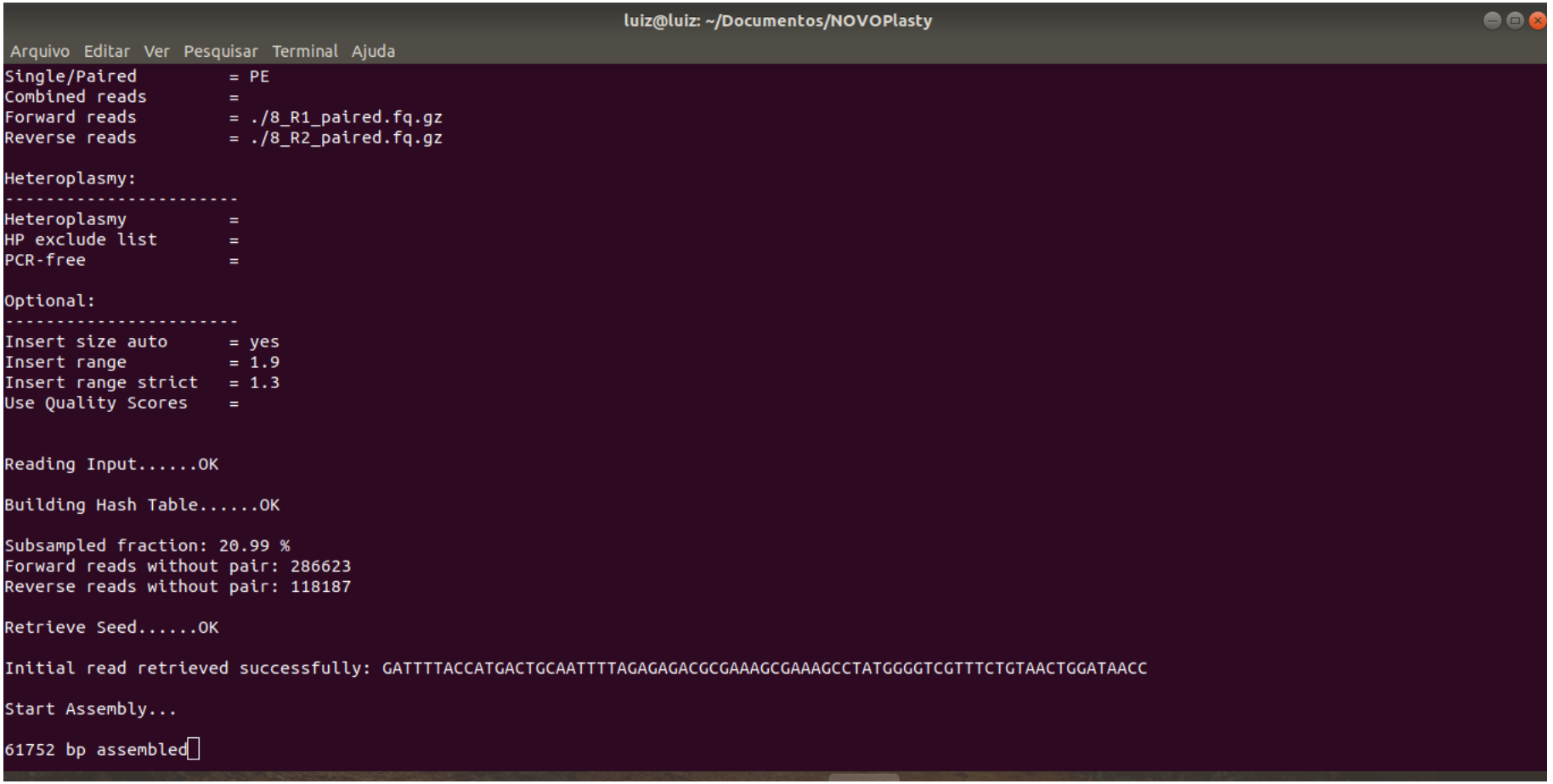
NOVOPLASTY

Rodando o programa:

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:



NOVOPLASTY

Arquivos de saída do programa:

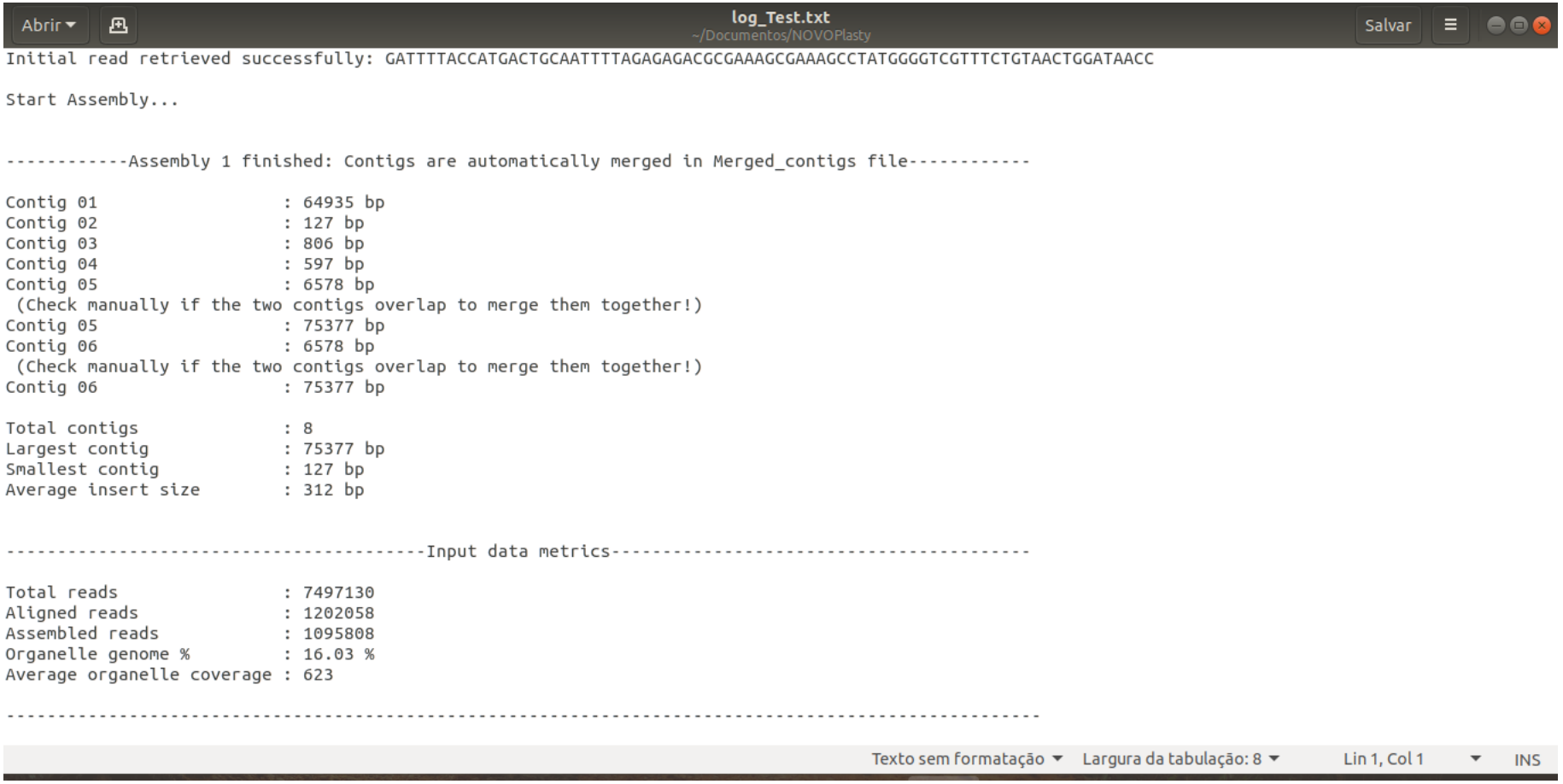
Contigs_1_Test.fasta: Arquivo contendo os contigs que foram gerados durante a montagem

Contigs_tmp_Test.txt: Arquivo backup

NOVOPLASTY

Arquivos de saída do programa:

log_Test.txt: Arquivo contendo todas as informações da analise.



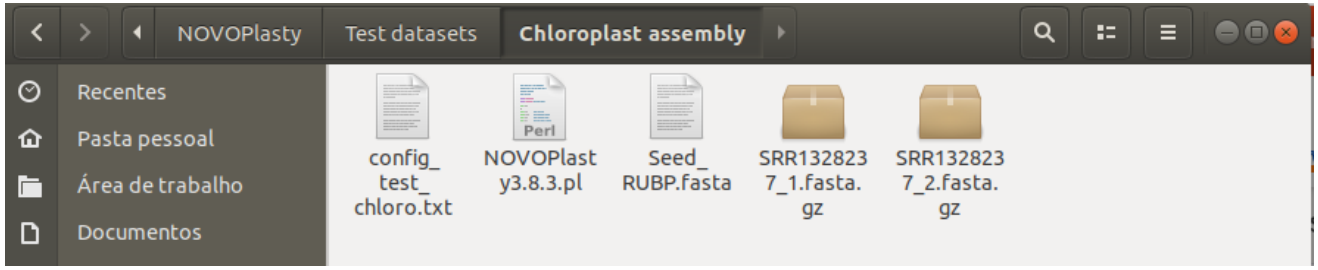
NOVOPLASTY

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

NOVOPLASTY

Rodando o programa com os arquivos de Teste:

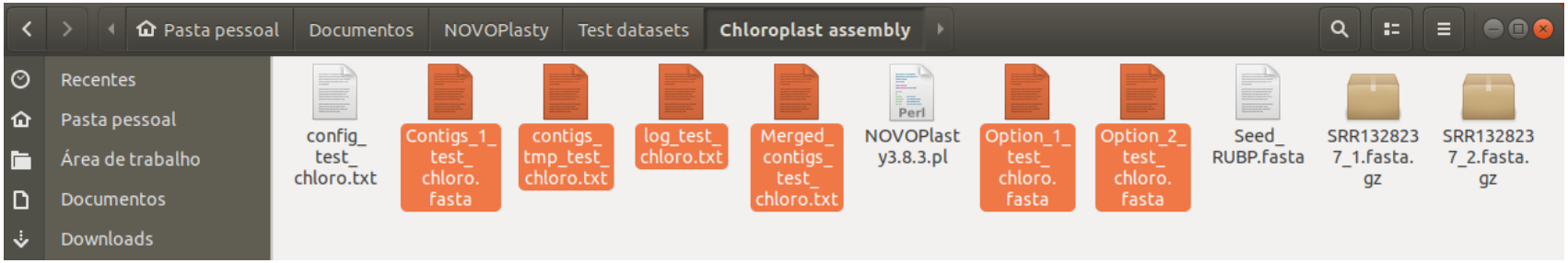


Para usar o arquivo de Teste: No terminal, acesse a pasta do NOVOPlasty, abra a pasta Test datasets e procure pela pasta chloroplast assembly. Para rodar a analise teste utilize o comando:

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

NOVOPLASTY

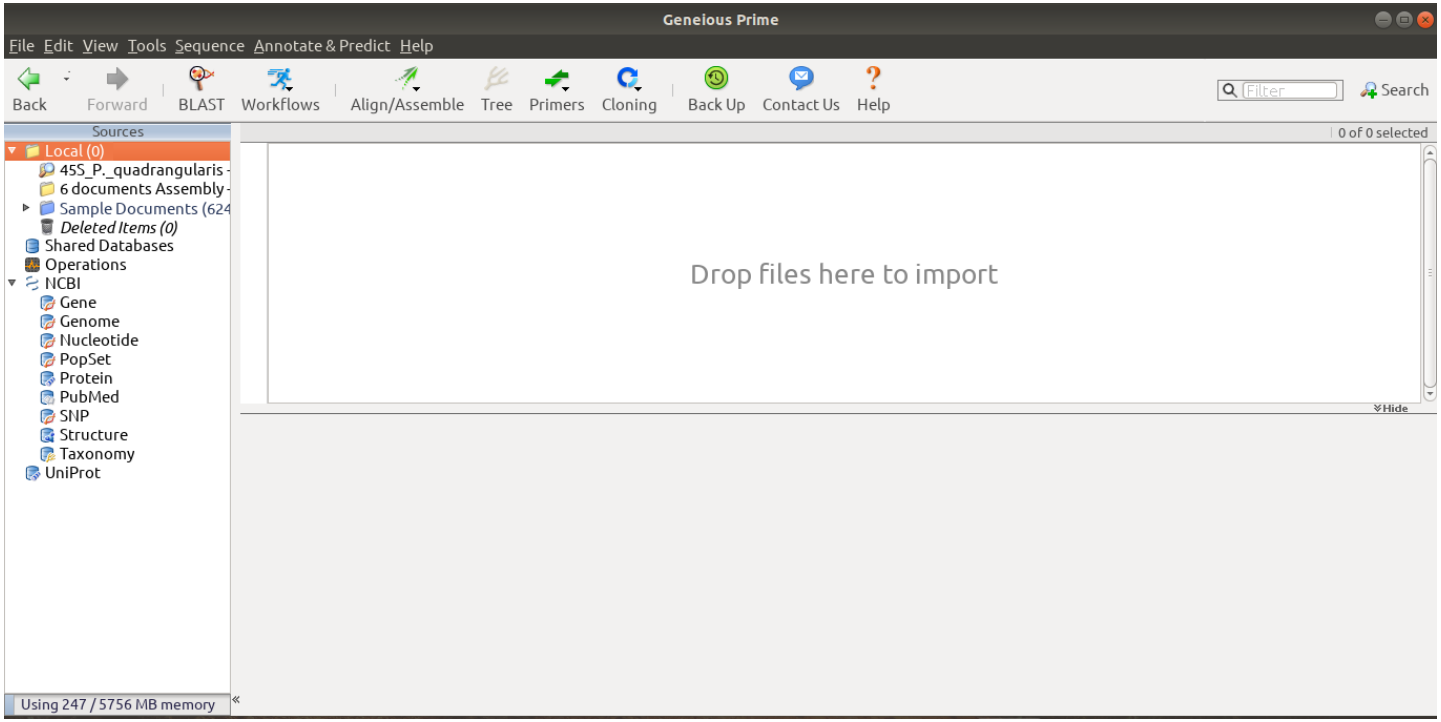
Arquivos de saída do programa:



COMO AVALIAR A MONTAGEM?

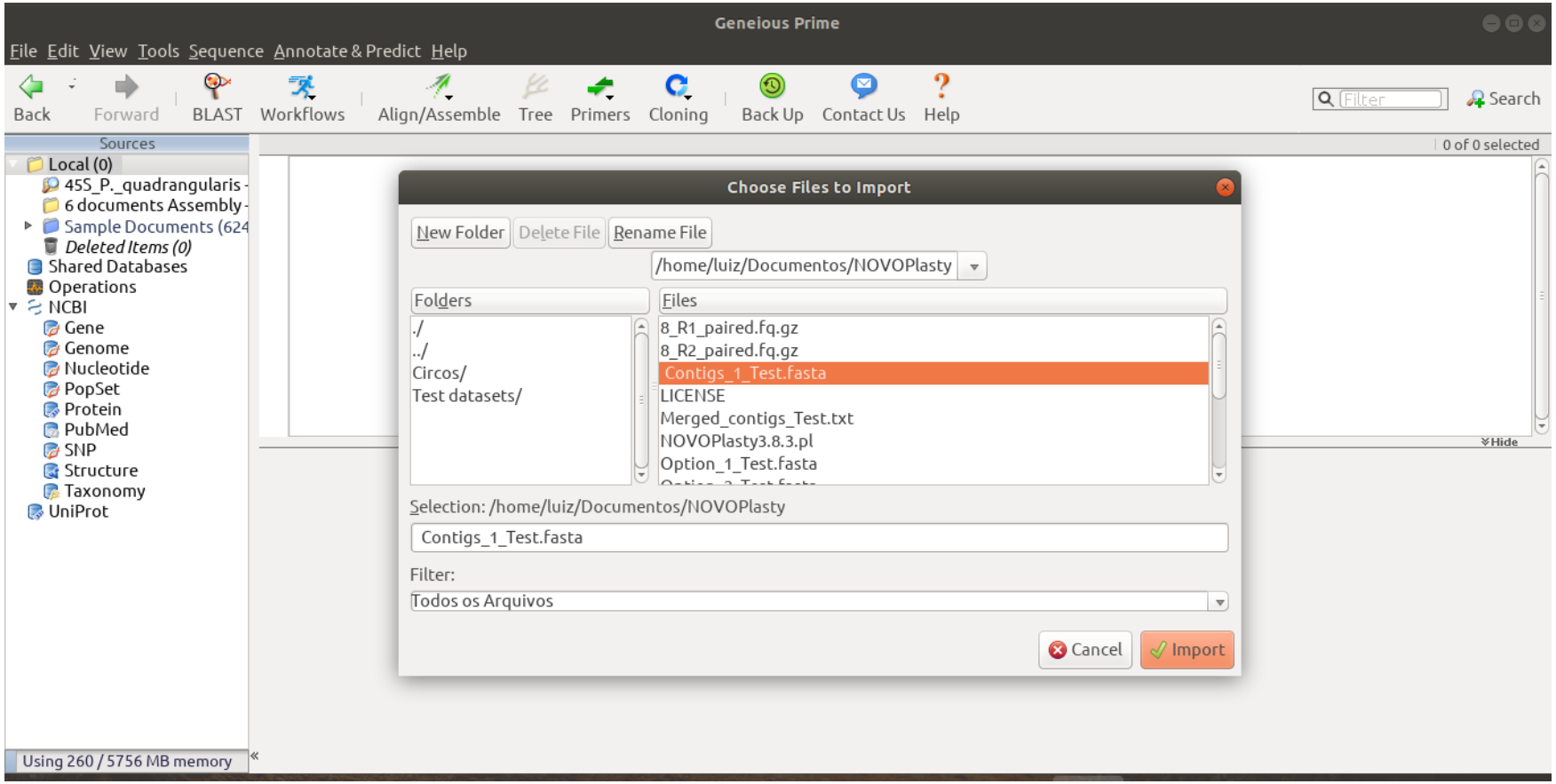
- Tamanho esperado do genoma
- Quantidade de contigs
- Média do tamanho dos contigs
- Valor de N50
- Quantidade de gaps

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



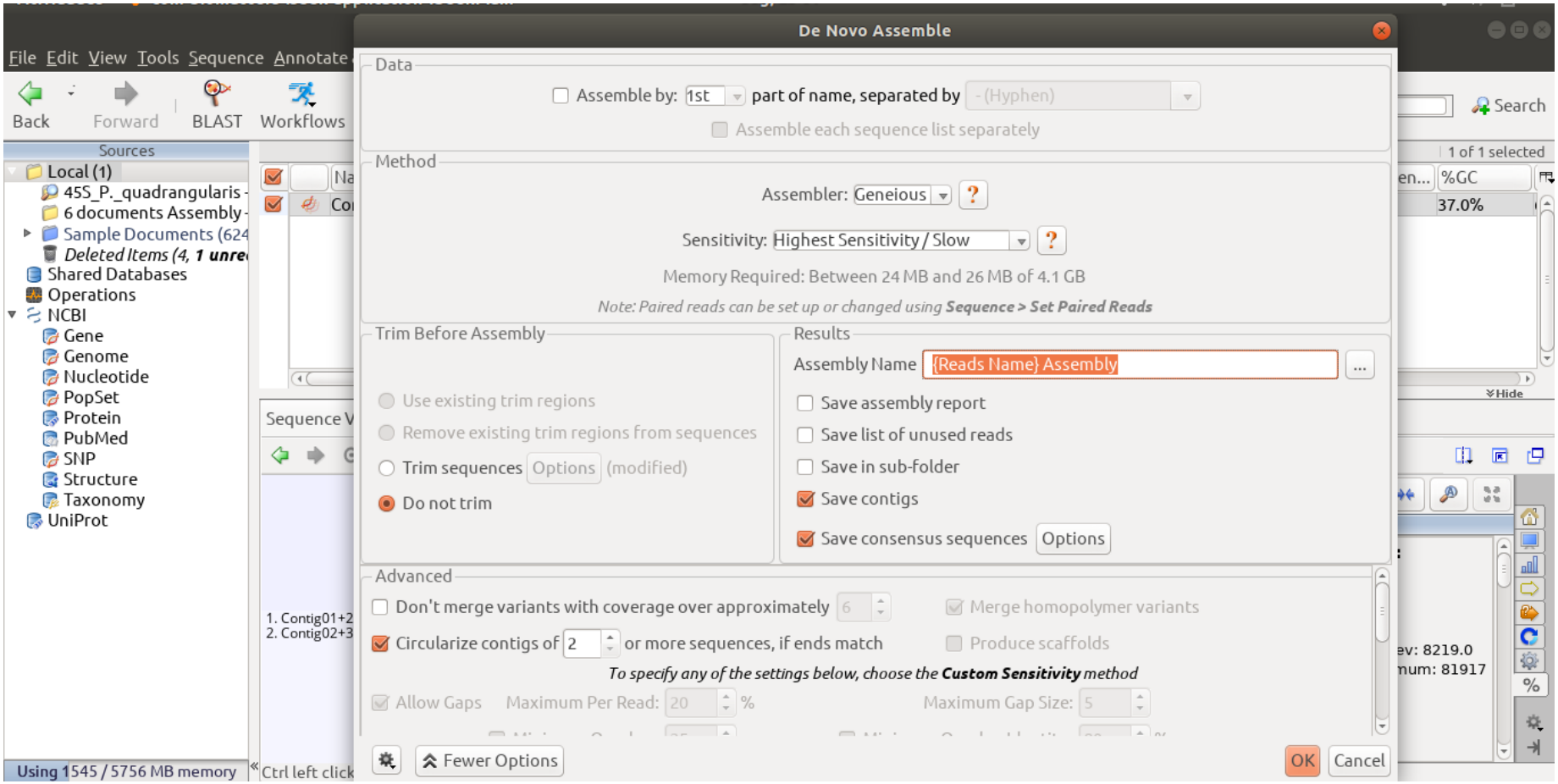
GENEIOUS

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



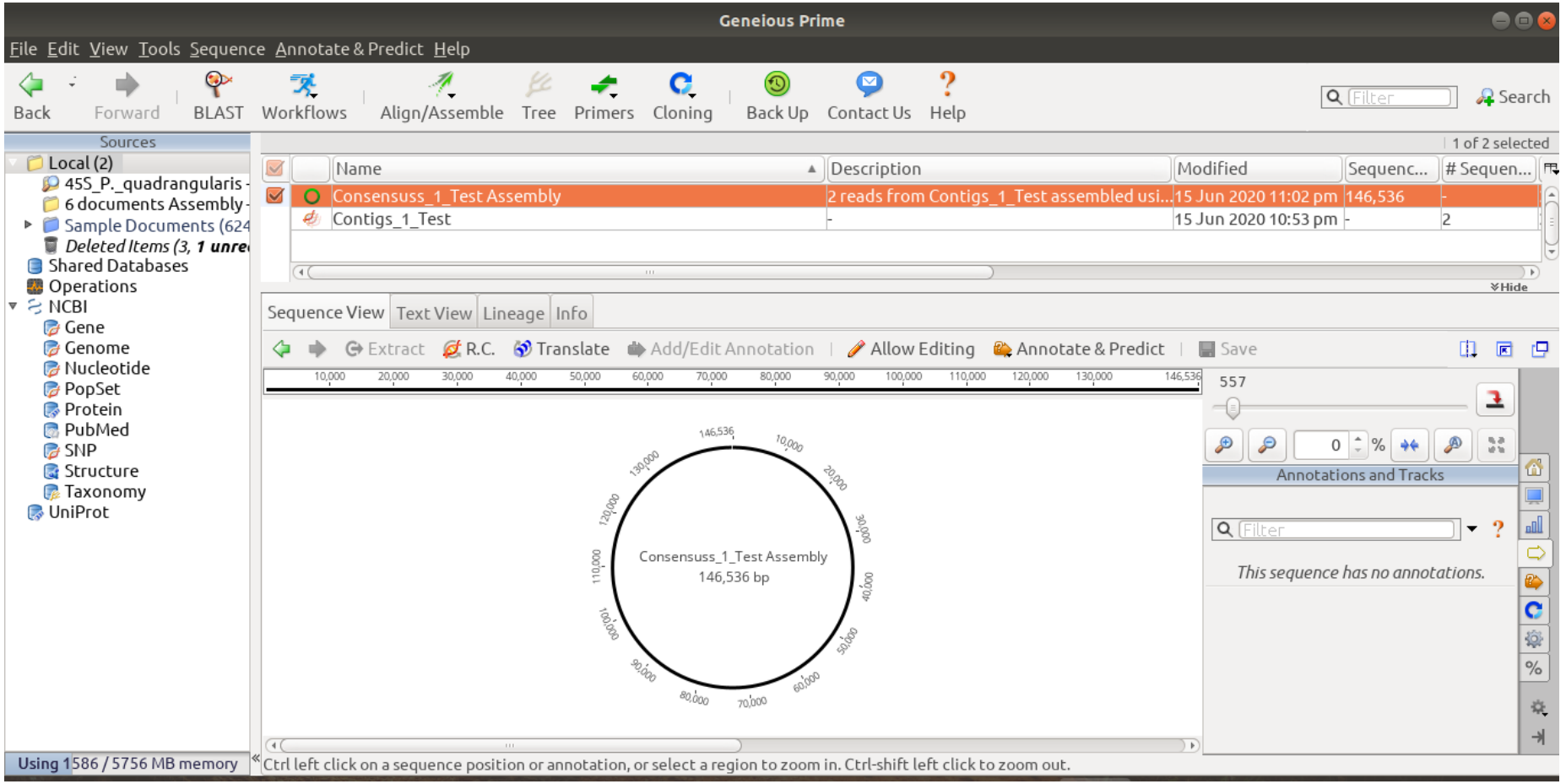
GENEIOUS

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



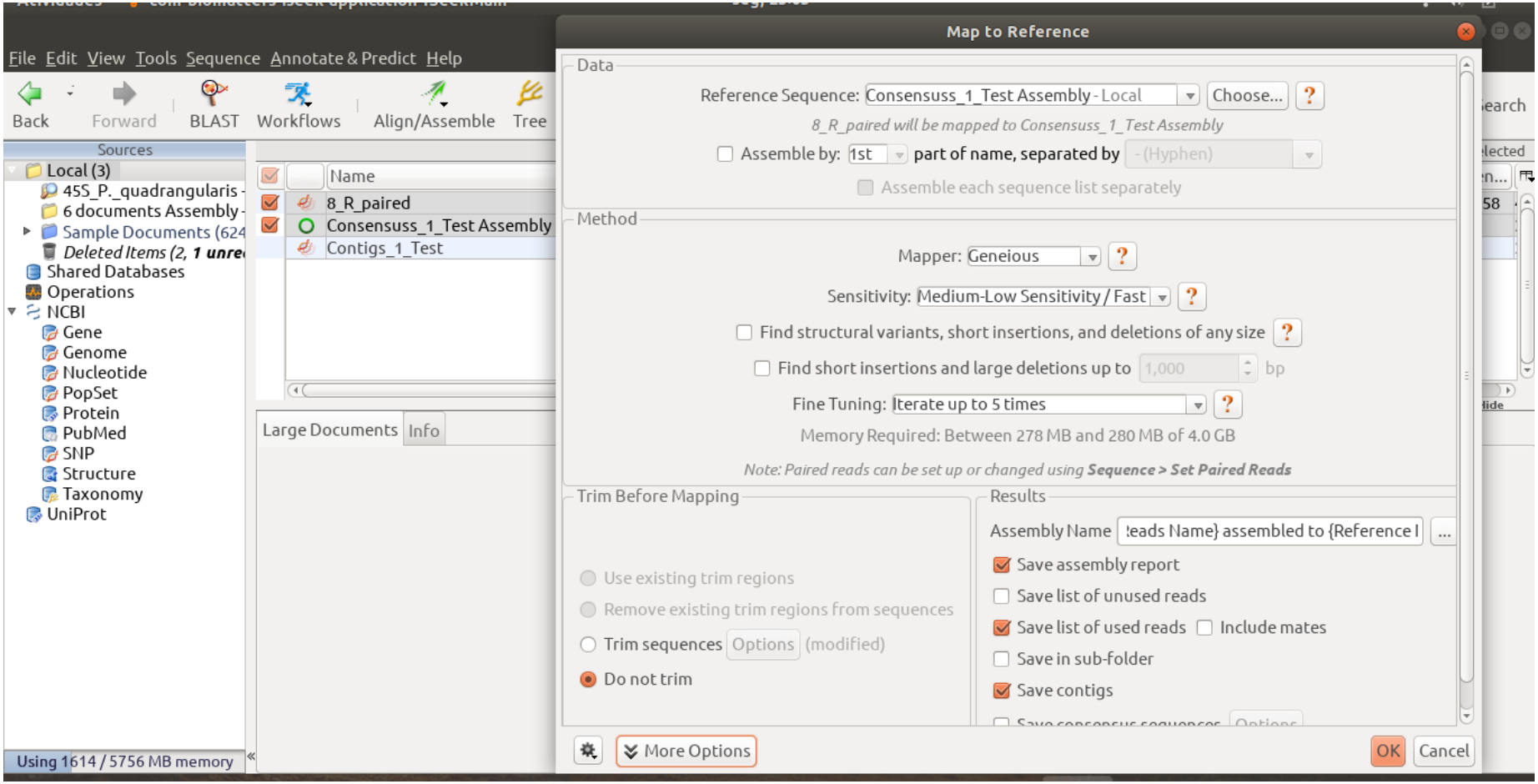
GENEIOUS

Resultados da montagem: Observar se houve circularização do genoma



GENEIOUS

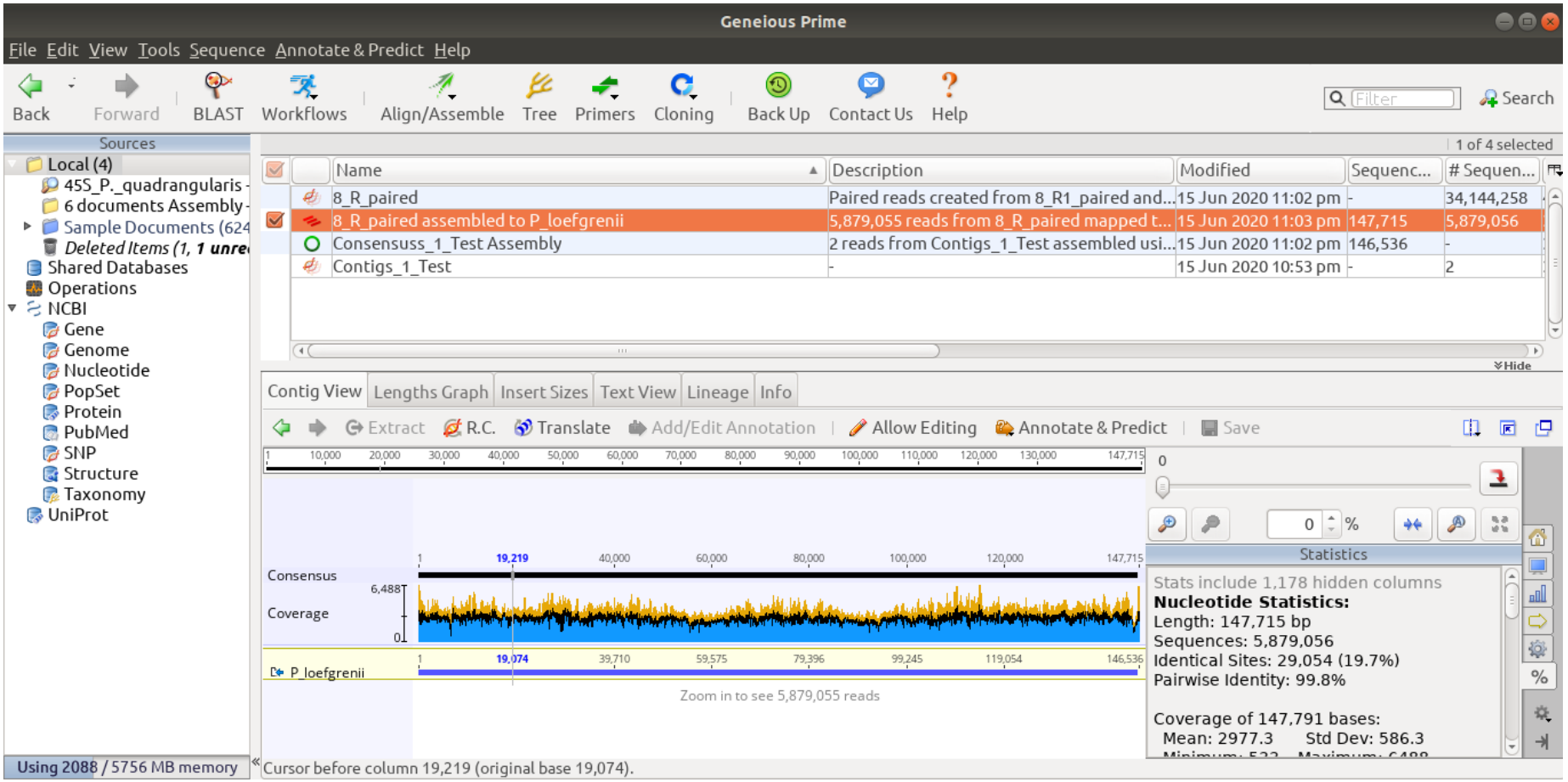
Mapear *reads* sobre a montagem: Fazer *upload* do arquivo Fastq com as *reads*. Abrir as opções em Align/Assemble e depois a função Map to Reference



GENEIOUS

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

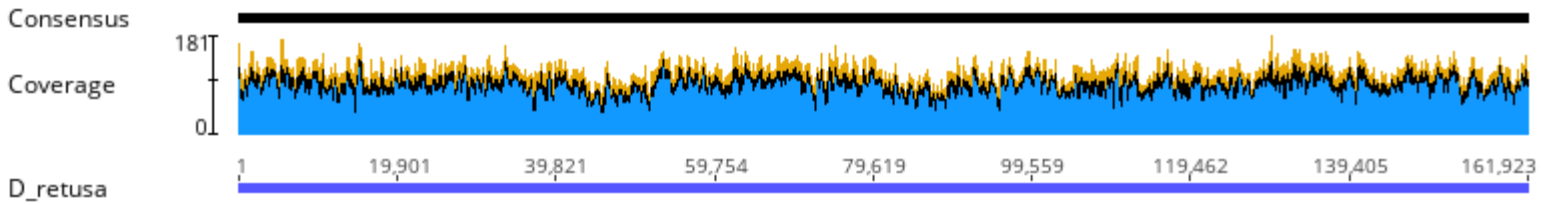
Esta função também permite obter a cobertura média do mapeamento



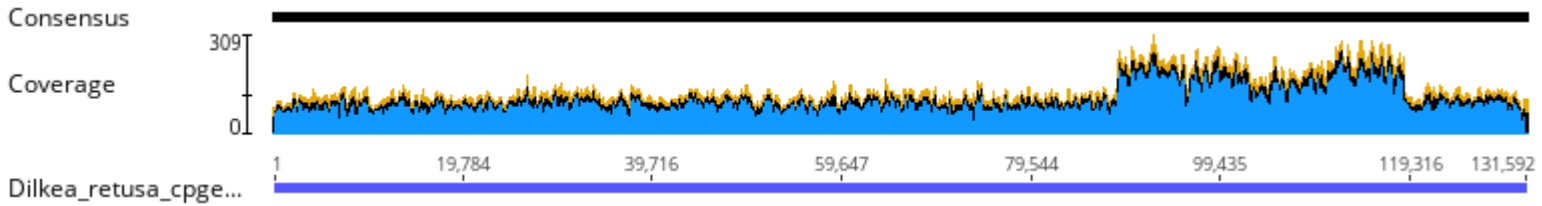
GENEIOUS

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

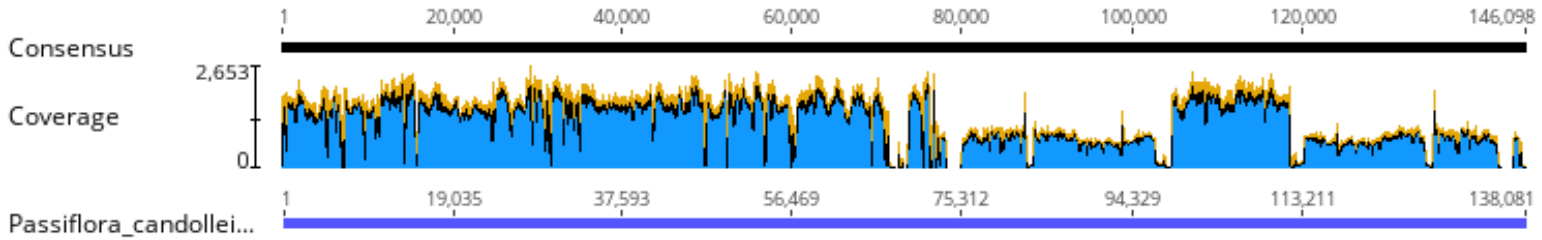
Exemplo: Montagem completa e mapeamento com cobertura contínua



Exemplo: Montagem com regiões erradas

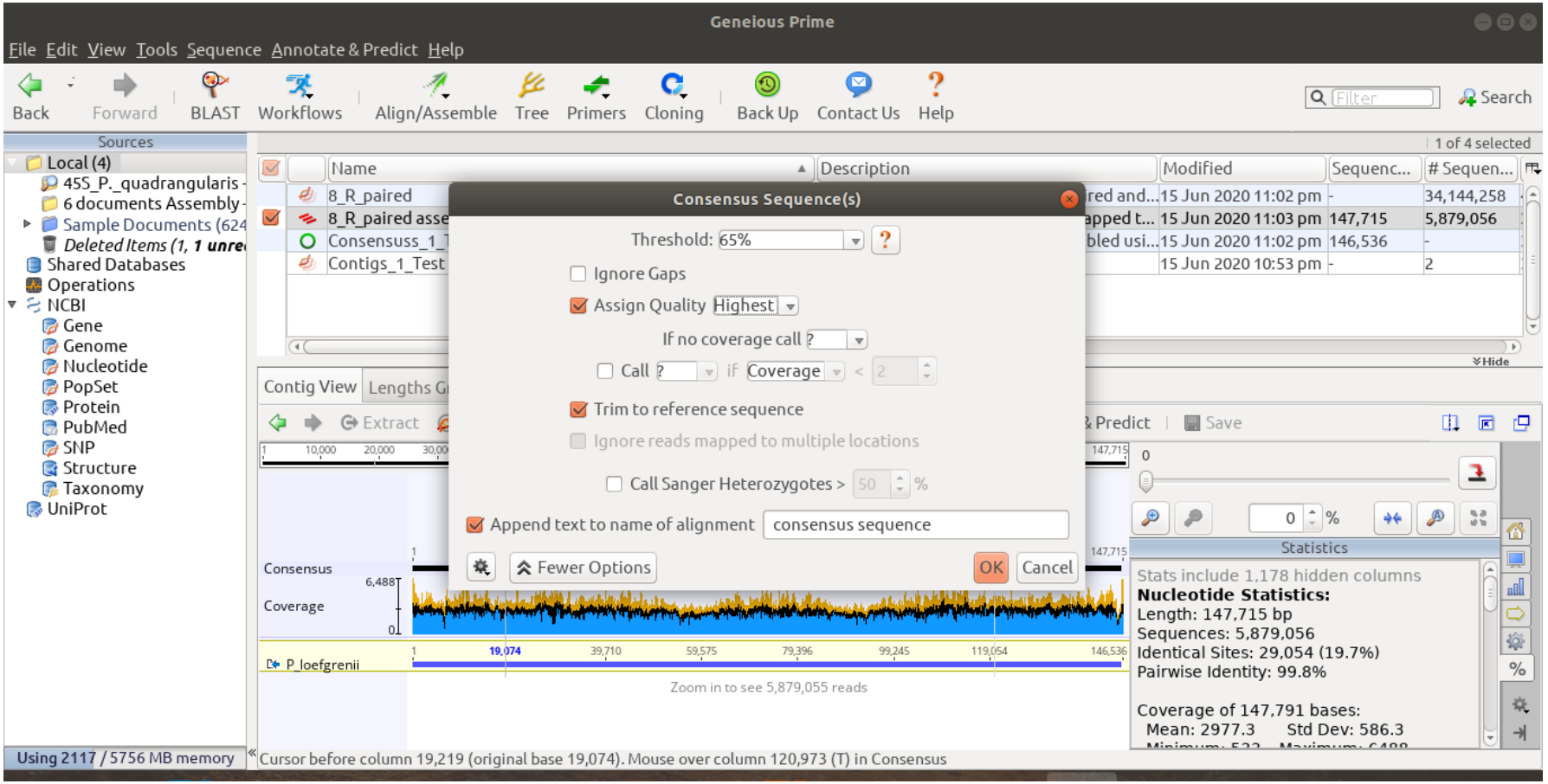


Exemplo: Montagem com regiões erradas



GENEIOUS

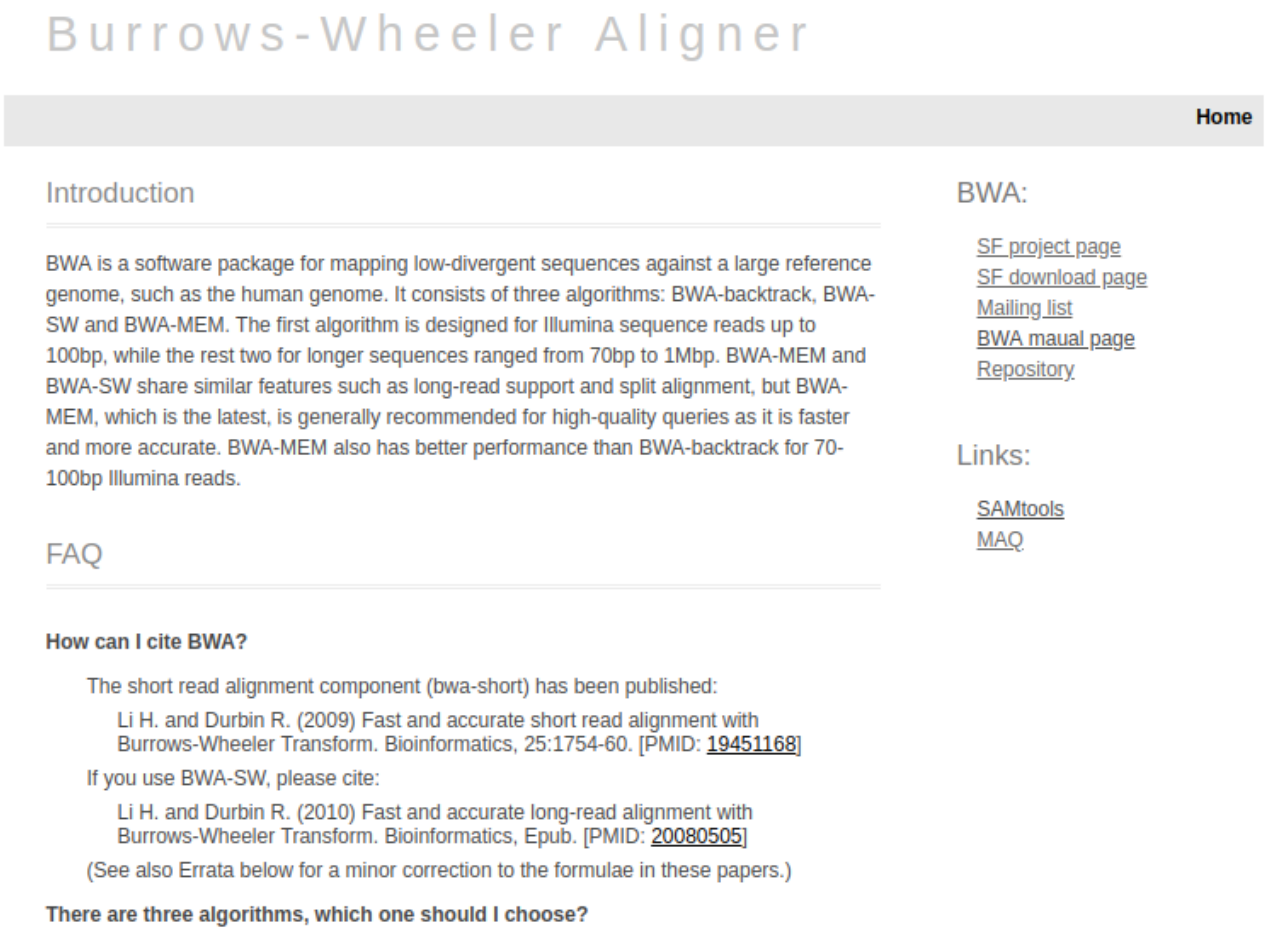
Obter sequência Consensus: Selecionar o arquivo de mapeamento e depois File / Export / Consensus sequence



BWA

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

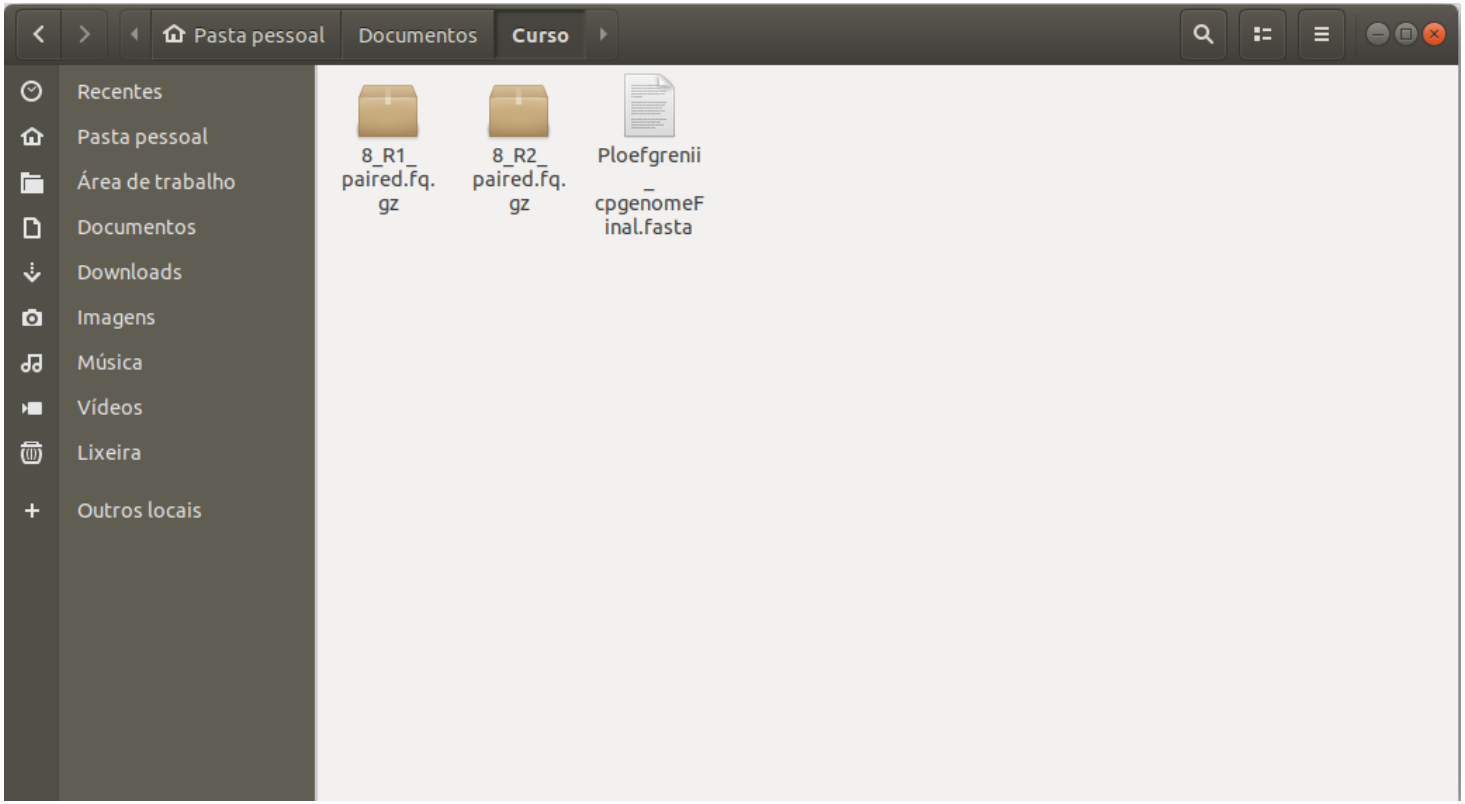
http://bio-bwa.sourceforge.net/



BWA

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

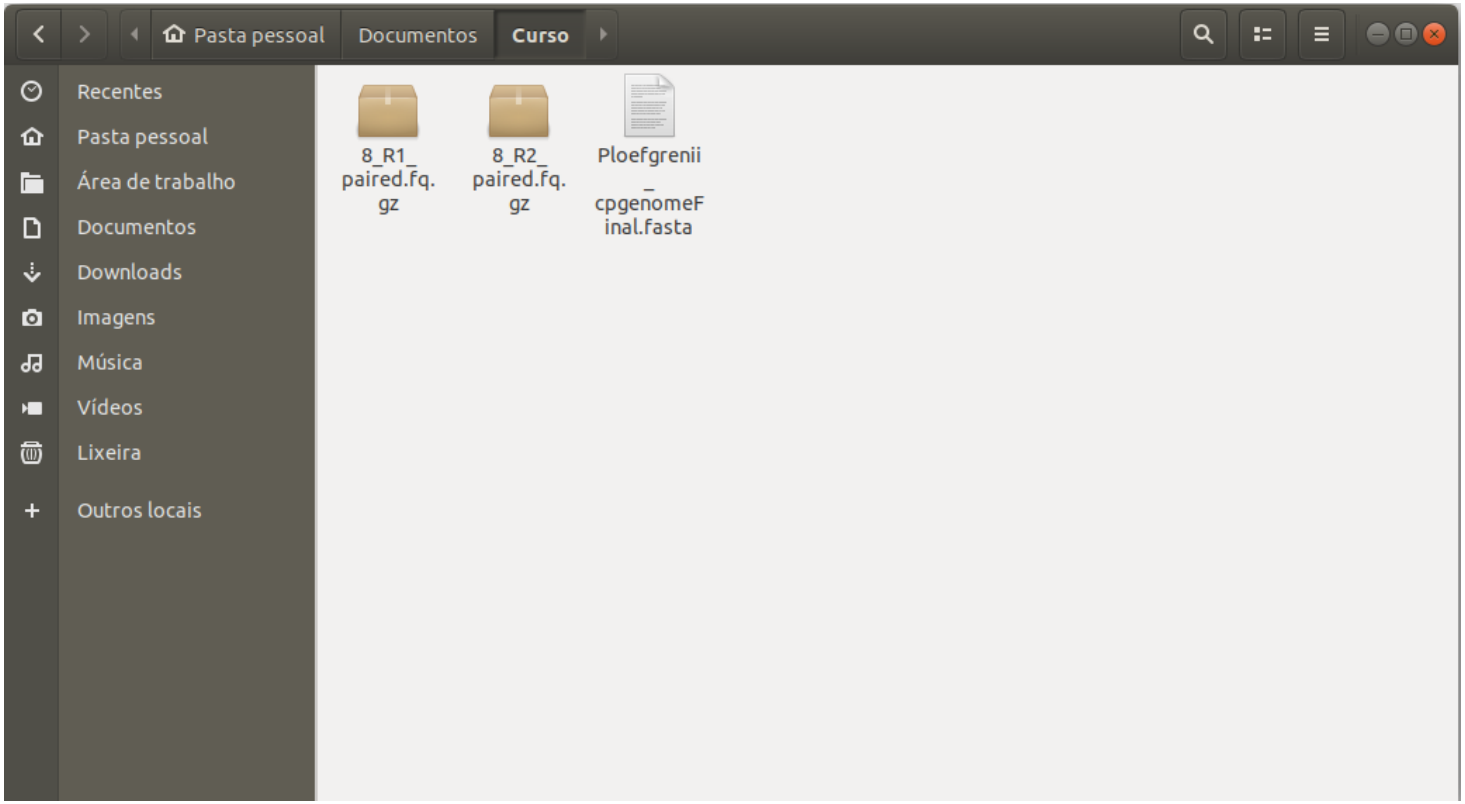
Após instalar o programa, criar uma pasta e incluir os arquivos .fastq contendo as *reads* e o arquivo .fasta com o genoma cp.



BWA

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

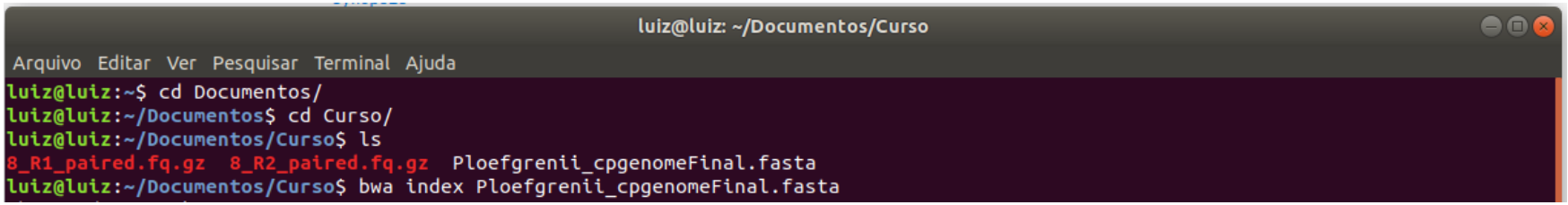
Após instalar o programa, criar uma pasta e incluir os arquivos .fastq contendo as *reads* e o arquivo .fasta com o genoma cp.

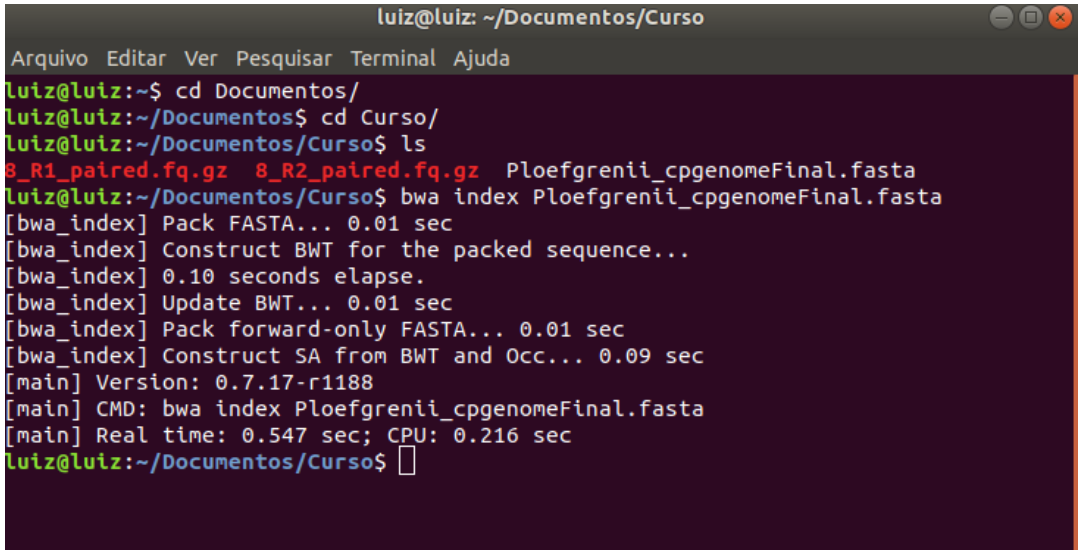


BWA

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



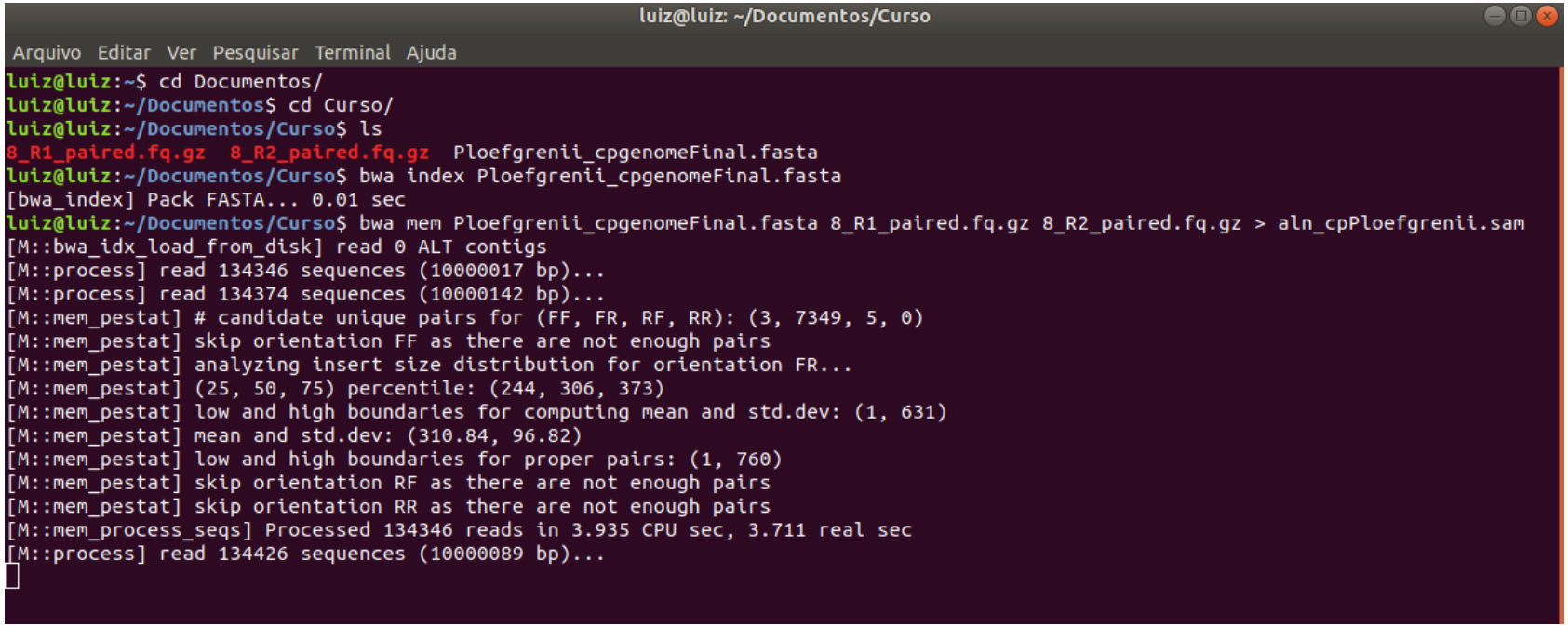


BWA

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:

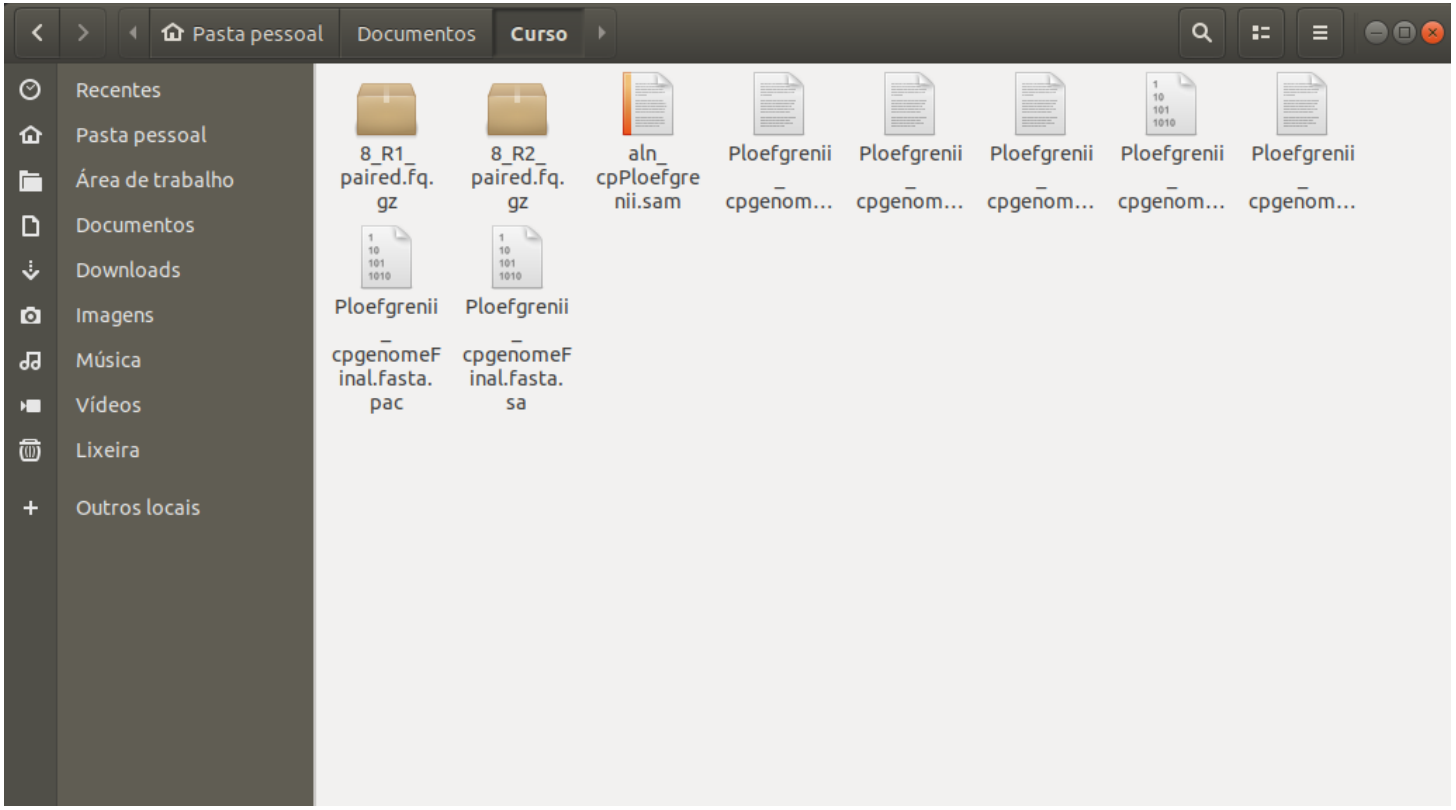
bwa mem arquivoreferência.fasta Arquivo_R1.fastq Arquivo_R2.fastq > Nomearquivodesaída**.sam**



BWA

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

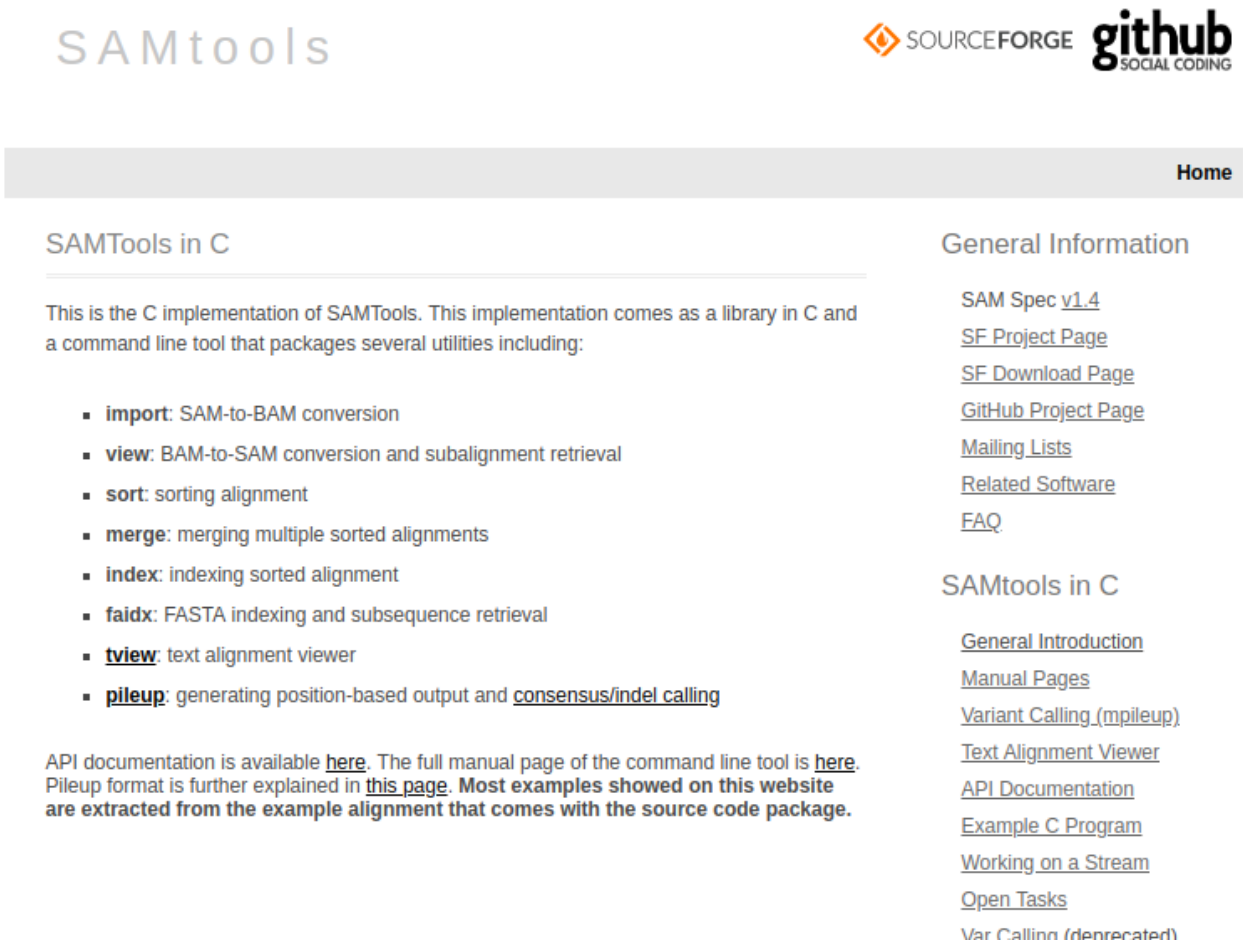
Arquivos de saída do mapeamento:
O arquivo .sam com as informações do mapeamento precisa ser convertido para o formato .bam



SAMTOOLS

Manipular arquivos em formato .sam

http://samtools.sourceforge.net/



SAMTOOLS

Converter o arquivo .sam em um arquivo .bam

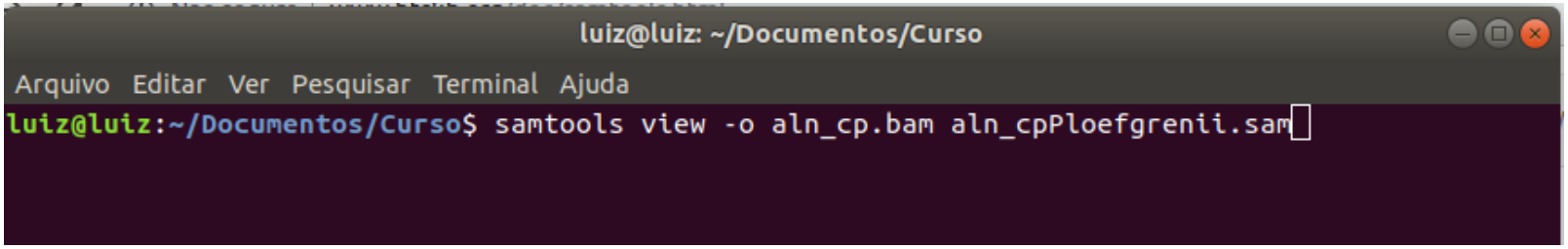
No terminal, entrar na pasta contendo os arquivos do mapeamento.

```
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
```

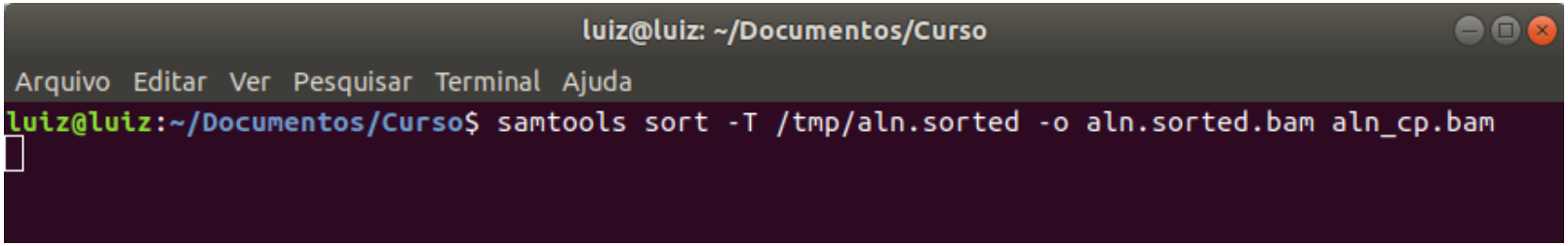
SAMTOOLS

Converter o arquivo .sam em um arquivo .bam

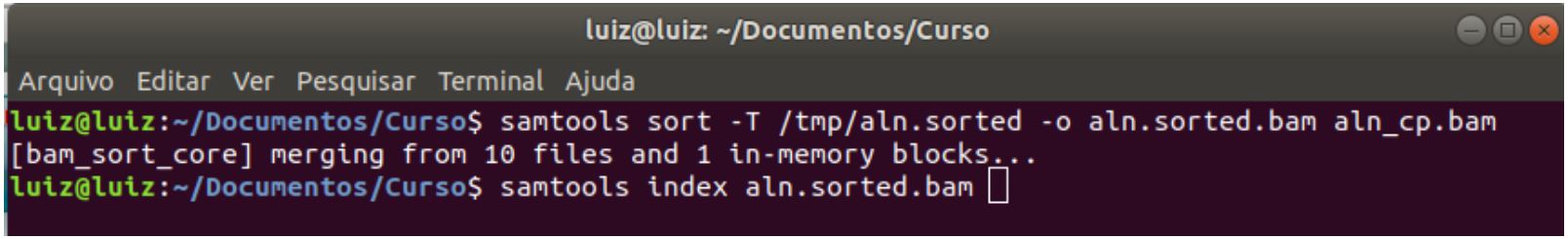
Na pasta contendo os arquivos, utilizar a opção samtools view para fazer a conversão do arquivo.



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



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



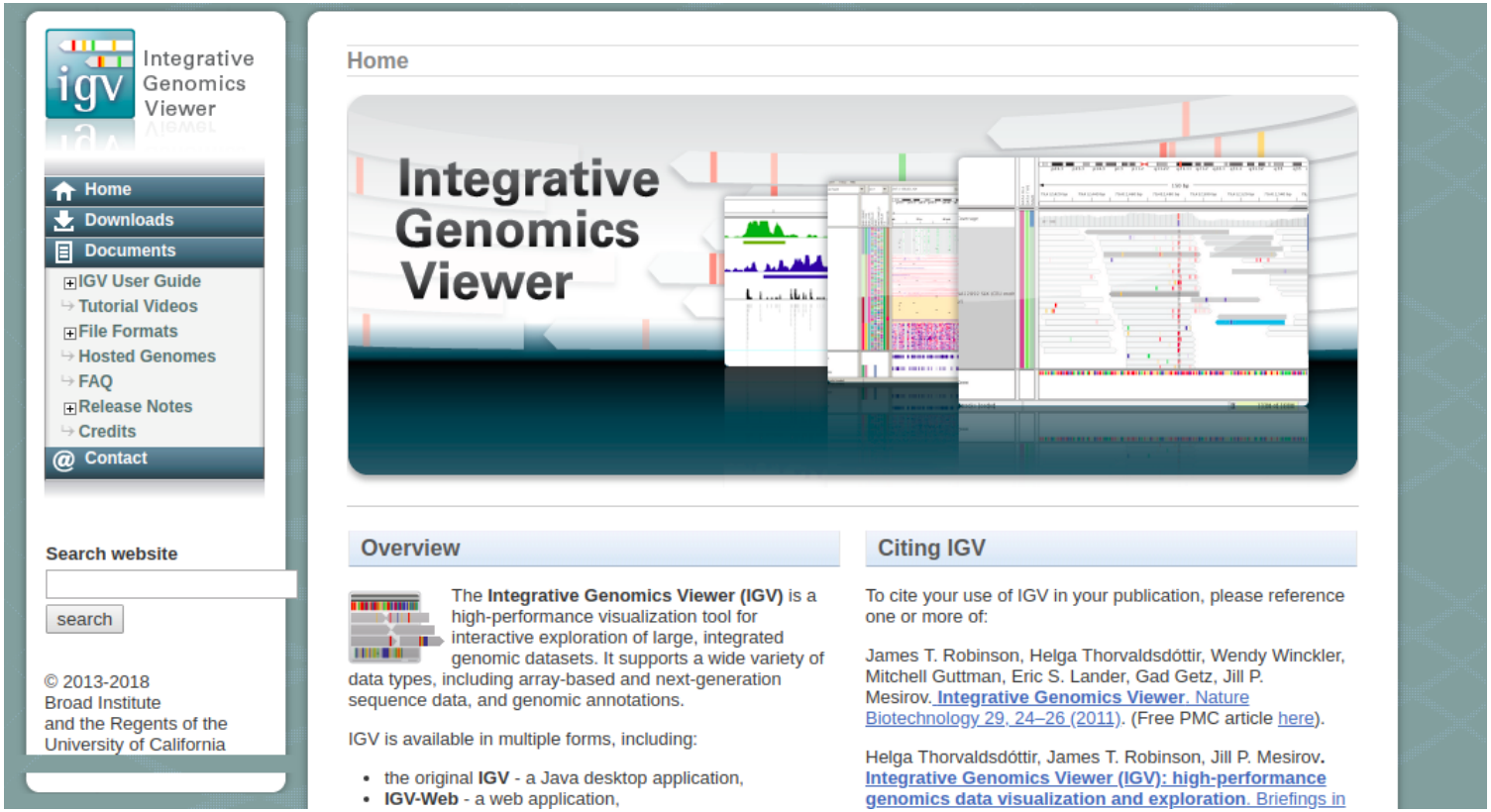
SAMTOOLS

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

INTEGRATIVE GENOMICS VIEWER (IGV)

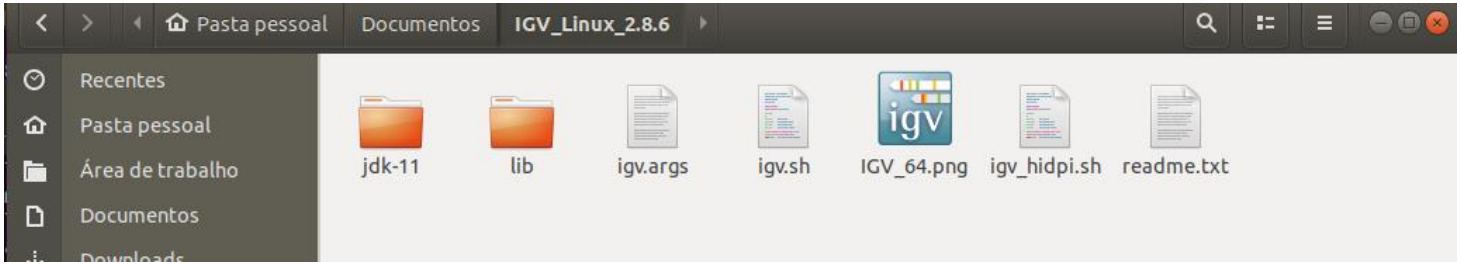
Para visualizar o resultado do mapeamento utilizar o IGV.

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

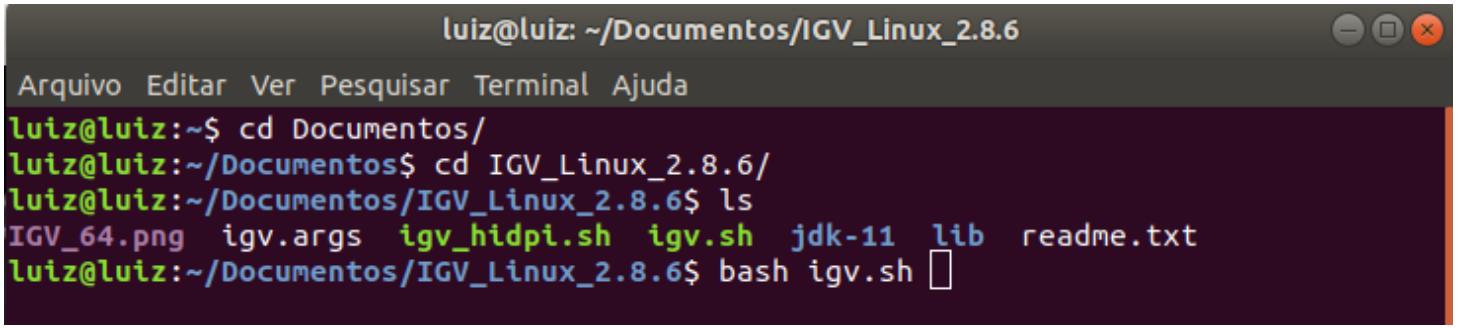


INTEGRATIVE GENOMICS VIEWER (IGV)

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

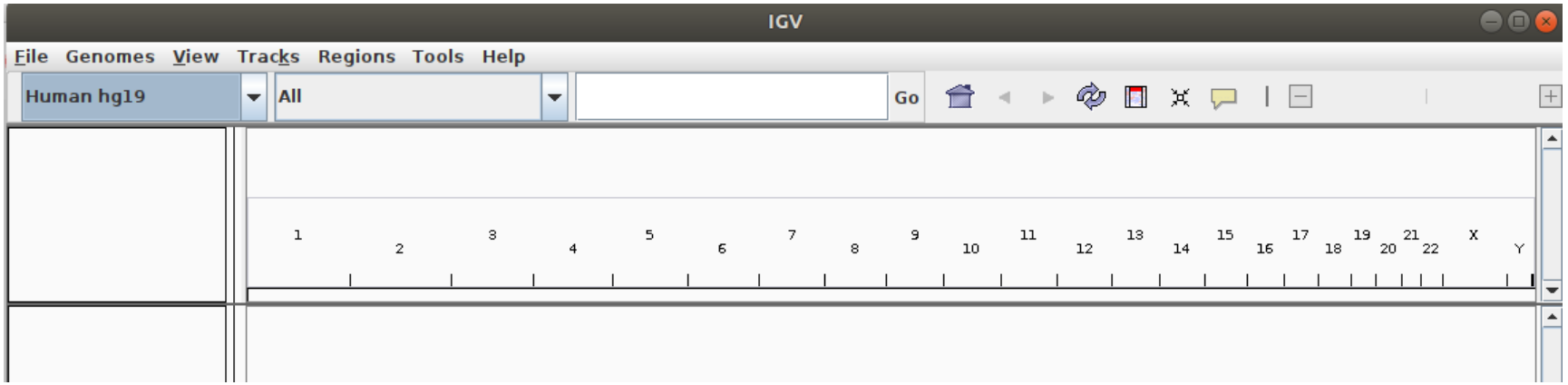


Para executar o programa, após localizar a pasta no terminal digitar: bash igv.sh

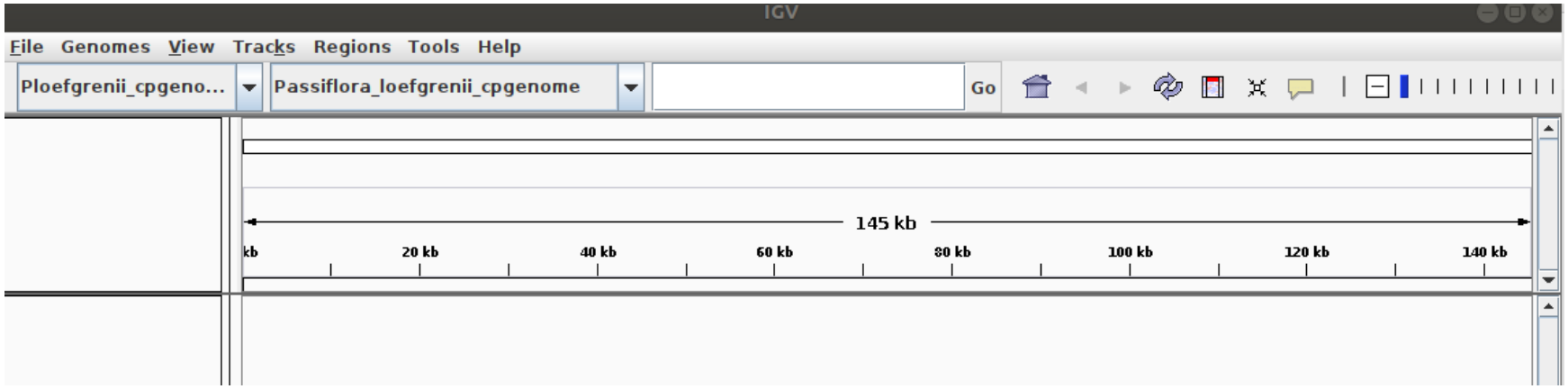


INTEGRATIVE GENOMICS VIEWER (IGV)

Após abrir o programa:

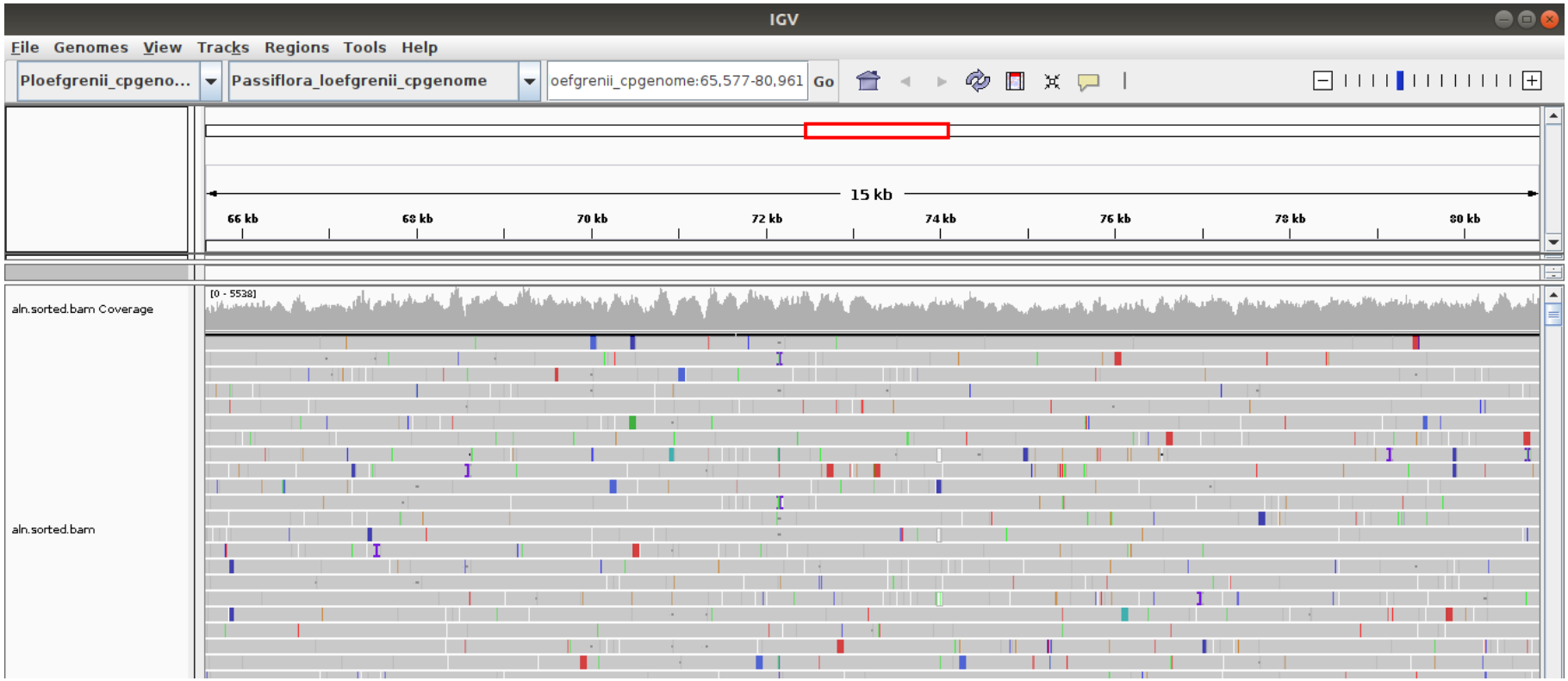


Para abrir a sequência referência clicar em Genomes e depois em Load Genome from File



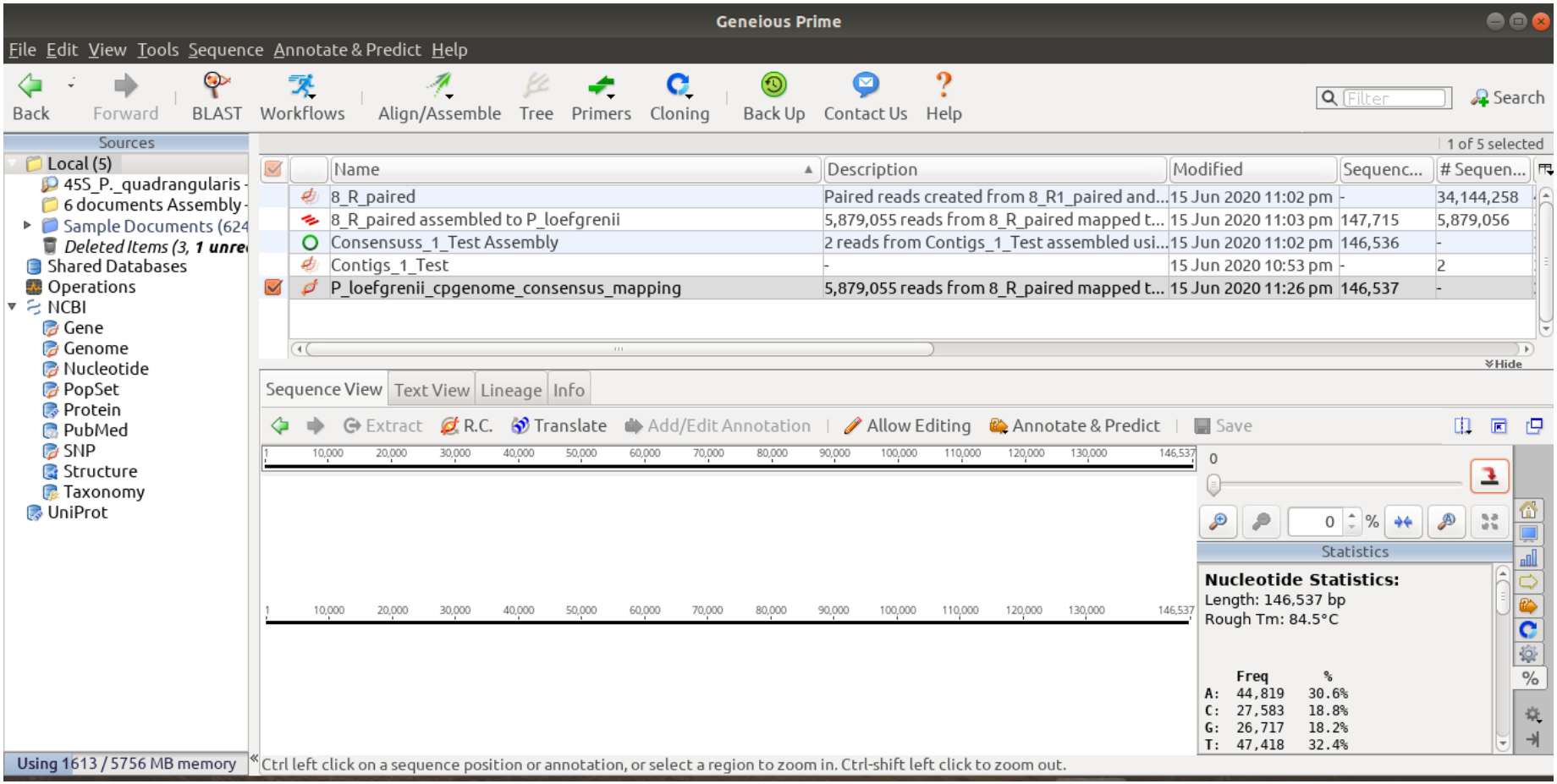
INTEGRATIVE GENOMICS VIEWER (IGV)

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



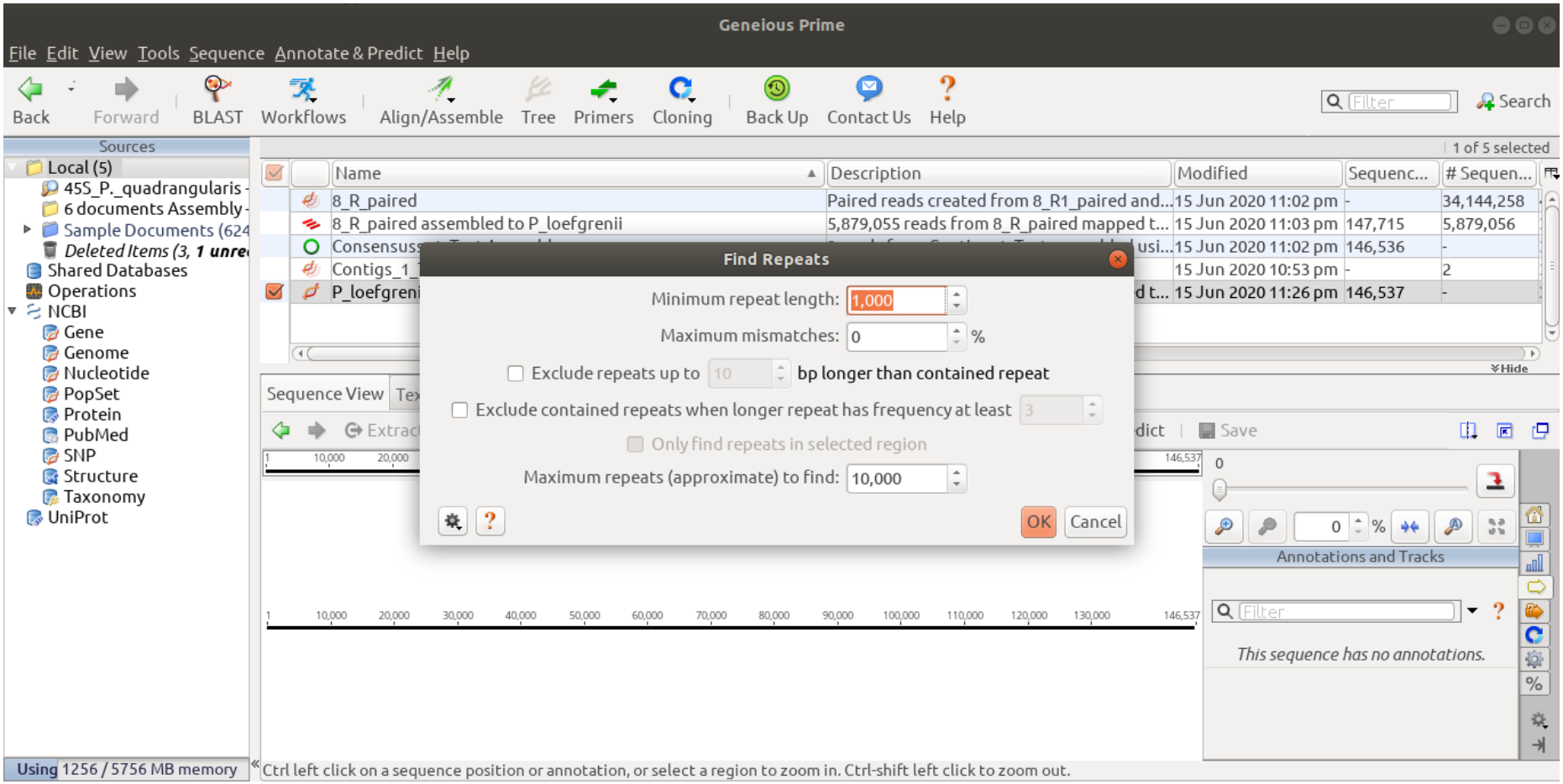
GENEIOUS

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



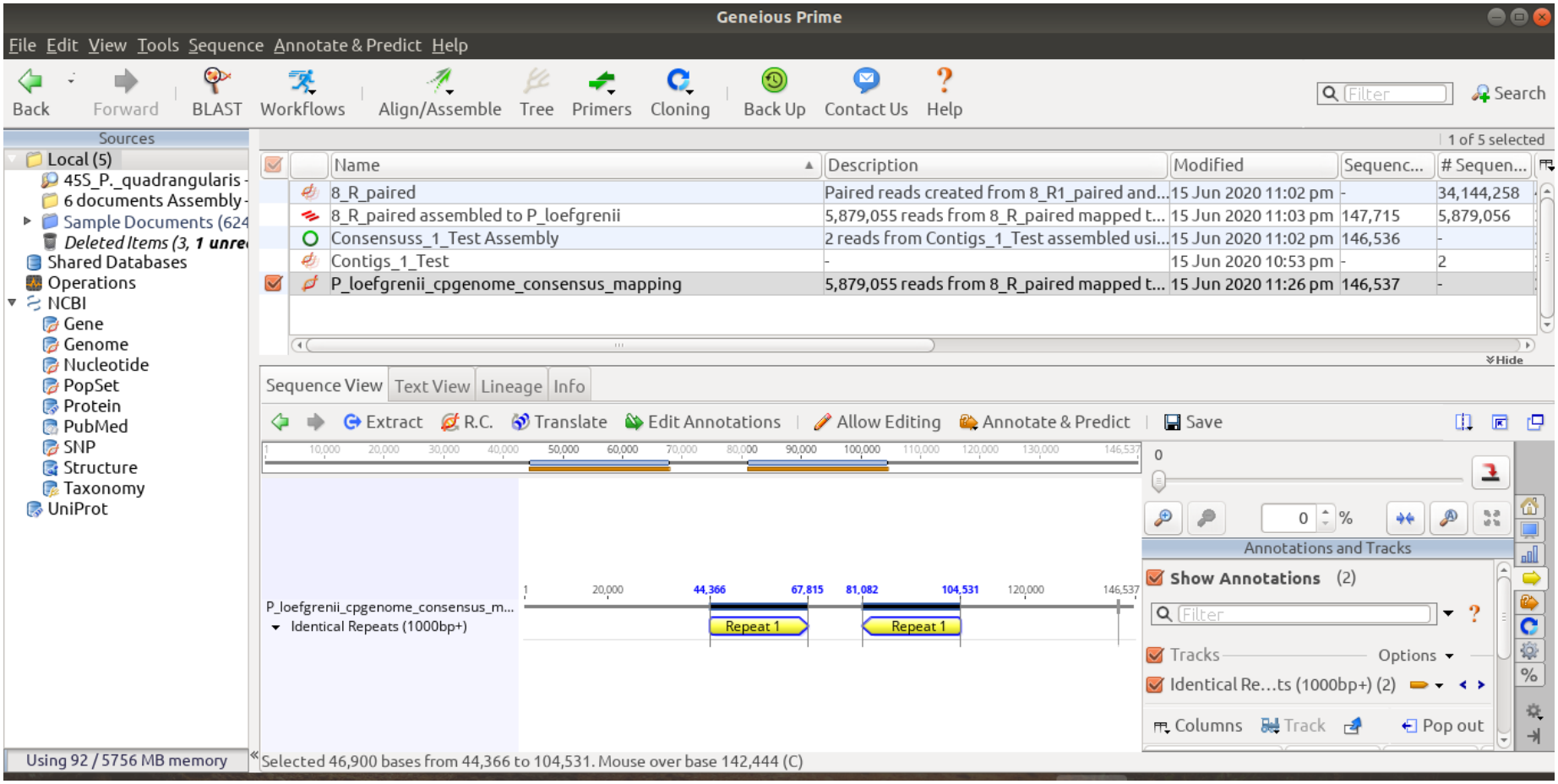
GENEIOUS

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



GENEIOUS

Tamanho, localização e orientação das regiões repetidas invertidas:

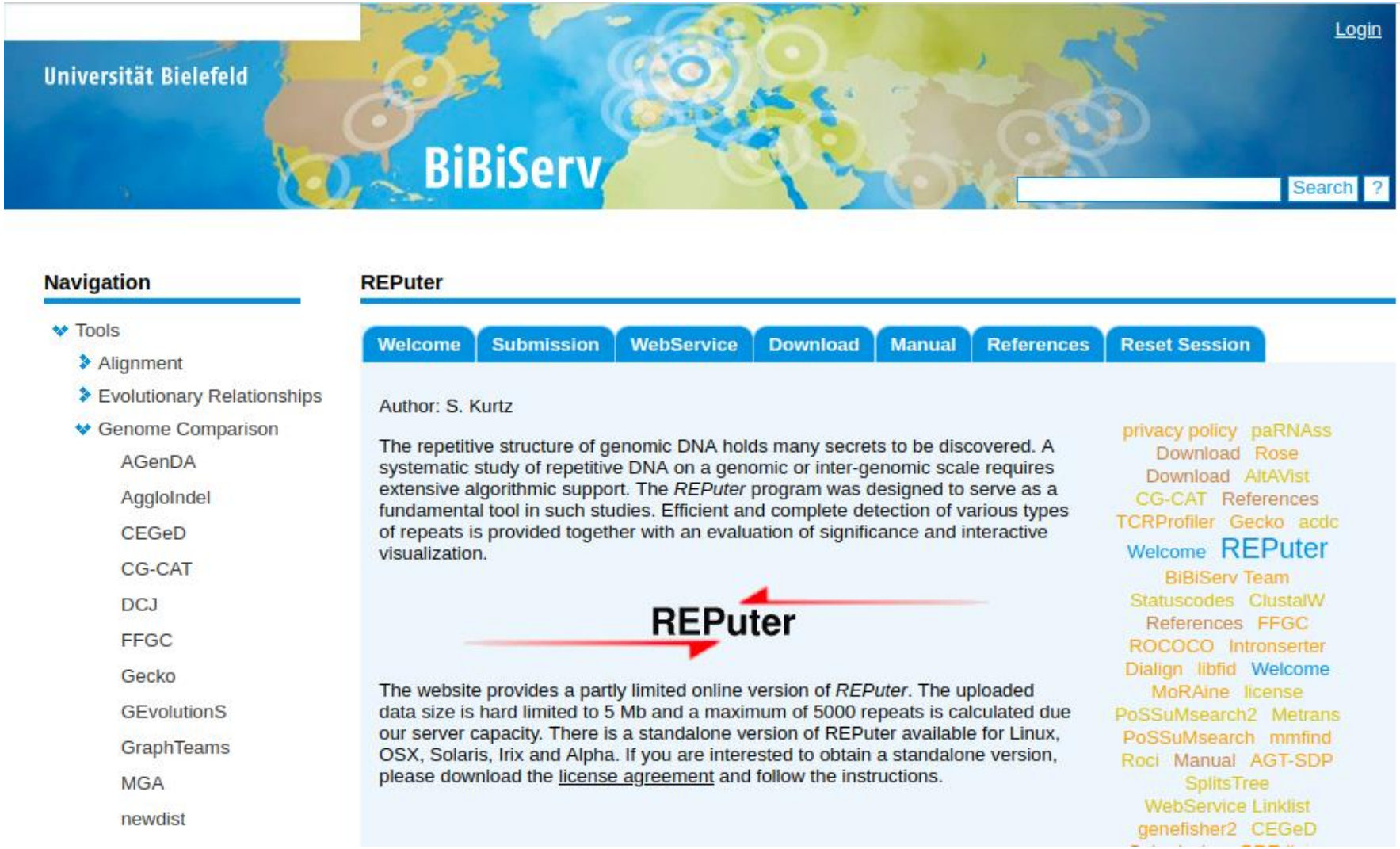


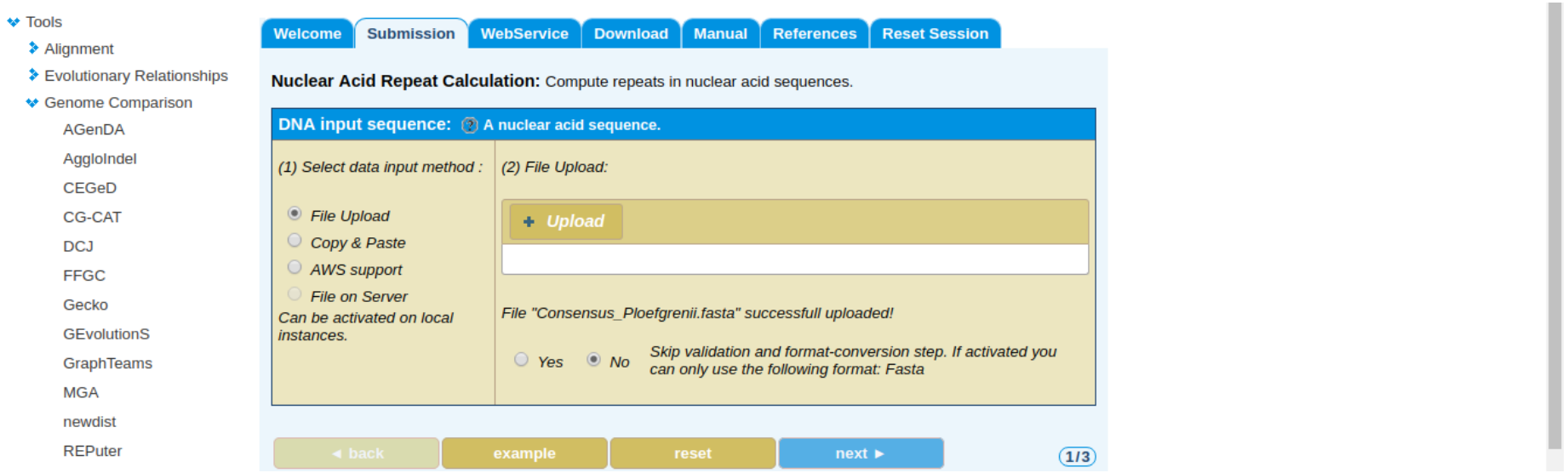
OBS: Para a anotação e depósito da sequência, o ideal após identificar a localização das IRs é alterar o início da sequência do genoma cp. Considere como início da sequência a primeira base após a segunda IR

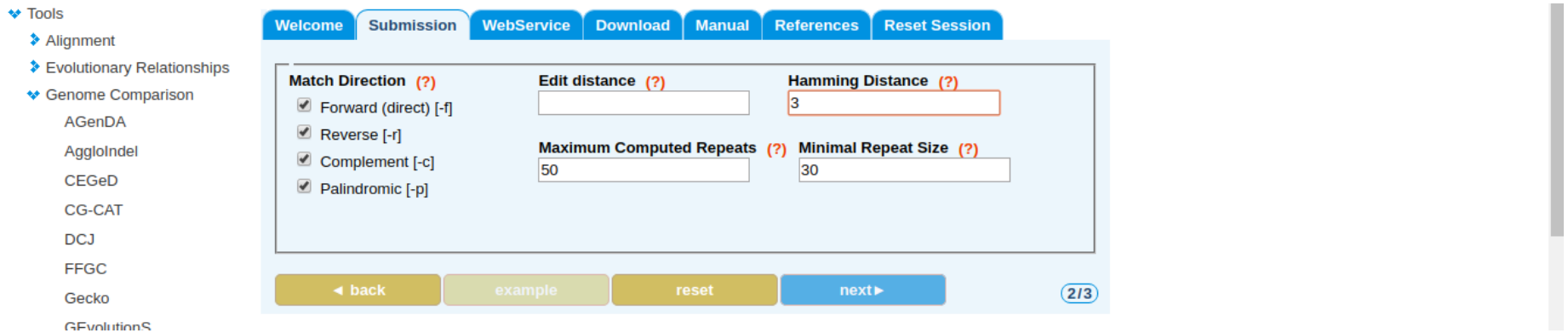
Desta forma o início da sequência será a primeira base da região LSC

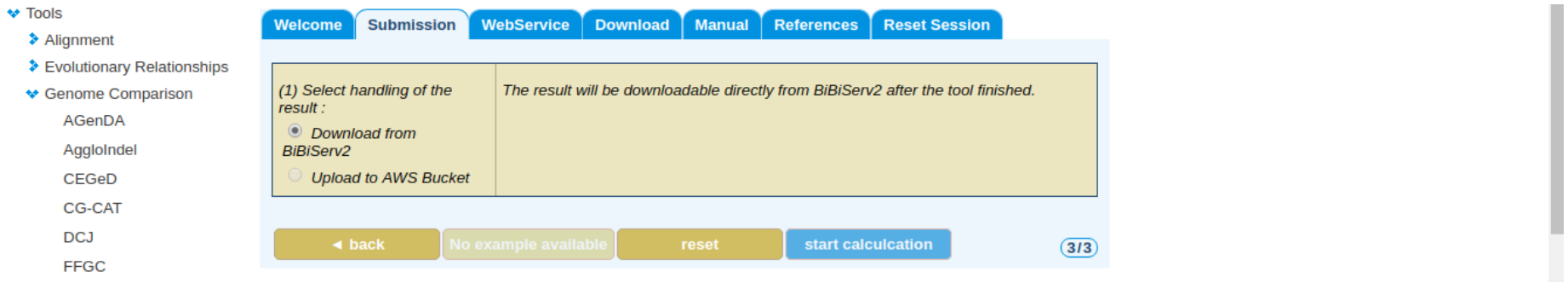
OBTER INFORMAÇÕES DAS REPETIÇÕES

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

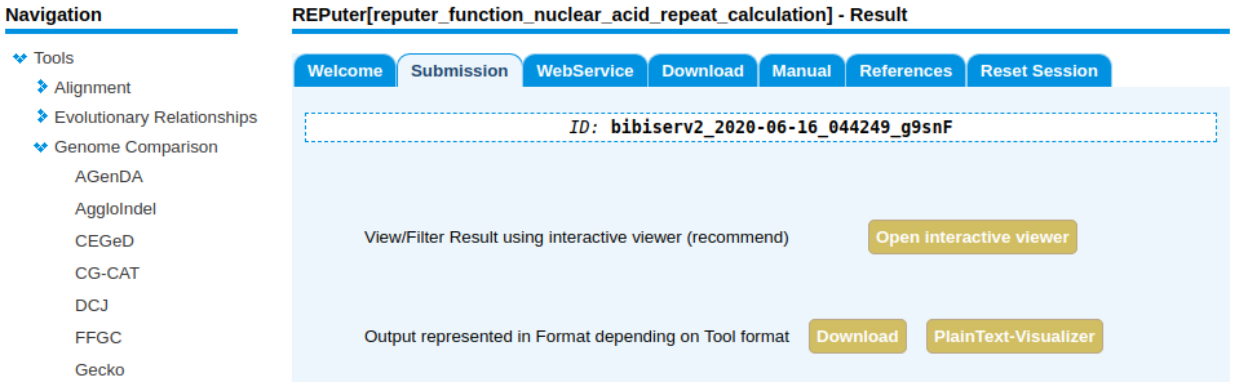



Tools
Alignment
Evolutionary Relationships
Genome Comparison
AGenDA
AggloIndel
CEGeD
CG-CAT
DCJ
FFGC
Gecko
GEvolutionS
GraphTeams
MGA
newdist
REPuter
Welcome
Submission
WebService
Download
Manual
References
Reset Session
Nuclear Acid Repeat Calculation: Compute repeats in nuclear acid sequences.
DNA 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
Welcome
Submission
WebService
Download
Manual
References
Reset Session
Match Direction (?)
Forward (direct) [-f]
Reverse [-r]
Complement [-c]
Palindromic [-p]
Edit distance (?)
Hamming Distance (?)
3
Maximum Computed Repeats (?)
50
Minimal Repeat Size (?)
30
◄ 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 calculcation
3/3

OBTER INFORMAÇÕES DAS REPETIÇÕES

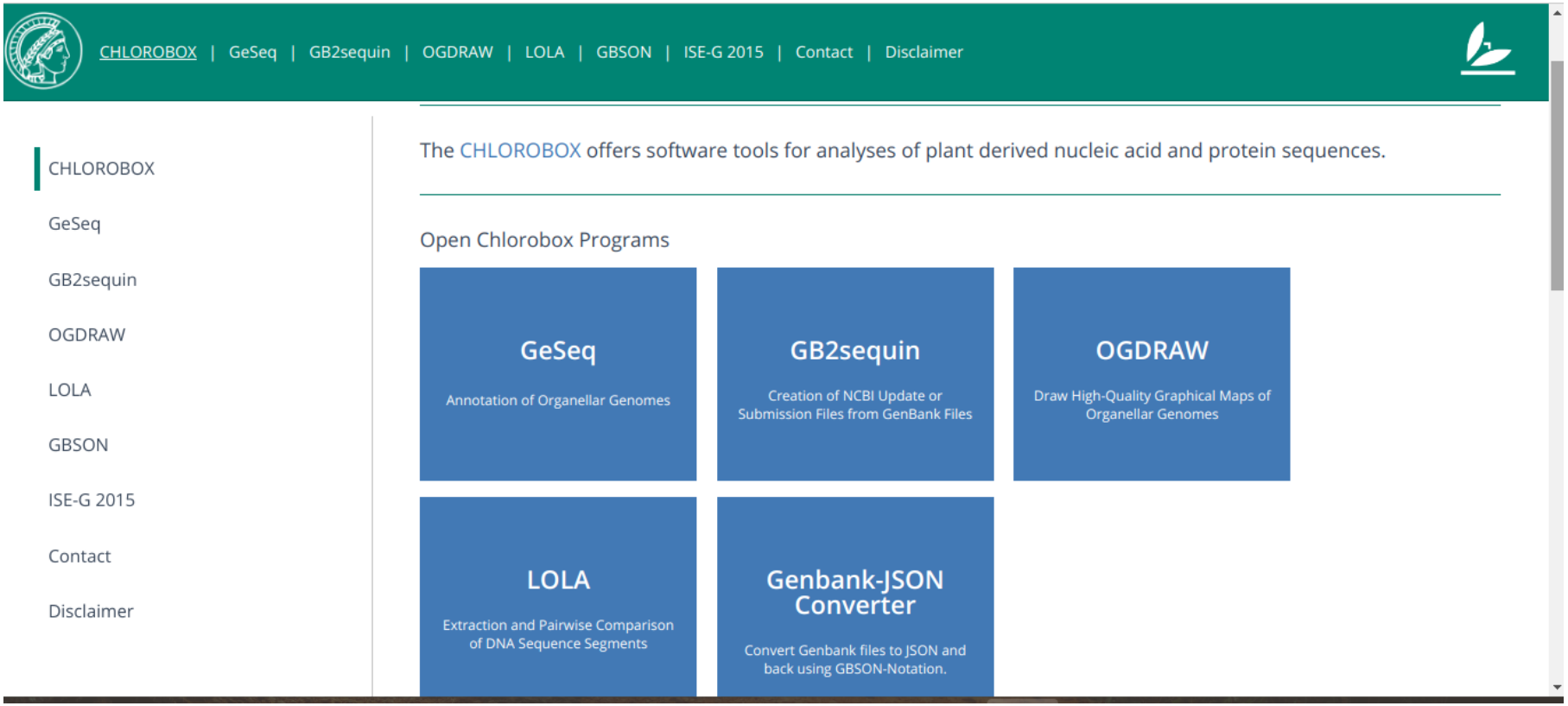



```
# /vol/bioapps/bin/repfind -c -f -p -r -l 30 -h 3 -best 50 /var/bibiserv2/anonymou
# 146537  3 30 /var/bibiserv2/anonymous/reputer/16/04/42/bibiserv2_2020-06-16_044
23450  44365 P 23450  81081   0 0.00e+00
 103 145913 F  103 145904  -1 1.81e-50
  97     48 F   97    156  -3 9.58e-43
  96    222 F   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

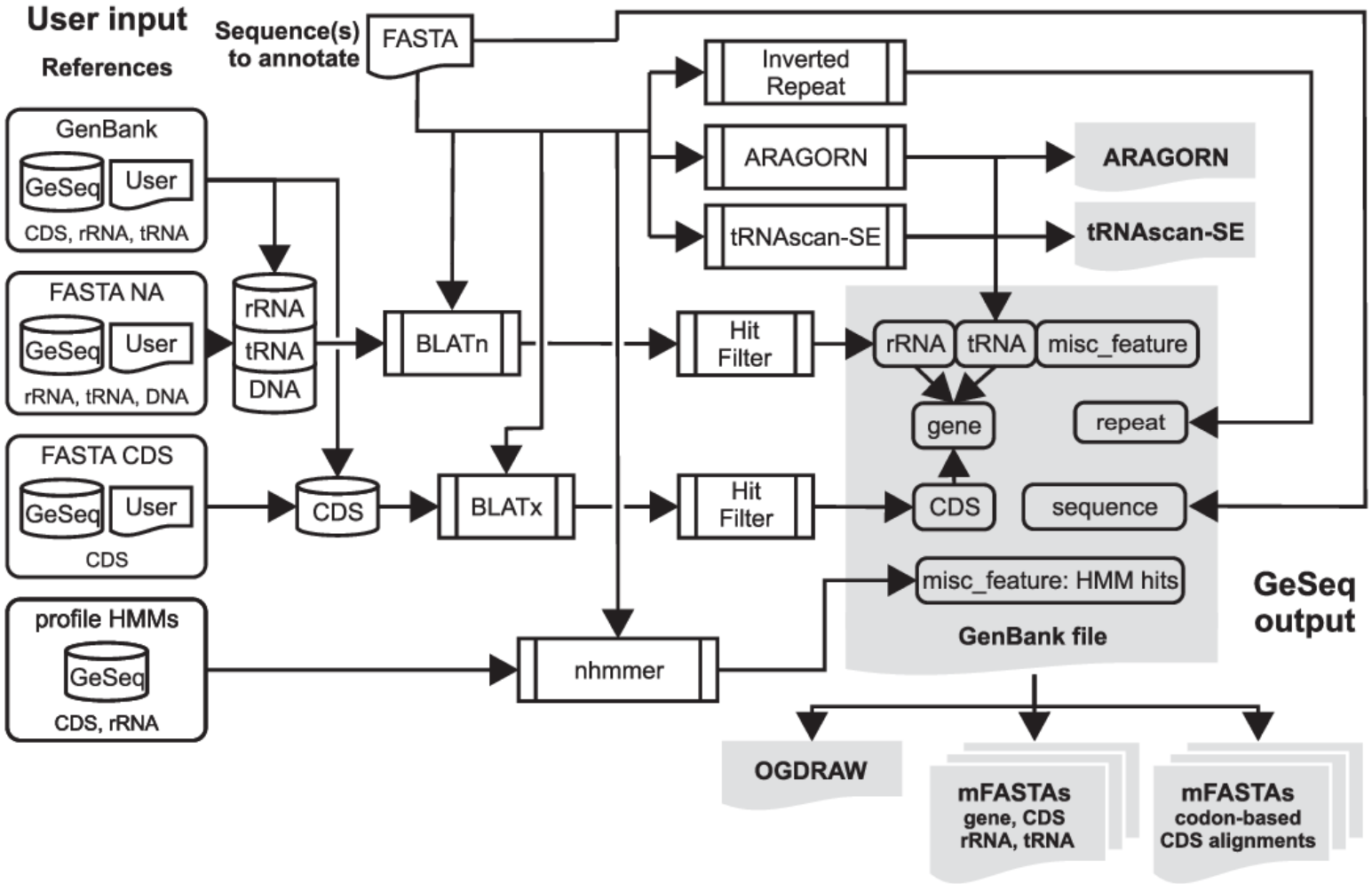
CHLOROBOX: Programas para análises com dados de genomas organelares
https://chlorobox.mpimp-golm.mpg.de/

Max Planck Institute of Molecular Plant Physiology



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

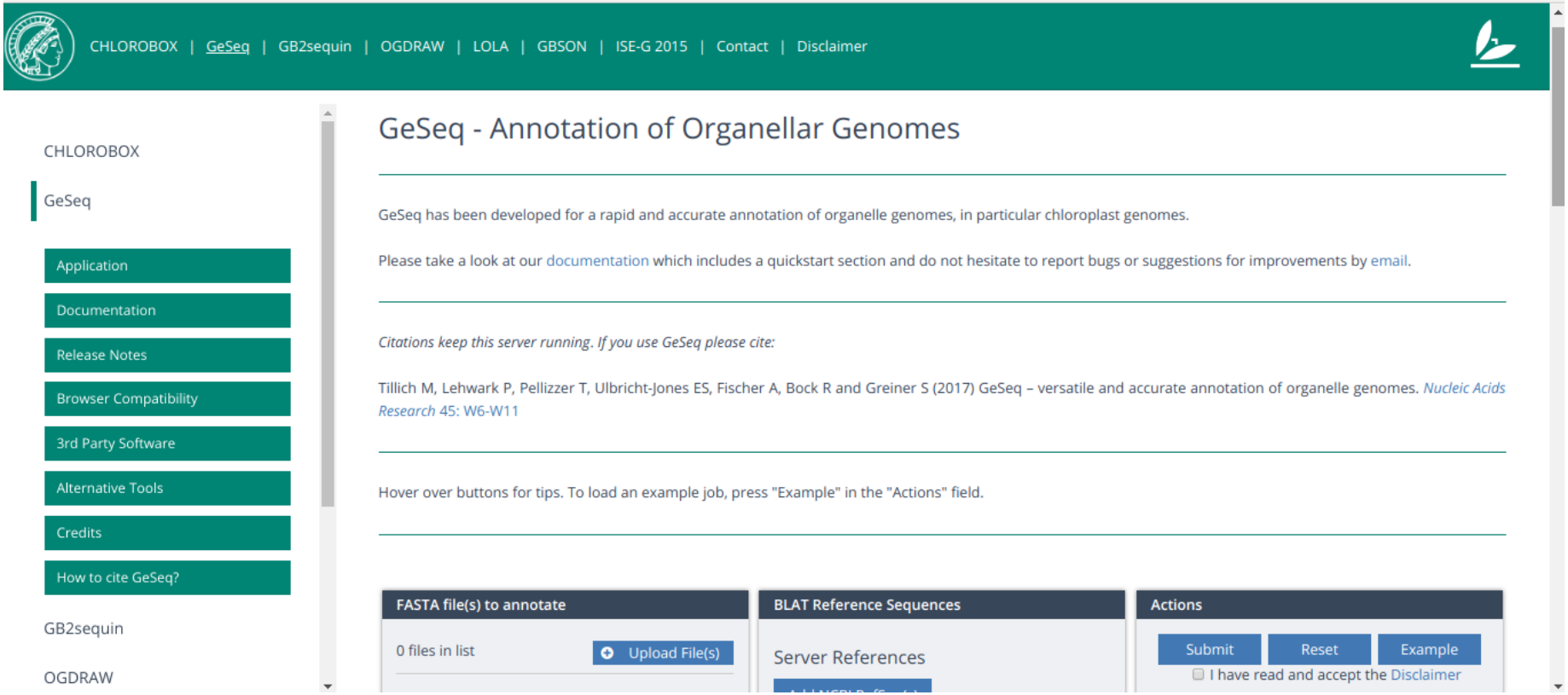
https://chlorobox.mpimp-golm.mpg.de/geseq.html

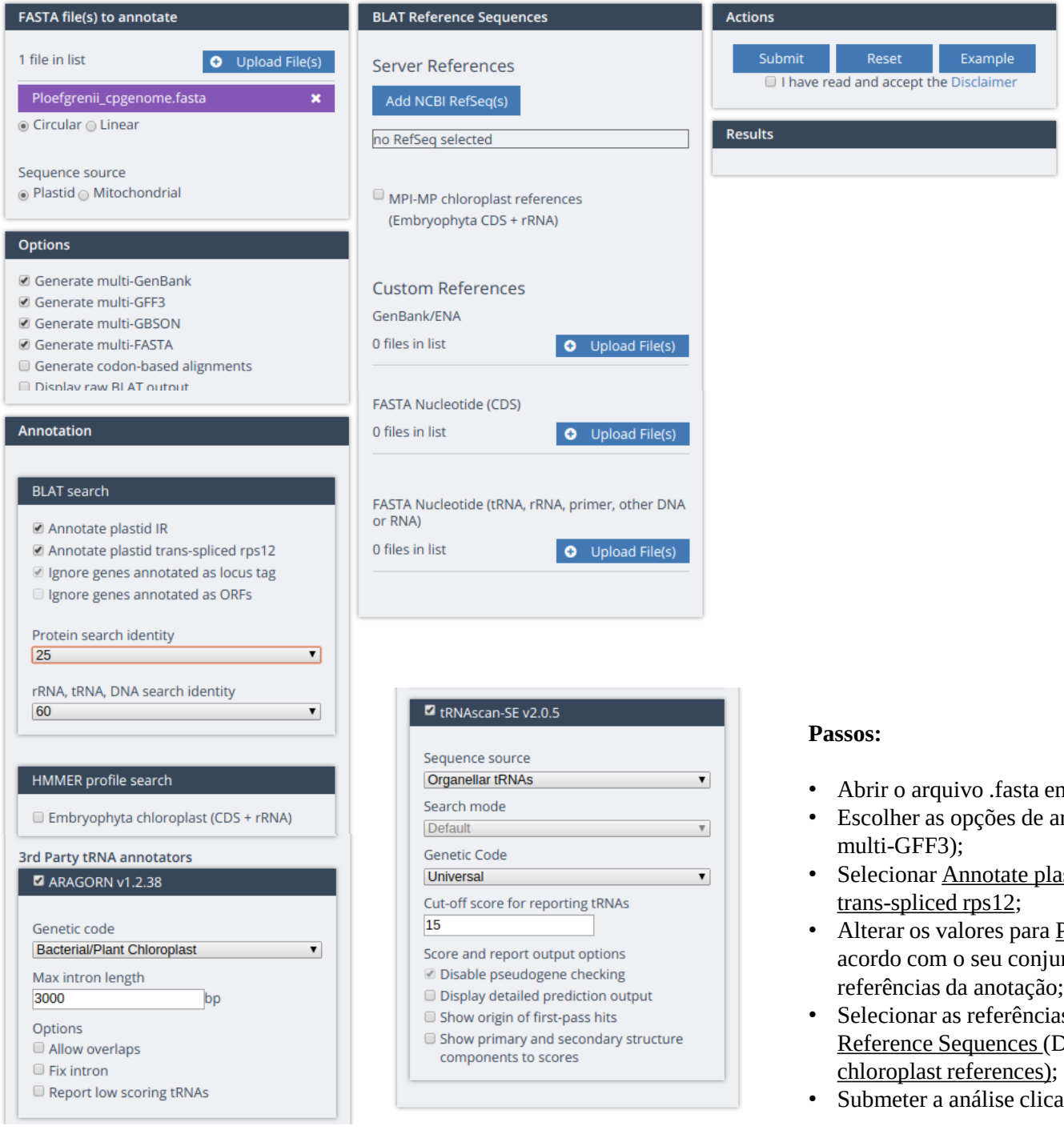


ANOTAÇÃO

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

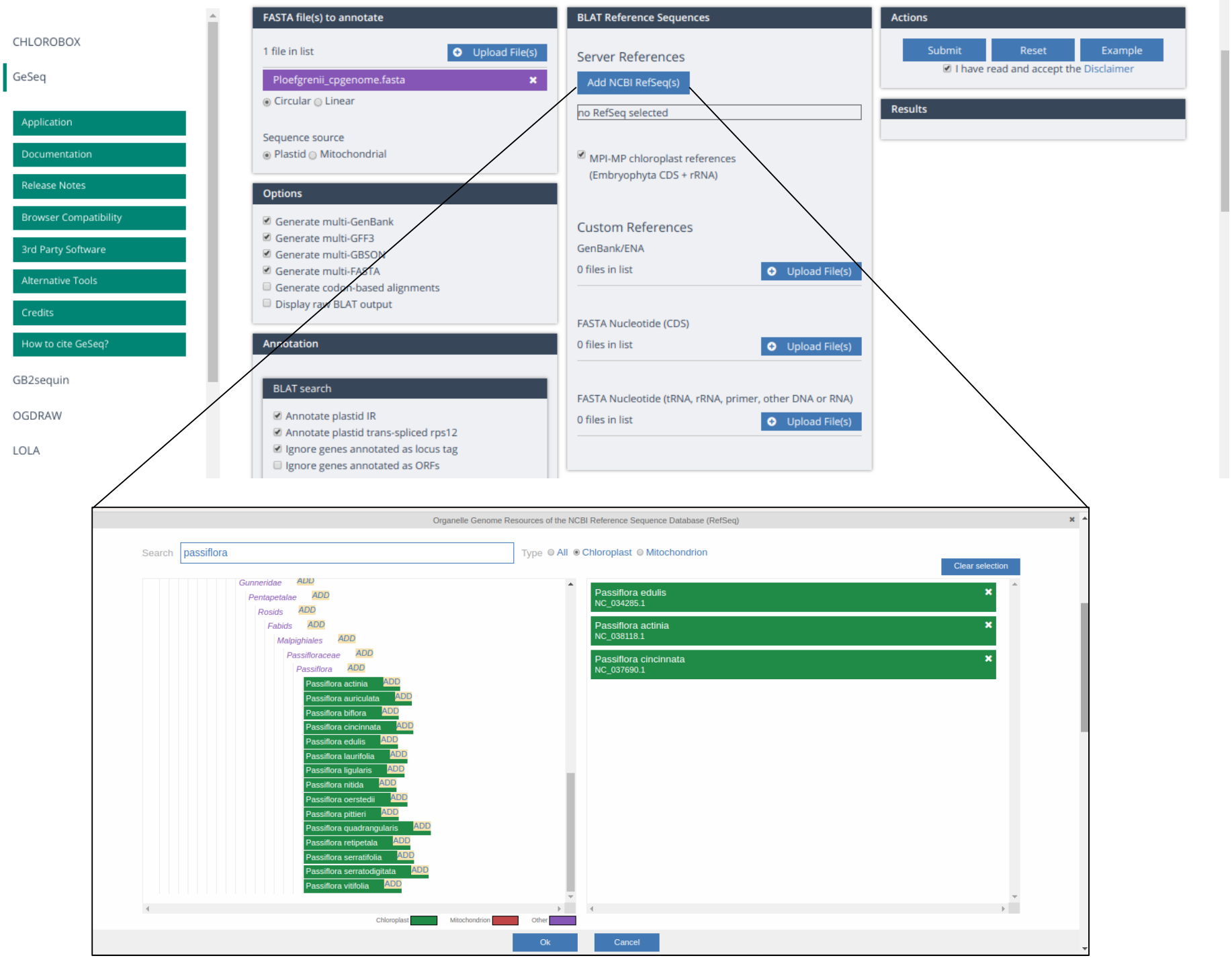
https://chlorobox.mpimp-golm.mpg.de/geseq.html



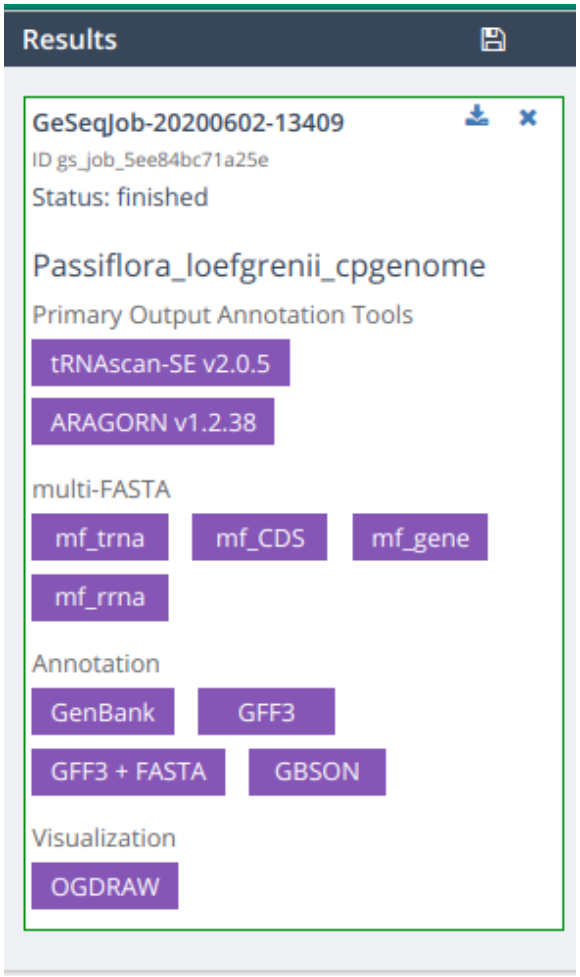


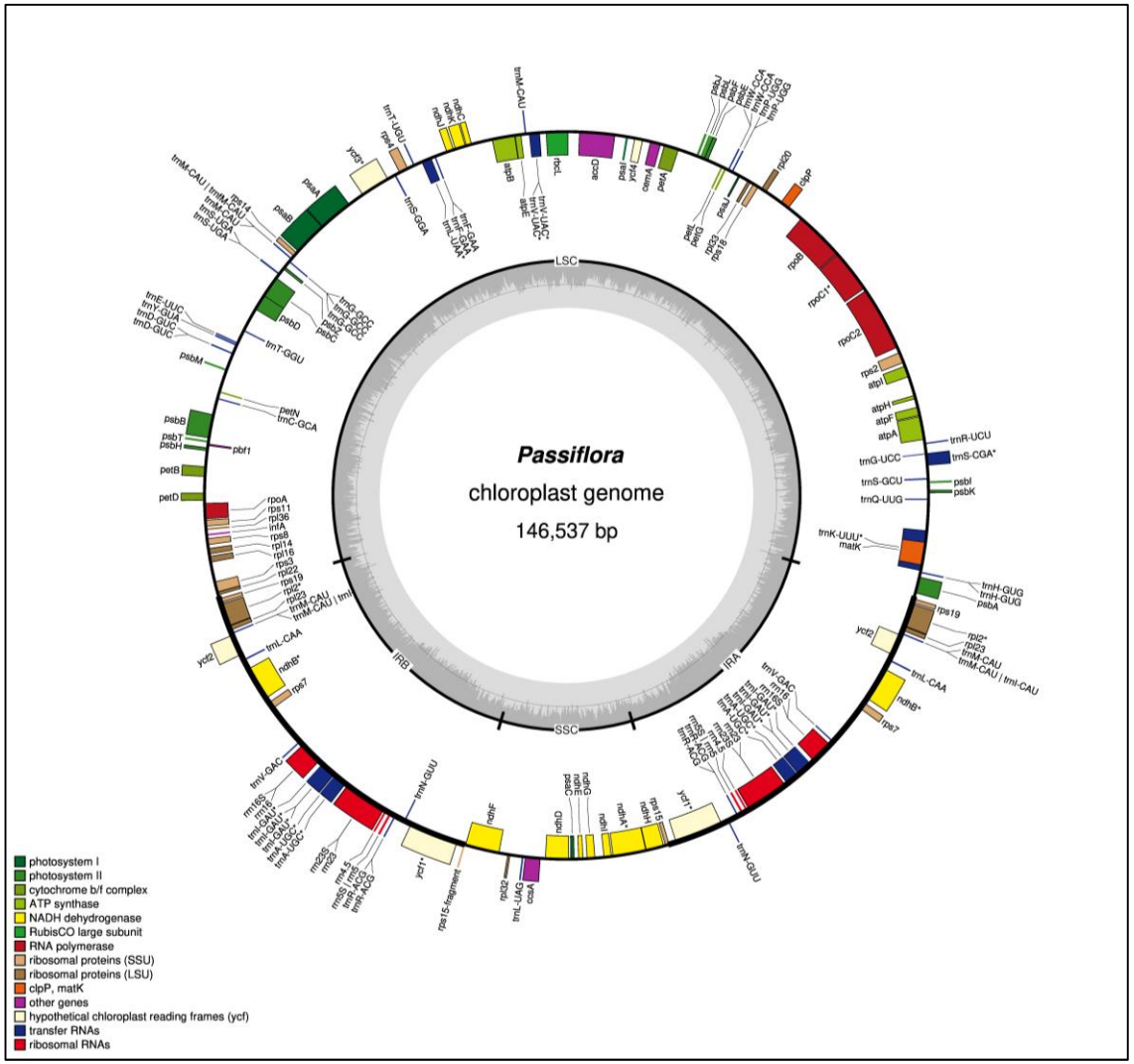
Passos:

- 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.


CHLOROBOX
GeSeq
Application
Documentation
Release Notes
Browser Compatibility
3rd Party Software
Alternative Tools
Credits
How to cite GeSeq?
GB2sequin
OGDRAW
LOLA
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
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
passiflora
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:





Resultados da anotação no Geseq:

File GeSeqJob-20200602-13409_Passiflora_loefgrenii_cpgenome_GenBank.gb

```
                     circular=true
            Program  RNA Editing disabled
            Submitted  16 Jun 2020 04:34:14 GMT
            Started  16 Jun 2020 04:34:17 GMT
            Finished  16 Jun 2020 04:35:30 GMT
FEATURES             Location/Qualifiers
     source          1..146537
                     /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, psl
                     score 99.3, coverage 100.00%, match 99.25%"
     CDS             229..1290
                     /gene="psbA"
                     /translation="MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTA
                     TSVFIIAFIAAPPVDIDGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASV
                     DEWLYNGGPYELIVLHFLLGVACYMGREWELSFRLGMRPWIAVAYSAPVAAAAAVFLI
                     YPIGQGSFSDGMPLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSL
                     VTSSLIRETTENESANEGYRFGQEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLA
                     AWPVIGIWFTALGISTMAFNLNGFNFNQSVVDSQGRVINTWADIINRANLGMEVMHER
                     NAHNFPLDLAAVEAPSTNG"
```

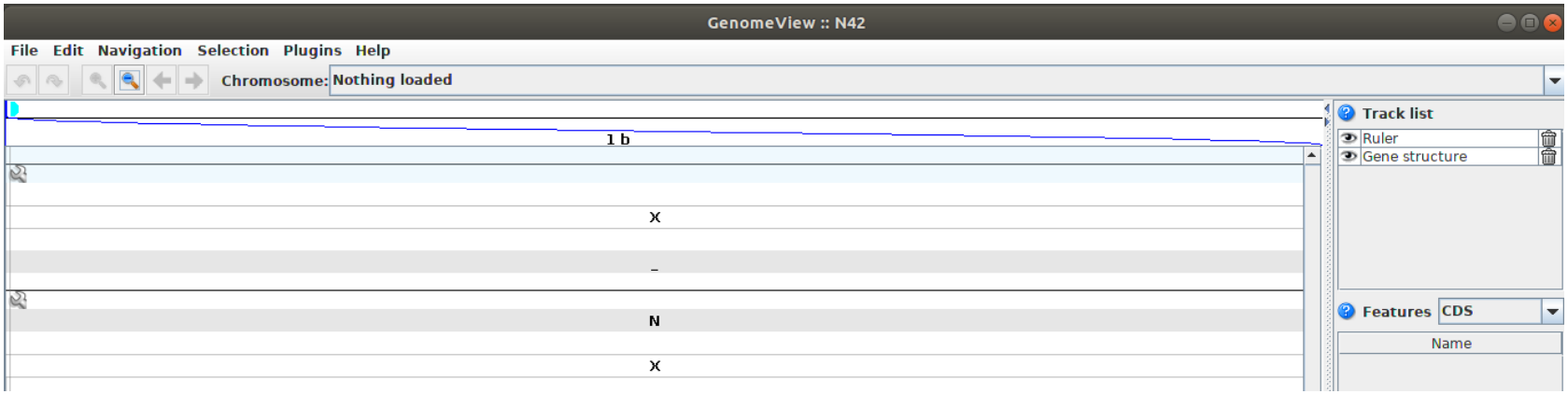
Arquivo .gff3 com os dados da anotação:

File GeSeqJob-20200602-13409_Passiflora_loefgrenii_cpgenome_GFF3.gff3

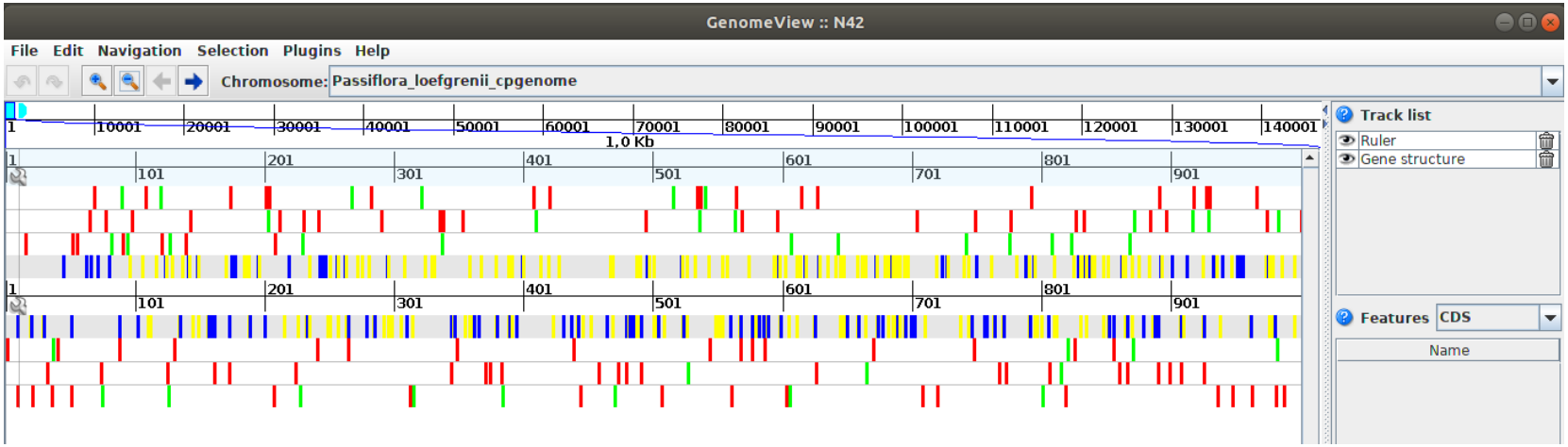
```
##gff-version 3
##source-version GeSeq 1.82
##sequence-region     Passiflora_loefgrenii_cpgenome       1      146537
Passiflora_loefgrenii_cpgenome       GeSeq  region 1      146537 .      +      0      Is_circular=true
Passiflora_loefgrenii_cpgenome       GeSeq  source 1      146537 .      +      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
```

CURADORIA MANUAL DA ANOTAÇÃO

GenomeView: https://genomeview.org/

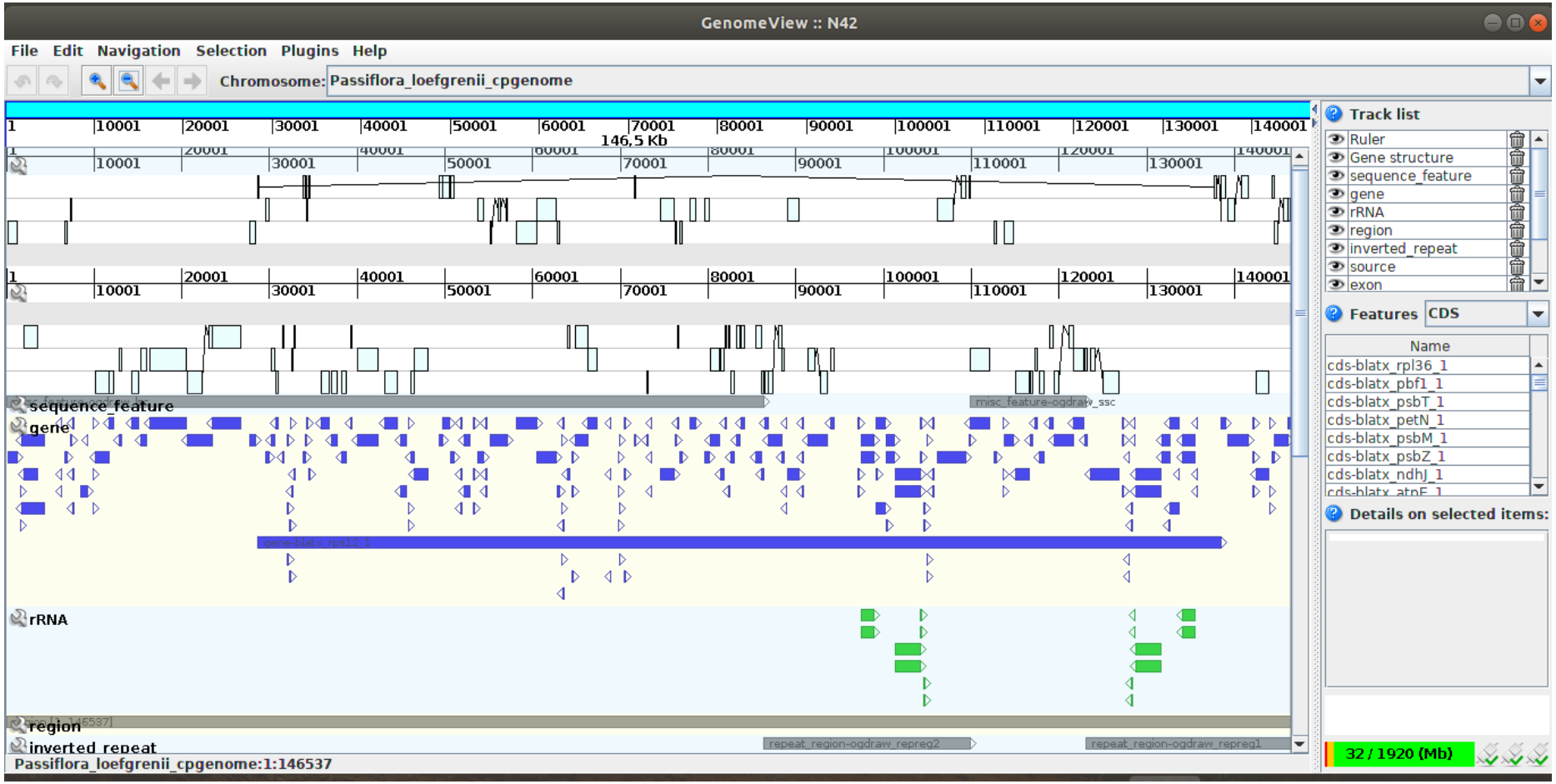


Primeiro passo: Abrir o arquivo fasta com a sequência do genoma cloroplastidial em File / Load data



CURADORIA MANUAL DA ANOTAÇÃO

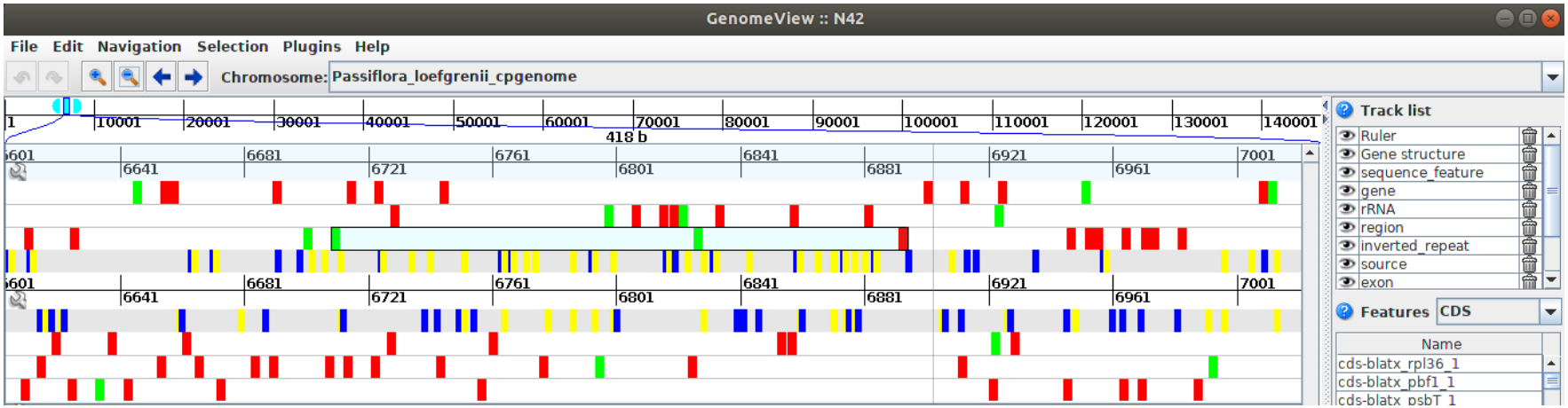
Abrir o arquivo .gff3 da anotação do genoma cloroplastidial em File / Load data



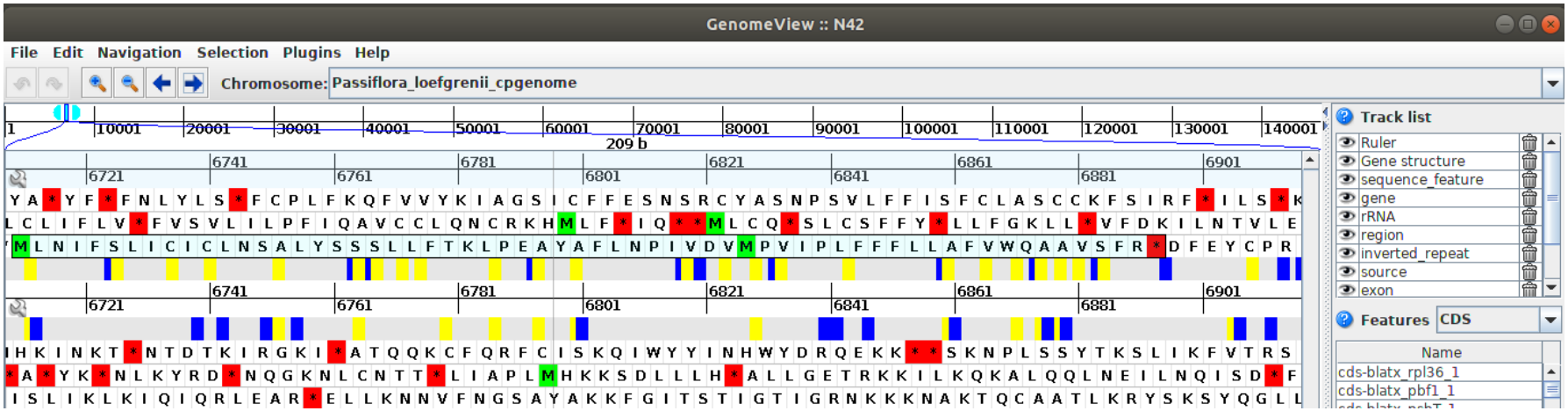
CURADORIA MANUAL DA ANOTAÇÃO

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

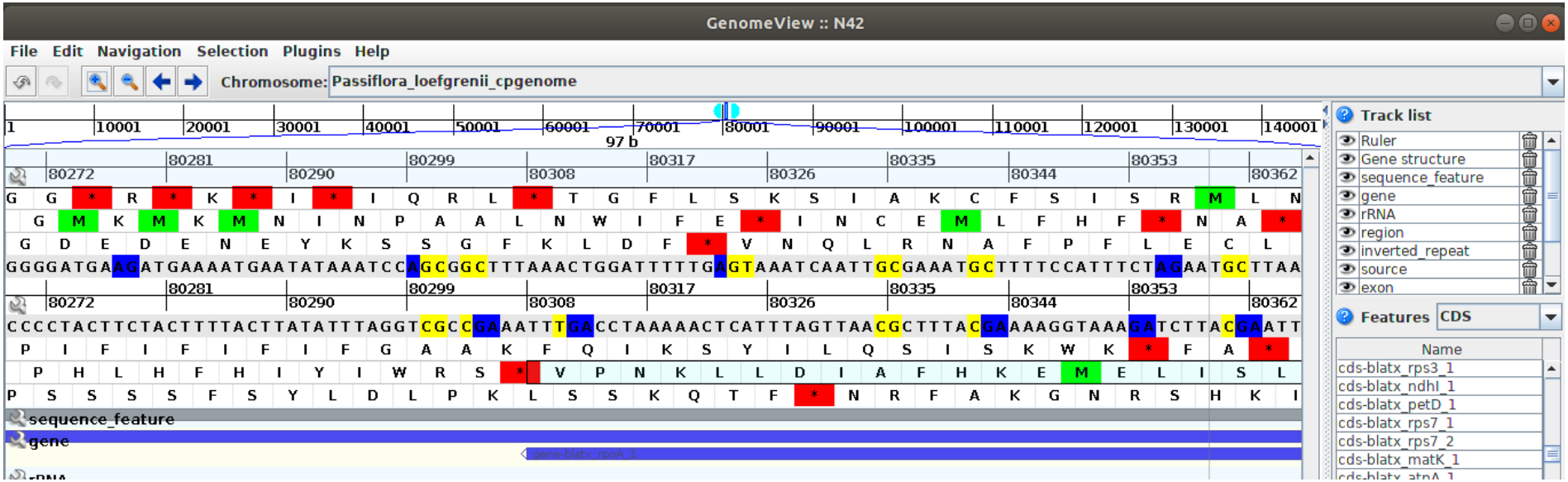
Exemplo: Estrutura do gene *psbK*:



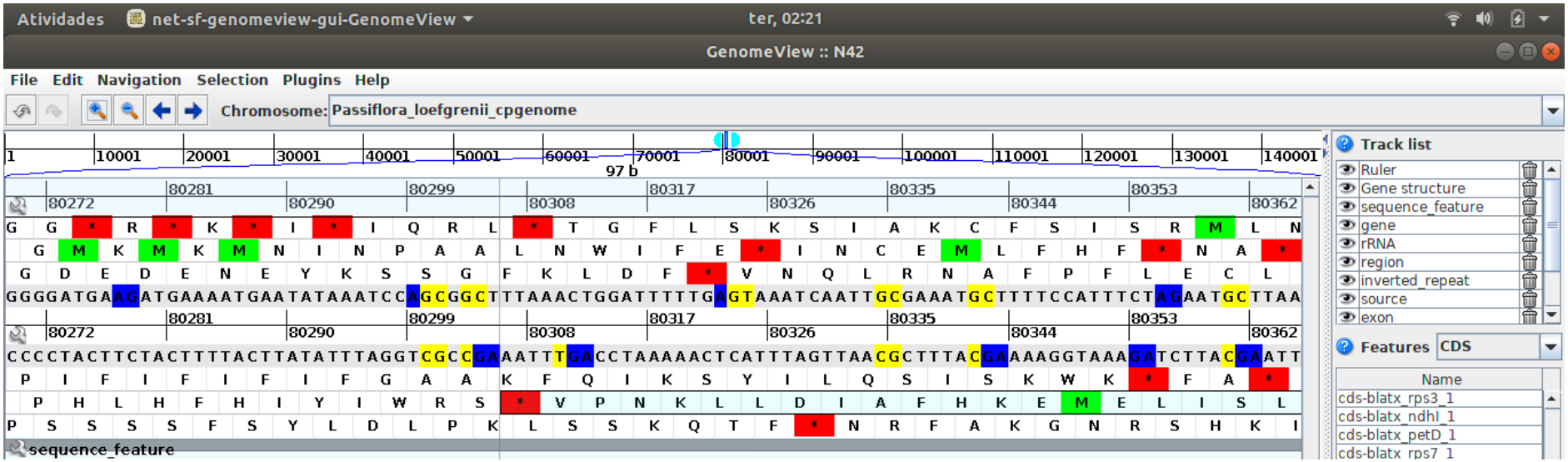
Exemplo: Sequência peptídica codificada pelo gene *psbK*:



Corrigir a anotação de sequências faltando *stop códons*:



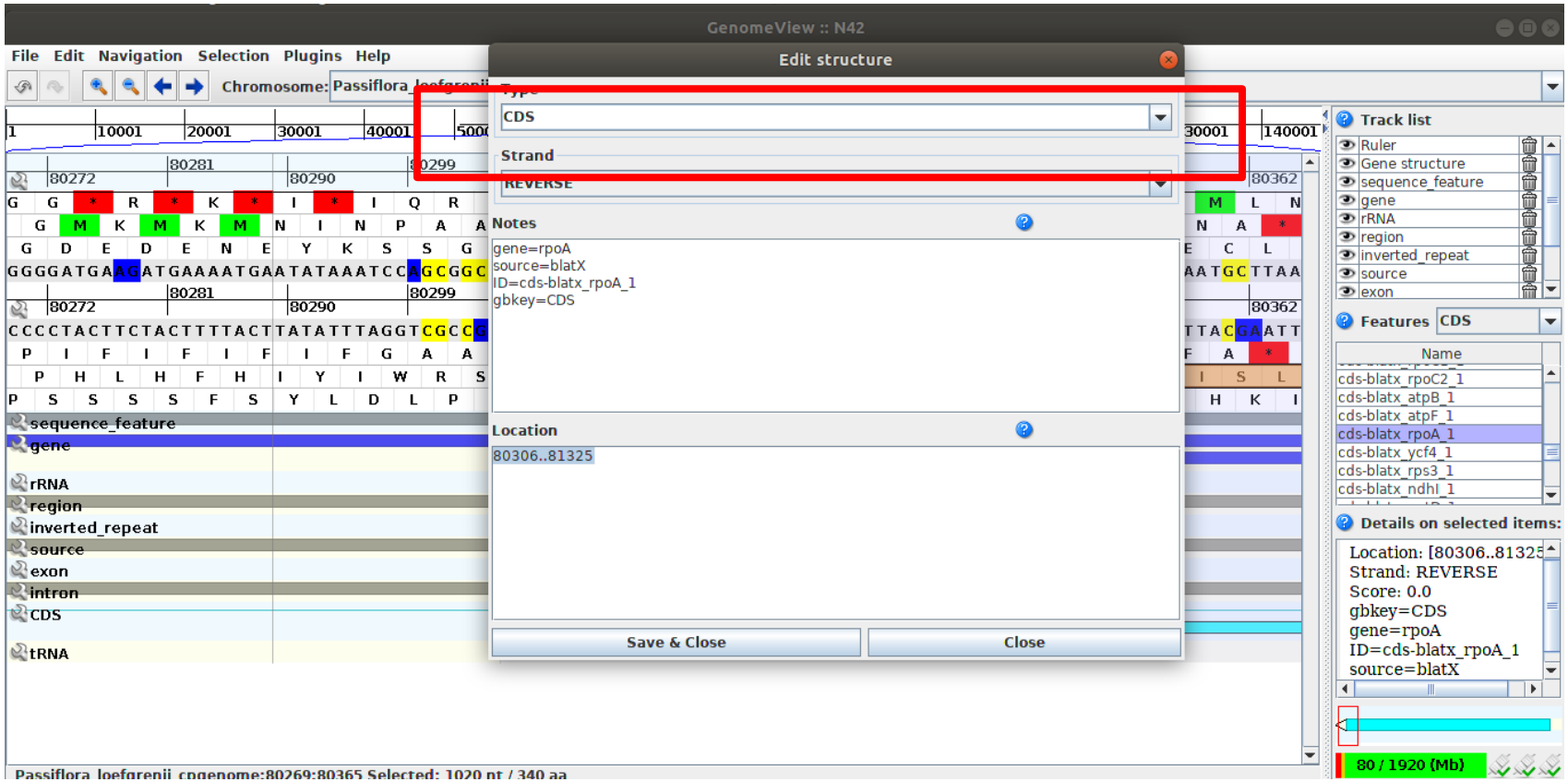
Exemplo: Completar a sequência adicionando na predição as bases que faltavam para o stop códon:



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

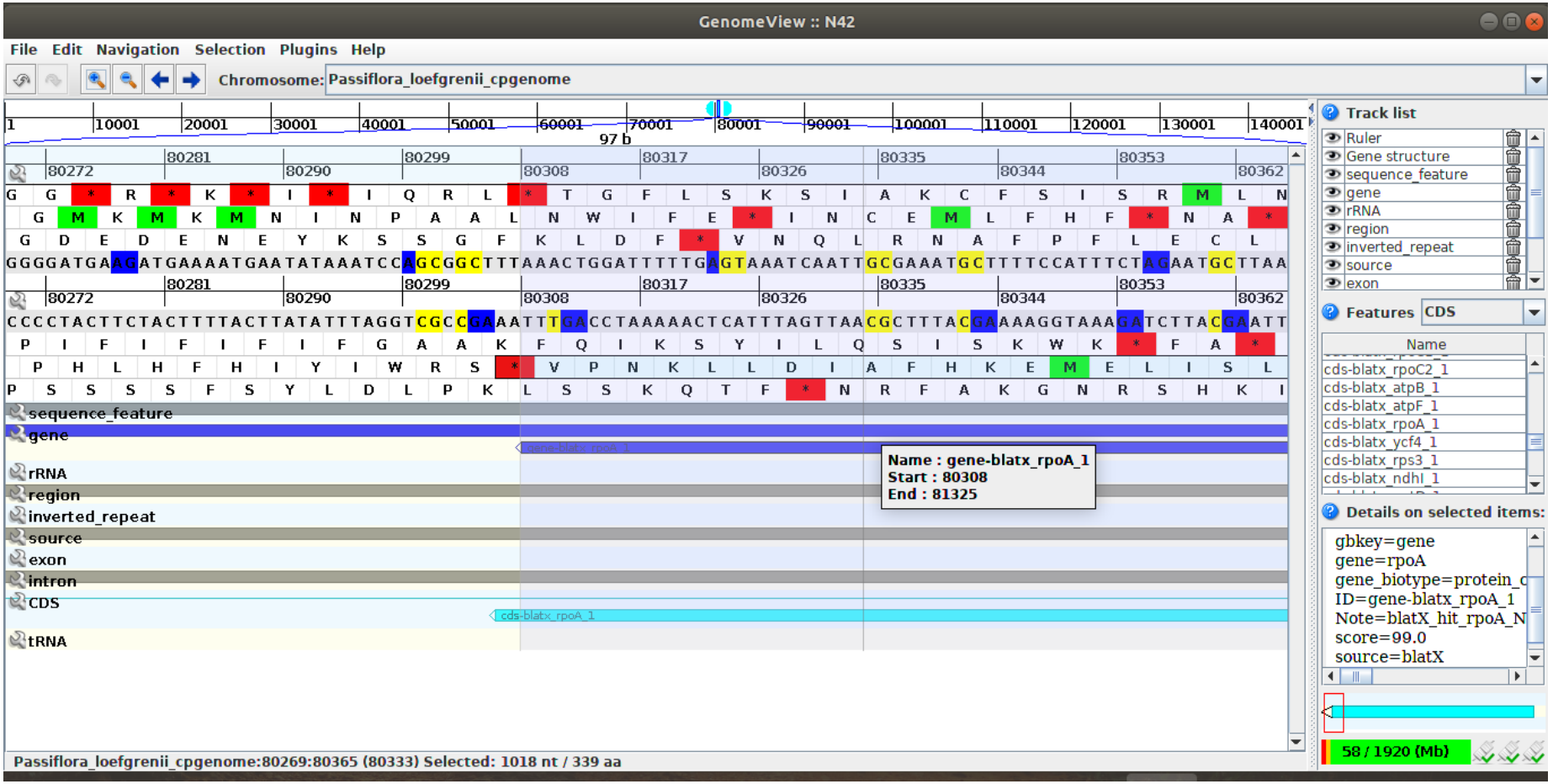
Na caixa Edit structure alterar as coordenadas das posições do gene na opção Location

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



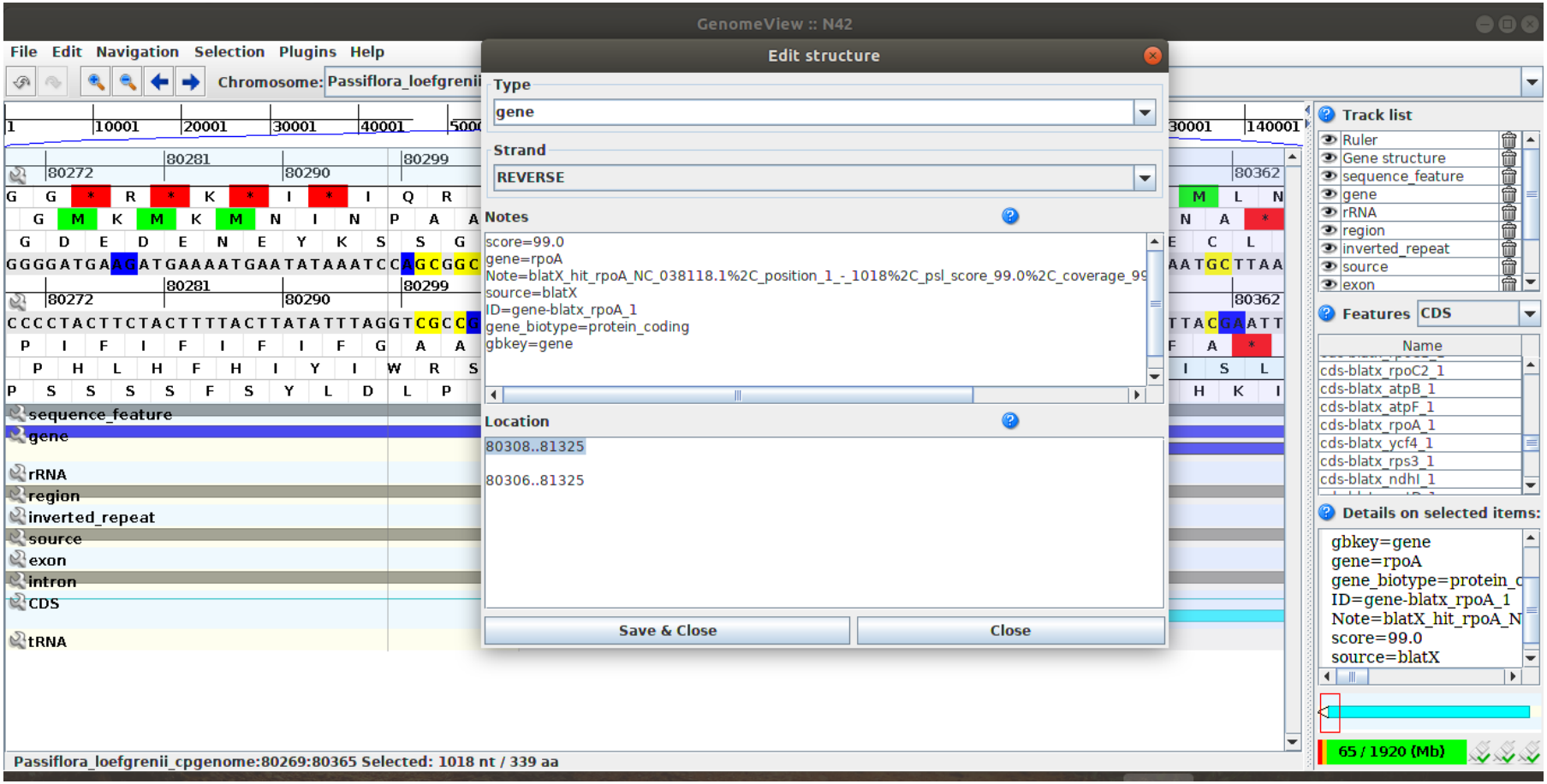
Obs: A primeira correção alterou apenas as posições da CDS

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



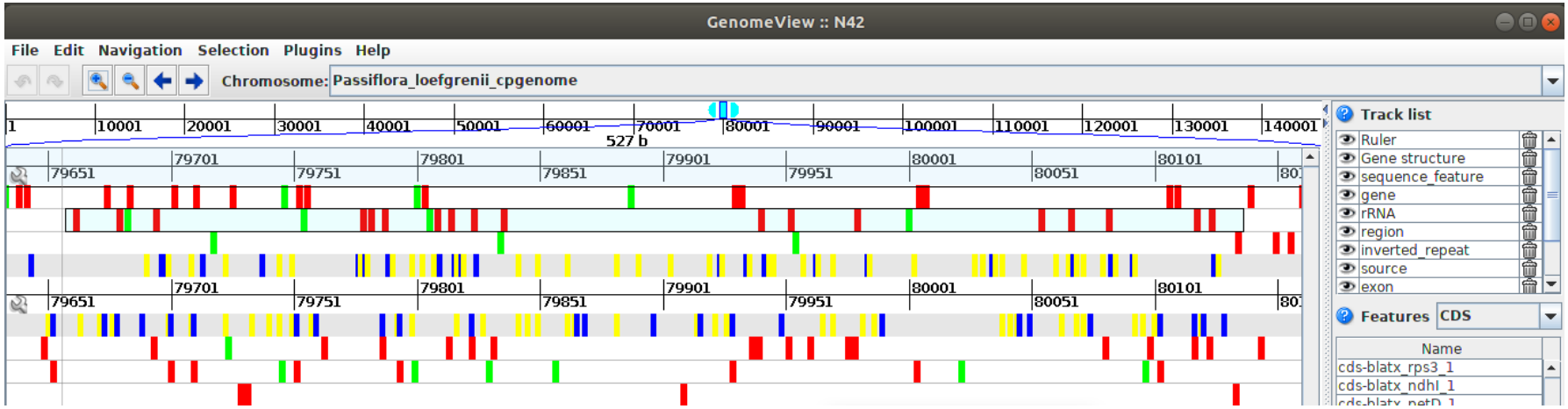
Na caixa Edit structure altere as coordenadas das posições do gene na opção Location

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

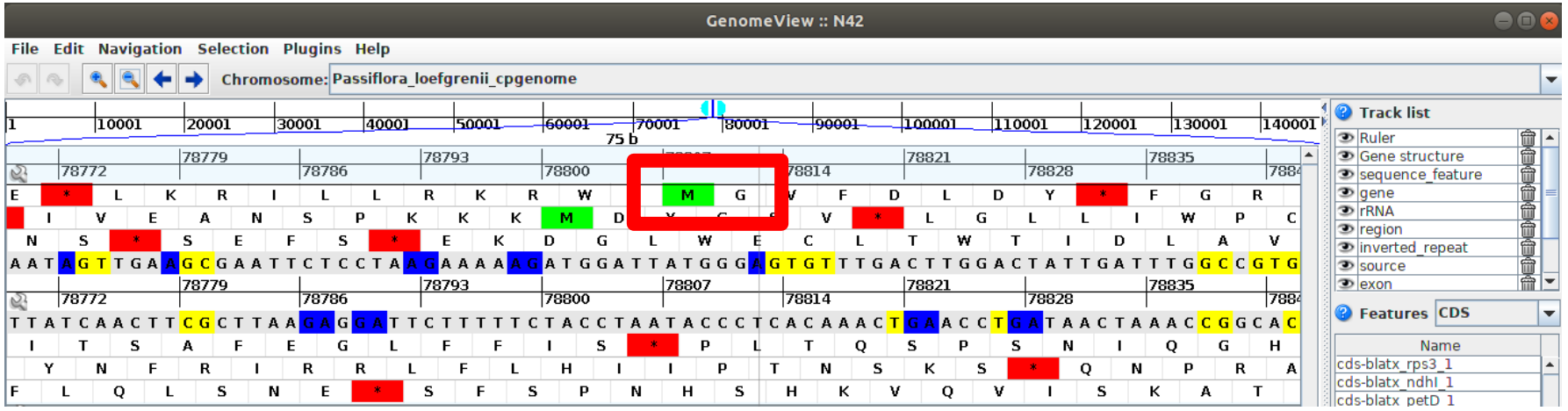


Corrigir a anotação de sequências contendo *stop códons* internos: Observar os outros quadros leitura para esta sequência, e procurar por aquele que não apresenta stop códons

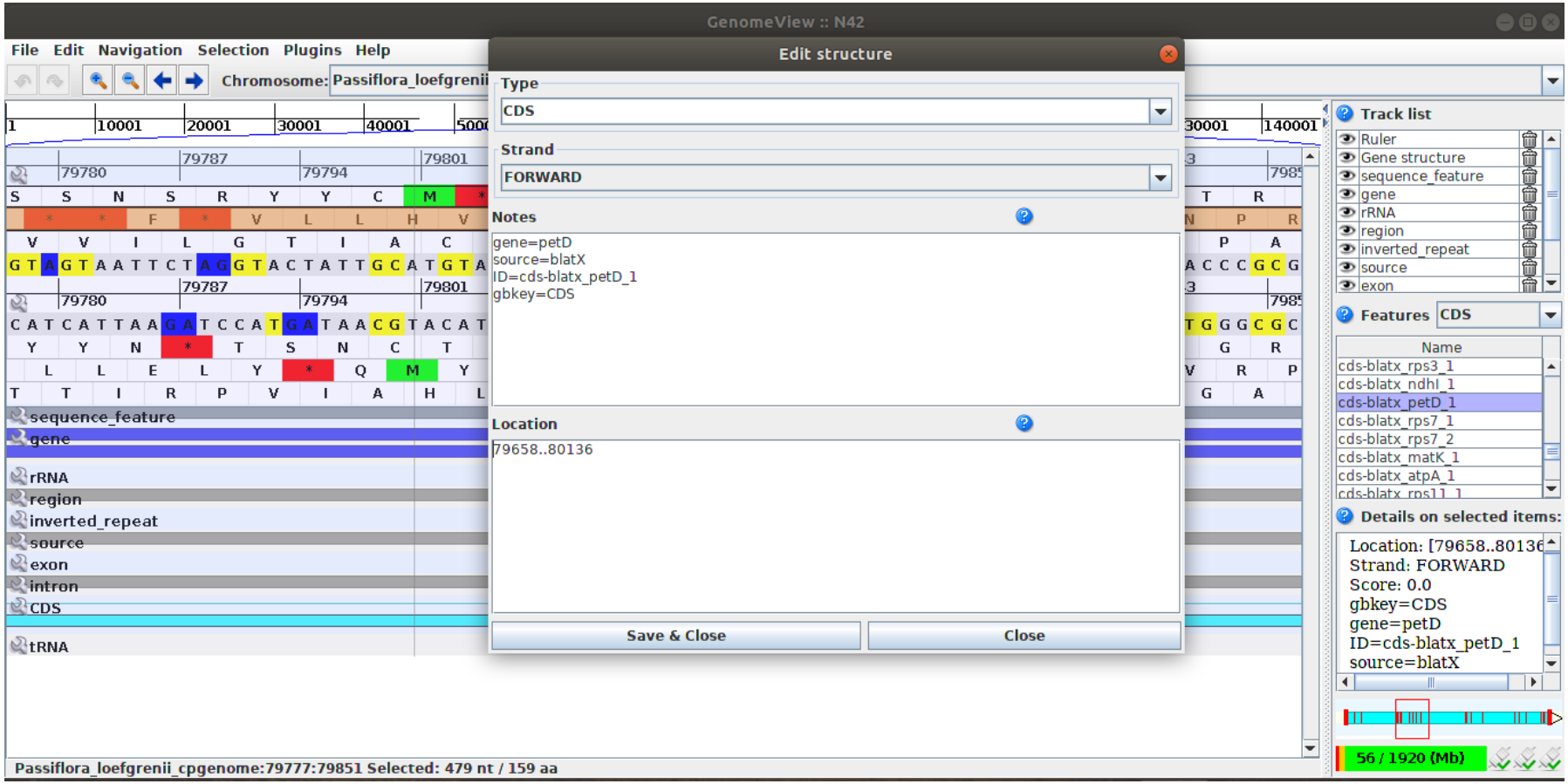
Exemplo: Estrutura predita do gene *petD* contendo stop códons:



Exemplo: Correção da anotação do gene *petD*:

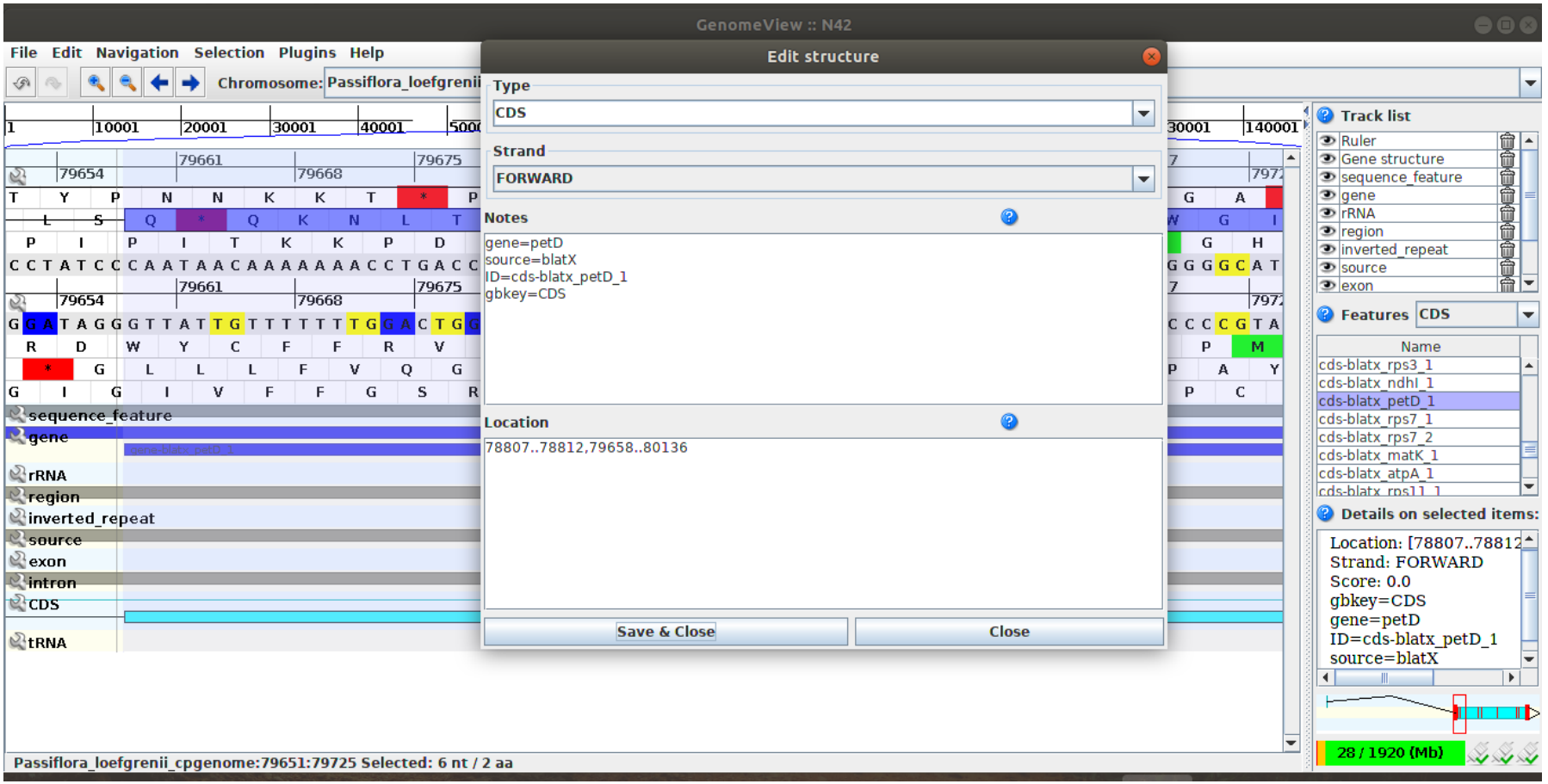


Para corrigir a anotação da estrutura de um gene: Clicar sobre a estrutura do gene, clicar na opção Edit na barra de ferramentas, e escolher a função Edit structure

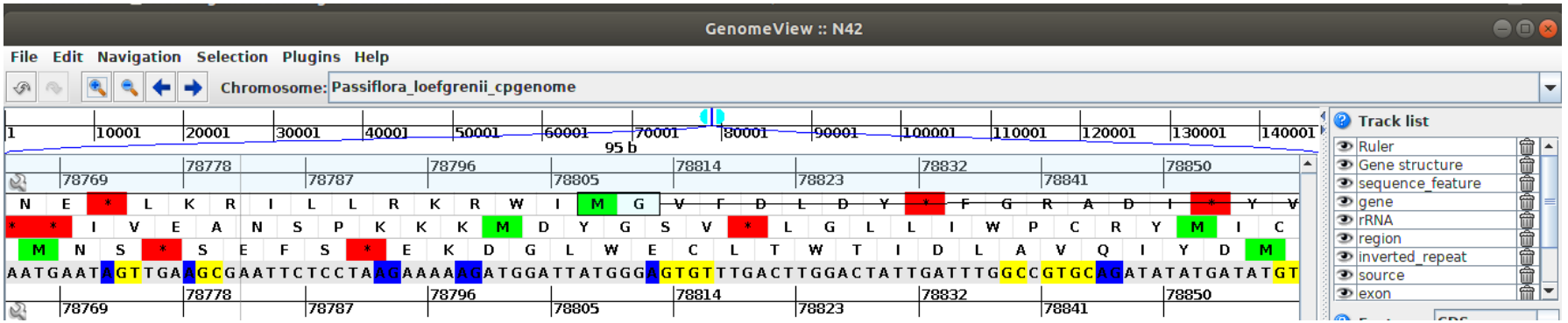


Para corrigir a anotação da estrutura de um gene: Na caixa Edit structure altere as coordenadas das posições do gene na opção Location

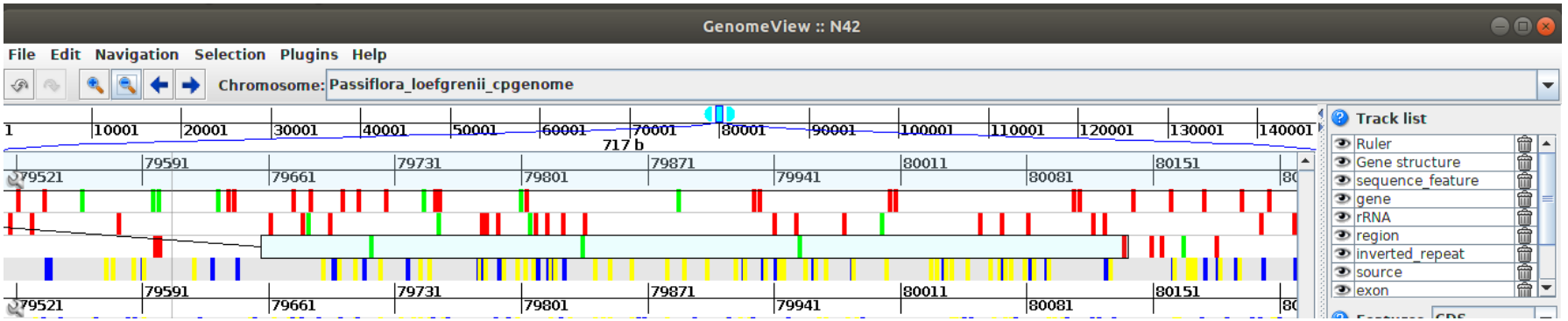
Exemplo: A localização do novo éxon para o gene *petD* foi incluída na opção Location:



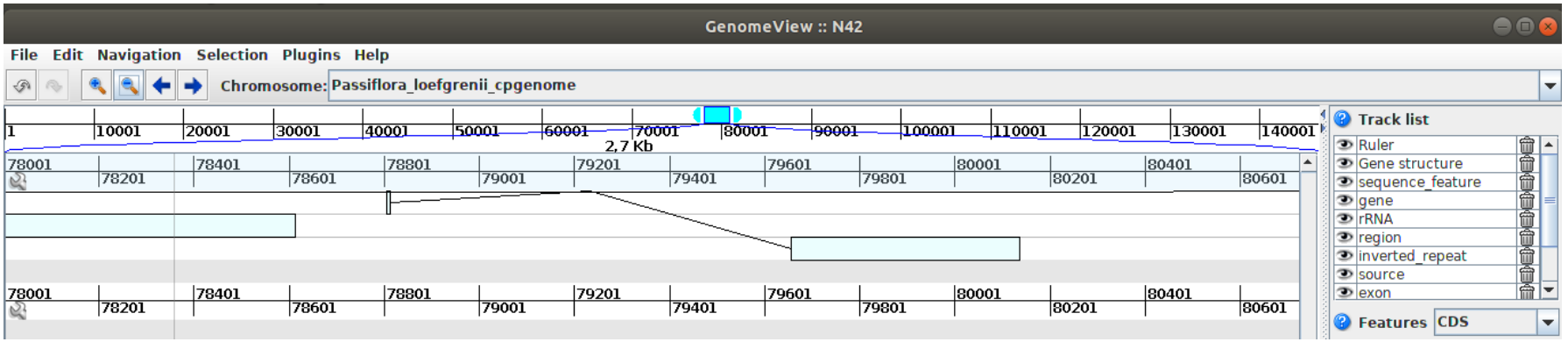
Exemplo: Primeiro éxon do gene *petD*:



Exemplo: Segundo éxon do gene *petD*:

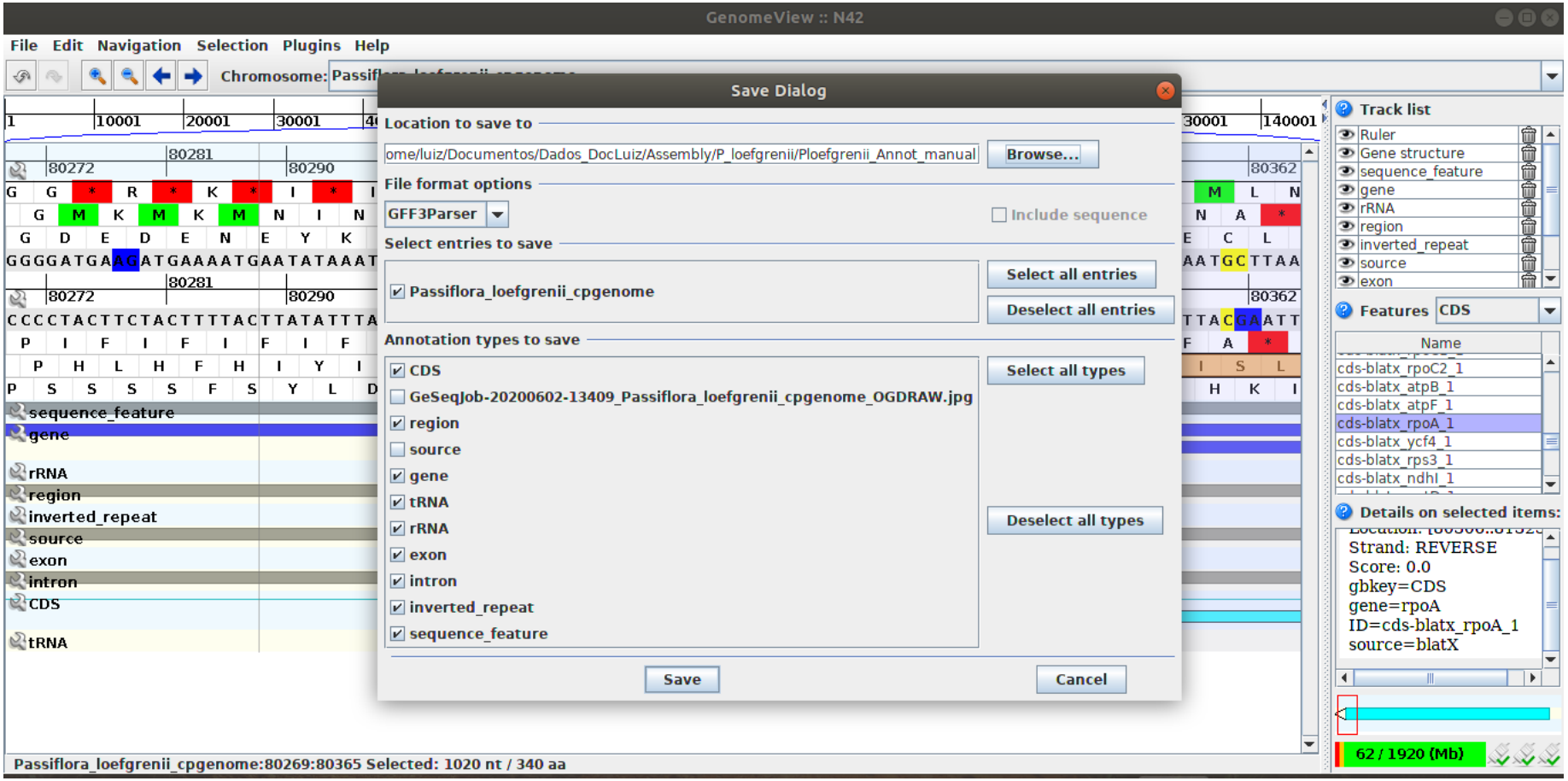


Exemplo: Estrutura completa do gene *petD*:



Conferiu a anotação de todos os genes?

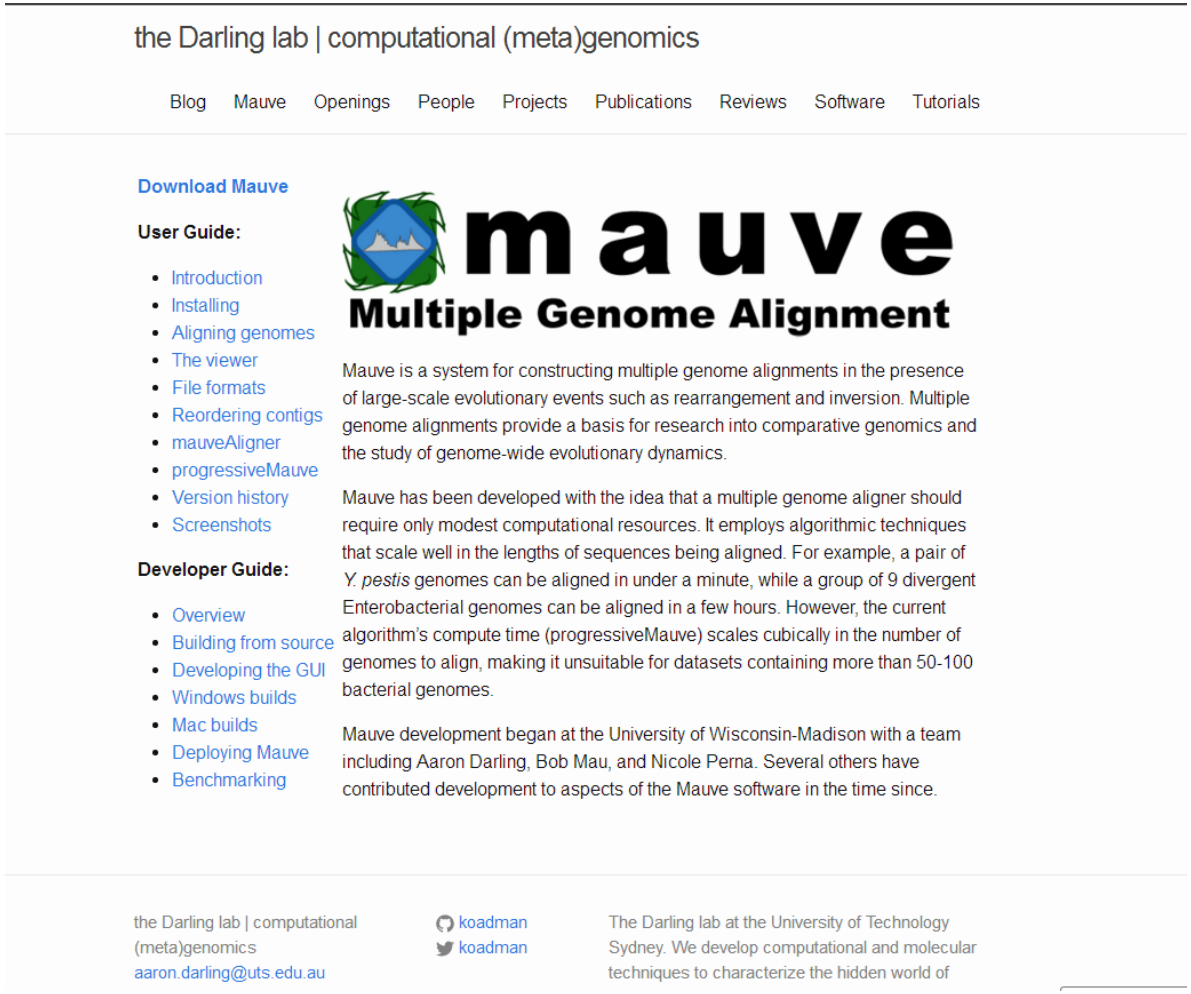
Para salvar o novo arquivo de anotação: Clique em File, e depois em Save annotation.



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

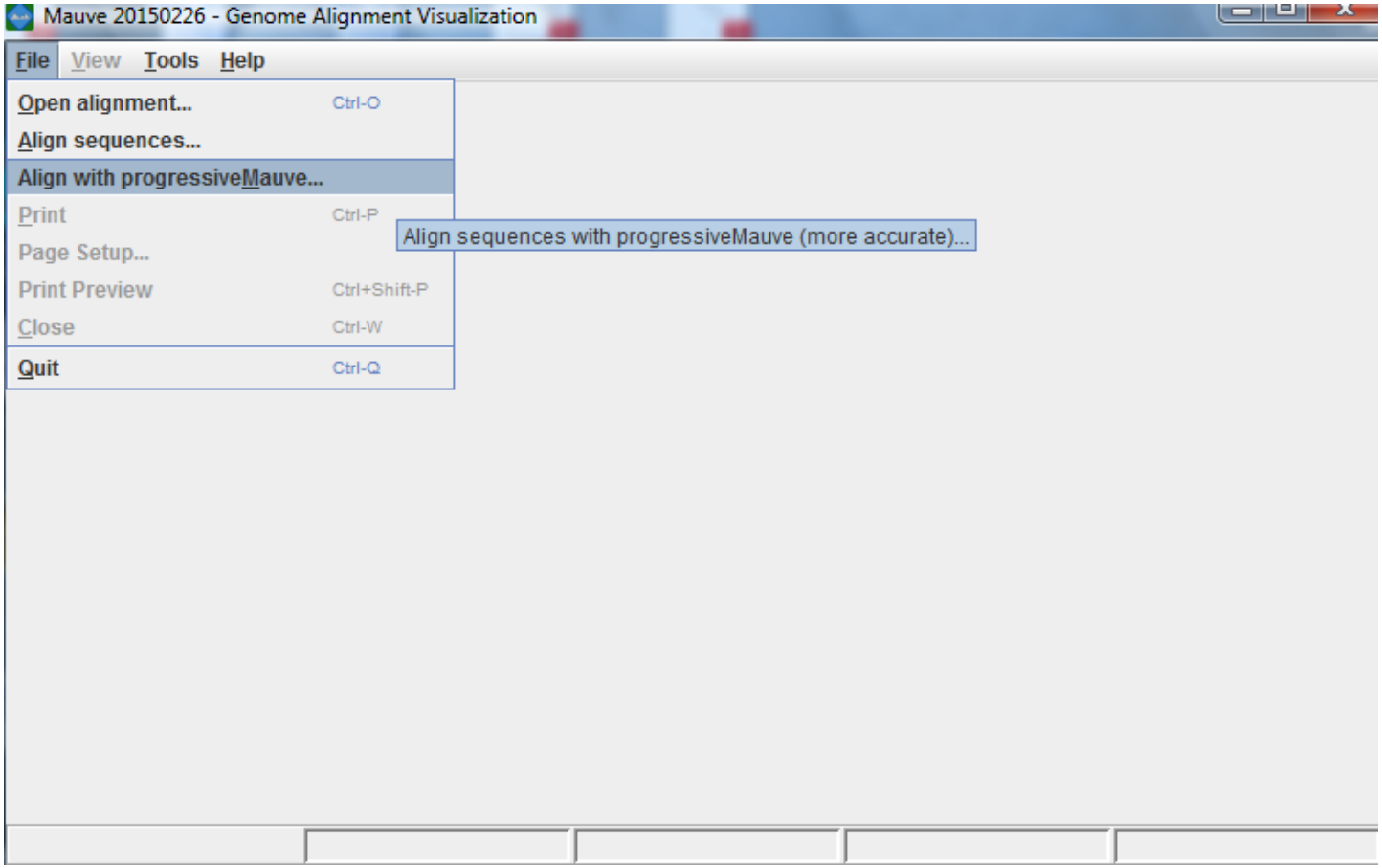
koadman
koadman

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

MAUVE

Alinhamento das sequências com a função progressiveMauve

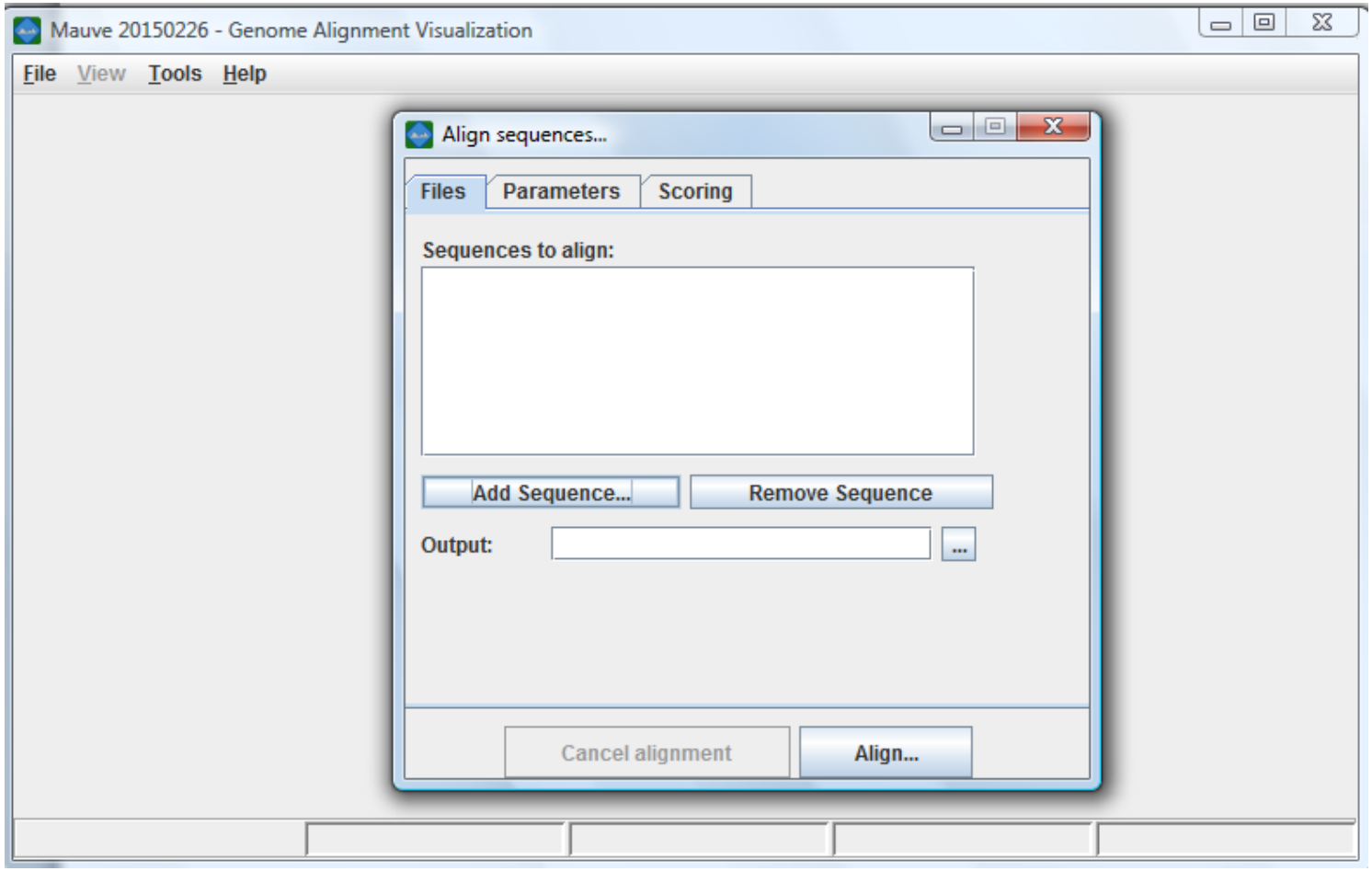
- Abrir o programa
- Clicar em File e escolher Align with progressiveMauve



MAUVE

Alinhamento das sequências com a função progressiveMauve

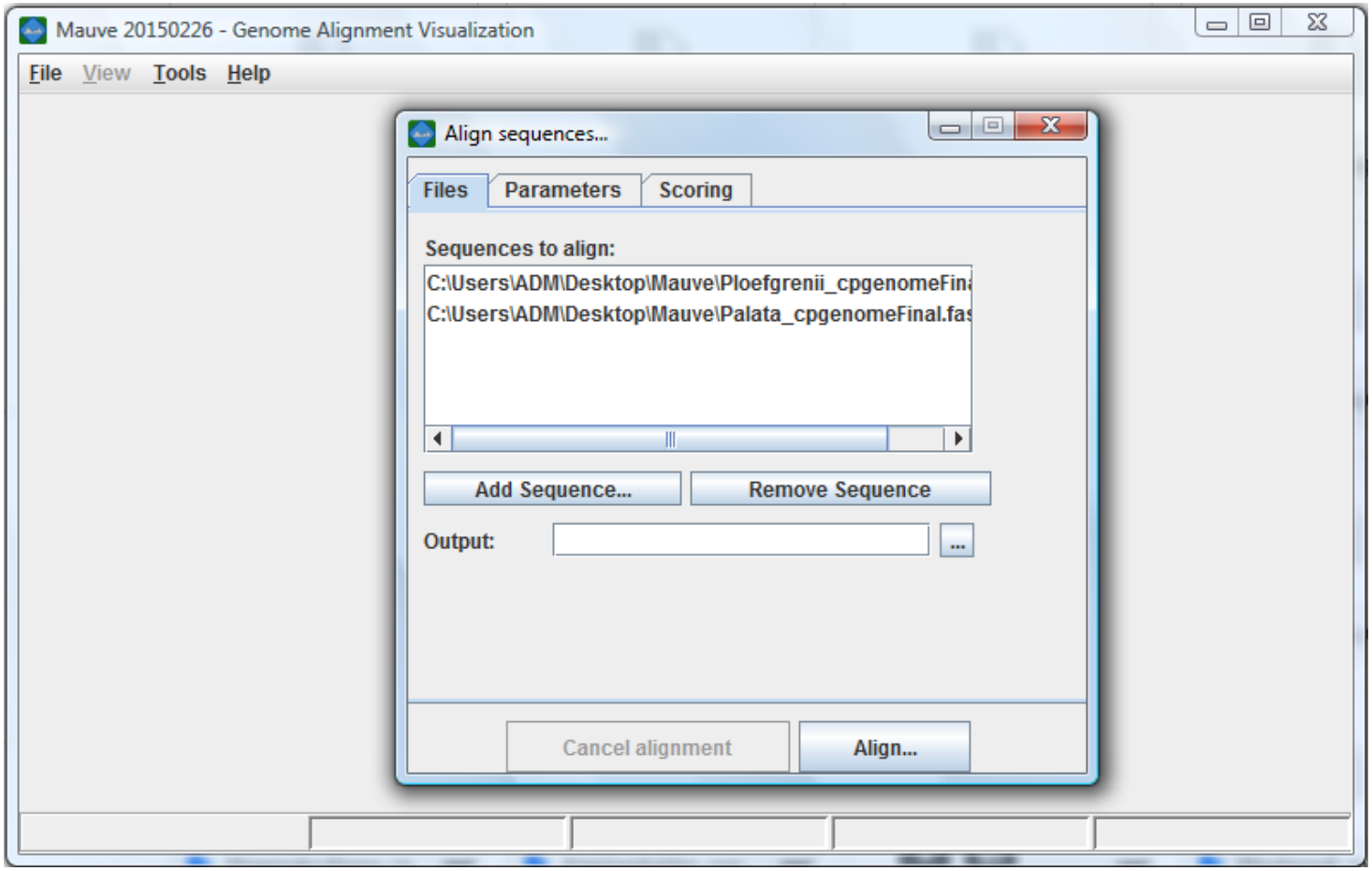
- Adicionar as sequências .fasta em Add Sequences



MAUVE

Alinhamento das sequências com a função progressiveMauve

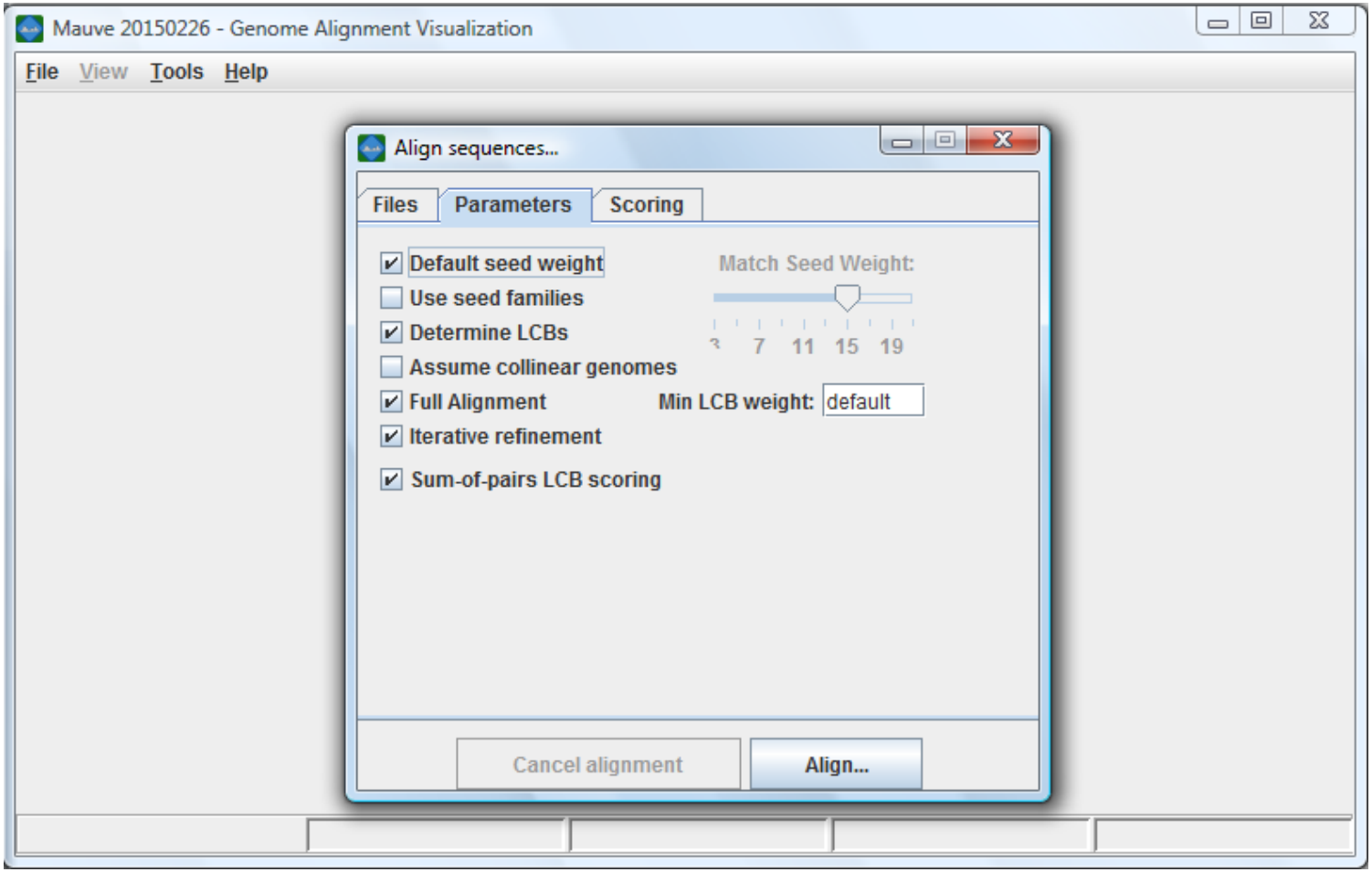
- Adicionar as sequências .fasta em Add Sequences



MAUVE

Alinhamento das sequências com a função progressiveMauve

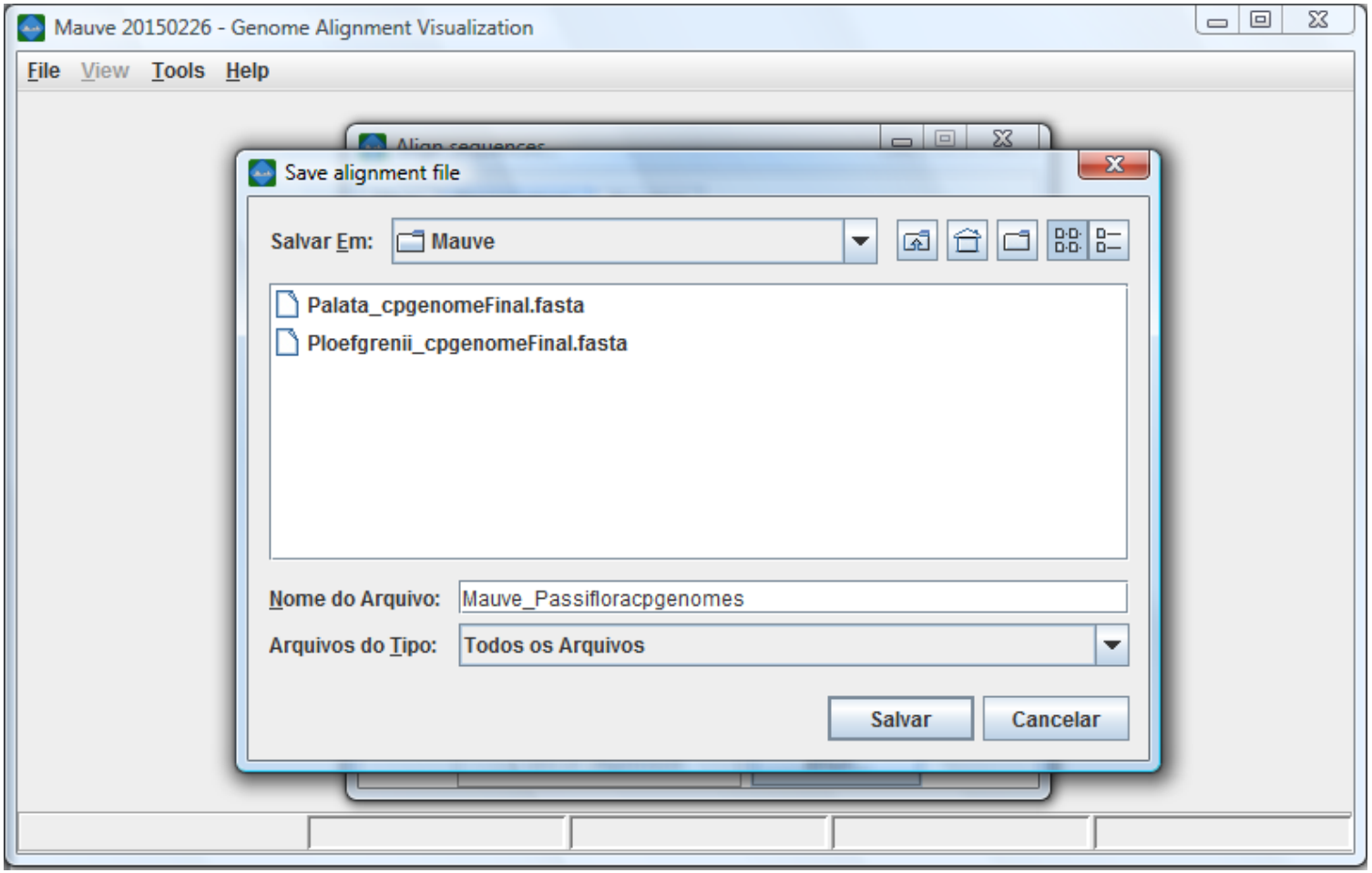
- Definir os parâmetros de alinhamento em Parameters



MAUVE

Alinhamento das sequências com a função progressiveMauve

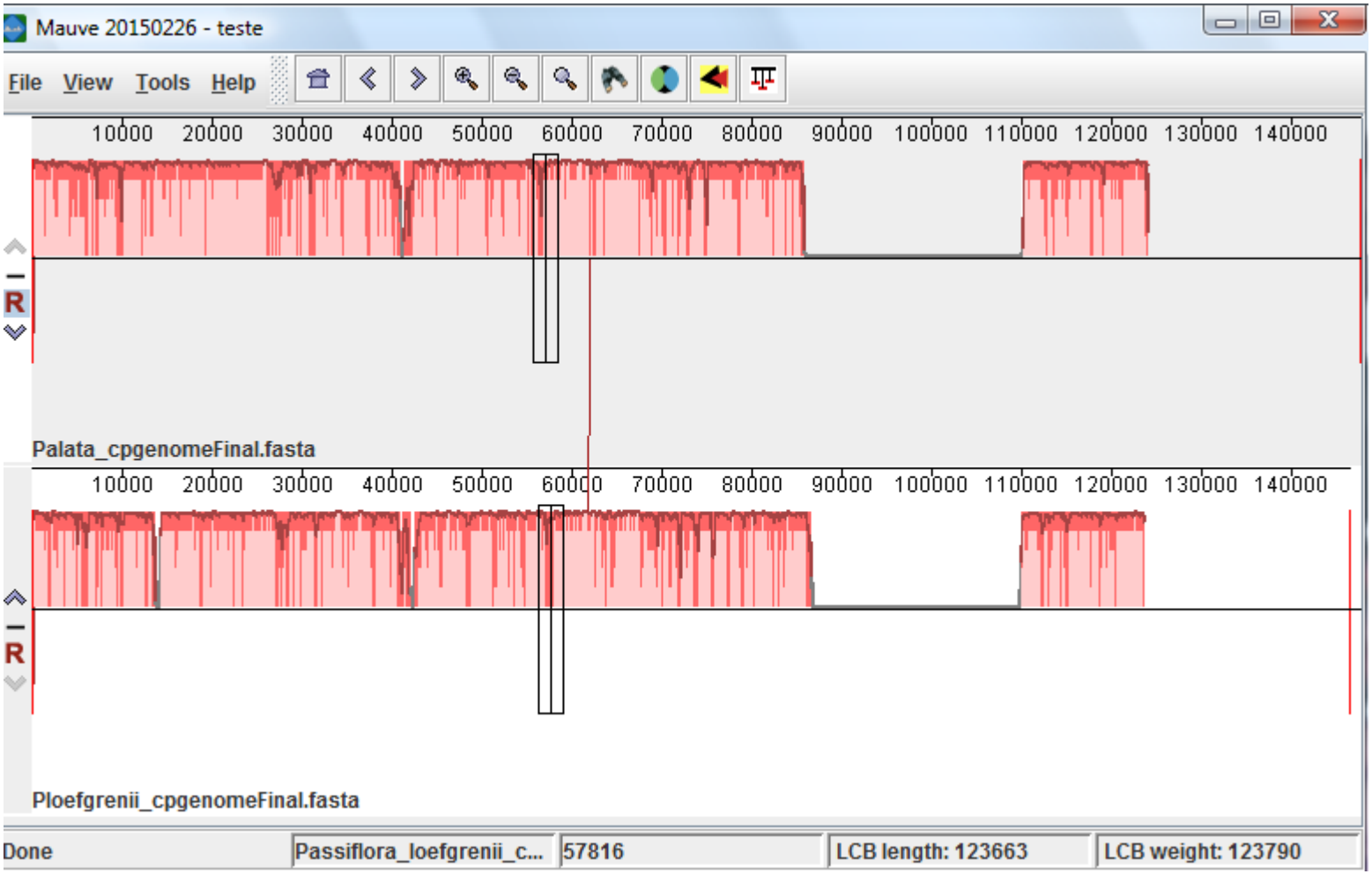
- Escolher uma pasta para salvar os arquivos de saída do programa.



MAUVE

Resultados

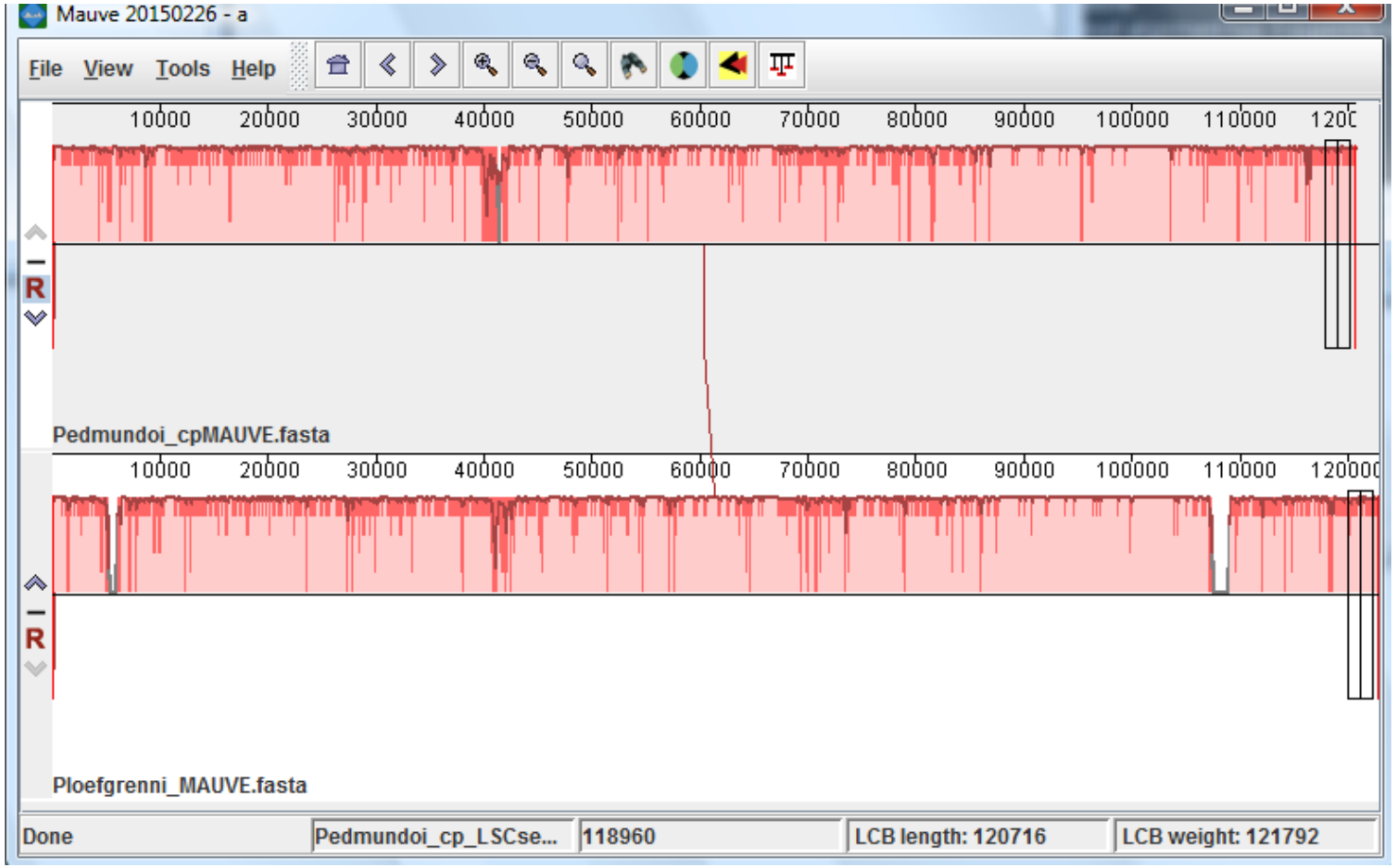
- Sintenia entre as sequências analisadas.



MAUVE

Resultados

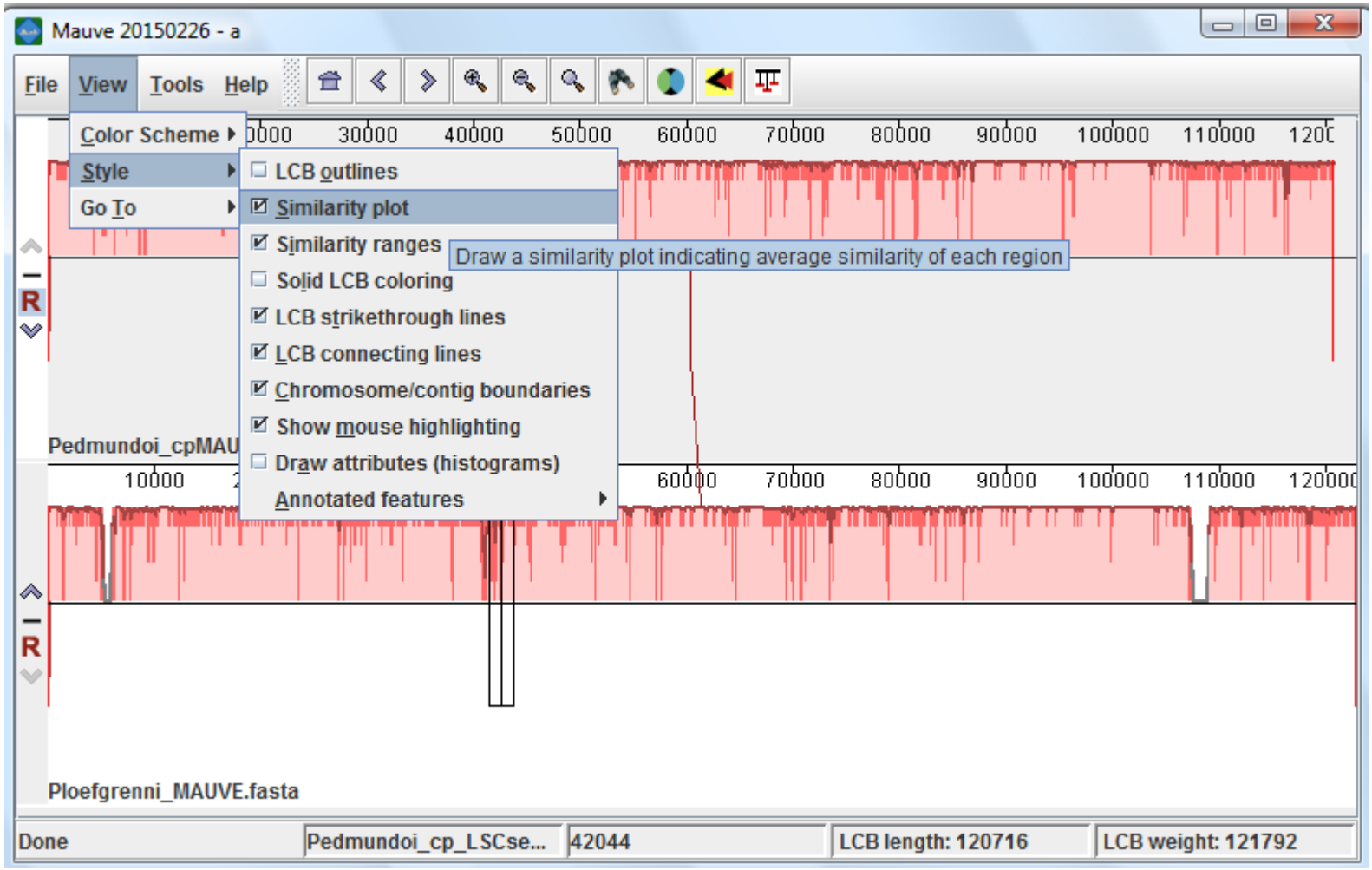
- Utilizando apenas uma das IRs.



MAUVE

Resultados

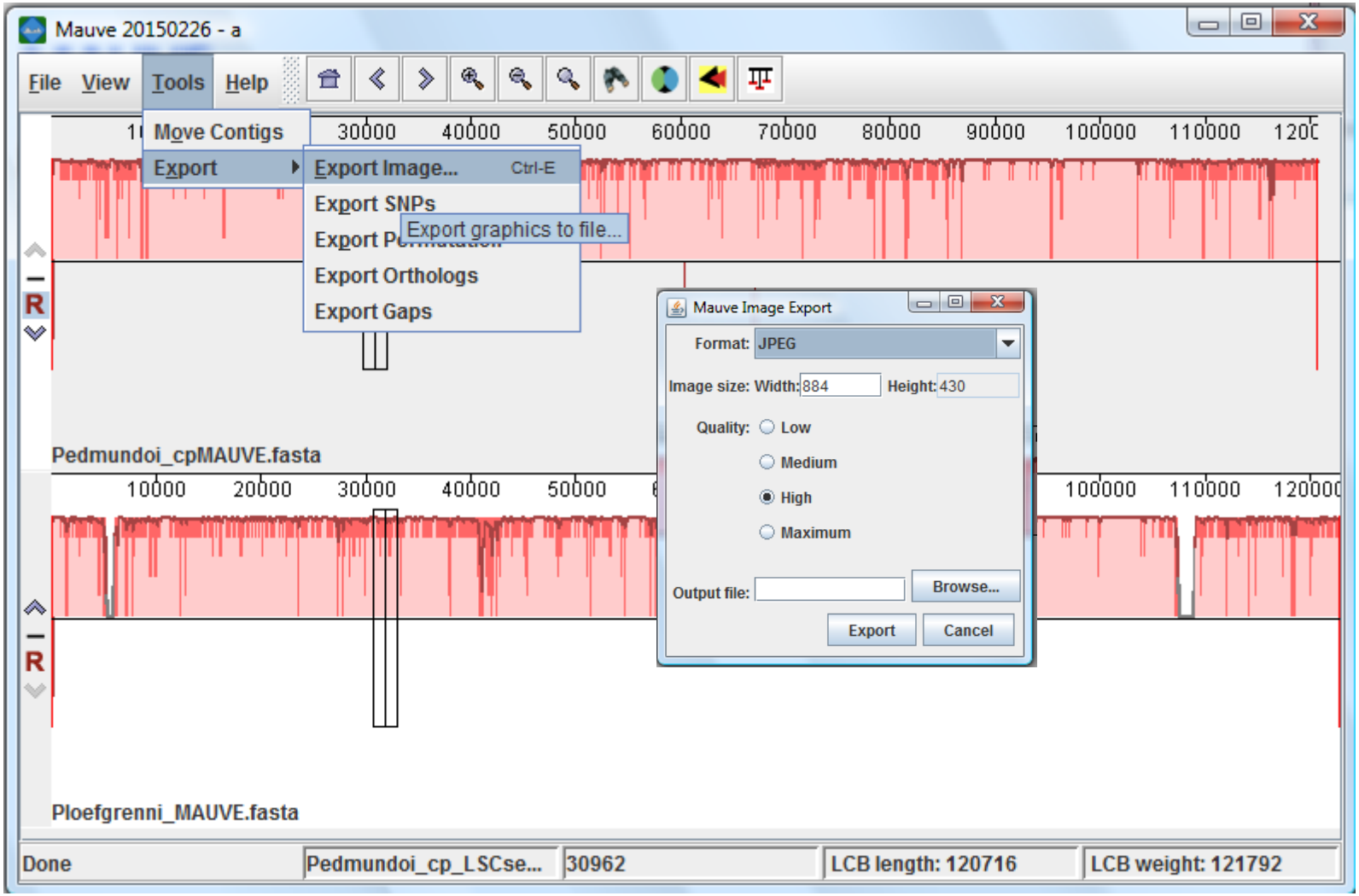
- Fazer alterações na imagem do alinhamento clicando em View e depois em Style.



MAUVE

Resultados

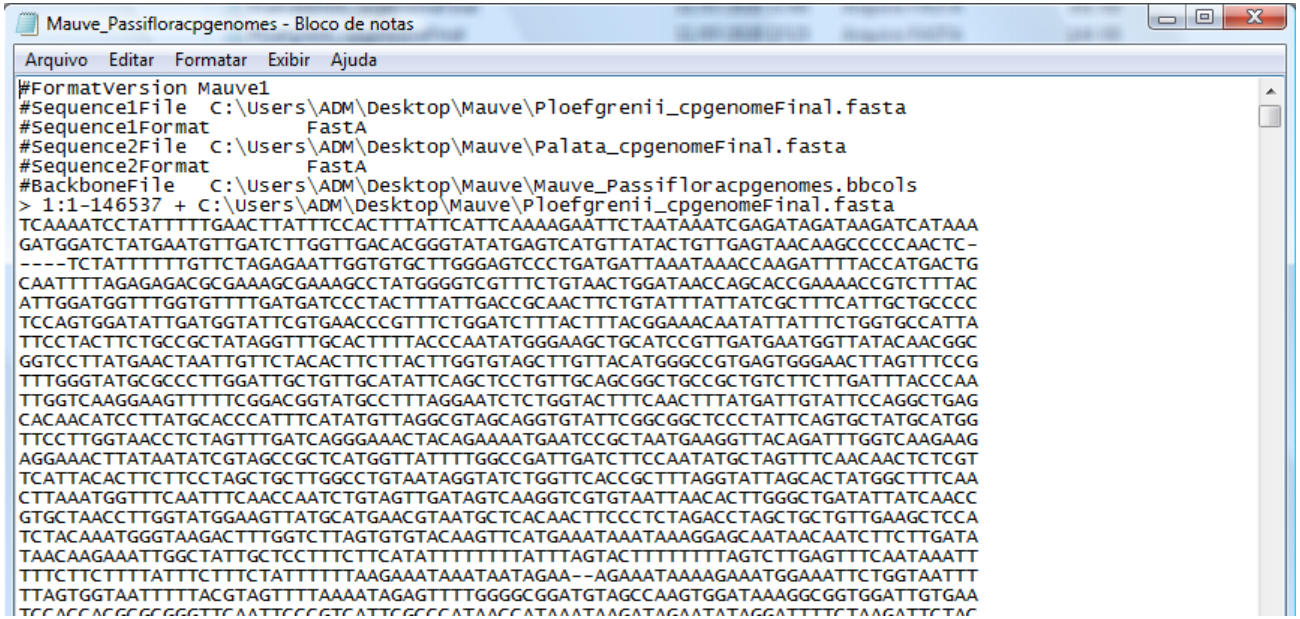
- Para salvar os arquivos de saída clicar em Tools e depois em Export.



MAUVE

Resultados

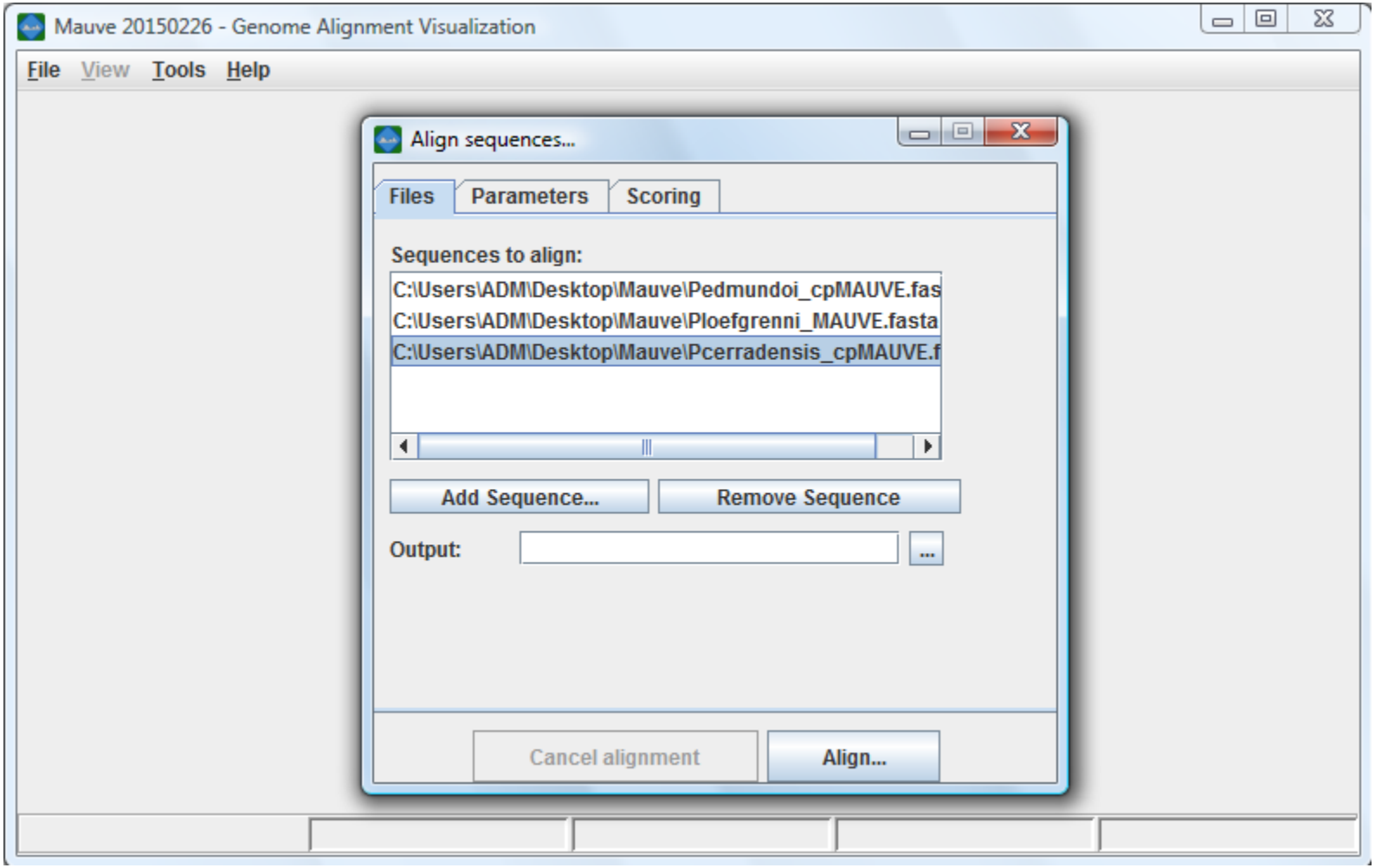
Nome	Modificado em	Tipo	Tamanho
Mauve_Passifloracpgenomes	11/07/2020 17:24	Arquivo	307 KB
Mauve_Passifloracpgenomes.backbone	11/07/2020 17:24	Arquivo BACKBONE	3 KB
Mauve_Passifloracpgenomes.bbcols	11/07/2020 17:24	Arquivo BBCOLS	2 KB
Mauve_Passifloracpgenomes.guide_tree	11/07/2020 17:24	Arquivo GUIDE_TR...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



MAUVE

Alinhamento das sequências com a função progressiveMauve

- Adicionar as sequências .fasta em Add Sequences



MAUVE

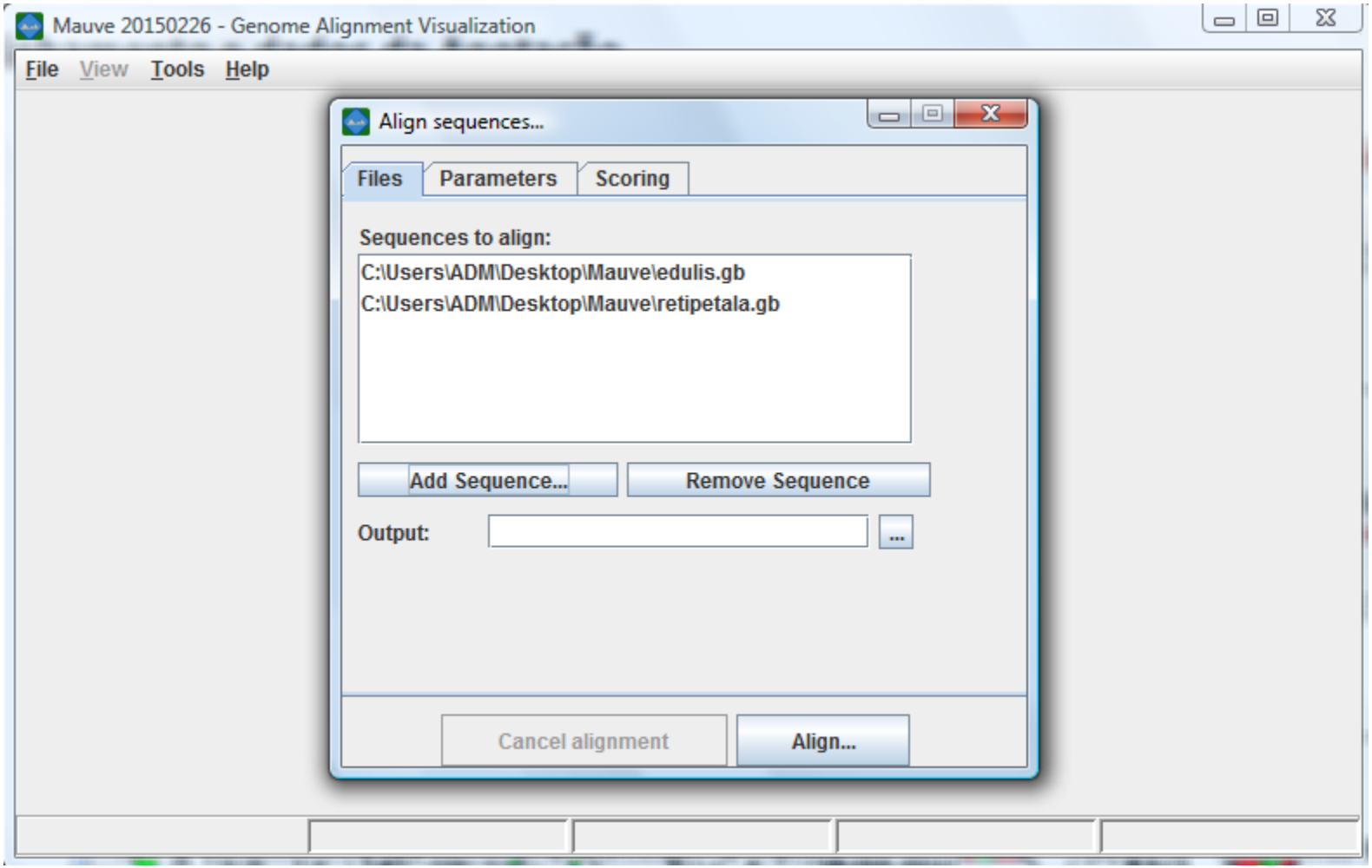
Resultados

- Rearranjos identificados entre os genomas cloroplastidiais.

MAUVE

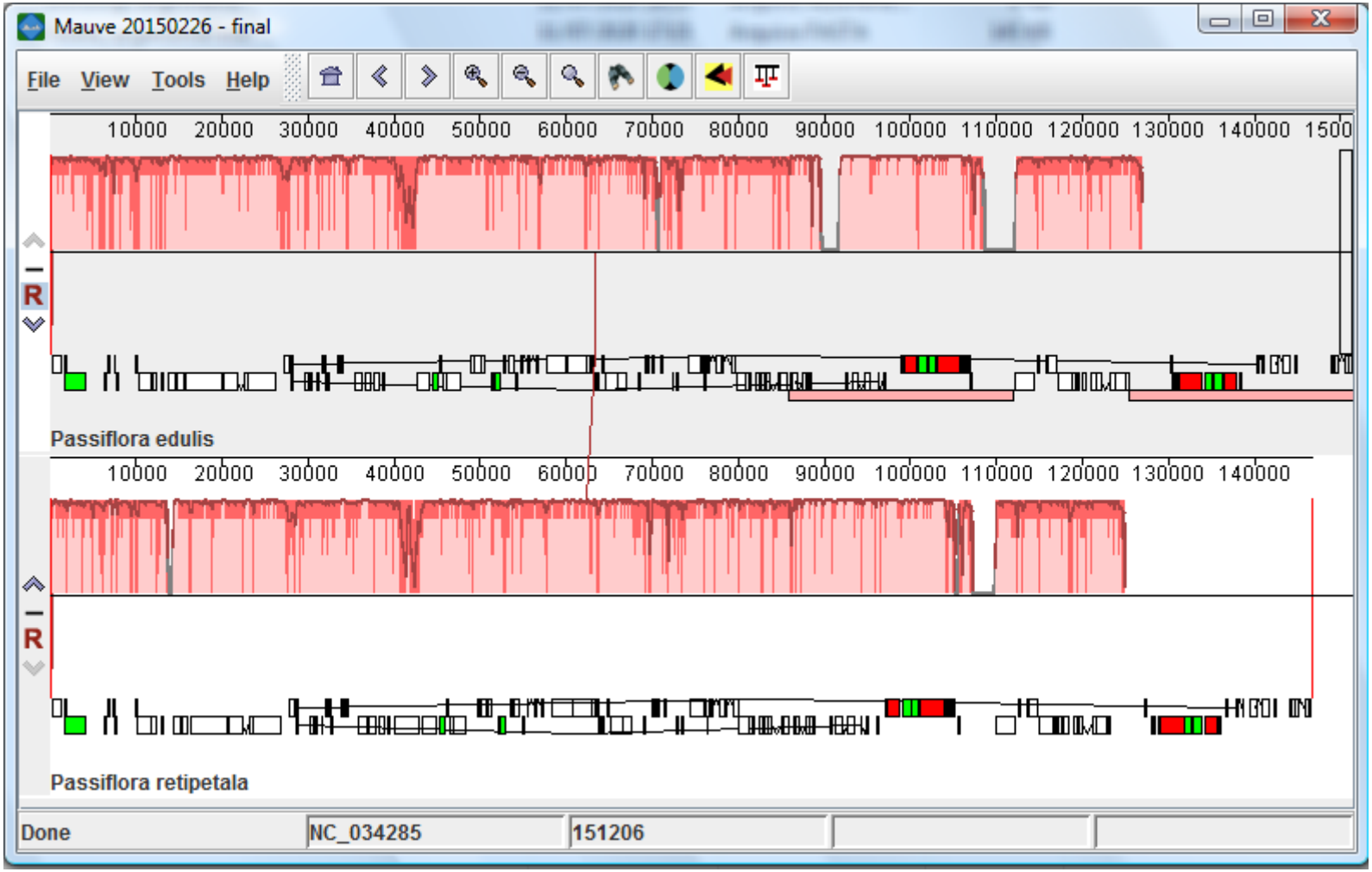
Alinhamento e dados da Anotação

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



MAUVE

Alinhamento e dados da Anotação



OBRIGADO!

luizcauz@usp.br