DBS
DIPARTIMENTO
DI BIOSCIENZE

Towards a reference genome for *Salvia pratensis*

UNIVERSITÀ DEGLI
STUDI DI MILANO

Paleni, Chiara[1]; Puricelli, Cristina[1]; Pieri, Camilla[1]; Manrique, Silvia[1]; De Paola, Larissa[1]; Bombarely, Aureliano[2]; Kater, Martin M. [1]; Lambertini, Carla[1]

[1] Botanical garden of Brera, Department of Biosciences, University of Milan, Via Brera 28, 20121 Milan, Italy
[2] Instituto de Biología Molecular y Celular de Plantas (IBMCP), CSIC-UPV, Calle Ingeniero Fausto Elio S/N, 46011 Valencia, Spain

1) BACKGROUND

The genus *Salvia* has a long history of human use. Out of more than 900 species in the genus, 25 can be found in the wild in Italy. *S. pratensis* (left) is one of the most common and is closely related to some endemic taxa with debated species rank (*S. ceratophylloides*, *S. saccardiana*, *S. haematodes*). In the Botanical Garden of Brera, we are interested in studying the distribution of genetic diversity in wild populations of *Salvia pratensis* and related taxa to inform species delineation and guide conservation efforts. We are also interested in what genes are responsible for the distinctive characteristics of *Salvia*, such as their flowers, the stamen mechanism or aromatic oil production. To reach these goals, we sequenced the whole genome of *Salvia pratensis*.

2) GOALS

1) Genome assembly of *Salvia pratensis*.

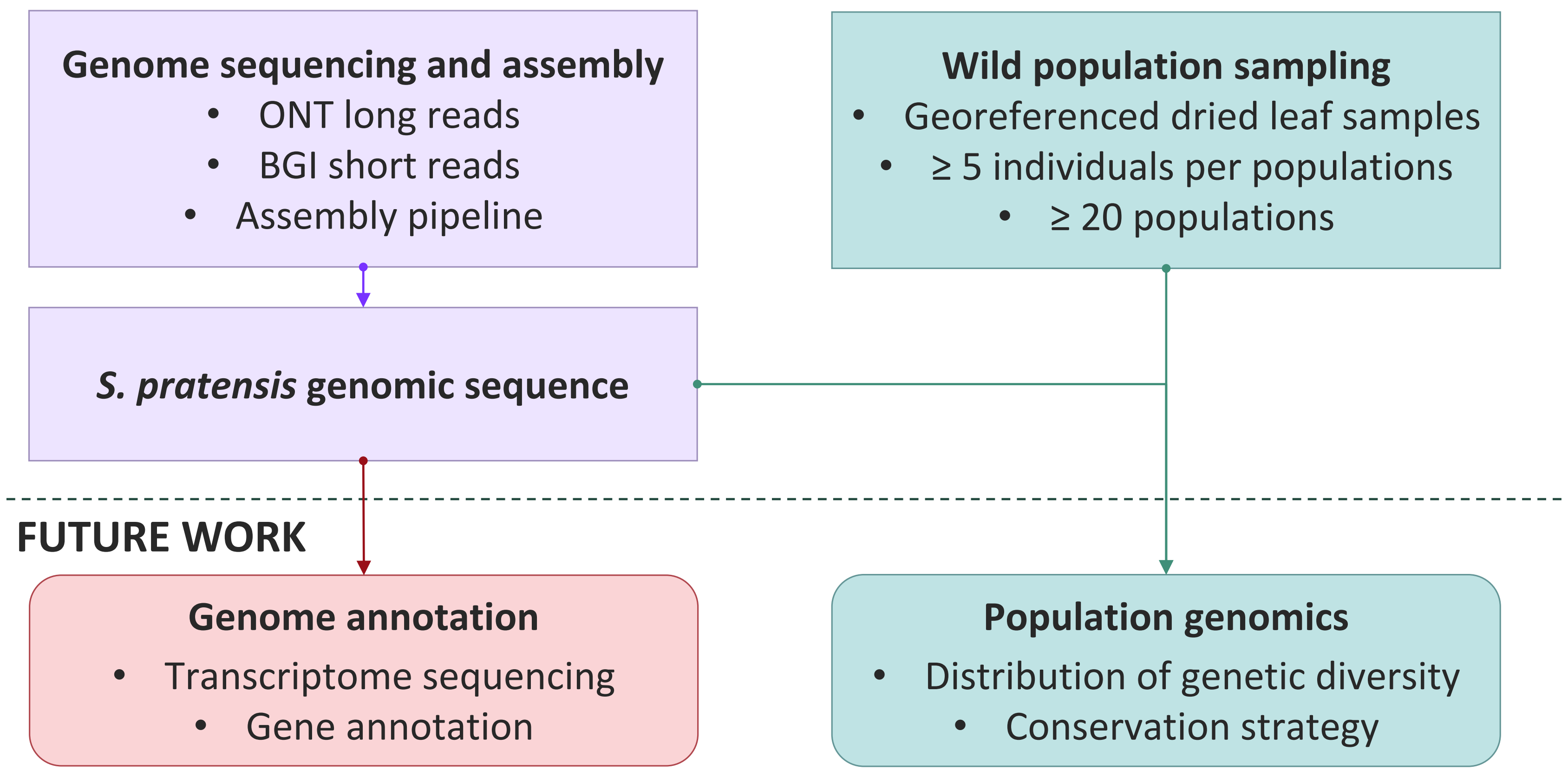
2) Functional annotation.

The genome sequence will be used to:

3) Identify endemic cryptic species in *S. pratensis* populations.

4) Find candidate genes involved in flower development.

3) METHODS AND WORK IN PROGRESS



Genome sequencing and assembly
- ONT long reads
- BGI short reads
- Assembly pipeline

Wild population sampling
- Georeferenced dried leaf samples
- ≥ 5 individuals per populations
- ≥ 20 populations

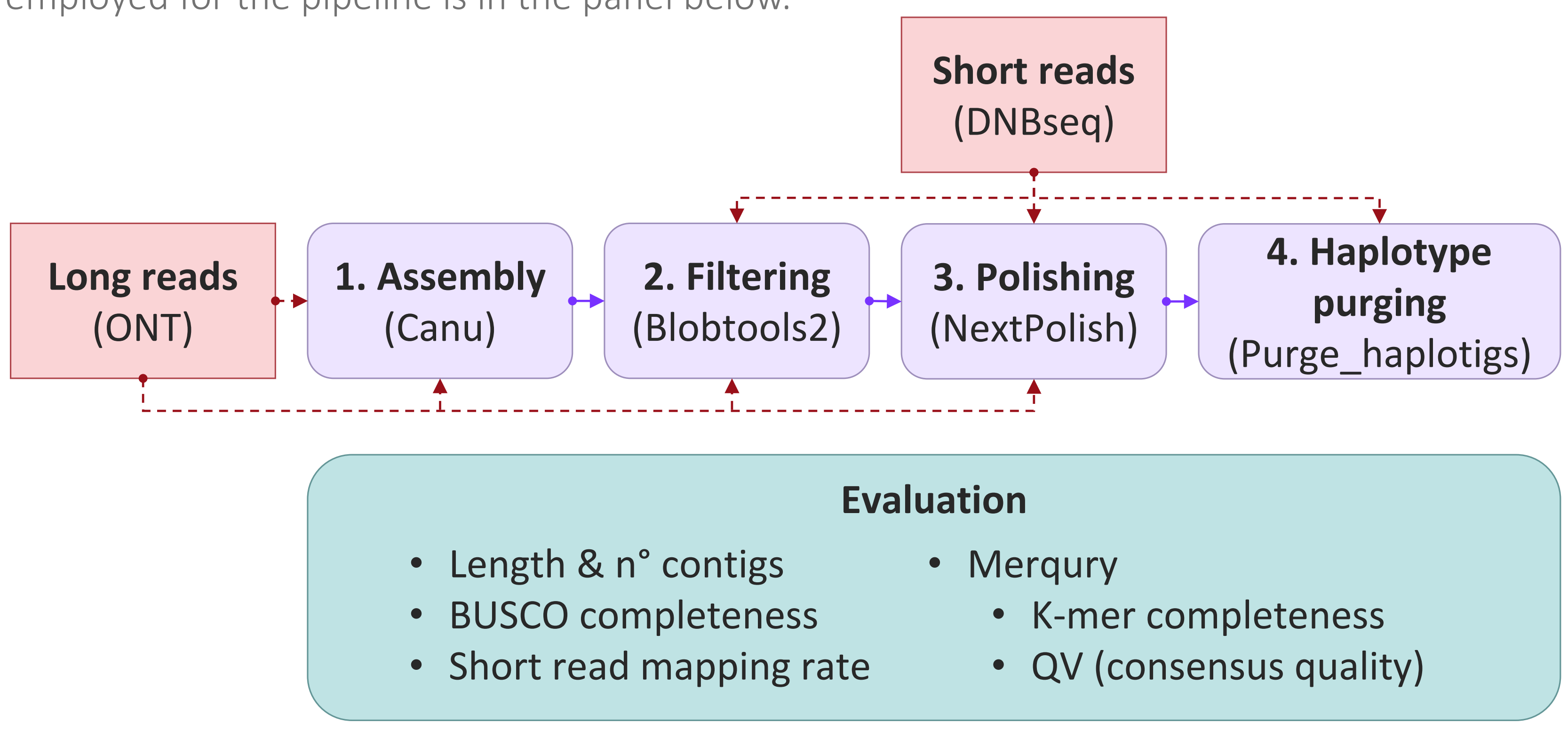
***S. pratensis* genomic sequence**

FUTURE WORK

Genome annotation
- Transcriptome sequencing
- Gene annotation

Population genomics
- Distribution of genetic diversity
- Conservation strategy

Assembly pipeline. Dashed lines connect data to software. Solid lines connect consecutive steps in the pipeline. Each step produces a version of the assembly which is evaluated according to the criteria described in the Evaluation step. Summary of sequencing data employed for the pipeline is in the panel below.



Short reads (DNBseq)

Long reads (ONT)

1. Assembly (Canu)

2. Filtering (Blobtools2)

3. Polishing (NextPolish)

4. Haplotype purging (Purge_haplotigs)

Evaluation
- Length & n° contigs
- BUSCO completeness
- Short read mapping rate
- Merqury
  - K-mer completeness
  - QV (consensus quality)

Summary of sequencing data employed in the Assembly pipeline above.

Short reads (BGI)	
Total Processed Reads (M)	173.43
Total Processed Bases (Gb)	25.91
Mean duplication (%)	10.08
Estimated Genome Coverage (X)	60

Long reads (ONT)	
Total Raw Reads (M)	2.36
Total Raw Bases (Gb)	74.96
Mean read quality	12.3
Mean read length (b)	31,665.4
Estimated Genome Coverage (X)	175

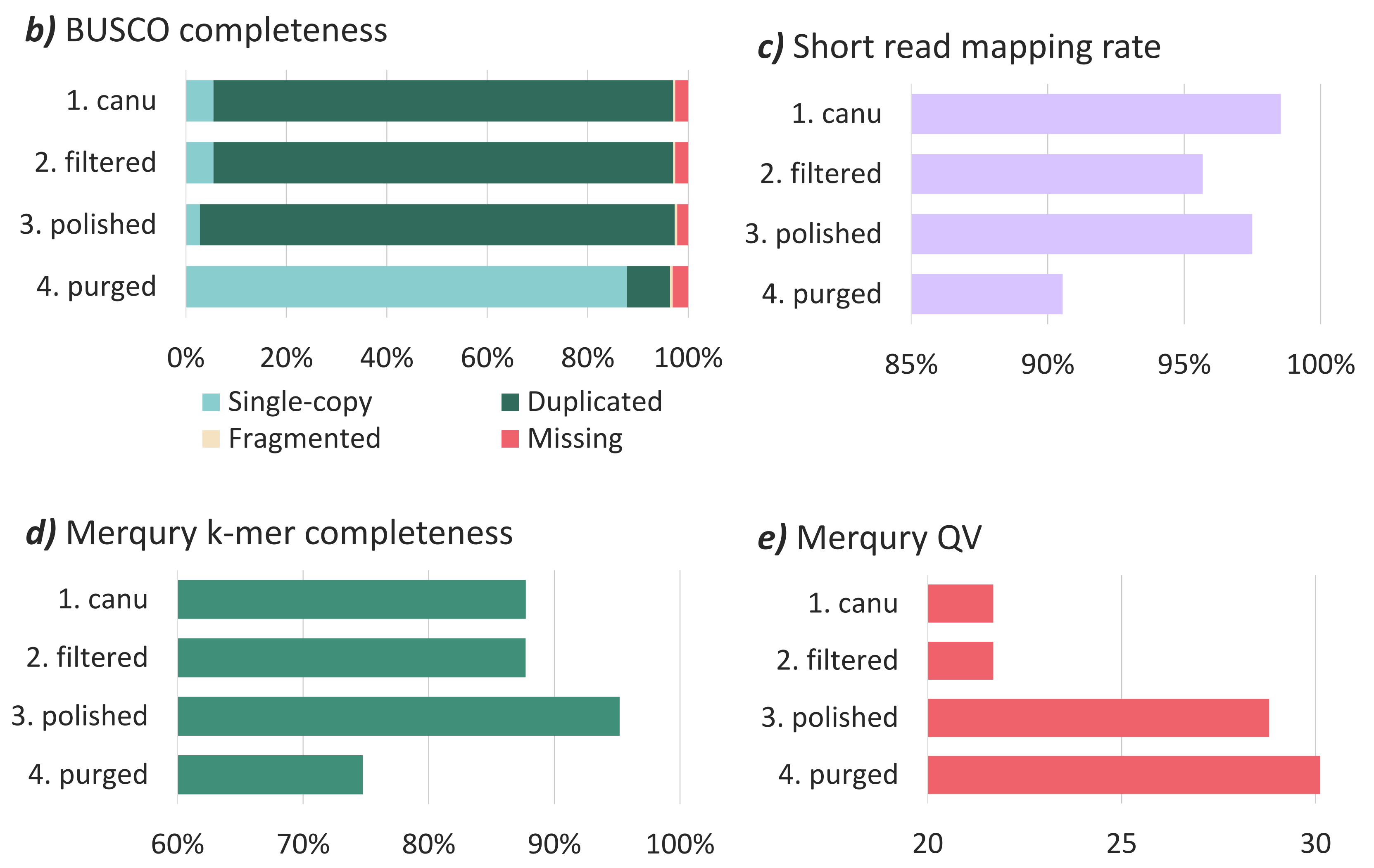
4) RESULTS

We assembled long reads (Canu) and after filtering (Blobtools2) and polishing (NextPolish) we obtained an assembly of length 877 Mb, comprised of 1,159 contigs (Statistics panel). This assembly length is twice the genome size estimation from previous experiments; BUSCO analysis revealed a high level of duplication. This issue may be caused by the assembler not collapsing haplotypes due to the heterozygosity rate of the individual. With haplotype purging (Purge_haplotigs) we reduced the assembly size and duplication rate, but the percentage of reads mapping on the assembly decreased. The consensus quality improved at each assembly step and the error rate of version 4 of the assembly is 0.10%. This is still far from reference quality and may cause issues in gene annotation. With the RepeatModeler-RepeatMasker pipeline on version 4 of the assembly we annotated 56% of the sequence as repetitive.

Statistics of versions 1 to 4 of the *S. pratensis* assembly (see Assembly pipeline).
a) length and contiguity, compared to genome size estimation from previous experiments.
b) BUSCO completeness evaluated against the eudicots_odb10 (2020-09-10) dataset.
c) mapping rate on the assembly of short reads from the same individual.
d-e) Merqury k-mer completeness (d) and consensus quality values (e), computed on reads from the same individual with k=19.

a)

Assembly version	Length (Mb)	n. contigs	N50 (Mb)	L50
1. canu	872	1,190	1.74	117
2. filtered	869	1,159	1.74	117
3. polished	877	1,159	1.76	117
4. purged	373	159	4.36	26
Flow cytometry GS estimate	430	-	-	-
K-mer analysis GS estimate	330 - 460	-	-	-



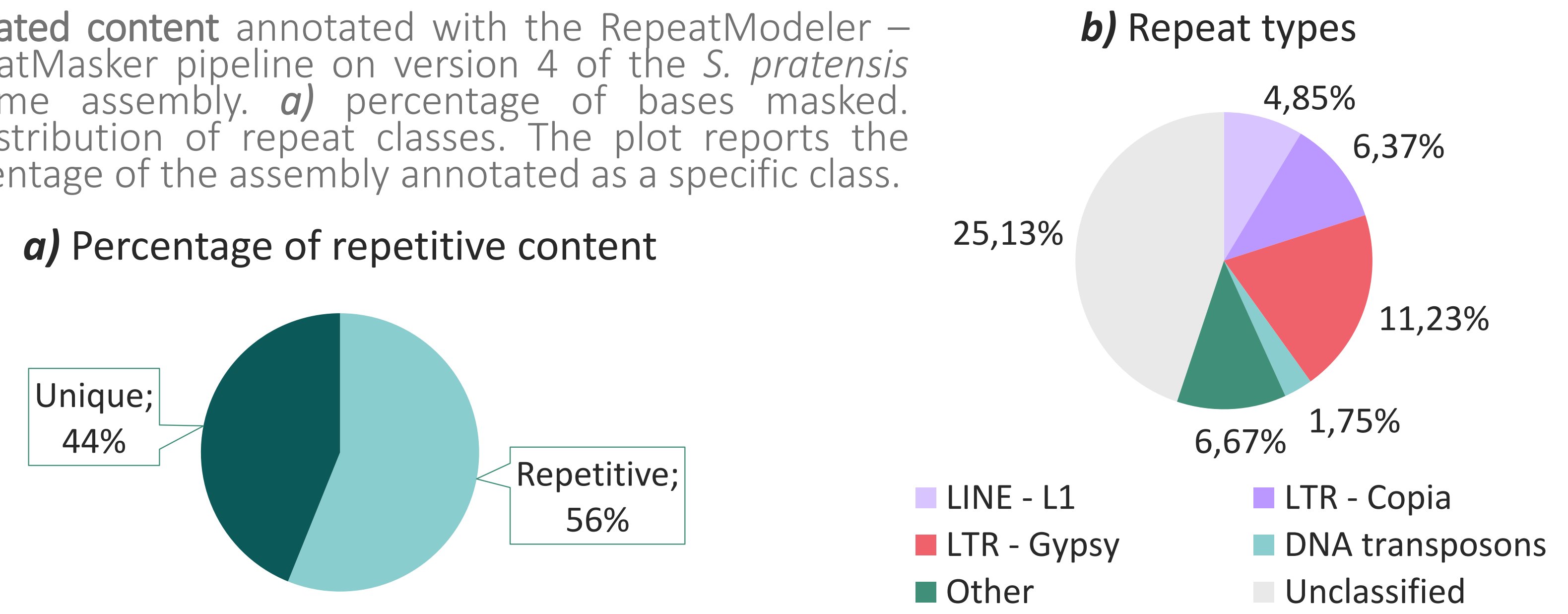
b) BUSCO completeness

c) Short read mapping rate

d) Merqury k-mer completeness

e) Merqury QV

Repeated content annotated with the RepeatModeler – RepeatMasker pipeline on version 4 of the *S. pratensis* genome assembly. *a)* percentage of bases masked. *b)* distribution of repeat classes. The plot reports the percentage of the assembly annotated as a specific class.



a) Percentage of repetitive content

b) Repeat types

5) FUTURE WORK

We have gathered a collection of samples from wild populations of *Salvia* thanks to our network of collaborators in Italy. We will use the assembled genome as reference to genotype the collection and perform a population genomics study to help clarify the species status of endemic taxa. Finally, we will sequence the transcriptome of different tissues and developmental stages of *S. pratensis* to annotate coding sequences and identify genes that may be involved in flower development, such as transcription factors of the MADS-box family.

References

S. saccardiana: Del Carratore *et al.*, 1999 (10.1080/11263509909381544); *S. haematodes:* Linnaeus C. Species Plantarum: 24, 1753; *S. ceratophylloides:* Arduino P. Animadversiorum Botanicorum Specimen Alterum. Ex Typographia Sansoniana: Venetis; 1764.

Blobtools2 v2.6.4 (10.1534/g3.119.400908); Canu v2.2 (10.1101/gr.215087.116); Purge_haplotigs v1.1.2 (10.1186/s12859-018-2485-7); NextPolish v1.4.0 (10.1093/bioinformatics/btz891); BUSCO v5.2.2 (10.1093/molbev/msab199); Merqury v1.3 (10.1186/s13059-020-02134-9); RepeatModeler, RepeatMasker via Dfam - TE Tools container v1.5 (10.1073/pnas.1921046117)