Plastid genome assembly

Tomáš Fér

Dept. of Botany, Charles University, Prague

June 2023

Basic approaches

- mapping to reference genes/introns/spacers (HybPhyloMaker)
  - a script available to prepare HPM-compatible reference from GenBank files
- mapping to full-genome reference (separate script using HPM output)
- *de novo* assembly/annotation – many pipelines
  - FastPlast
  - GetOrganelle
  - NOVOplasty
  - ORG.asm
  - …

Freudenthal et al. 2020: *A systematic comparison of chloroplast genome assembly tools.* Genome Biology 21: 254.

Reference mapping

- prepare a reference from GenBank file using ‘extractGBplastome.sh’ (https://raw.githubusercontent.com/tomas-fer/scripts/master/extractGBplastome.sh)

```
gene            complement(101946..102413)
                /gene="rps7"
                /locus_tag="LK123_pgp023"
                /db_xref="GeneID:68666775"
CDS             complement(101946..102413)
                /gene="rps7"
                /locus_tag="LK123_pgp023"
                /codon_start=1
                /transl_table=11
                /product="ribosomal protein S7"
                /protein_id="YP_010219699.1"
                /db_xref="GeneID:68666775"
                /translation="MSRRGTAEEKTAKSDPIYRNRLVNMLVNRILKHGKKSLAYQIIY
                RAMKKIQQKTETNPLSVLRQAIRGVTPDIAVKARRVSGSTHQVPIEIGSTQGKALAIR
                WLLGASRKRPGRNMAFKLSSELVDAAKGSGDAIRKKEETHRMAEANRAFAHFR"
gene            105205..105276
                /gene="trnV-GAC"
                /locus_tag="LK123_pgt025"
                /db_xref="GeneID:68666776"
tRNA            105205..105276
                /gene="trnV-GAC"
                /locus_tag="LK123_pgt025"
                /product="tRNA-Val"
                /db_xref="GeneID:68666776"
gene            105504..106994
                /gene="rrn16"
                /locus_tag="LK123_pgr008"
                /db_xref="GeneID:68666777"
rRNA            105504..106994
                /gene="rrn16"
                /locus_tag="LK123_pgr008"
                /product="16S ribosomal RNA"
                /db_xref="GeneID:68666777"
```

```
>109_109_trnMxCAU_tRNA
acctacttaactcagcggttagagtattgctttcatacggc
gggagtcattggttcaaatccaatagtaggta
>110_110_trnMxCAU-atpE_non
gaacttattagataccgcagtcaatggtatctaataagttt
ttatacacatttgattttagtaatatttttttttgtatcttt
>111_111_atpE_CDS
ttatgaaatggcattgatagcctctactcgtgtcctagccc
gtcggagagctagatttgcctcaattgtttgtctttttcct
tcagcttttctcaaagcagcttccgctatttcaagagtttg
>230_230_4.5S_rRNA
gaaggtcacggcgagacgagccgtttatcattacgataggt
gtcaagtggaagtgcagtgatgtatgcagctgaggcatcct
aacagaccggtagacttgaac
```

- use the reference within HybPhyloMaker

De novo assembly

- extract chloroplast derived reads from whole genome or enriched dataset
- produce the assembly, ideally fully circular
- identify LSC, SSC and IRs
- annotate the assembly
  - DOGMA (https://dogma.ccbb.utexas.edu/)
  - GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html)
  - Plastid Genome Annotator (PGA) (https://github.com/quxiaojian/PGA)

Fast-Plast

- https://github.com/mrmckain/Fast-Plast
- read trimming (Trimmomatic)
- read extraction – mapping to reference (Bowtie 2)
- de novo assembly using de Bruijn graphs (SPAdes)
- iterative seed-based assembly to close gaps of contigs with low coverage (afin)
- check for gene content (blast+)
- identification of quadripartite structure and proper ordering of LSC, SSC and IRs
- coverage analysis (jellyfish)

Fast-Plast

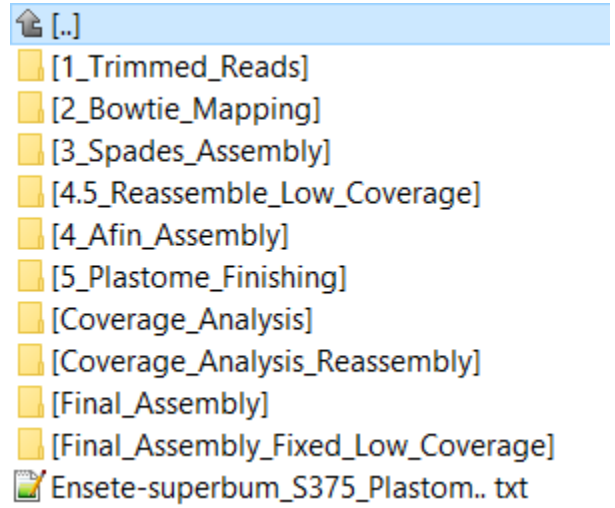
- input files – gzipped FASTQ files (PE or SE)
- specify plant order for plastome reads retrieval (FastPlast includes more than 1,000 plastomes to create bowtie2 index)
- `--name` – prefix to all files

```
perl fast-plast.pl \
-1 genus-species_code_R1.fastq.gz \
-2 genus-species_code_R2.fastq.gz \
--name genus-species_code \
--bowtie_index Zingiberales \
--coverage_analysis
```

Fast-Plast results

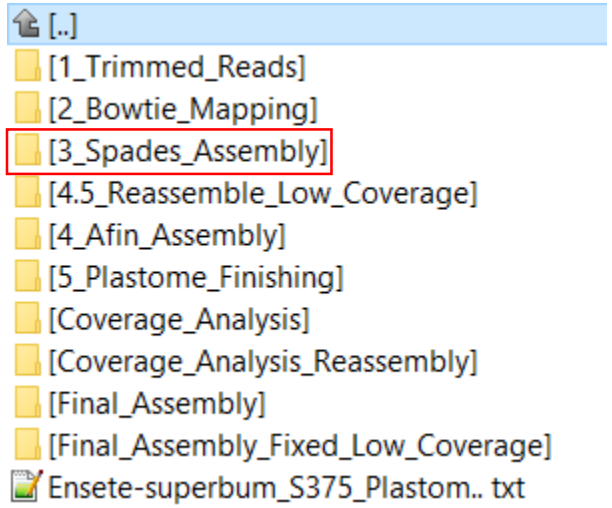
- input files – gzipped FASTQ files (PE or SE)
- specify plant order for plastome reads retrieval (FastPlast includes more than 1,000 plastomes to create bowtie2 index)
- `--name` – prefix to all files

```
perl fast-plast.pl \
-1 genus-species_code_R1.fastq.gz \
-2 genus-species_code_R2.fastq.gz \
--name genus-species_code \
--bowtie_index Zingiberales \
--coverage_analysis
```



Fast-Plast assembly

- SPAdes (https://cab.spbu.ru/software/spades/)
- de Bruijn graph assembler



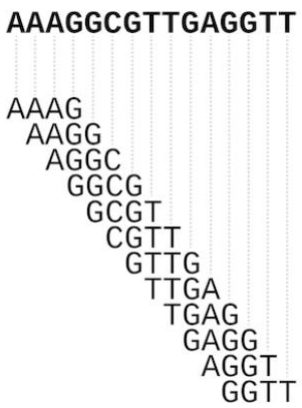
de Bruijn graphs

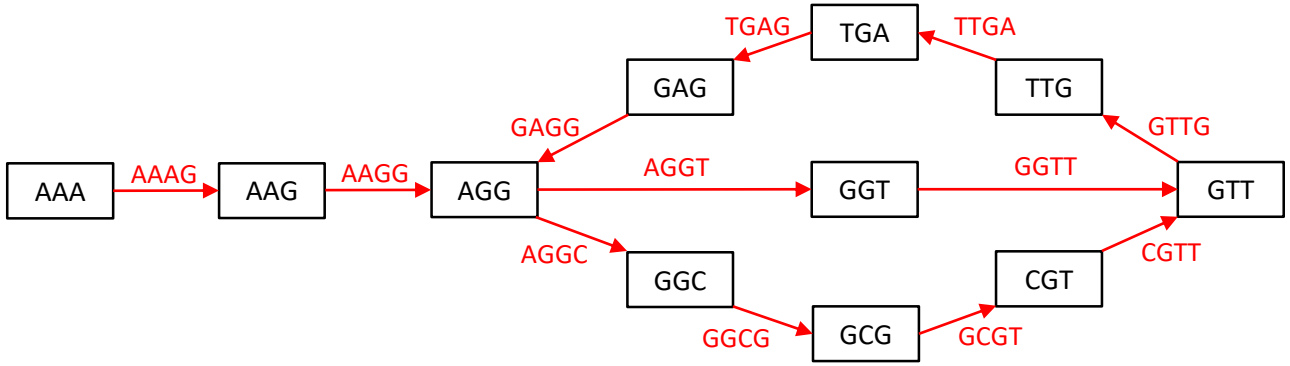
- network made up of nodes and edges (directed multigraph)
- these comes from the overlaps between k-mers

k-mers – a (DNA) molecule of the length *k*

- every possible (k-1)-mer is assigned to a node
- edges are all possible k-mers
- connect nodes by a directed edge if there is a k-mer whose
- prefix (i.e., all position except the last one) is the former node
- suffix (i.e., all position except the first one) is the latter node
- Eulerian cycle in the graph (Eulerian walk) – visits each edge exactly ones

k-mer = 4





single or multiple Eulerian walk possible? Why?

de Bruijn graphs

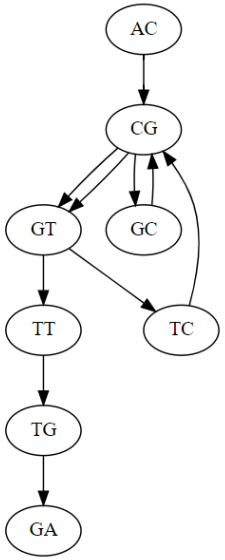
- requirements for straightforward graph
  - all *k*-mers present in the genome sequenced (gaps in sequencing lead to fragmented graphs)
  - all *k*-mers are error-free (error correction possible)
  - each *k*-mer appears at most once in the genome (different coverage requires normalization)
  - genome consists of a single circular chromosome

- play with *k*-mers and graphs using this Jupyter Notebook by B. Langmead https://colab.research.google.com/drive/1pQu9tJZ9RNpk8AaL2ThEYXol3lu7Rw34
- experiment with different *k*-mer settings

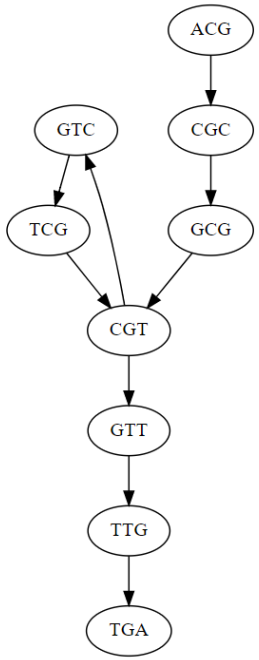
de Bruijn graphs

ACGCGTCGTTGA

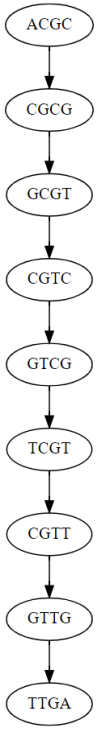
k=3



k=4



k=5



All graphs with Eulerian path

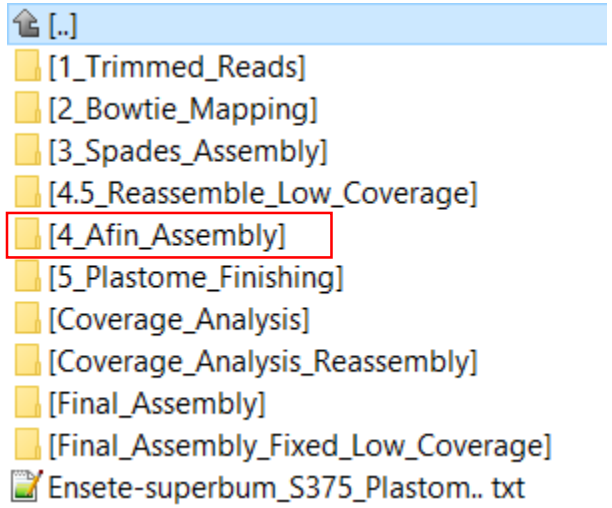
- all nodes (except first and last) are balanced (i.e., # incoming edges = # outcoming edges)
- starting and ending nodes are semibalanced

Fast-Plast assembly

- SPAdes (https://cab.spbu.ru/software/spades/)
- de Bruijn graph assembler

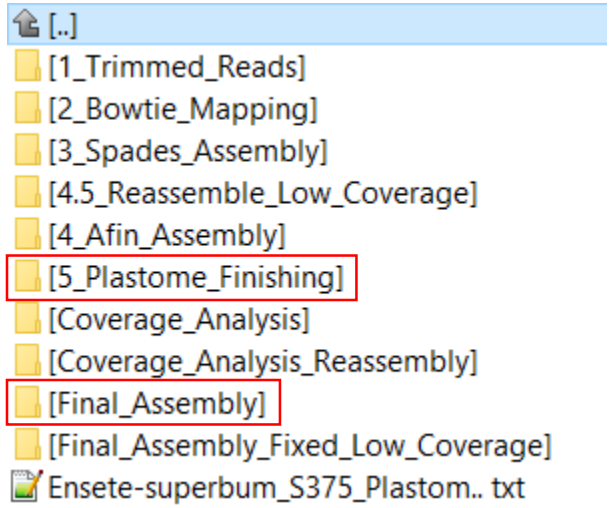
- afin assembly on SPAdes contigs
  - iterative seed-based assembly
  - for closing gaps

- scaffolding with SSPACE (https://github.com/nsoranzo/sspace_basic)
  - if more than one contig found after afin
  - contig extension/scaffolding using PE reads



Fast-Plast finishing

- identification of genes in the assembly (blast)
  - gene composition of the assembly
- identification/orientation of LSC, SSC and IRs
  - full sequences
  - sequences split into 4 pieces



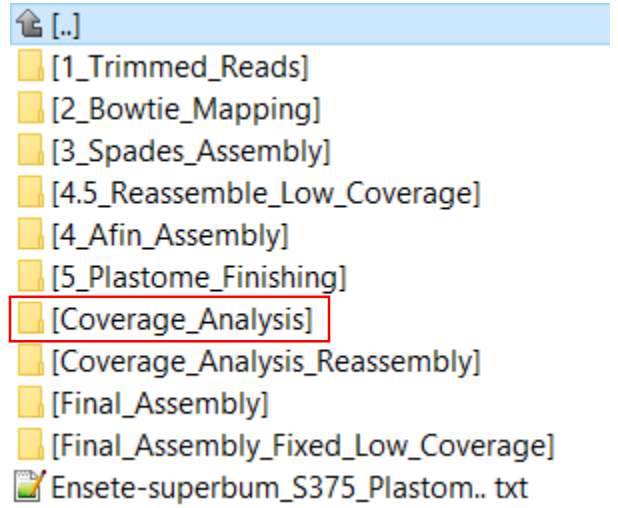
Fast-Plast coverage analysis

- k-mer coverage (Jellyfish)
- sudden coverage ‘doubling’ identifies IR boundary

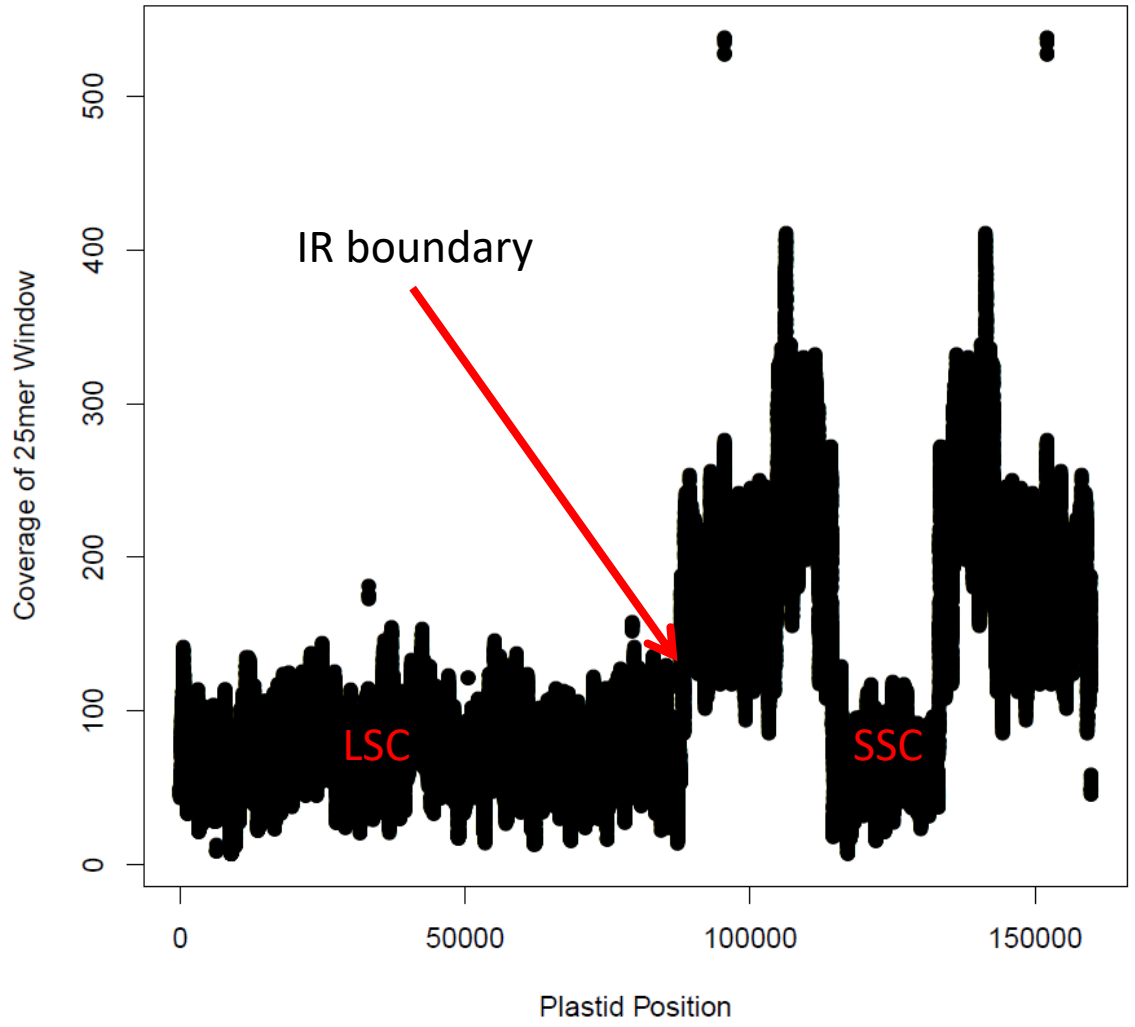
79096	ACTTTACTCCTTTTTTTTTTACATT	79095	110
79097	CTTTACTCCTTTTTTTTTTACATTT	79096	115
79098	TTTACTCCTTTTTTTTTTACATTTT	79097	114
79099	TTACTCCTTTTTTTTTTACATTTTT	79098	114
79100	TACTCCTTTTTTTTTTACATTTTTT	79099	112
79101	ACTCCTTTTTTTTTTACATTTTTTA	79100	114
79102	CTCCTTTTTTTTTTACATTTTTTAT	79101	114
79103	TCCTTTTTTTTTTACATTTTTTATT	79102	112
79104	CCTTTTTTTTTTACATTTTTTATTT	79103	112
79105	CTTTTTTTTTTACATTTTTTATTTT	79104	118
79106	TTTTTTTTTTACATTTTTTATTTTC	79105	196
79107	TTTTTTTTTACATTTTTTATTTTCA	79106	199
79108	TTTTTTTTACATTTTTTATTTTCAA	79107	201
79109	TTTTTTTACATTTTTTATTTTCAAT	79108	207
79110	TTTTTTACATTTTTTATTTTCAATT	79109	208
79111	TTTTTACATTTTTTATTTTCAATTT	79110	205
79112	TTTTACATTTTTTATTTTCAATTTA	79111	205
79113	TTTACATTTTTTATTTTCAATTTAA	79112	199
79114	TTACATTTTTTATTTTCAATTTAAA	79113	198
79115	TACATTTTTTATTTTCAATTTAAAG	79114	204
79116	ACATTTTTTATTTTCAATTTAAAGA	79115	209
79117	CATTTTTTATTTTCAATTTAAAGAT	79116	212
79118	ATTTTTTATTTTCAATTTAAAGATT	79117	209
79119	TTTTTTATTTTCAATTTAAAGATTG	79118	207

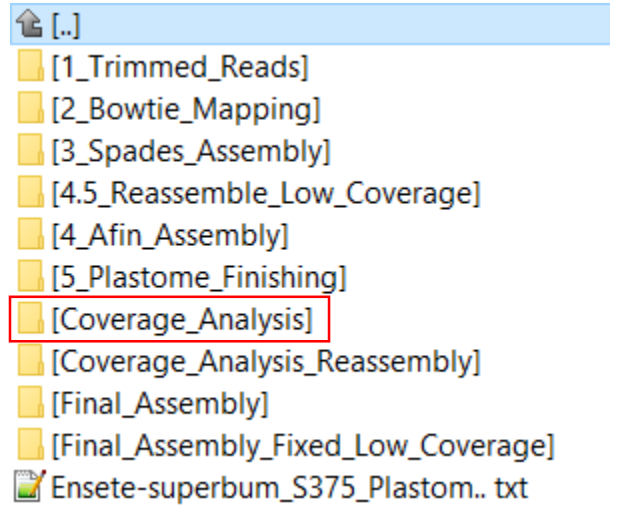
end of LSC

beginning of IR



Fast-Plast coverage analysis

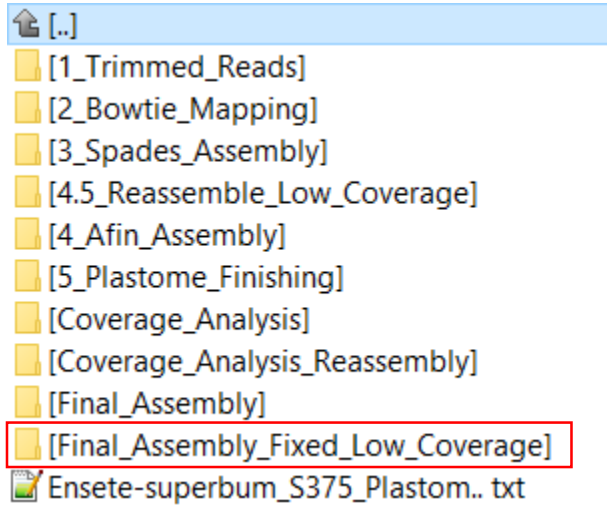




Fast-Plast reassembly

- if regions with low coverage were identified

- low coverage regions removed
- contig broken into pieces
- reassembly from afin step
- coverage analysis of reassembly



Fast-Plast final results

Sample: Ensete-superbum_S375

Fast-Plast Version: Fast-Plast v.1.2.8

Total Cleaned Pair-End Reads: 1095096

Total Cleaned Single End Reads: 11531

Total Concordantly Mapped Reads: 199582

Total Non-concordantly Mapped Reads: 1273

Total Chloroplast Genome Length: 159320

Large Single Copy Size: 79105

Inverted Repeat Size: 34607

Small Single Copy Size: 11001

Minimum Coverage Used for Verification: 31.9663802410244

VALUES BELOW FROM REASSEMBLED PLASTOME

Total Chloroplast Genome Length: 162818

Large Single Copy Size: 74743

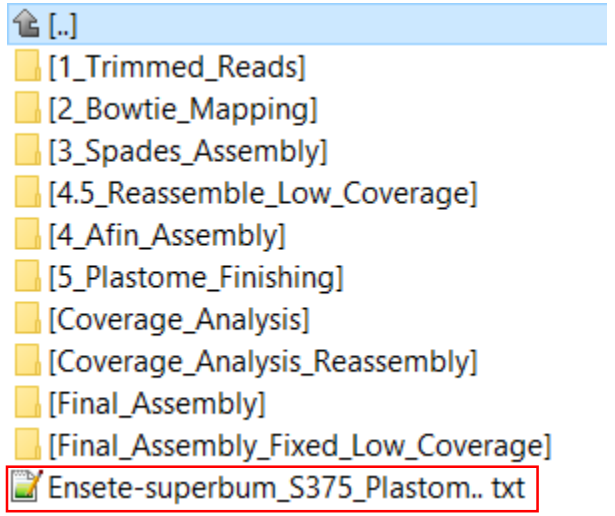
Inverted Repeat Size: 38537

Small Single Copy Size: 11001

Average Large Single Copy Coverage: 125

Average Inverted Repeat Coverage: 330

Average Small Single Copy Coverage: 109



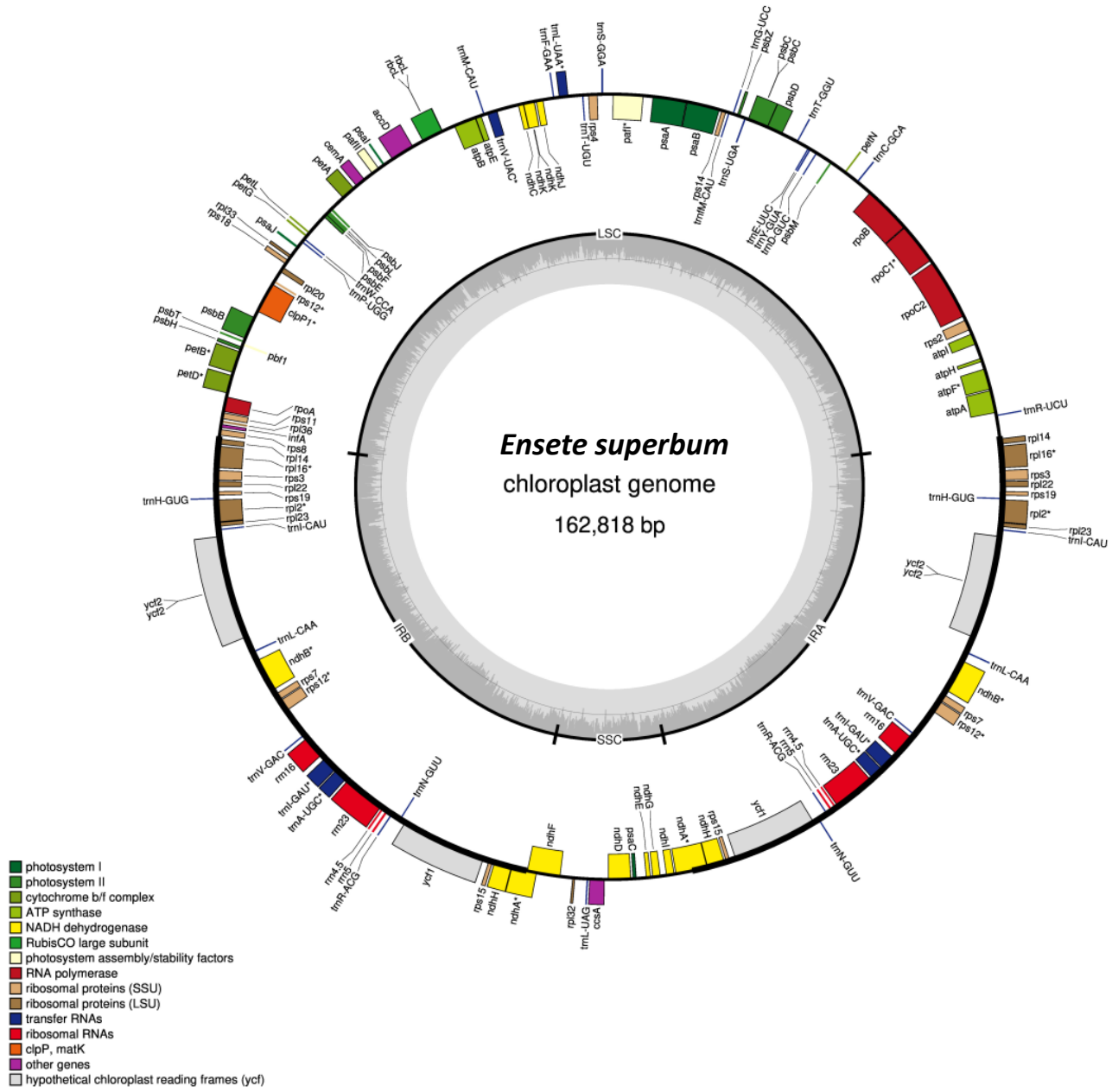
Plastome annotation

- MPI-MP CHLOROBOX (https://chlorobox.mpimp-golm.mpg.de/index.html)
- GeSeq
  - upload FASTA file to annotate, reference possible
  - CDS, tRNA, rRNA
  - diverse tRNA annotators (ARAGORN, ARWEN, tRNAscan-SE)
  - annotation support
    - Chloë (https://chloe.plastid.org/)
    - Mfannot (https://megasun.bch.umontreal.ca/RNAweasel/)
- results
  - output from primary annotation tools
  - annotation – GenBank, GFF3, GBSON
  - visualization - OGDRAW

Plastome annotation

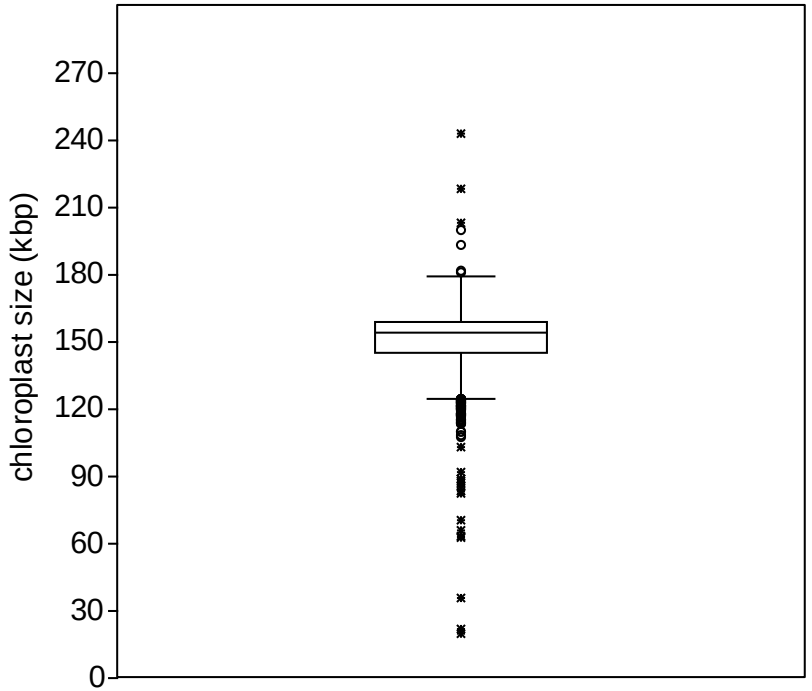
- OGDRAW - Draw Organelle Genome Maps
- https://chlorobox.mpimp-golm.mpg.de/OGDraw.html

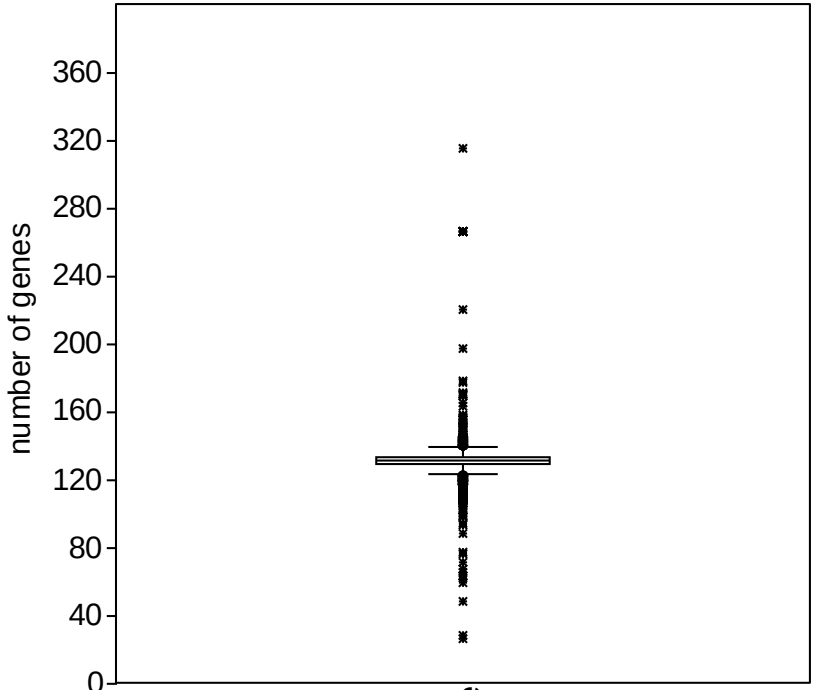
- from GenBank file



Chloroplast size, number of genes

- cca 1,700 sequenced plastomes (land plants)
- size 150 kbp (19 – 243)
- 131 (26 – 315) genes: 84 proteins, 8 rRNA, 37 tRNA





https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/