Towards a reference genome for *Salvia pratensis*

DIPARTIMENTO DI BIOSCIENZE

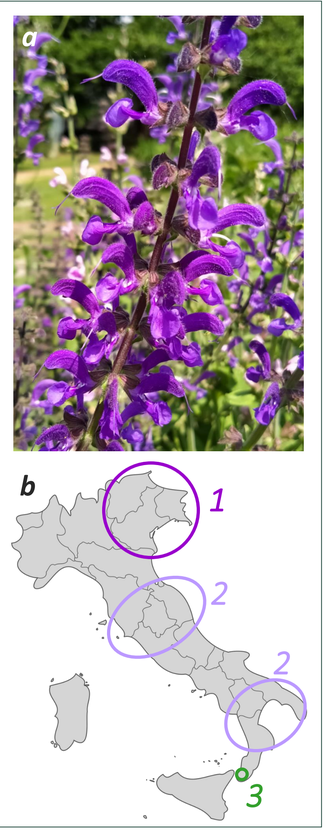
UNIVERSITÀ DEGLI STUDI DI MILANO

Paleni, Chiara[1]; Puricelli, Cristina[1]; Pieri, Camilla[1]; Manrique, Silvia[2]; De Paola, Larissa[1]; Bombarely, Aureliano[3]; 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 Universitario de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Camino de Vera S/N, 46022 Valencia, Spain
[3] Instituto de Biología Molecular y Celular de Plantas, Universitat Politècnica de València, 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* is one of the most common and is closely related to some endemic taxa with debated species rank. In the Botanical Garden of Brera, we are 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 genes responsible for the distinctive flower morphology and pollination mechanisms of *Salvia pratensis*. To reach these goals, we sequenced the whole genome of *Salvia pratensis* and sampled wild populations across Italy.



Salvia pratensis and related species. ***a)*** *S. pratensis* flowers (Città Studi Botanical garden, Milan, May 2022). ***b)*** Known range of the endemic Italian Salvia taxa mentioned in the text: ***1)*** *S. saccardiana;* ***2)*** *S. haematodes;* ***3)*** *S. ceratophylloides*. *S. pratensis* is commonly found in most of the Northern and Central regions. *S. ceratophylloides* grows in an extremely limited range and was considered extinct in the wild between 1997-2008.

2) GOALS

1) Genome assembly and annotation of *Salvia pratensis*.

2) Genomic diversity study on wild populations.

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

Genome sequencing and assembly
- Oxford Nanopore long reads
- DNBSEQ PE short reads
- PacBio HiFi reads

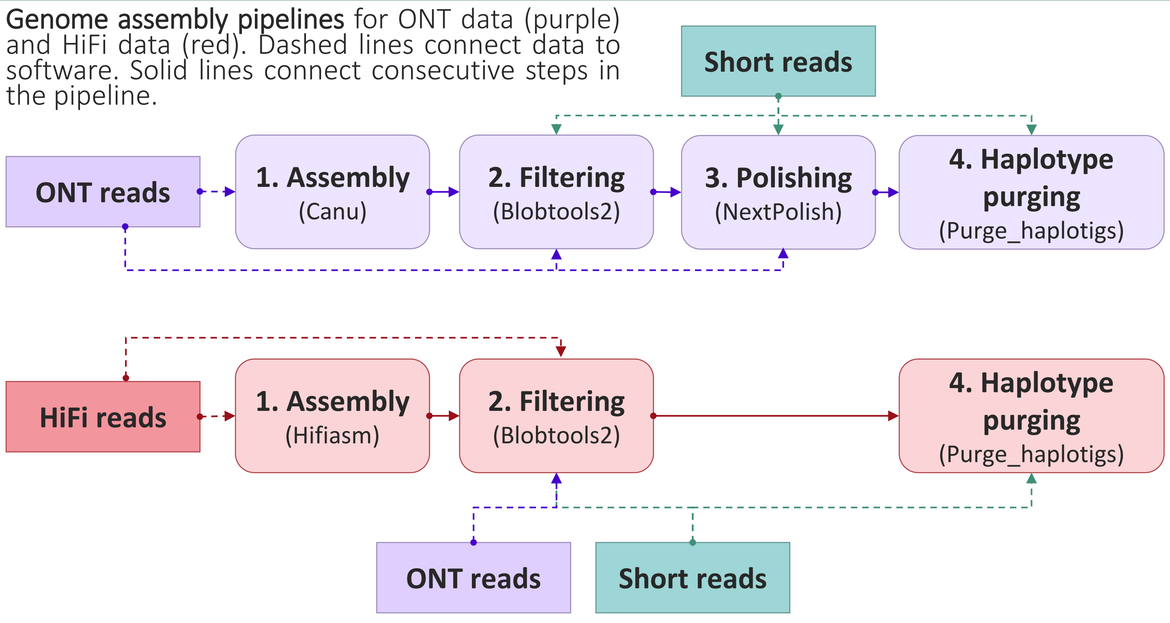
Wild population sampling
- Georeferenced dried leaf samples
- ≥ 5 individuals per populations
- ≥ 20 populations

***S. pratensis* genomic sequence**

Flower developement (WIP)
- Transcriptome sequencing
- RNA-FISH

Population genomics (WIP)
- Distribution of genetic diversity
- Conservation strategy

Genome assembly pipelines for ONT data (purple) and HiFi data (red). Dashed lines connect data to software. Solid lines connect consecutive steps in the pipeline.



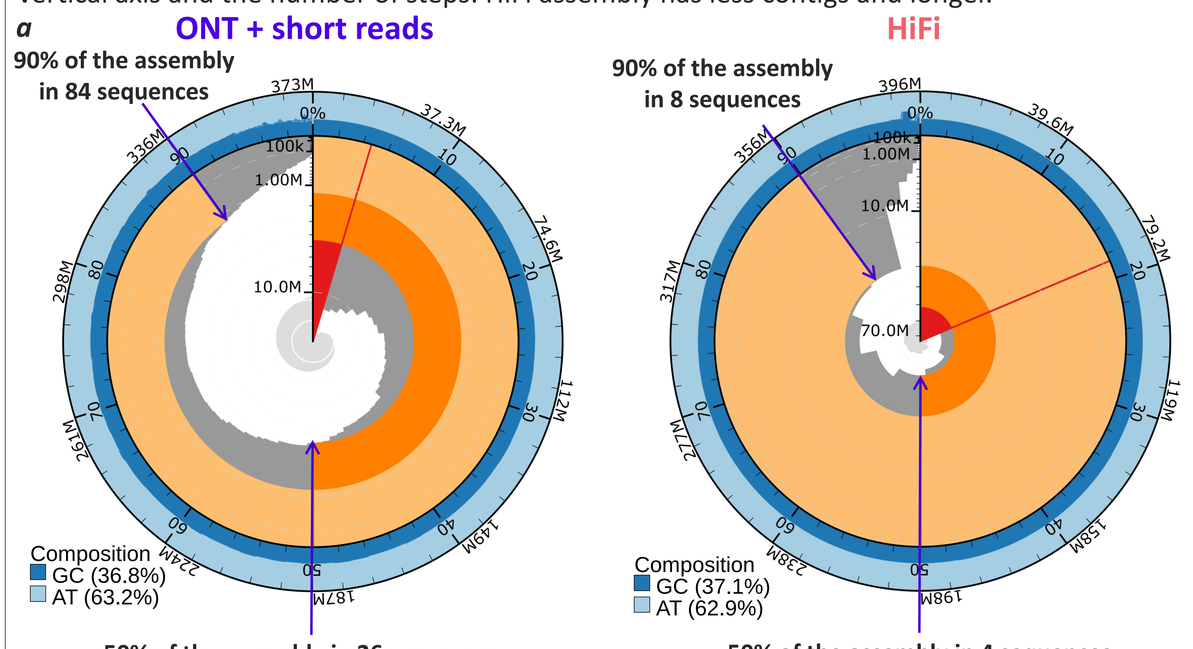
Summary of sequencing data employed in the Assembly pipeline above.

	Short reads (BGI)	Long reads (ONT)	HiFi reads (PacBio)
Total Reads (M)	173.43	2.36	3.16
Total Bases (Gb)	25.91	74.96	28.76
Mean read length (b)	2x150	31,665	9,078
Est. Genome Coverage (X)	60	175	70
Read N50 (kb)	-	42	12

4) RESULTS

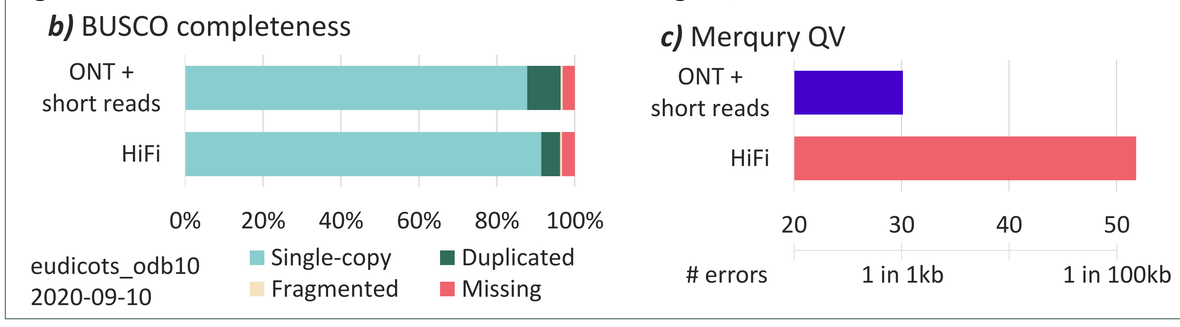
We first used Oxford Nanopore reads for assembly, and then a HiFi read set. HiFi reads had a large positive impact on assembly contiguity and quality when compared to ONT for *S. pratensis*, a non-model plant with high heterozygosity (1.5% according to k-mer analysis). Even without polishing, necessary in the ONT assembly because of low read quality, the HiFi assembly shows higher agreement with short read data (shown by the high QV score, corresponding to a low error rate of the assembly). Improvements are likely due to a better ability to separately assemble haplotypes with more accurate HiFi reads and the introduction of chimeras mixing the two haplotypes in the ONT assembly during polishing. The Hifiasm assembler has issues in regions containing repeats of rDNA, causing several short contigs (12 after purging, mean length: 42kb). These may come from organellar DNA either free or from nuclear insertions and might be solved by scaffolding the contigs with ONT reads, which are much longer. A better assembly facilitates gene annotation and read mapping.

Genome assembly results comparison. ***a)*** Snail plot (Blobtools2) comparing the ONT assembly (left) to the HiFi assembly (right). Each of the "steps" of the plot, in dark gray, is one of the sequences (or contigs) comprising the assembly, in order of length (represented by the height of the step). Assembly statistics and color legend at the bottom. Note the different scale of the vertical axis and the number of steps: HiFi assembly has less contigs and longer.



Assembly statistics	ONT + short	HiFi
# contigs	159	33
Sequence length (Mb)	373	396
Longest sequence (Mb)	17.3	73.7
L50 | N90(Mb)	26 | 4.36 Mb	4 | 50.9 Mb
L90 | N90 (Mb)	84 | 1.34 Mb	8 | 29.6 Mb

b) BUSCO analysis checks for presence of highly conserved, single-copy genes in the assembly. A high percentage of missing or duplicated BUSCOs indicates issues. ***c)*** Assembly consensus quality as evaluated by Merqury k-mer QV (QV= -10Log(Error Rate)). Merqury compares short sequences (k-mers) in the assembly to those of the high quality, short read set. A high agreement and low rate of misassemblies lead to a high QV.



5) POPULATION SAMPLING

Thanks to a network of collaborators, we have gathered a collection of *S. pratensis* leaf samples dried in silica gel (right: sampled locations) and work is in progress to extract DNA and prepare a Genotyping-By-Sequencing library. However, several regions are not well represented, and we may repeat the sampling campaign during the flowering season to better cover Central and Southern Italy.

6) FUTURE WORK

We will use the assembled genome as reference to genotype the wild collection and perform a population genomics study to help clarify the species status of endemic taxa. 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 MADS-box transcription factors.

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. Distribution data from the Portal to the Flora of Italy (http:/dryades.units.it/floritaly) and Spampinato et al., 2019 (10.5772/intechopen.84905). 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); Hifiasm v0.16.1-r375 (10.1038/s41592-020-01056-5); BUSCO v5.2.2 (10.1093/molbev/msab199); Merqury v1.3 (10.1186/s13059-020-02134-9; 10.1038/s41587-022-01261-x). Sequencing performed by: BGI Tech Solutions (short reads); KeyGene; Next Generation Sequencing Platform, University of Bern.