# How to find the mitochondrial genome in fungal assemblies.

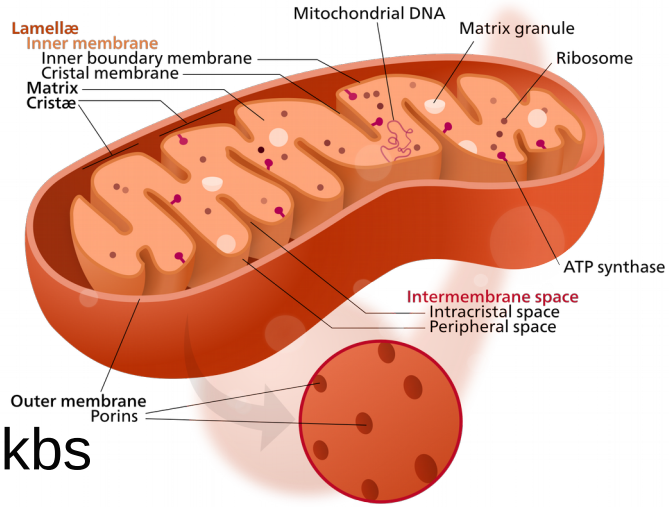
13.10.2017

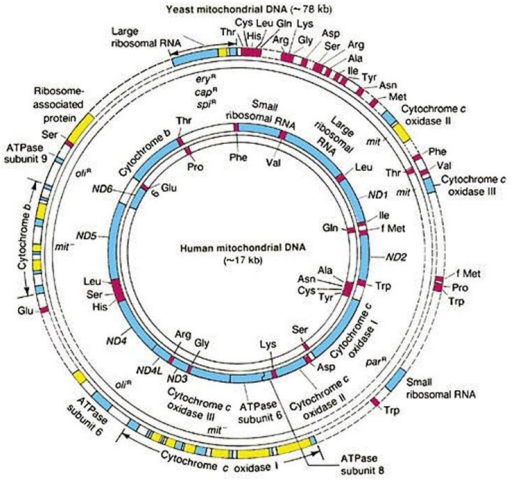
Marie Lataretu

Bioinformatics/High Throughput Analysis, Manja Marz

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

# Fungal mtDNA




- ca. 16-110 kbs
- Mostly assumed to be circular

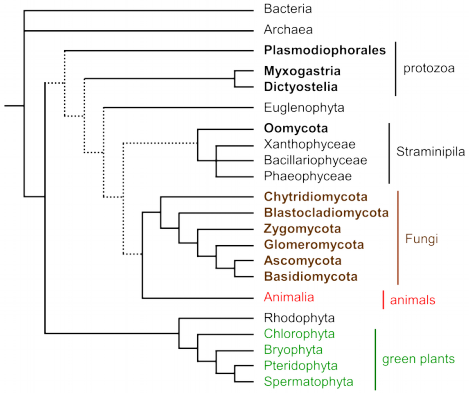

Ntertili M., Kirmitzoglou I., Kouvelis V.N., Promponas V.J., Typas M.A.
MitoFun: A Curated Resource of Complete Fungal Mitochondrial Genomes. Submitted. 2013.

# Fungal mtDNA

- Highly variability of the gene order
- Recombination
- Large intergenic regions
- Highly variable intron numbers

Aguileta G, de Vienne DM, Ross ON, et al.
High Variability of Mitochondrial Gene Order among Fungi. Genome Biology and Evolution. 2014.

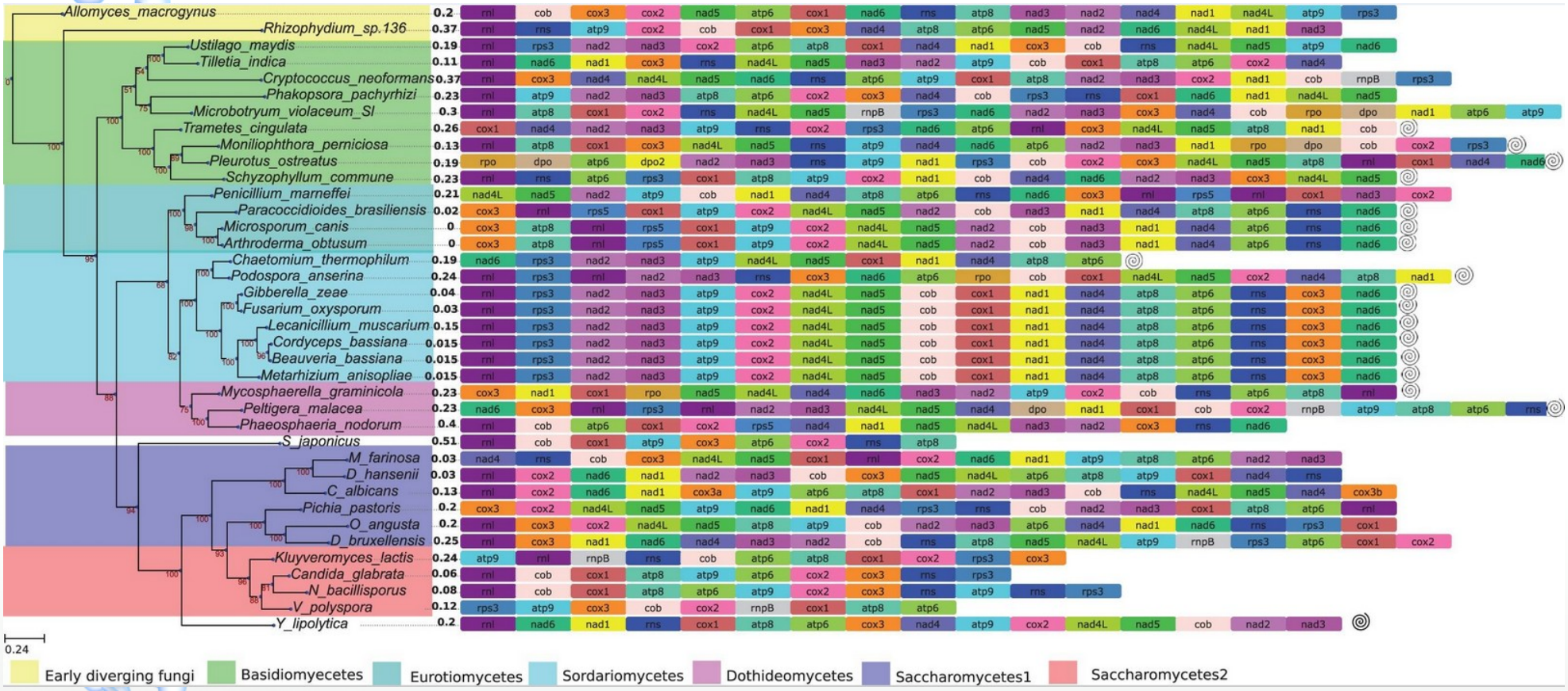
# Why mtDNA?



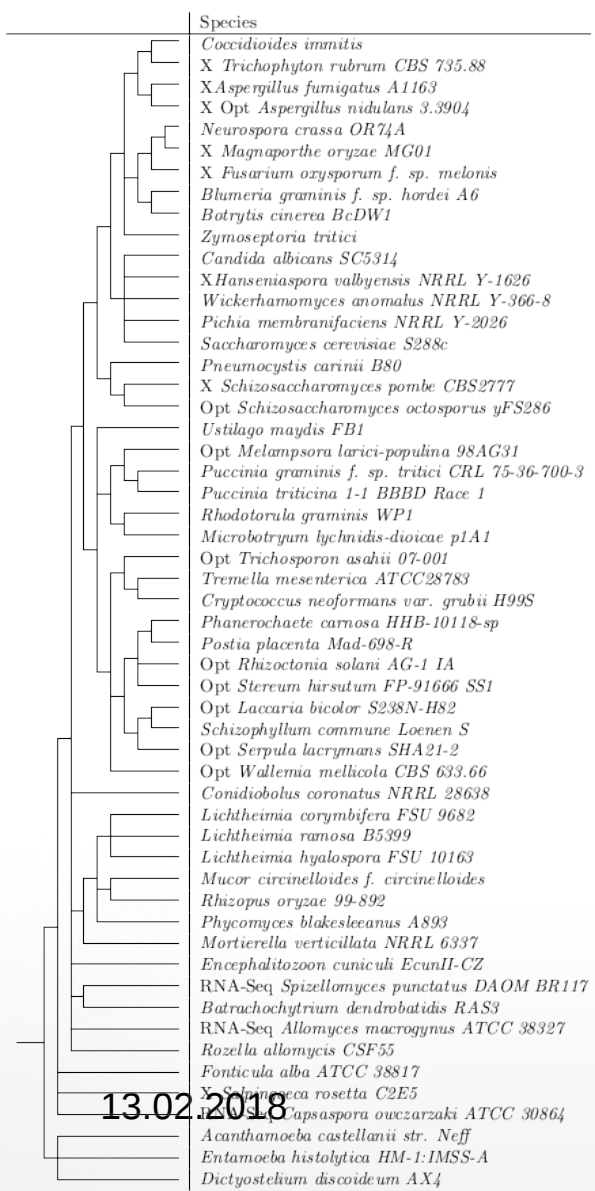


Georg Hausner. 6 - Fungal Mitochondrial Genomes, Plasmids and Introns.
In Dilip K. Arora and George G. Khachatourians, editors, Fungal Genomics,
Volume 3 of Applied Mycology and Biotechnology, pages 101 – 131. Elsevier, 2003.

# mt gene order
# according to GenBank annotation




Aguileta G, de Vienne DM, Ross ON, et al. High Variability of Mitochondrial Gene Order among Fungi. Genome Biology and Evolution. 2014.




# Data

- NGS data of 33 fungi
- 8 mt reference genomes

# Approach

- Which assembler to use?

# Approach

- Which assembler to use?

JR-Assembler MIRA ABySS v2


# Approach

- Which assembler to use?

SOAPdenovo2

Cap3

SSPACE

# Approach

- Which assembler to use?
- How to find “the mt contig”?

# Approach

- Which assembler to use?
- How to find “the mt contig”?
  - `BLAST` against refseq data
    - 19 feature proteins
    - rrnL, rrnS

# Approach

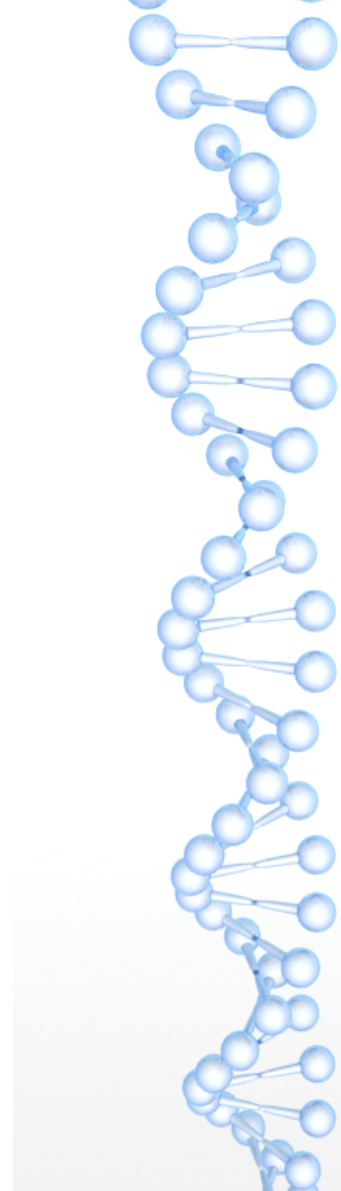
- Which assembler to use?
- How to find “the mt contig”?
  - `BLAST` against refseq data
    - 19 feature proteins
    - rrnL, rrnS
  - Read coverage of the contig

Salmon

# Approach

- Which assembler to use?
- How to find “the mt contig”?
  - `BLAST` against refseq data
    - 19 feature proteins
    - rrnL, rrnS
  - Read coverage of the contig

**HISAT2**



## BLAST result

## HISAT2 result

1)

# BLAST result

# HISAT2 result

1) Discard BLAST hits with

- Same genus
- Identity < 70 %
- E-value > $10^{-10}$

# BLAST result

# HISAT2 result

1) Discard BLAST hits with
   - Same genus
   - Identity < 70 %
   - E-value > $10^{-10}$
2) BLAST features
   - Number of blast hits
   - Unique set of hits on contig
   - $\frac{nt\ covered\ by\ blast\ hits}{contig\ length}$

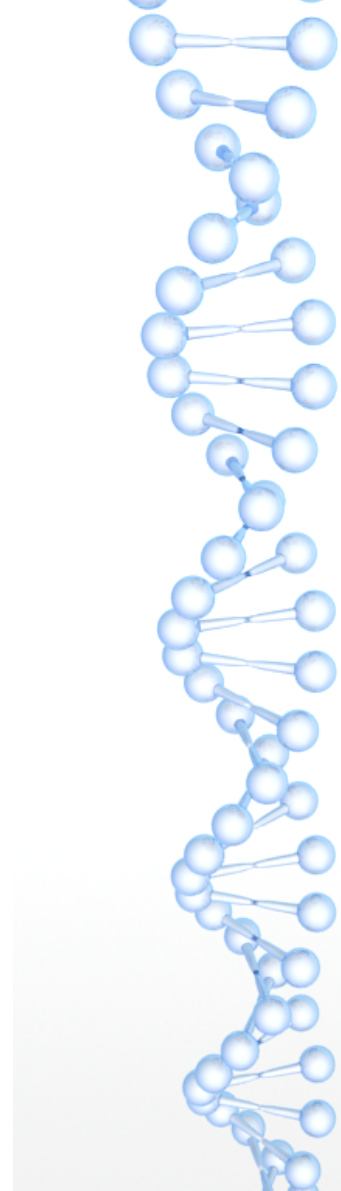
## BLAST result

1) Discard BLAST hits with
   - Same genus
   - Identity < 70 %
   - E-value > $10^{-10}$
2) BLAST features
   - Number of blast hits
   - Unique set of hits on contig
   - $\frac{nt\ covered\ by\ blast\ hits}{contig\ length}$

## HISAT2 result

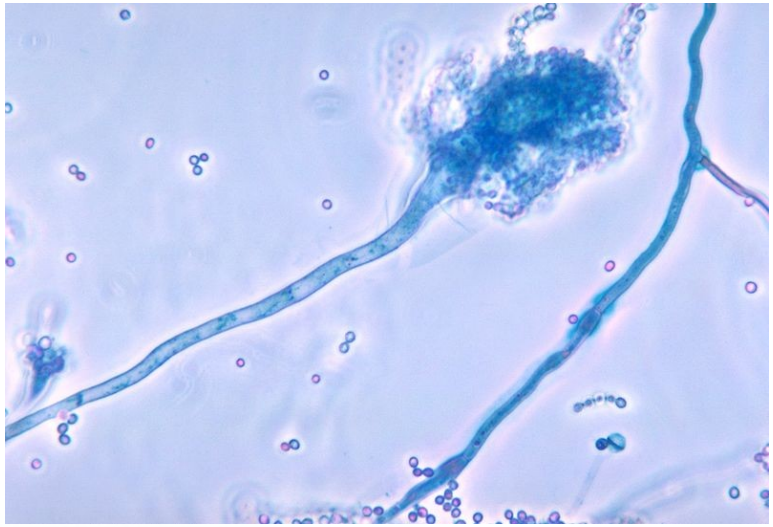
- Normalized mean read coverage of the contig

  → 95$^{th}$ percentile?

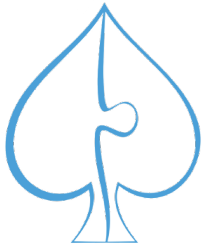


# Example with mt reference: *Aspergillus fumigatus*

# Example with mt reference: *Aspergillus fumigatus*



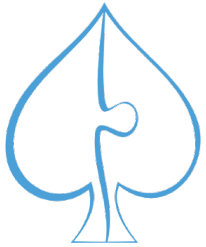
# Example with mt reference: *Aspergillus fumigatus*



Cap3

SSPACE

# Example with mt reference: *Aspergillus fumigatus*

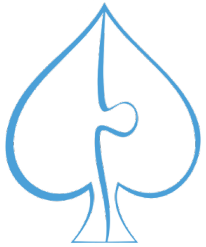


Cap3

SSPACE

1) `BLAST` features
2) Read coverage of contigs

# Example with mt reference: *Aspergillus fumigatus*



Cap3

SSPACE

1) BLAST features
2) Read coverage of contigs
3) Comparison with reference mtDNA

} Pick “the mt contig”

# Contigs filtered by BLAST hits

| assembly_name |
|---|
| spades |
| spades |
| spades_cap |
| spades_cap |
| spades_sspace |
| spades_sspace |
| spades_cap_sspace |
| spades_cap_sspace |

# Contigs filtered by BLAST hits

| assembly_name | #contigs | #contigs_filtered | contig_name | contig_length |
|---|---|---|---|---|
| spades | 1185 | 2 | NODE_59 | 163686 |
| spades | 1185 | 2 | NODE_157 | 30872 |
| spades_cap | 512 | 2 | NODE_59 | 163686 |
| spades_cap | 512 | 2 | NODE_157 | 30872 |
| spades_sspace | 873 | 2 | scaffold54 | 289931 |
| spades_sspace | 873 | 2 | scaffold128 | 30872 |
| spades_cap_sspace | 453 | 2 | scaffold54 | 289931 |
| spades_cap_sspace | 453 | 2 | scaffold137 | 30872 |

# Contigs filtered by BLAST hits

| assembly_name | #contigs | #contigs_filtered | contig_name | contig_length |
|---|---|---|---|---|
| spades | 1185 | 2 | NODE_59 | 163686 |
| spades | 1185 | 2 | NODE_157 | 30872 |
| spades_cap | 512 | 2 | NODE_59 | 163686 |
| spades_cap | 512 | 2 | NODE_157 | 30872 |
| spades_sspace | 873 | 2 | scaffold54 | 289931 |
| spades_sspace | 873 | 2 | scaffold128 | 30872 |
| spades_cap_sspace | 453 | 2 | scaffold54 | 289931 |
| spades_cap_sspace | 453 | 2 | scaffold137 | 30872 |

# BLAST features

| assembly_name | contig_name | contig_length |
|---|---|---|
| spades | NODE_59 | 163686 |
| spades | NODE_157 | 30872 |
| spades_sspace | scaffold54 | 289931 |
| spades_sspace | scaffold128 | 30872 |

# BLAST features

| assembly_name | contig_name | contig_length | #feature_prot | #rrnL | #rrnS | covered_by_blast/ Contig_length | unique_blast_set |
|---|---|---|---|---|---|---|---|
| spades | NODE_59 | 163686 | 25 | 0 | 0 | 0,00071 | yes |
| spades | NODE_157 | 30872 | 770 | 117 | 696 | 0,44024 | yes |
| spades_sspace | scaffold54 | 289931 | 25 | 0 | 0 | 0,0004 | yes |
| spades_sspace | scaffold128 | 30872 | 770 | 117 | 696 | 0,44024 | yes |

# Read coverage of contigs

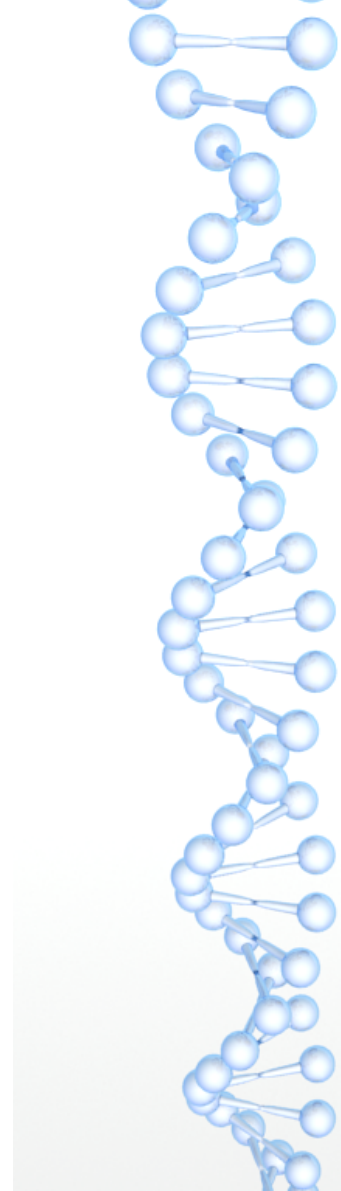
| assembly_name | contig_name | contig_length |
|---|---|---|
| spades | NODE_59 | 163686 |
| spades | NODE_157 | 30872 |
| spades_sspace | scaffold54 | 289931 |
| spades_sspace | scaffold128 | 30872 |

# Read coverage of contigs

| assembly_name | contig_name | contig_length | avg_read_cov | in_95th_percentile_read_cov |
|---|---|---|---|---|
| spades | NODE_59 | 163686 | 117,87152 | no |
| spades | NODE_157 | 30872 | 3400,40393 | yes |
| spades_sspace | scaffold54 | 289931 | 110,97829 | no |
| spades_sspace | scaffold128 | 30872 | 3237,71197 | yes |

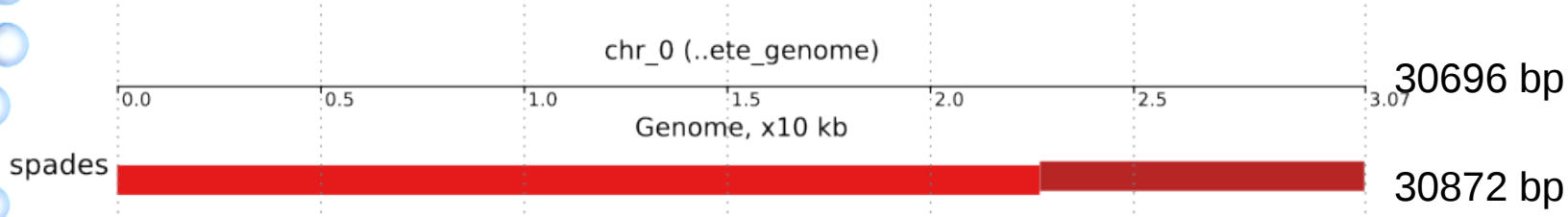
# Read coverage of contigs

| assembly_name | contig_name | contig_length | avg_read_cov | in_95th_percentile_read_cov |
|---|---|---|---|---|
| spades | NODE_59 | 163686 | 117,87152 | no |
| spades | NODE_157 | 30872 | 3400,40393 | yes |
| spades_sspace | scaffold54 | 289931 | 110,97829 | no |
| spades_sspace | scaffold128 | 30872 | 3237,71197 | yes |

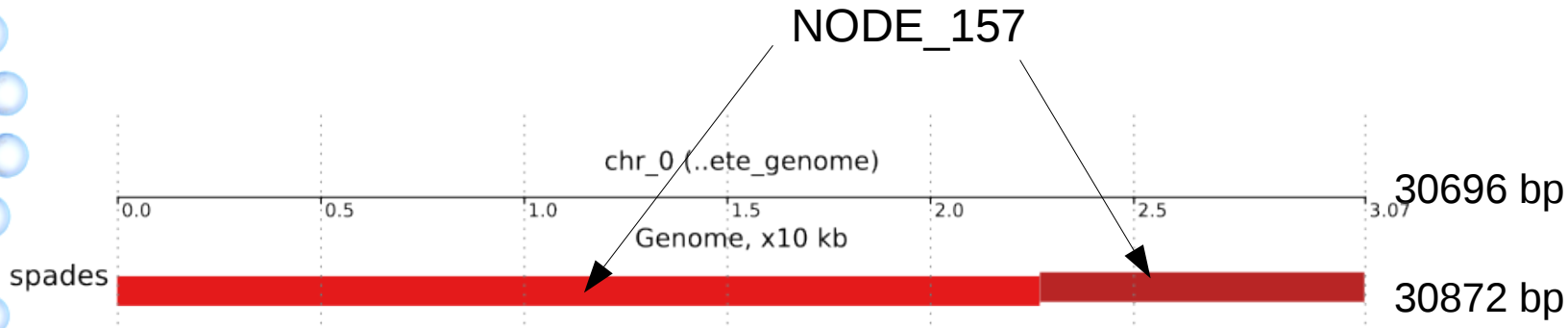


# QUAST and MAUVE alignment

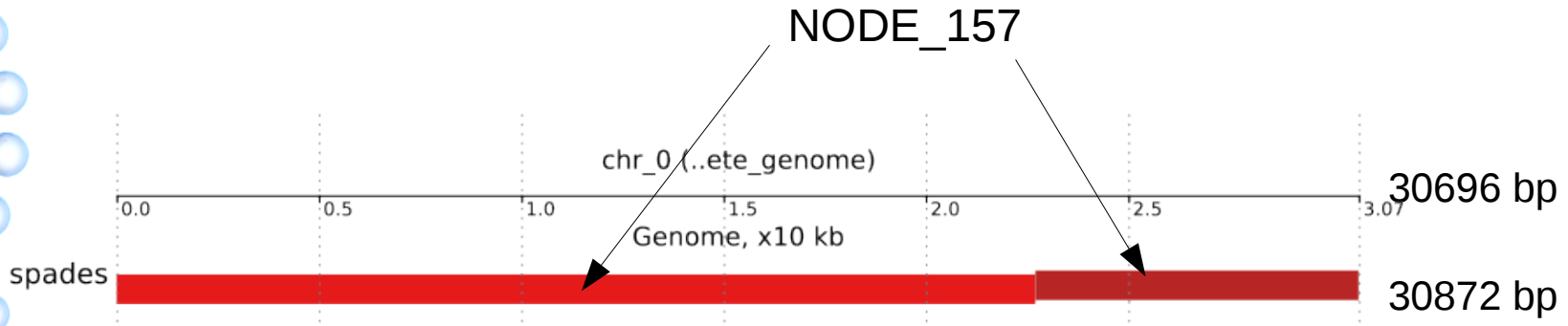
# QUAST and MAUVE alignment

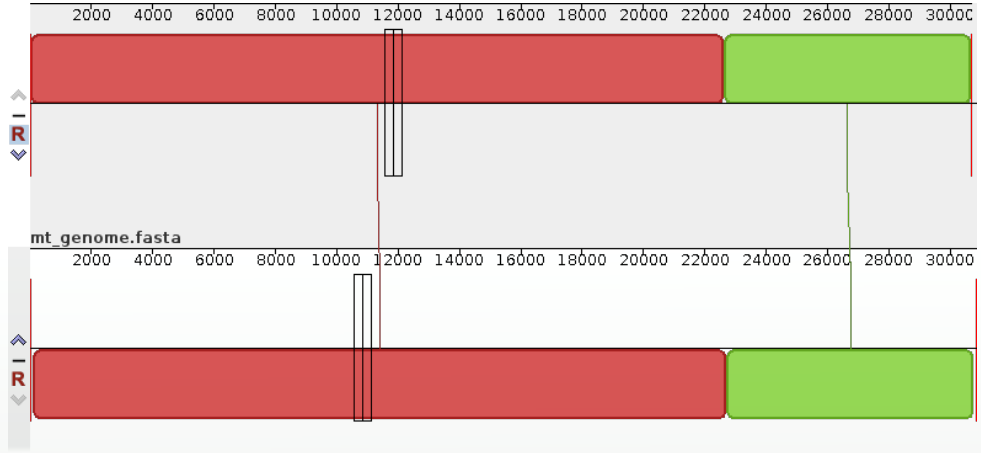


# QUAST and MAUVE alignment



# QUAST and MAUVE alignment



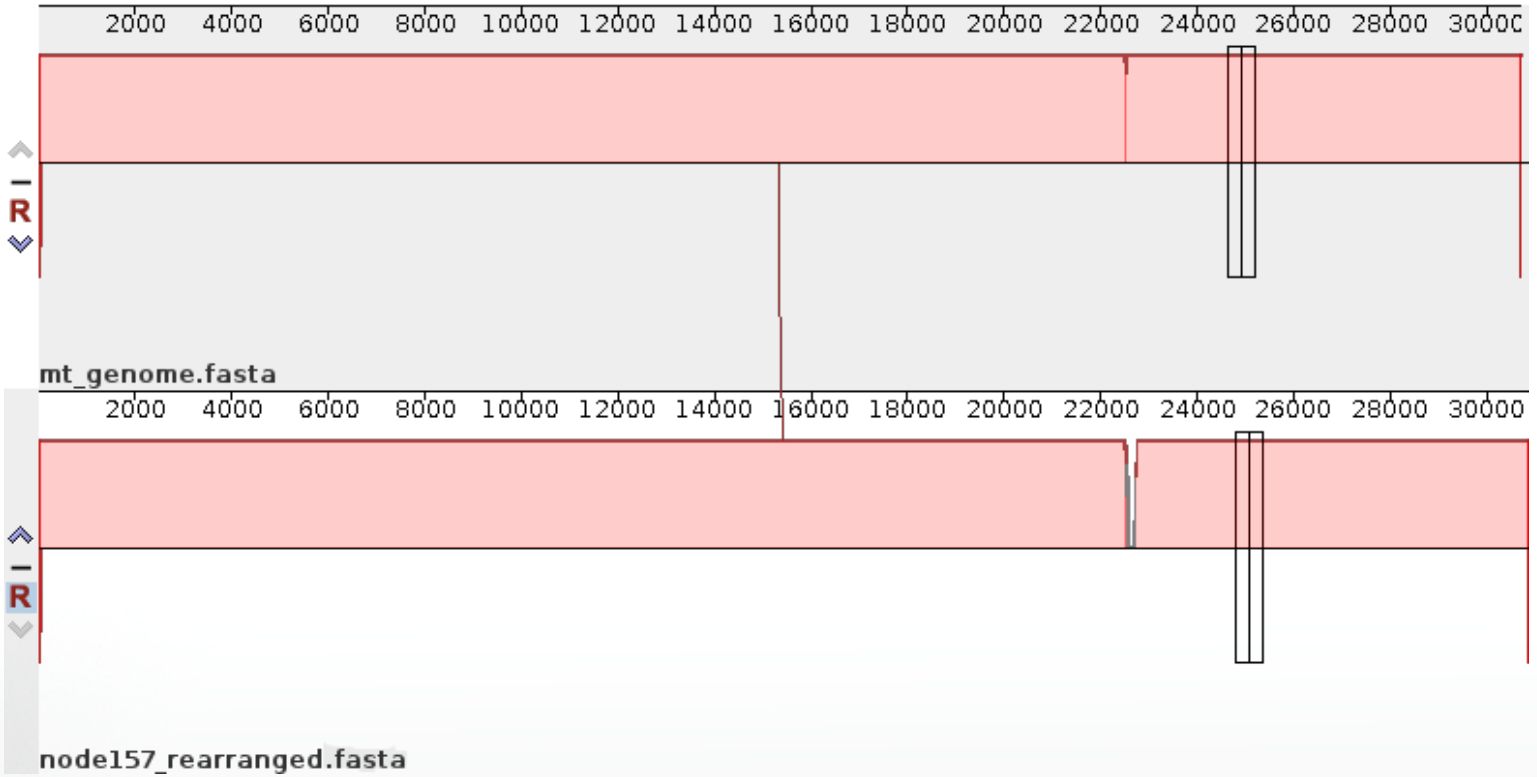



# BLAST

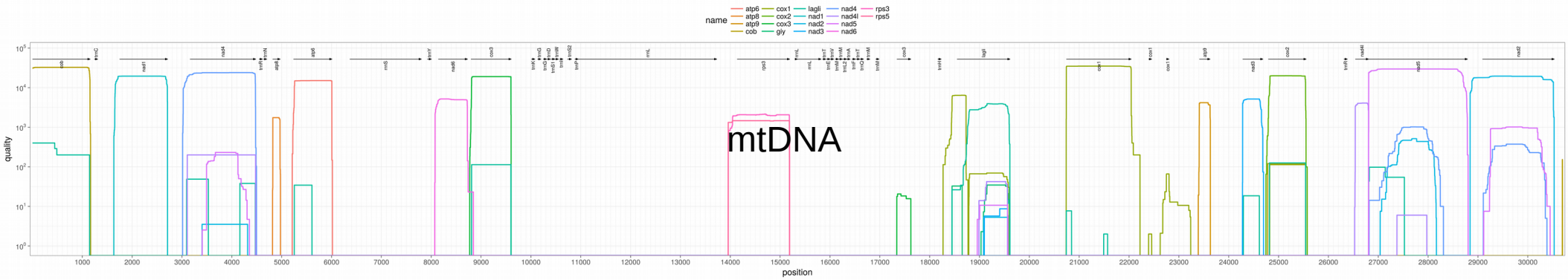
- `SPAdes` vs. reference mtDNA

| qseqid | sseqid | pident | qstart | qend | sstart | send | evalue | bitscore |
|---|---|---|---|---|---|---|---|---|
| NODE_157 | mtRef | 100.00 | 1 | 22704 | 22704 | 1 | 0.0 | 41927 |
| NODE_157 | mtRef | 100.00 | 22705 | 30872 | 30696 | 22529 | 0.0 | 15084 |
| ... | | | | | | | | |

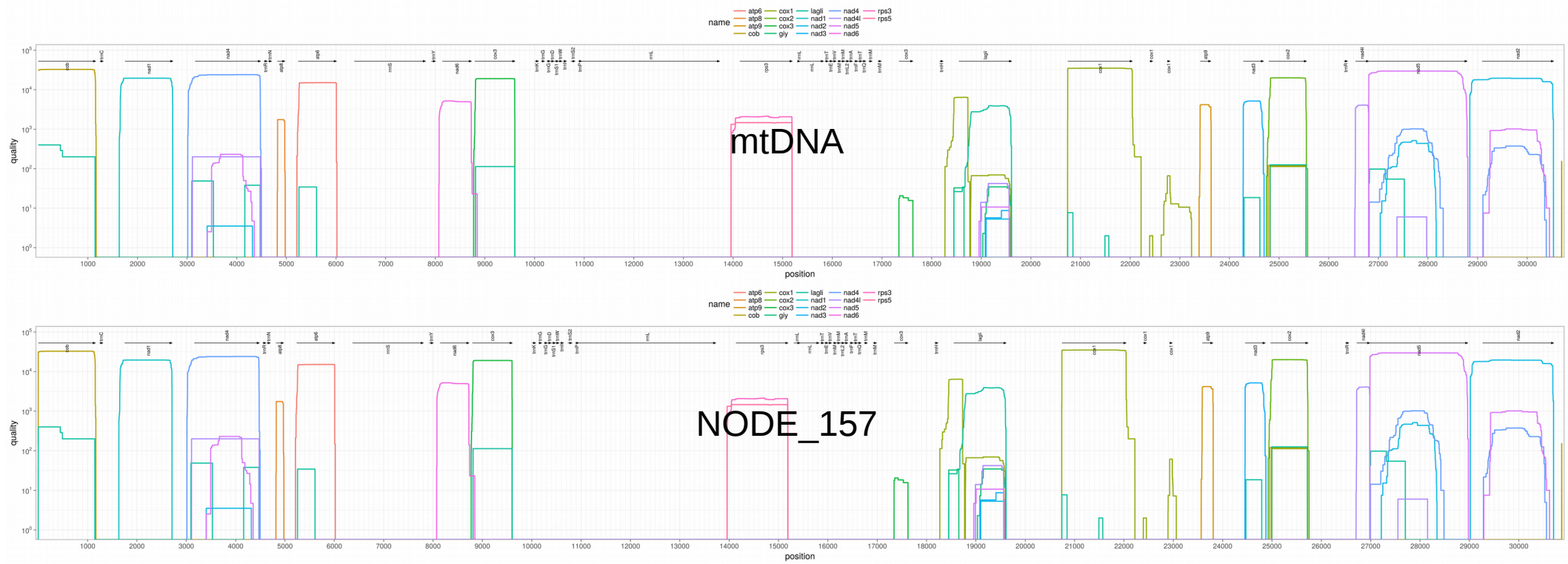
# mauve alignment



# MITOS2 annotation

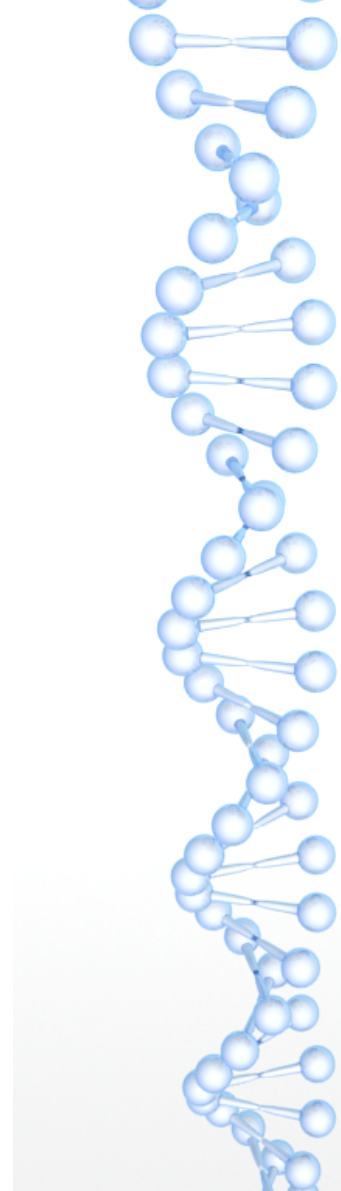


# MITOS2 annotation



# Stats

- NGS data for 33 species, 8 reference mtDNA
  - 26 passed the pipeline
    - 19 `MITOS2` annotation
      - 9 good
      - 7 so-so
      - 3 fail



# Methodical problems

# Methodical problems

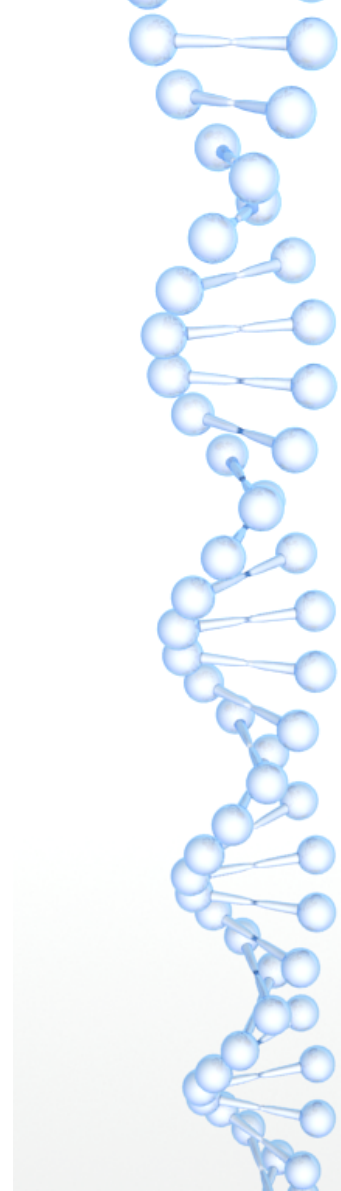
- Is it rigth to discard low quality `BLAST` hits?

# Methodical problems

- Is it rigth to discard low quality `BLAST` hits?
- If several mt contigs, which order?

# Methodical problems

- Is it rigth to discard low quality `BLAST` hits?
- If several mt contigs, which order?
- Cutting of contigs necessary?



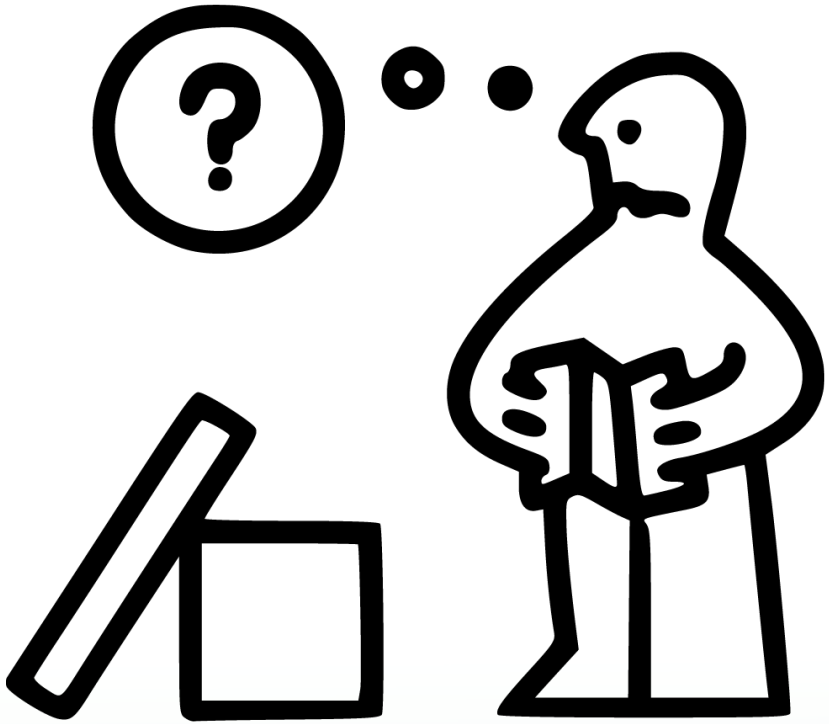
# Thanks for your attention!

# Thanks for your attention!

## And thanks to

Konstantin Riege

Matthias Bernt



# References


- Ntertili M., Kirmitzoglou I., Kouvelis V.N., Promponas V.J., Typas M.A. (2013). MitoFun: A Curated Resource of Complete Fungal Mitochondrial Genomes. Submitted
- Aguileta G, de Vienne DM, Ross ON, et al. High Variability of Mitochondrial Gene Order among Fungi. Genome Biology and Evolution. 2014;6(2):451-465. doi:10.1093/gbe/evu028.
- Georg Hausner. 6 - Fungal Mitochondrial Genomes, Plasmids and Introns. In Dilip K. Arora and George G. Khachatourians, editors, Fungal Genomics, Volume 3 of Applied Mycology and Biotechnology, pages 101 – 131. Elsevier, 2003.

# Technical problems

- Wrong species label
- Paired end reads, that are not named properly

# FASTQ header

- Header in 1st file

```
@HWI-ST377:113:C02FUABXX:2:1101:1169:2171_1:Y:0:CAGATC/1
@HWI-ST377:113:C02FUABXX:2:1101:1149:2180_1:Y:0:CAGATC/1
@HWI-ST377:113:C02FUABXX:2:1101:1236:2190_1:N:0:CAGATC/1
```

- Header in 2nd file

```
@HWI-ST377:113:C02FUABXX:2:1101:1169:2171_2:Y:0:CAGATC/2
@HWI-ST377:113:C02FUABXX:2:1101:1149:2180_2:Y:0:CAGATC/2
@HWI-ST377:113:C02FUABXX:2:1101:1236:2190_2:N:0:CAGATC/2
```

# FASTQ header

- Header in 1st file

```
@HWI-ST377:113:C02FUABXX:2:1101:1169:2171_1:Y:0:CAGATC/1
@HWI-ST377:113:C02FUABXX:2:1101:1149:2180_1:Y:0:CAGATC/1
@HWI-ST377:113:C02FUABXX:2:1101:1236:2190_1:N:0:CAGATC/1
```

- Header in 2nd file

```
@HWI-ST377:113:C02FUABXX:2:1101:1169:2171_2:Y:0:CAGATC/2
@HWI-ST377:113:C02FUABXX:2:1101:1149:2180_2:Y:0:CAGATC/2
@HWI-ST377:113:C02FUABXX:2:1101:1236:2190_2:N:0:CAGATC/2
```

# Technical problems

- Wrong species label

# Species label

- ENA entry

| Study accession | Sample accession | Secondary sample accession | Experiment accession | Run accession | Tax ID | Scientific name | Instrument model | Library layout | FASTQ files (FTP) | FASTQ files (Galaxy) | Submitted files (FTP) | Submitted files (Galaxy) | NCBI SRA file (FTP) | NCBI SRA file (Galaxy) | CRAM Index files (FTP) | CRAM Index files (Galaxy) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRJEB2977 | SAMEA2061183 | ERS236451 | ERX303996 | ERR331067 | 559306 | Ustilago maydis FB1 | Illumina HiSeq 2000 | PAIRED | File 1 File 2 | File 1 File 2 | BAM File 1 | BAM File 1 | File 1 | File 1 | | |

# Species label

- ENA entry

| Study accession | Sample accession | Secondary sample accession | Experiment accession | Run accession | Tax ID | Scientific name | Instrument model | Library layout | FASTQ files (FTP) | FASTQ files (Galaxy) | Submitted files (FTP) | Submitted files (Galaxy) | NCBI SRA file (FTP) | NCBI SRA file (Galaxy) | CRAM Index files (FTP) | CRAM Index files (Galaxy) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRJEB2977 | SAMEA2061183 | ERS236451 | ERX303996 | ERR331067 | 559306 | Ustilago maydis FB1 | Illumina HiSeq 2000 | PAIRED | File 1 File 2 | File 1 File 2 | BAM File 1 | BAM File 1 | File 1 | File 1 | | |

# Species label

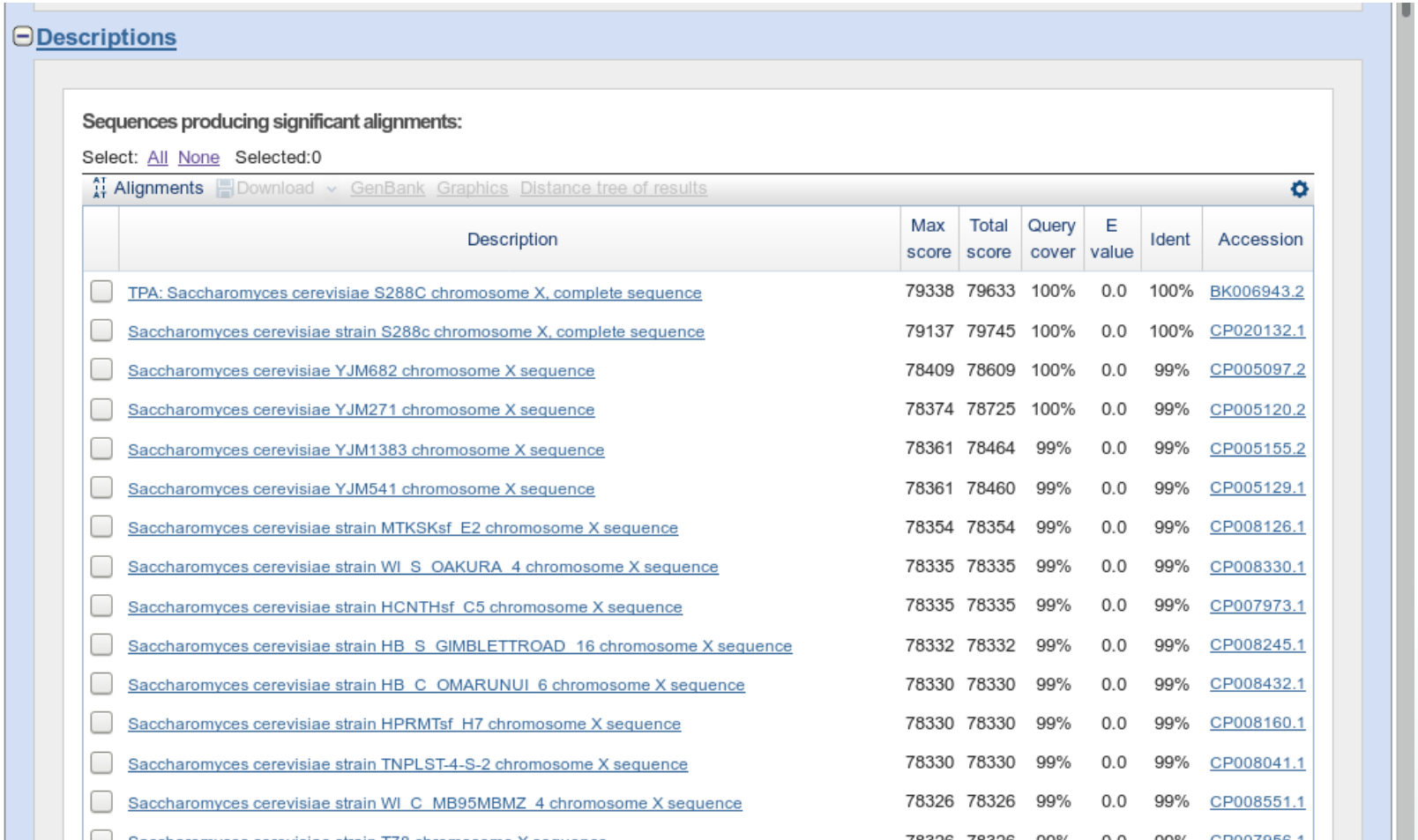
- ENA entry

| Study accession | Sample accession | Secondary sample accession | Experiment accession | Run accession | Tax ID | Scientific name | Instrument model | Library layout | FASTQ files (FTP) | FASTQ files (Galaxy) | Submitted files (FTP) | Submitted files (Galaxy) | NCBI SRA file (FTP) | NCBI SRA file (Galaxy) | CRAM Index files (FTP) | CRAM Index files (Galaxy) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRJEB2977 | SAMEA2061183 | ERS236451 | ERX303996 | ERR331067 | 559306 | Ustilago maydis FB1 | Illumina HiSeq 2000 | PAIRED | File 1 File 2 | File 1 File 2 | BAM File 1 | BAM File 1 | File 1 | File 1 | | |

- `BLAST`: assembly vs. reference mtDNA

| | |
|---|---|
| soapdenovo2k91.tab | 0 bytes |
| soapdenovo2k91_cap.tab | 0 bytes |
| soapdenovo2k91_cap_sspace.tab | 0 bytes |
| soapdenovo2k91_sspace.tab | 0 bytes |
| spades.tab | 0 bytes |
| spades_cap.tab | 0 bytes |
| spades_cap_sspace.tab | 0 bytes |
| spades_sspace.tab | 0 bytes |

# Species label



## Descriptions

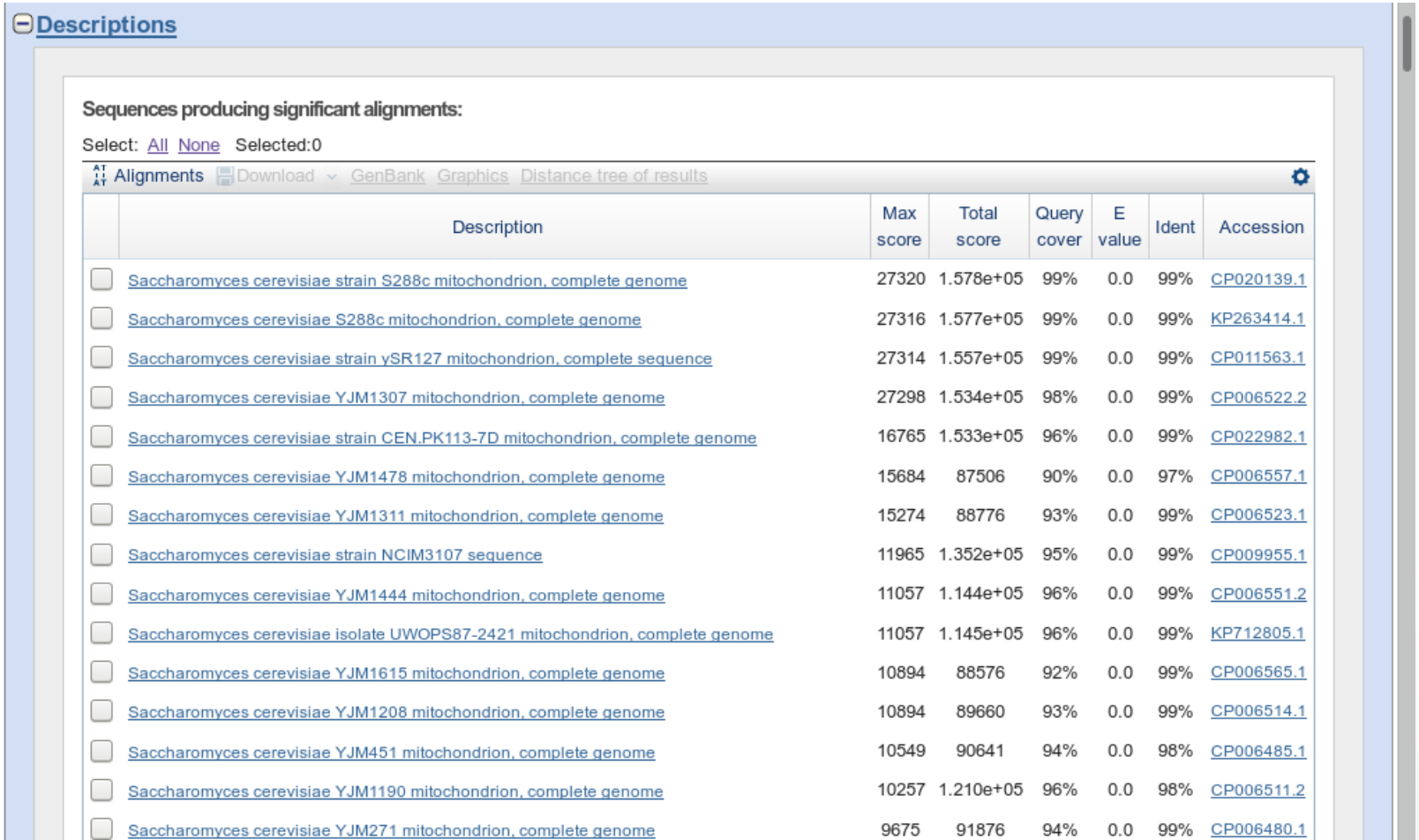
**Sequences producing significant alignments:**

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| TPA: Saccharomyces cerevisiae S288C chromosome X, complete sequence | 79338 | 79633 | 100% | 0.0 | 100% | BK006943.2 |
| Saccharomyces cerevisiae strain S288c chromosome X, complete sequence | 79137 | 79745 | 100% | 0.0 | 100% | CP020132.1 |
| Saccharomyces cerevisiae YJM682 chromosome X sequence | 78409 | 78609 | 100% | 0.0 | 99% | CP005097.2 |
| Saccharomyces cerevisiae YJM271 chromosome X sequence | 78374 | 78725 | 100% | 0.0 | 99% | CP005120.2 |
| Saccharomyces cerevisiae YJM1383 chromosome X sequence | 78361 | 78464 | 99% | 0.0 | 99% | CP005155.2 |
| Saccharomyces cerevisiae YJM541 chromosome X sequence | 78361 | 78460 | 99% | 0.0 | 99% | CP005129.1 |
| Saccharomyces cerevisiae strain MTKSKsf_E2 chromosome X sequence | 78354 | 78354 | 99% | 0.0 | 99% | CP008126.1 |
| Saccharomyces cerevisiae strain WI_S_OAKURA_4 chromosome X sequence | 78335 | 78335 | 99% | 0.0 | 99% | CP008330.1 |
| Saccharomyces cerevisiae strain HCNTHsf_C5 chromosome X sequence | 78335 | 78335 | 99% | 0.0 | 99% | CP007973.1 |
| Saccharomyces cerevisiae strain HB_S_GIMBLETTROAD_16 chromosome X sequence | 78332 | 78332 | 99% | 0.0 | 99% | CP008245.1 |
| Saccharomyces cerevisiae strain HB_C_OMARUNUI_6 chromosome X sequence | 78330 | 78330 | 99% | 0.0 | 99% | CP008432.1 |
| Saccharomyces cerevisiae strain HPRMTsf_H7 chromosome X sequence | 78330 | 78330 | 99% | 0.0 | 99% | CP008160.1 |
| Saccharomyces cerevisiae strain TNPLST-4-S-2 chromosome X sequence | 78330 | 78330 | 99% | 0.0 | 99% | CP008041.1 |
| Saccharomyces cerevisiae strain WI_C_MB95MBMZ_4 chromosome X sequence | 78326 | 78326 | 99% | 0.0 | 99% | CP008551.1 |

# Species label



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| Saccharomyces cerevisiae strain S288c mitochondrion, complete genome | 27320 | 1.578e+05 | 99% | 0.0 | 99% | CP020139.1 |
| Saccharomyces cerevisiae S288c mitochondrion, complete genome | 27316 | 1.577e+05 | 99% | 0.0 | 99% | KP263414.1 |
| Saccharomyces cerevisiae strain ySR127 mitochondrion, complete sequence | 27314 | 1.557e+05 | 99% | 0.0 | 99% | CP011563.1 |
| Saccharomyces cerevisiae YJM1307 mitochondrion, complete genome | 27298 | 1.534e+05 | 98% | 0.0 | 99% | CP006522.2 |
| Saccharomyces cerevisiae strain CEN.PK113-7D mitochondrion, complete genome | 16765 | 1.533e+05 | 96% | 0.0 | 99% | CP022982.1 |
| Saccharomyces cerevisiae YJM1478 mitochondrion, complete genome | 15684 | 87506 | 90% | 0.0 | 97% | CP006557.1 |
| Saccharomyces cerevisiae YJM1311 mitochondrion, complete genome | 15274 | 88776 | 93% | 0.0 | 99% | CP006523.1 |
| Saccharomyces cerevisiae strain NCIM3107 sequence | 11965 | 1.352e+05 | 95% | 0.0 | 99% | CP009955.1 |
| Saccharomyces cerevisiae YJM1444 mitochondrion, complete genome | 11057 | 1.144e+05 | 96% | 0.0 | 99% | CP006551.2 |
| Saccharomyces cerevisiae isolate UWOPS87-2421 mitochondrion, complete genome | 11057 | 1.145e+05 | 96% | 0.0 | 99% | KP712805.1 |
| Saccharomyces cerevisiae YJM1615 mitochondrion, complete genome | 10894 | 88576 | 92% | 0.0 | 99% | CP006565.1 |
| Saccharomyces cerevisiae YJM1208 mitochondrion, complete genome | 10894 | 89660 | 93% | 0.0 | 99% | CP006514.1 |
| Saccharomyces cerevisiae YJM451 mitochondrion, complete genome | 10549 | 90641 | 94% | 0.0 | 98% | CP006485.1 |
| Saccharomyces cerevisiae YJM1190 mitochondrion, complete genome | 10257 | 1.210e+05 | 96% | 0.0 | 98% | CP006511.2 |
| Saccharomyces cerevisiae YJM271 mitochondrion, complete genome | 9675 | 91876 | 94% | 0.0 | 99% | CP006480.1 |