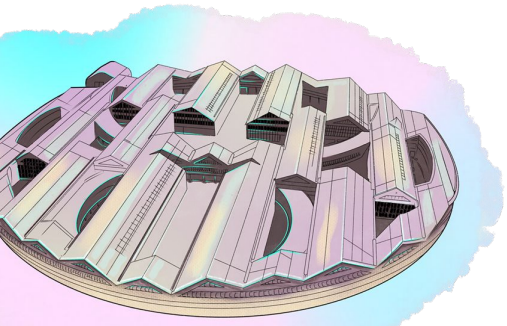
Identification of unknown plant

NGS final project

20.03.2023



Authored by:
Ekaterina Kashuk
Alisa Fedorenko
Leonid Sidorov

Skoltech

Study objectives

For our project we had paired end sequencing data of an **unknown** plant genome.

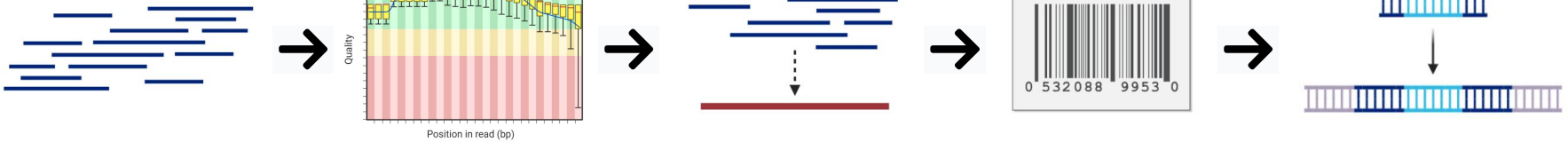
We had to do ***de novo*** genome **assembly**

But the data **is not sufficient** for the assembly of the complet **nuclear** genome, **but** sufficient for **high-copy** genomic **segments**

Thus, our tasks were:

1. To do *de novo* **assembly** for these segments;
2. Using them to **identify plant** specie with DNA-based identification (**DNA barcoding**).

The project workflow



Paired-end Illumina reads (101-101 bp)

Quality control and trimming reads

De novo genome assembly by SPAdes

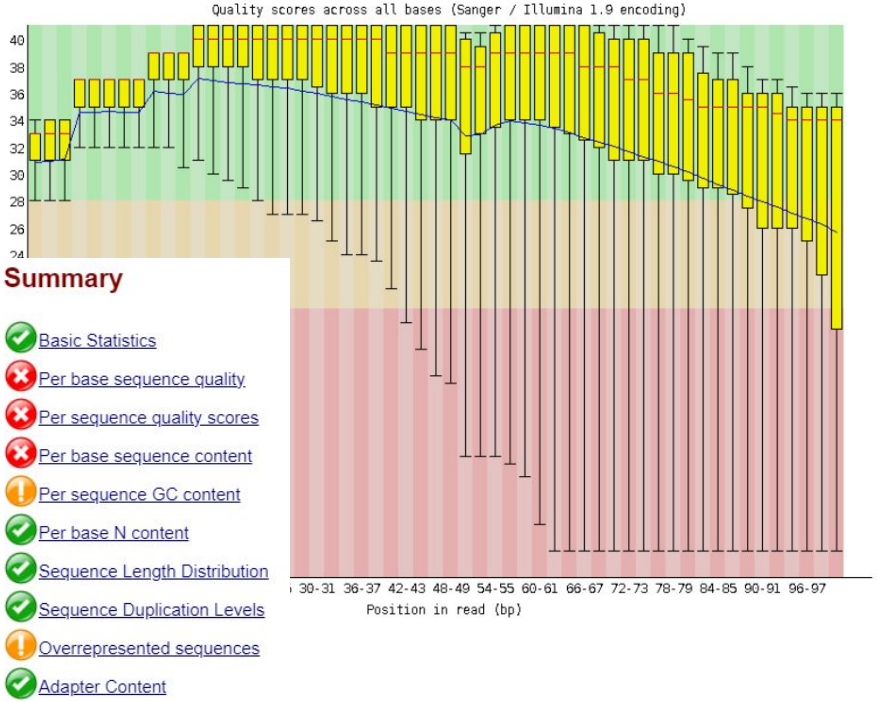
DNA barcoding for species identification

Searching for homologous sequences

Quality control

- Quality control was provided with FastQC software, no adapter was found;
- Trimming was done via fastp with default parameters: reads shorter than 50 were discarded, window size was equal to 4, threshold for quality was equal to 20

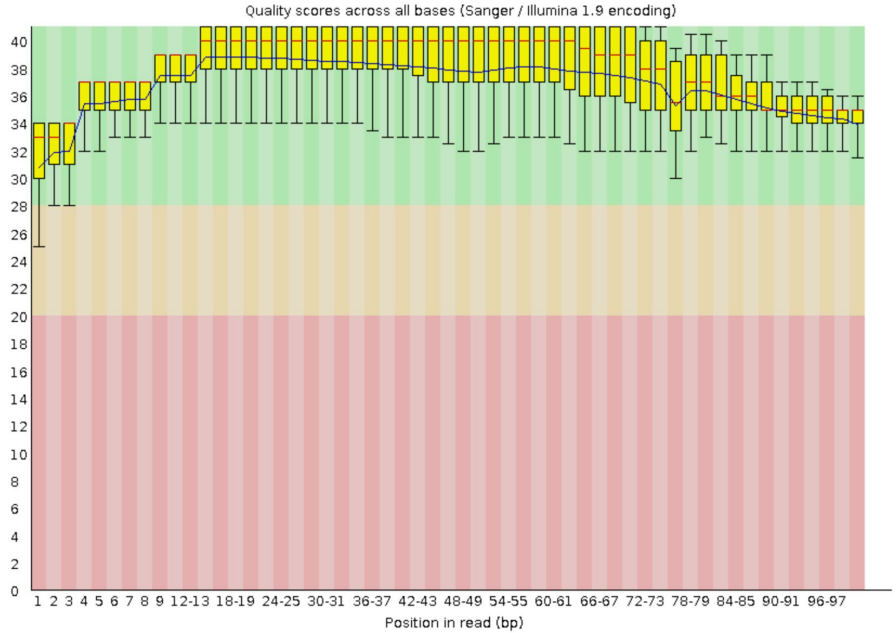
Statistics before the trimming

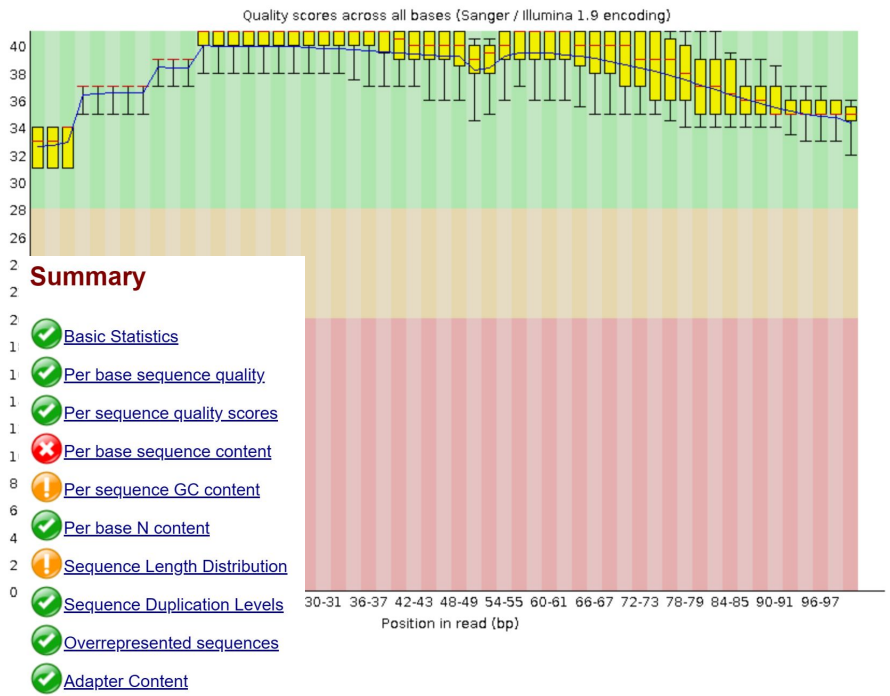


Skoltech

Quality control

- Per base sequence quality was improved;
- Per sequence quality scores were also improved



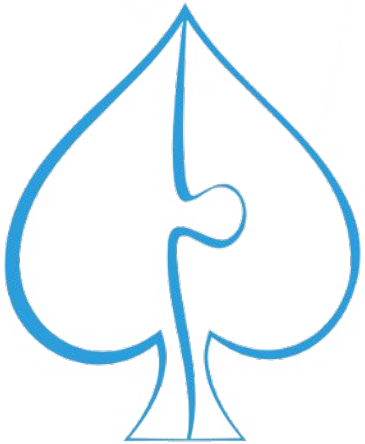


Statistics after the trimming

De novo genome assembly by SPAdes

- For the assembly we used SPAdes software;
- This tool utilizes de Brujin graphs;
- To obtain the statistics we used SeqKit

As we did not assemble the whole nuclear genome, we proceeded with the assembled fragments

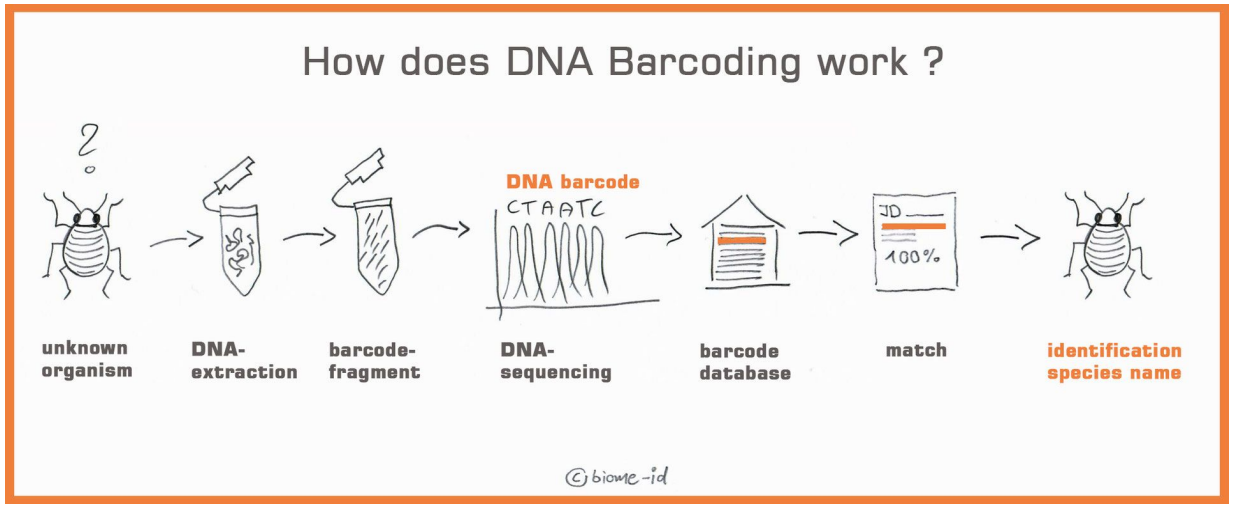


type	num_seqs	sum_len	min_len	avg_qlen	max_len	Q1	Q2	Q3	sum_gap	N50	Q20(%)	Q30(%)	GC(%)
DNA	16,38	5,833,439	56	356.1	12,038	220	253	359	0	356	0	0	31.58

Skoltech

DNA barcoding for species identification

- DNA barcoding is a method used to **identify species**. It works by analysing a **specific region of DNA** (DNA barcode). The sequence of this DNA barcode is then **compared to a reference** library which contains information of many species linked to their barcodes.



DNA barcoding

Skoltech

The issue: there is no universal barcode candidate for identification of all plant groups

Marker	Genomic source	Type	Hits
ITS1 + ITS2 +5.8SrRNA	Nuclear	Transcribed spacers and 5.8S gene	**1**
matK	Plastid	Protein-coding	0
psbK-1	Plastid	Protein-coding	0
rbcL	Plastid	Protein-coding	0
rpoC1	Plastid	Protein-coding	0
trnH-psbA	Plastid	Intergenic spacer	0
trnL	Plastid	Intron	0
cox1	Mitochondria	Protein-coding	0
16S rRNA + 23S rRNA + 4.5S rRNA	Plastid (Rhizanthella gardneri)	Corresponding rRNA genes	**4**

DNA barcoding

Skoltech

The advantage: achieving maximum species discrimination

Marker	Genomic source	Type	Hits
ITS1 + ITS2 +5.8S rRNA	Nuclear	Transcribed spacers and 5.8S gene	**1**
matK	Plastid	Protein-coding	0
psbK-1	Plastid	Protein-coding	0
rbcL	Plastid	Protein-coding	0
rpoC1	Plastid	Protein-cod	0
trnH-psbA	Plastid	Inte sp	
trnL	Plastid	I	
cox1	Mitochondria	Prote	
16S rRNA + 23S rRNA + 4.5S rRNA	Plastid (Rhizanthella gardneri)	Corre rRNA genes	

The two-locus core barcode:
rbcL (ribulose 1, 5-bisphosphate carboxylase/oxygenase large subunit) + **matK** (maturase K)

DNA barcoding

Skoltech

The issue: the discriminating ability of these markers has been found be very low

Marker	Genomic source	Type	Hits
ITS1 + ITS2 +5.8S rRNA	Nuclear	Transcribed spacers and 5.8S gene	**1**
matK	Plastid	Protein-coding	**0**
psbK-1	Plastid	Protein-coding	0
rbcL	Plastid	Protein-coding	**0**
rpoC1	Plastid	Protein-cod	0
trnH-psbA	Plastid	Inte sp	
trnL	Plastid	I	
cox1	Mitochondria	Prote	
16S rRNA + 23S rRNA + 4.5S rRNA	Plastid (Rhizanthella gardneri)	Corre rRNA genes	

The two-locus core barcode:
rbcL (ribulose 1, 5-bisphosphate carboxylase/oxygenase large subunit) + **matK** (maturase K)

DNA barcoding

Skoltech

The issue: there is no universal combination for identification of all plant groups

Marker	Genomic source	Type	Hits
ITS1 + ITS2 +5.8S rRNA	Nuclear	Transcribed spacers and 5.8S gene	**1**
matK	Plastid	Protein-coding	0
rbcL-1	Plastid	Protein-coding	0
rbcL	Plastid	Protein-coding	0
rpoC1	Plastid	Protein-coding	0
trnH-psbA	Plastid	Intergenic spacer	0
trnL	Plastid	Intron	0
cox1	Mitochondria	Protein-coding	0
16S rRNA + 23S rRNA + 4.5S rRNA	Plastid (Rhizanthella gardneri)	Corresponding rRNA genes	**4**

Nuclear internal transcribed spacer (ITS)

Searching for homology regions

Making blast database from assembly

Checking homology of barcodes with local blastn

Found sequences were checked via blastn to identify species

Searching for homology regions

```
Score = 220 bits (119),  Expect = 1e-57
Identities = 155/172 (90%), Gaps = 3/172 (2%)
```

Sequence (homologue of marker ITS1 + ITS2 + 5.8S rRNA):

CTGTAAGCTAAACATGACTCTCGGCAATGGATATCTCGGCTCCCGCATCGATGAAGAACGCAGCGAA
ATGCGATACGTGGTGCGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCA
AGGCCCTTAGGCCAAGGGCACGCCTGCCTGGGCGTCA

Epipogium aphyllum isolate JAE small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence

Sequence ID: MK450410.1 Length: 803 Number of Matches: 1

Range 1: 337 to 507 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
309 bits(342)	9e-80	171/171(100%)	0/171(0%)	Plus/Plus

```
Query  1    CTGTAAGCTAAACATGACTCTCGGCAATGGATATCTCGGCTCCCGCATCGATGAAGAACG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  337  CTGTAAGCTAAACATGACTCTCGGCAATGGATATCTCGGCTCCCGCATCGATGAAGAACG  396

Query  61   CAGCGAAATGCGATACGTGGTGCGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAAC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  397  CAGCGAAATGCGATACGTGGTGCGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAAC  456

Query  121  GCAAGTTGCGCCCAAGGCCCTTAGGCCAAGGGCACGCCTGCCTGGGCGTCA  171
            |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  457  GCAAGTTGCGCCCAAGGCCCTTAGGCCAAGGGCACGCCTGCCTGGGCGTCA  507
```

Searching for homology regions

```
Score = 588 bits (318),  Expect = 2e-167
Identities = 367/391 (94%), Gaps = 1/391 (0%)
```

Sequences (homologue of marker 16S rRNA + 23S rRNA + 4.5S rRNA):

AGTGGGAGGGCCACCGATCAACGGATAAAAGTTACTCTAGGGATAACAGGCTGATCTTCGCCGAGAGTTCACATC
GACGGGAAGGTTTGGCACCTCGATGTCGGCTCTTCGCCACCTGGGGCTGAAGTGTGTTCCAAGGGTTGGGCTGT
TCGCCCATTAAAGCGGTACGTGAGCTGGGTTCAGAACGTCGTGAGACAGTTCGGTCCATATCCGGTGTGGGCGCT
AGAGCATTGAGGGGTATTTTCCCTAGTACGAGAGGACCGGGAAGGACGCACCTCTGGTGTACCAGTTATCGTGCC
TACGGTAAATGCTGGGTAGCTAAGTGCGGGGTGGATAACTGCTGAAAGCATATAAGTAGTAAGCCCACCCCAAGA
TGAGTGCTCTCCTATT

Epipogium aphyllum plastid, complete genome

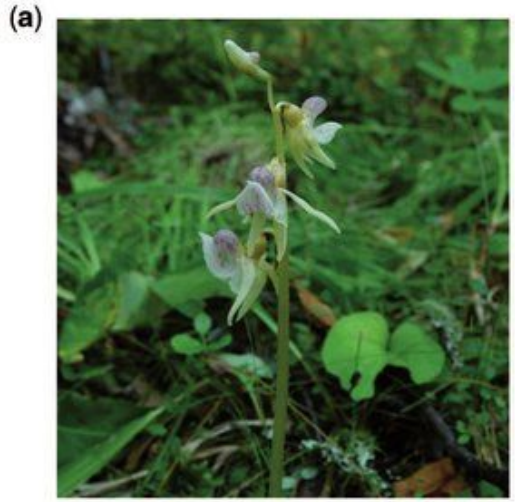
Sequence ID: NC_026449.1 **Length:** 30650 **Number of Matches:** 2

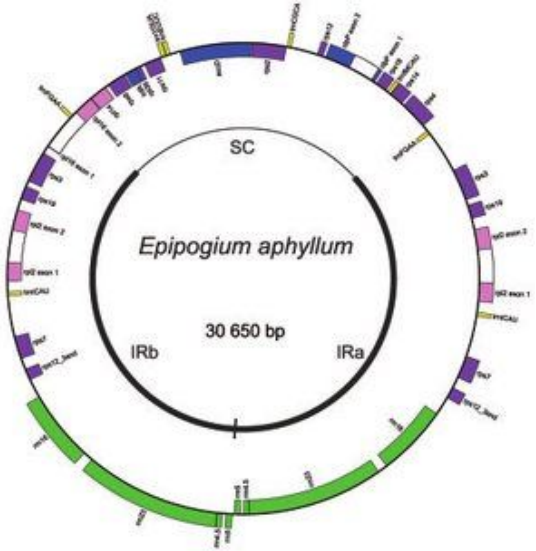
Range 1: 18619 to 19007 GenBank Graphics ▼ Next Match ▲

Score	Expect	Identities	Gaps	Strand
693 bits(768)	0.0	387/389(99%)	0/389(0%)	Plus/Plus

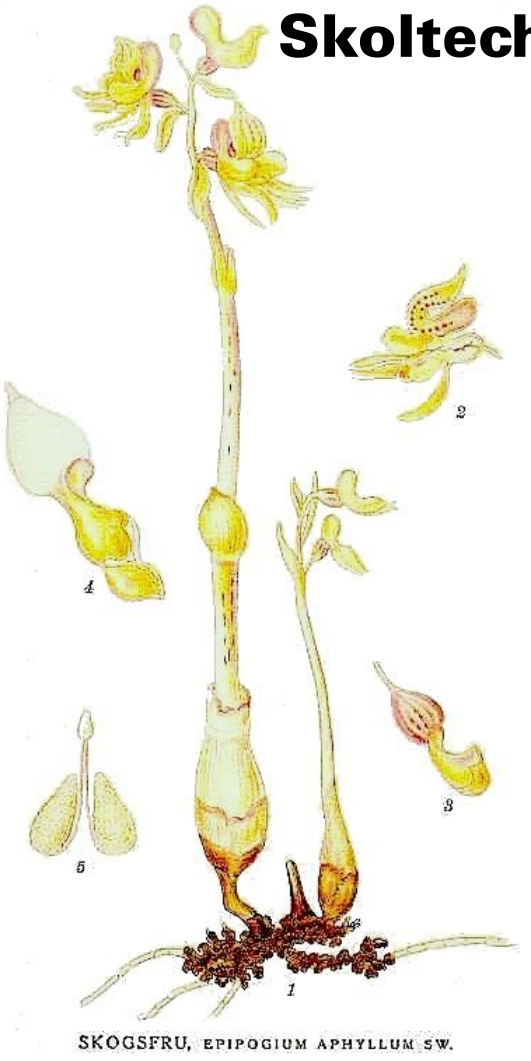
Epipogium aphyllum or *Ghost orchid*

- obligate mycoheterotrophs (or epiparasites) that obtain nutrients from mycorrhizal networks





General view and plastid genome map of Epipogium aphyllum
https://doi.org/10.1093/gbe/evv019



Skoltech

Thank you for your attention!

Link to Github repository

References

1. Nevill, P.G., Zhong, X., Tonti-Filippini, J. et al. Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. Plant Methods 16, 1 (2020). https://doi.org/10.1186/s13007-019-0534-5

2. Zeng, CX., Hollingsworth, P.M., Yang, J. et al. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods 14, 43 (2018). https://doi.org/10.1186/s13007-018-0300-0

3. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PLoS One. 2011;6(5):e19254. doi: 10.1371/journal.pone.0019254. Epub 2011 May 26. PMID: 21637336; PMCID: PMC3102656.