UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

PON RICERCA E INNOVAZIONE

REACT EU

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI BIOSCIENZE

PHD SCHOOL IN MOLECULAR AND CELLULAR BIOLOGY

XXXVII CYCLE

Genomics for wild and medicinal plant diversity studies

Chiara Paleni

R13553

Scientific advisor: Prof. Martin Kater

A.A. 2023/2024

The doctoral scholarship was co-funded with resources in the frame of the PON REACT-EU financing Program
Action IV.5 – Doctorates on green topics
This thesis reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Table of contents

ABSTRACT **1**

AIM OF THE PROJECT **2**

INTRODUCTION **4**

1. Genomics for biodiversity and the role of botanical gardens **4**

2. Reference genome generation **5**

2.1. K-mer spectrum analysis *8*

2.2. Genome assembly *8*

2.3. Polishing *10*

2.4. Scaffolding and gap filling *10*

2.5. Assembly quality evaluation *11*

3. Genome annotation **13**

3.1. Repeat annotation *13*

3.2. Gene annotation *14*

3.3. Functional annotation *15*

4. Reduced representation sequencing and population genomics **17**

4.1. Reduced representation sequencing *17*

4.2. Population genomics *19*

4.3. Summary statistics *20*

4.5. Non-parametric methods *21*

4.6. Parametric methods *21*

5. The genus *Salvia* **23**

5.1. Stamen architecture and taxonomy *23*

5.2. Salvia in Italy *26*

6. The MADS-box family of transcription factors **33**

7. *Althaea officinalis* **35**

8. Other species – *Campanula thyrsoides* and *Solanum* **37**

RESULTS AND DISCUSSION **38**

1. The *Salvia pratensis* genome **39**

1.1. Genome assembly *39*

1.2. Transcriptome sequencing *51*

1.3. Genome annotation *52*

1.4. Functional annotation *57*

1.5. Whole genome duplications and gene family analysis *58*

2. Flower morphology and MADS-box genes **61**

2.1. Preliminary observations of Salvia pratensis flowers *61*

2.2. MADS-box family annotation and expression *66*
3. Species interrelationships among Italian *Salvia* **72**
3.1. Sampling campaign *72*
3.2. GBS library preparation *72*
3.3. Sequencing data evaluation *77*
3.4. Phylogenetic relationships and ploidy of Italian Salvia *82*
3.5. Salvia pratensis and endemics *88*
4. *Althaea officinalis* **96**
4.1. Evaluation of phenotypic and genetic variability *96*
4.2. Evaluation of self-pollinating efficiency *98*
4.3. Genome sequencing *99*
4.4. Transcriptome sequencing *100*
4.5. Pool sequencing *100*
5. Other species **101**
Campanula thyrsoides subsp. carniolica *101*
Horticultural landraces *101*
CONCLUSIONS AND FUTURE PERSPECTIVES **104**
ACKNOWLEDGEMENTS **105**
MATERIALS AND METHODS **106**
Plant samples **106**
Salvia pratensis genome sequencing *106*
Genotyping-by-Sequencing *106*
Althaea officinalis *109*
Flower morphology and measurements **109**
Flow cytometry **109**
Nucleic acid extraction **110**
Whole genome sequencing *110*
Transcriptome sequencing *110*
Optimization of DNA extraction from silica-dried leaf for GBS *111*
GBS library preparation **114**
Sequencing, read quality evaluation and trimming **119**
k-mer analysis **120**
***Salvia pratensis* genome assembly and annotation** **120**
ONT assembly *120*
HiFi assembly *121*
Organelle genomes *121*
Genome annotation **122**
Repeats *122*

Genes *122*
Whole genome duplication, synteny and orthology analysis *123*
MADS-box factors **123**
Gene family analysis *123*
Qualitative transcriptional profiling *124*
Population genomics of *Salvia* **124**
Other reference genomes used **125**
Server details **125**
REFERENCES **126**
APPENDIX **143**

Abstract

Genomic resources are a fundamental asset in investigating genetic diversity, evolutionary history and adaptive potential of populations. During this doctorate, we aimed at developing genomic resources for non-model, medicinal plant species in order to facilitate research in biodiversity genomics, flower development, and genetic improvement, focusing on two plant species: *Salvia pratensis* and *Althaea officinalis*. In one part of the project, a good quality, annotated genome assembly was obtained for *Salvia pratensis*. The genome was used to study flower development related genes of the MADS-box family of transcription factors and as a reference to perform species delimitation between *S. pratensis* and three closely related taxa which are endemic to Italy or have an otherwise restricted range. A phylogeny for Italian *Salvia* species was also obtained. Genetic variation data was obtained with Genotyping-By-Sequencing (GBS). In the second part of the project, in collaboration with Aboca S.p.A. Società Agricola, we set the basis for future studies aimed at improving Aboca's collection of *Althaea officinalis*, a medicinal plant used for many natural products, to increase activity of its root extracts. Experimental fields were set up in Aboca and Milan to obtain plants for genome size estimation and to test the viability of self-pollinated seeds. Finally, a protocol was set up in Aboca to efficiently obtain root extracts from individual plants, in order to have a measure of the desired phenotype. This is essential to verify how much variation is present in the collection and in the future link it to the plant's genotype. The results from this project set the stage for improving the *ex situ* collection of the *Salvia* genus in the Botanical Garden of Brera, give support to delimitation of endemic Italian species, and set the stage for genetic improvement of Aboca's collection, potentially with integration of wild genotypes.

Aim of the project

The project aimed at developing genomic resources for non-model, medicinal plant species in order to advance the research in biodiversity genomics, flower development, and genetic improvement, focusing on two plant species: *Salvia pratensis* and *Althaea officinalis*. Side projects were carried out on other species: *Campanula thyrsoides*, *Solanum lycopersicum* and *Solanum tuberosum*.

The first study was the sequencing, assembly and annotation of the *S. pratensis* genome. Different bioinformatic approaches for genome assembly starting from long reads were applied and compared to select the best protocol, and the genome was annotated with functional elements employing transcriptomic data. The second study focused on the molecular mechanisms related to flower development and male sterility in *S. pratensis*. The genome was used to identify genes of the MADS-box transcription factor family and investigate their expression profiles during flower development. The third study focused on the study of genetic diversity within native species of *Salvia* in Italy, with the long-term goal of improving the *ex situ* collections of botanical gardens to better represent the biodiversity of the Italian peninsula. A wide collection of dried leaf samples from natural populations was collected thanks to a network of collaborators all throughout Italy, coordinated by the Botanical Garden of Brera. The collection was analysed with the Genotyping-By-Sequencing (GBS) method to obtain genome-wide genetic variation data at a reasonable cost from hundreds of samples. The DNA extraction methods and GBS protocol had to be optimized for *Salvia*. The dataset was analysed with population genomic and phylogenomic methods to investigate two main research questions: a) species delimitation and population structure in endemic *Salvia pratensis* taxa (*S. saccardiana, S. bertolonii, S. haematodes, S. ceratophylloides* within the *Salvia pratensis* group (Pignatti, 2017) sometimes considered subspecies of *S. pratensis,* sometimes independent species) ; b) genome-wide phylogeny of all Italian *Salvia*.

A part of the study in collaboration with Aboca SpA Società Agricola, a company producing natural products for healthcare, based in Aboca (AR), and Pistrino (PG), was focused on developing genetic resources of a plant of medicinal interest, *Althaea officinalis*. Root extracts from this plant have adhesive properties to human mucous membranes which are useful for producing cold syrups and other products from completely natural extracts. *A. officinalis* crops have not been domesticated and show diversity and heterogeneity in the amount and efficacy of the root extract. The aim of the study was therefore to identify genotypes with a higher adhesive activity to improve Aboca's production of root extracts with this activity, potentially by identifying markers to screen the most productive genotypes in crops and wild populations. First, we decided to screen Aboca crops for differences in root extract activity among genotype and test the heritability of the phenotype, and subsequently to characterize the most efficient individuals with genomic (pooled sequencing) and transcriptomic data (RNAseq). Aboca has no previous experience of any genome-wide study on plants.

Three side projects in collaboration with the University of Pavia concern the conservation of plant species and genetic resources. The aim was to investigate genetic diversity in 1) *Campanula thyrsoides* subsp. *carniolica* in Parco Nazionale delle Dolomiti Bellunesi, a monocarpic species with extremely small and declining populations, and 2) in local collections of historical and heirloom varieties of tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*) within the project "La RAVA e la FAVA" – ConseRvAzione, VAlorizzazione e caratterizzazione morFologicA e genetica di VArietà ortive "minori" locali" of Region Lombardy.

Introduction

1. Genomics for biodiversity and the role of botanical gardens

With the reduction in cost of sequencing technologies, the value that genomic resources can provide to all fields of plant research, even those historically less funded such as wild biodiversity studies and conservation, has become evident (Supple & Shapiro, 2018; Theissinger et al., 2023; Hogg, 2023). With rising extinction rates and habitat degradation, there has been a push to sequence as much of the Earth's biodiversity as soon as possible (Blaxter et al., 2022). Developing genomic resources for a species of interest can benefit research in a multitude of ways and several projects have been dedicated to sequencing eukaryotic biodiversity, from the Earth Biogenome Project (Lewin et al., 2018), to the Darwin Tree of Life (The Darwin Tree of Life Project Consortium, 2022), the Vertebrate Genome Project (Rhie et al., 2021), the 10K Plants (S. Cheng et al., 2018), the European Reference Genome Atlas (Cartney et al., 2024), and many more.

Botanical gardens, that were born for the cultivation of medicinal plants, had an important role in this research. Most high throughput sequencing techniques require a large amount of starting material, which usually refers to young, undamaged leaves, highlighting the value of having living plants in order to develop a genome reference or any large scale sequencing project requiring high quality DNA (Zuntini et al., 2024). As such, botanical gardens are playing and will continue to play a central role in plant genomics research. Botanical gardens are not only living museums, tourist destinations, and sites for educational activities, but they are research centers dedicated to biodiversity conservation and restoration (Botanic Gardens Conservation International, n.d.) with collections that extend beyond medicinal plants. Indeed, many important research projects are carried out by large Botanical Gardens all over the world, with major examples being the Kew Gardens and the Millennium Seed Bank. As botanical gardens transition to the genomic era, also the Botanical Garden of Brera has dedicated efforts into genomics for improving the biodiversity of its collections according to genetic concepts (Master theses: Pieri, 2019; Paleni, 2021).

Genetic diversity, meaning differences in DNA sequences between individuals, populations, or species, is one of the three components of biodiversity (Biodiversity Convention). Small population sizes resulting from habitat destruction, bottlenecks, or self-pollination in *ex situ* collections can be detrimental for the maintenance of population genetic diversity. Low genetic diversity can cause a variety of negative effects, from lower adaptive potential to changing environments, to inbreeding depression, to large-scale effects on the entire ecosystem and loss of ecosystem services (Stange et al., 2021). These factors highlight how important it is to assess the genetic diversity of a species or a population for any kind of biodiversity study (Bernatchez et al., 2024) and is fundamental for repopulation projects or *ex situ* collections.

Next Generation sequencing (NGS) opened new doors to the field of conservation genetics. Where genetic diversity was previously studied through mitochondrial, nuclear or plastid markers, or via targeted conserved

genes, population and species evolution can now be studied at unprecedented levels of resolution (Supple & Shapiro, 2018). Each gene tells a different story of evolution, and by investigating up to thousands of independent markers we can get a clearer picture of phylogenetic relationships between different species, populations or cultivars. Reference genomes are a great asset in biodiversity studies. Reference genomes serve as a reference to interpret sequencing data and can be used to map genetic variation within genes and regulatory elements. The comparison of multiple genomes from different species can help identify the major factors that play a role in evolution such as gene family expansion and contraction, genome duplication, and transposable element activity (Formenti, Theissinger, et al., 2022).

Plant genomic research was kickstarted by model organisms such as *Arabidopsis thaliana*, but was quickly applied to economically important crops, such as rice, wheat, maize, and tomato. Genetics and selective breeding have historically been used to obtain more productive plants or plants with desired traits. Nowadays, genomics plays a role in improving crop resistance to climate changes or pathogens (Pourkheirandish et al., 2020). Several genomic approaches are routinely applied to crops: for example, the study of genes responsible for stress response or higher yields can identify target genes for marker-assisted selection (MAS); genome editing of selected genes can accelerate artificial selection of plants with desirable traits; the traits of desirable phenotypes can be linked to genetic markers with Genome Wide Association Studies (GWAS). As a result of selective breeding, domesticated crops tend to be genetically uniform. For example, in cultivated tomato, only a few genetic variants are responsible for the wide range of shapes, flavours, and colours of the fruit (Blanca et al., 2021). However, as previously stated, low genetic diversity has detrimental effect on the adaptive potential of plants, exposing crops to the risk of being wiped out by pathogens: grape, potato and bananas were affected by these issues historically. One way to increase genetic diversity of crop species could be to introduce diversity from Crop Wild Relatives (CWR), meaning, relatives of domesticated species which still live in the wild. These plants have higher levels of genetic diversity and are generally more tolerant to stresses and pathogens (Pourkheirandish et al., 2020).

2. Reference genome generation

While the first sequenced genomes were the result of monumental effort and collaboration between multiple institutes, it is now possible to obtain a reference sequence for an organism of interest at a relatively achievable price for many research groups. A reference genome is a representation, as complete as possible, of the genomic DNA of an organism. The process of reconstructing this sequence from High-Throughput sequencing data (**Figure 1**) is called **genome assembly**. Long read sequencing technologies are instrumental in all genome assemblies (Rhie et al., 2021), but especially in plant genomes. Plant genomes present a very high percentage of repetitive sequences, following whole genome duplication (WGD) events, explosion of transposable elements, and polyploidy (Heslop-Harrison et al., 2022; Wendel et al., 2016). Plants in general are more tolerant to large-scale genomic modifications than animals are, resulting in more complex genomes

and inheritance patterns. The range of genome sizes, chromosome numbers, and ploidy in plants is extremely large, with genomes going from 12.56 Mbp (millions of base pairs) (Pellicer & Leitch, 2020) to 160.45 Gbp (billions of base pairs) (Fernández et al., 2024) and several levels of ploidy present even within the same species or in intercrossing species complexes. As such, having cytological data (chromosome number and ploidy) and genome size estimation from wet lab experiments prior to genome sequencing greatly helps in defining the methods and amount of sequencing necessary as well as evaluate critically the assembly results.

Currently, the standard methods for genome assembly include the following steps, ideally performed on high quality genomic DNA from the same individual.

1. **K-mer spectrum analysis** with short read sequencing.
2. **Assembly** from long read sequencing, either PacBio HiFi or Oxford Nanopore technologies (ONT).
3. **Scaffolding**, for example using optical maps, ultra-long ONT reads, or HiC.
4. **Assembly quality evaluation**.

The main points of each of these steps will now be briefly discussed. See **Figure 2** for an overview of three of the most commonly employed sequencing technologies (Illumina short read sequencing, Oxford Nanopore Technologies long read sequencing, PacBio HiFi sequencing) and **Table 1** for a list of software tools that are employed for all steps illustrated.

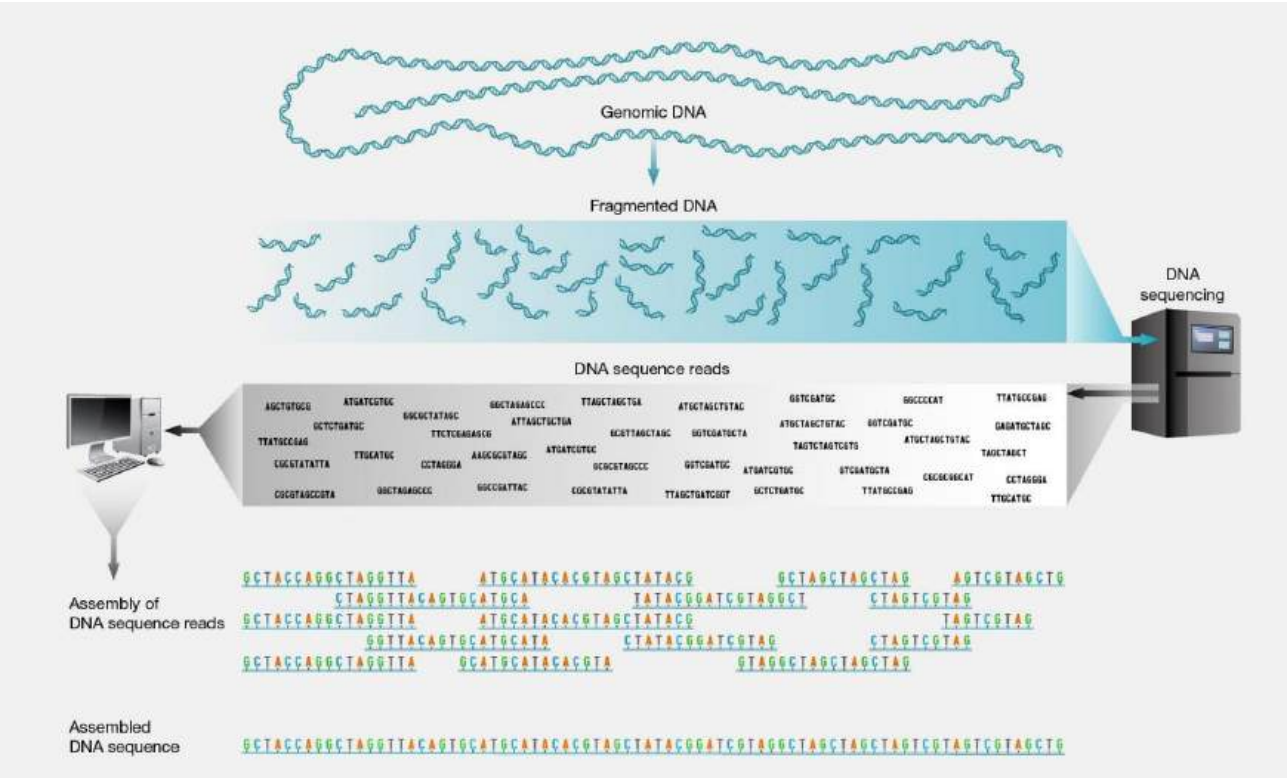


Figure 1. Overview of the shotgun sequencing approach employed by Second and Third generation sequencing. The genomic DNA is fragmented into smaller molecules and the sequence of those molecules is determined. The overall genomic sequence is then reconstructed computationally either de novo or by comparison with a known genome sequence (=mapping).
Image credits: https://www.genome.gov/genetics-glossary/Shotgun-Sequencing

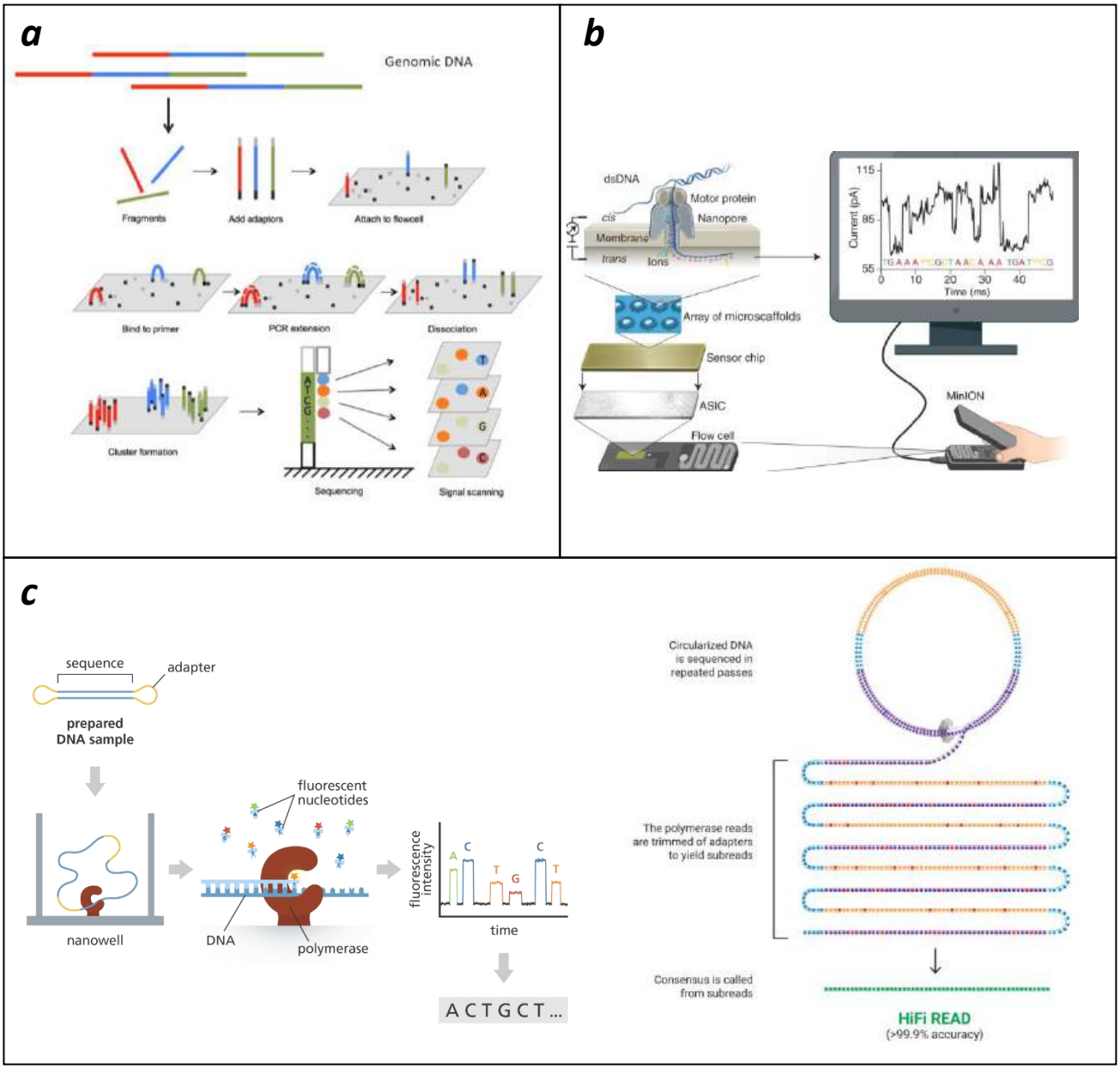


Figure 2. Overview of most used sequencing technologies. ***a)*** Illumina Sequencing By Synthesis (SBS). The fragments from genomic DNA are hybridized to a chip where bridge amplification occurs to create many copies of the same fragments. The sequence of the fragment is determined thanks to the use of DNA polymerase using fluorescent nucleotides. The fluorescent signal emitted by each cluster of molecules at each step is recorded and converted to sequence data. Illumina reads range from 50 to 300 bp. ***b)*** Oxford Nanopore Technologies (ONT) sequencing. Native DNA or RNA molecules pass through a pore protein. A difference in potential is applied to the two sides of the membrane and the current flowing through the pore is affected by the sequence of the DNA passing through it. Electric signal is decoded and converted to sequence data. ONT reads can reach up to tens or hundreds of megabases in length. ***c)*** PacBio HiFi sequencing. The DNA fragments are circularized and SBS is carried out in nanowells by one molecule of polymerase on one molecule of DNA at a time. The same fragment is sequenced tens of times thanks to the circularization and a consensus is called from all the reads, allowing a very high accuracy. HiFi read length ranges in the tens of kilobases.
Image credits:
https://doi.org/10.1038/s41587-021-01108-x
https://doi.org/10.5772/61657
https://www.yourgenome.org/theme/what-is-pacbio-smrt-sequencing/
https://www.pacb.com/blog/understanding-accuracy-in-dna-sequencing/

2.1. K-mer spectrum analysis

Several properties of the genome can be investigated prior to assembly through k-mer spectrum analysis of the unassembled, usually short-read set. K-mers are all subsequence of length k of a longer biological sequence (in this case, DNA). The k-mer spectrum is a histogram that represents the frequency of k-mers in the read set and it will have a peak corresponding to the average frequency of a K-mer, which depends on the sequencing depth, as well as other peaks due to sequencing error, heterozygosity and repeats (**Figure 3**) (Jenike et al., 2024). Following k-mer counting, several tools fit a statistical model on this spectrum to estimate several properties such as the genome size, the rate of heterozygosity, the percentage of repetitive sequences and the ploidy.

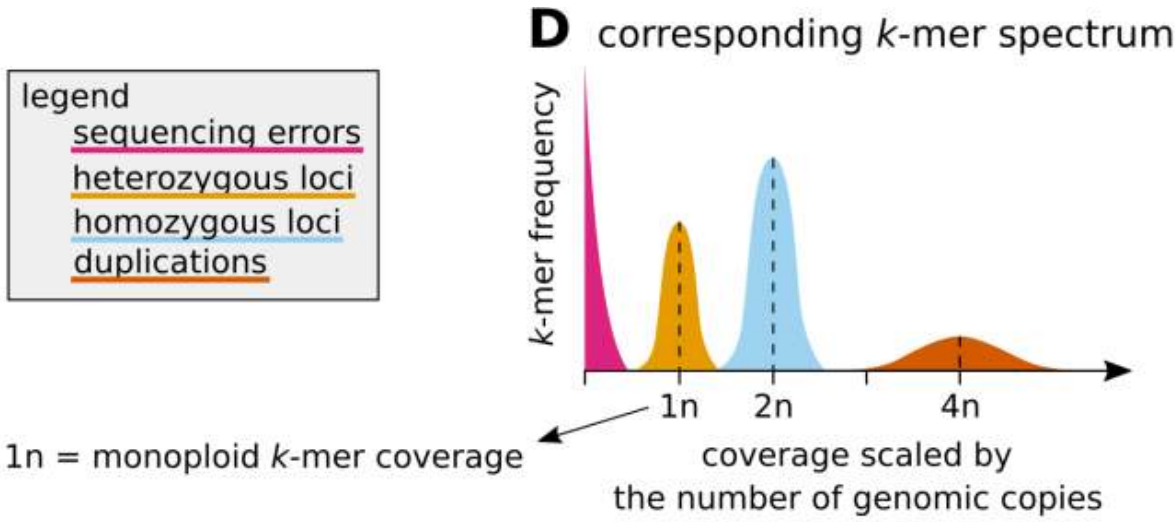


Figure 3. K-mer spectrum example. 1n: haploid peak; 2n: diploid peak; 4n: repetitive peak. Colors in the histogram correspond to the different origins of k-mers (legend on the left).
Image credits:10.48550/ARXIV.2404.01519

2.2. Genome assembly

Genome assembly is the process of reconstructing the whole genomic sequence of an individual starting from the sequencing reads. Almost like solving a puzzle, we are trying to reconstruct the bigger picture starting from small fragments. Most assemblers attempt to represent the maternal and paternal copies of each chromosome as a singular sequence. These are called collapsed or pseudo-haploid assemblies.

There are two main family of methods for genome assembly: De-Brujin Graphs (DBG) and Overlap-Layout-Consensus (OLC) (Z. Li et al., 2012). Both methods are graph-based; the sequence data is arranged into a graph according to overlapping sequences between reads (OLC) or consecutive k-mers (DBG). DBGs were generally faster, less computationally expensive and recommended for short read data, while OLC was used for long reads with high error rates. An optimal path is then found for traversing this graph, which is a representation of the original genomic sequence.

In the past, genome assemblies were routinely obtained from either exclusively short read sets or from hybrid approaches combining long and short reads. These methods generally allowed for assemblies either at the contig or scaffold level. Nowadays, most eukaryote genome sequencing projects rely mainly on long reads and usually employ short reads for polishing, evaluation, or scaffolding (HiC). Long reads are almost essential for reconstructing plant genomes, given the high frequency of repeated sequences, and most assemblers are based on evolutions of the OLC algorithm.

Traversing the graph is complicated by the presence of errors in the read set, repetitive sequences or heterozygous sequences (Miller et al., 2010). The problem of error rates is solved either by decomposing the reads in k-mers, which allows excluding low frequency k-mers from the graph, or performing read correction prior to assembly, which is instrumental in the case of long, high error rate reads such as ONT. Repeats can only be solved by having reads longer than the repeated sequences or long-range positional information, such as the information coming from paired-end or mate paired short reads. High rates of heterozygosity, which can be expected when sequencing a non-model, non-domesticated plant organism, complicate assembly, as they may lead to assembling reads from the two haplotypes separately and thus artificially increasing the rates of repetitive sequences (Guiglielmoni et al., 2021) (**Figure 4**). To solve this issue, haplotype purging approaches were developed, which aim at discarding duplicated sequences. However, the haploid assembly is still, in most cases, a mix of haplotypes coming from both parents (haplotype switching) (X. Zhang et al., 2020). Both repeats and heterozygosity may be solved by the assembler by breaking up the sequence into smaller sequences (**Figure 4**).

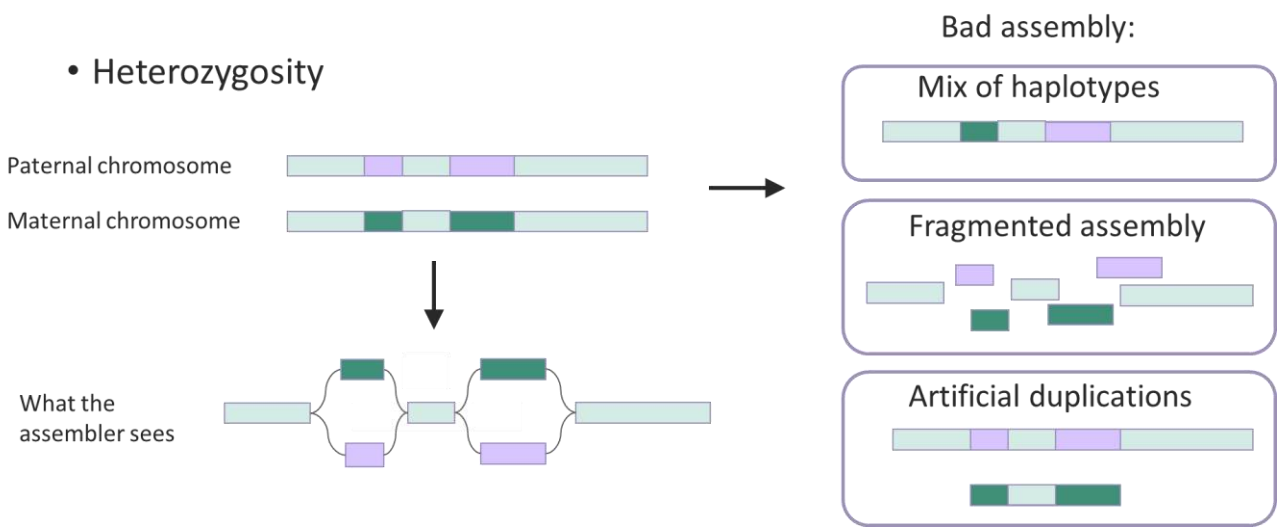


Figure 4. Example of issues in assembling highly heterozygous genomes. Left: the presence of homozygous (light green) and heterozygous (violet or dark green) regions in a genomic sequence causes the development of "bubbles" in the assembly graph, where the same homozygous region is linked to both heterozygous regions. Right: To solve these structures, in absence of long-range information or information on which haplotypes are maternal and which maternal, the assembler may either choose randomly which haplotypes to keep (top), fail to solve the structure and split the sequence in different shorter contigs (center), or wrongly duplicate the homozygous regions (bottom).

2.3. Polishing

Assembly can be followed by polishing, in which accurate, generally short reads are mapped on the assembly generated from error-prone long reads and a consensus is called to correct any differences between the read set and the assembly. ONT reads from older instruments tend to have high error rates and perform poorly in homopolymer repeats and correction with Illumina data can highly increase the accuracy of the assembly by correcting single nucleotide errors and small indels. However, correcting a pseudohaploid assembly implies that reads coming from both copies of a homologous chromosome will be mapped on the same sequence. As such, the result of polishing may switch between different variants coming from the two different heterozygous sequences and cause haplotype switching and chimeras, meaning, sequences that are not present in the original read set and consequently in the genome.

2.4. Scaffolding and gap filling

Each contiguous sequence resulting from a genome assembly process is called a contig. When starting assembly, it is expected that the number of contigs will be much higher than the actual chromosome number of the individual. Long range information is needed to organize the contigs into longer scaffolds or chromosomes. However, excessive fragmentation is a sign of a poor-quality assembly. With either sequencing data from the parents or chromosomal scaffolding data many assemblers are now able to produce fully phased, chromosome-level assemblies, which are a closer representation of the real DNA sequence present in the nucleus. Nevertheless, while chromosome-level assemblies are now achievable, full phasing requires higher sequencing depth.

Scaffolding refers to orienting and organizing the contigs in larger sequences called scaffolds. In a scaffold, approximate distance and orientation between contigs can be determined, but not the exact sequence between them (J. Luo et al., 2021). This can be due for example to the presence of regions that are either not accurately sequenced by the sequencing technology or not assembled by the assembler. Uncalled bases are represented with “N” characters in the sequence. After scaffolds have been determined, it is possible to try and utilize the sequencing data, for example, by remapping the reads on the scaffolds, to fill some of those uncalled bases (gap filling). Scaffolding can be achieved with long-distance information, such as what can be obtained by paired-end or mate-paired short reads, long or ultra-long reads (>100kb), proximity ligation approaches (HiC), or genetic and optical mapping. Scaffolding can also be achieved by comparison with a reference genome of the same or extremely closely related species. In particular, HiC is an instrumental technique in reaching fully phased assemblies, especially in polyploids, and HiC information can be used directly during assembly, for example, by Hifiasm. However, for many projects in which the genome is not the end goal, a pseudohaploid assembly with a low level of fragmentation can still be adequate.

SOFTWARE	REFERENCE	NOTES
k-mer counting		
Jellyfish	(Marçais & Kingsford, 2011)	
Meryl	(Rhie et al., 2020)	
k-mer spectrum analysis		
GenomeScope	(Ranallo-Benavidez et al., 2020)	
FindGSE	(H. Sun et al., 2018)	
SmudgePlot	(Ranallo-Benavidez et al., 2020)	Ploidy estimation
Short read / hybrid assembly		
SpaDES	(Bankevich et al., 2012)	
SOAPdenovo2	(R. Luo et al., 2012)	
MaSurCa	(Zimin et al., 2013)	
Unicycler	(Wick et al., 2017)	
Long read assembly		
Hifiasm	(H. Cheng et al., 2021)	HiFi only; Haplotype phasing with HiC
Canu / HiCanu /TrioCanu	(Koren et al., 2017); (Koren et al., 2018)	HiFi/ONT/Haplotype phasing with parental data
Flye	(Kolmogorov et al., 2019)	
Shasta	(Shafin et al., 2020)	
Polishing		
Polca	(Zimin & Salzberg, 2020)	
Pilon	(Walker et al., 2014),	
Medaka	(Oxford Nanopore Technologies Ltd., 2017/2018)	
NextPolish	(Hu et al., 2020).	
Haplotype purging		
Purge_haplotigs	(Roach et al., 2018)	
Purge_dups	(Guan et al., 2020)	
Scaffolding		
SSPACE	(Boetzer et al., 2011)	PE / Long reads
SCOP	(M. Li et al., 2019)	PE
LINKS	(Warren et al., 2015;	Long reads
SLR	Luo et al., 2019)	Long reads
ALLMAPS	(Tang et al., 2015)	Genetic mapping
OMGS	(W. Pan et al., 2020)	Optical mapping
RagTag	(Alonge et al., 2022)	Reference-based
Salsa	(Ghurye et al., 2019),	HiC
YaHS	(Zhou et al., 2023),	HiC
3D-DNA	(Dudchenko et al., 2017)	HiC

Table 1. List of some of the available software tools for all steps of genome assembly, split by the passage in the assembly pipeline they can be used for. Software: software name. Reference: paper on its development. Notes: additional notes on its use or the data type it employs.

2.5. Assembly quality evaluation

Ultimately, a genome assembly is a simplified model of the real genetic makeup of an organism. While we may keep getting closer and closer to it, there is not ground truth we can compare to. Moreover, genome assembly is computationally expensive, and most software are heuristic, meaning, the complete space of possibilities is not explored, and the results depend closely on the parameters chosen for execution. During a genome assembly project, several different assembly software tools and parameters are tested and the

results of each are compared with several metrics (Angel et al., 2018, p. 20). These metrics can roughly be split in three categories: contiguity, completeness and accuracy. Additionally, it is crucial to remove from the assembly potential contaminant sequences that may have been present in the read set. See **Table 2** for a brief explanation of the most used quality evaluation metrics.

METRIC	NOTES/SOFTWARE
Contiguity	Quast (Gurevich et al., 2013), FastaSeqStats (Bombarely, n.d.)
# contigs	Number of sequences in the assembly
Largest contig	Length of longest fragment
L50	Minimum number of contigs whose length sums up to 50% of the assembly size
N50	Length of L50th contig when sorted longest to shortest
Completeness	
Total length	Assembly length (should be close to the haploid genome size, unless fully phased). QUAST, FastaSeqStats
BUSCO or OMARK completeness %	Presence of highly conserved, single-copy genes (close to 100% is better). BUSCO (Manni et al., 2021), OMArk (Nevers et al., 2024)
k-mer complenetess %	Compare assembly k-mer spectrum to short-read library k-mer spectrum (close to 100% is generally better, unless highly heterozygous). Merqury (Rhie et al., 2020), Merfin (Formenti, Rhie, et al., 2022)
Read remapping	Percentage of reads that map on the assembly; percentage of assembly with non-zero coverage. Any mapping tool + Bedtools (Quinlan, 2014)
Duplication %	Presence of duplicated sequenced in the assembly. Can be estimated with gene sequences (BUSCO, OMArk) or k-mer spectrum analysis (Merqury)
Accuracy	
QV	Accuracy on the single base, computed as -10 $\log_{10}$(error rate). Can be computed with reference-free k-mer spectrum analysis by counting assembly-only k-mers. Merqury (Rhie et al., 2020), Merfin (Formenti, Rhie, et al., 2022)
# misassemblies, # mismatches	Large or small differences in sequence between the assembly and a reference genome (not applicable for *de novo*). QUAST; any mapping + variant calling tool

Table 2. Metrics for genome assembly evaluation (left) and short explanation (right) along with some of the software tools employed for their estimation.

Contiguity refers to the level of fragmentation of the assembly. An ideal reference genome contains a number of sequences equal to the number of chromosomes. Generally, the number of contiguous sequences (contigs) is much higher, and the length of each individual sequence can range from a few kb to tens of Mb, depending on the sequencing techniques and the assembler. Several metrics were determined to evaluate contiguity (Angel et al., 2018); see table Contiguity.

Completeness refers to how much of the genome is represented in the assembly. See table Completeness for a list of methods to assess completeness. The simplest approaches rely on comparing the assembly size to the expected genome size or check the mapping rate of sequencing reads on the assembly. Another family of methods instead relies on scanning the genome sequence for conserved genes among organisms of the same lineage in the expected number of copies. A fourth family of approaches is reference-free k-mer spectrum analysis. The assembly can be split into k-mers and for each unique k-mer the number of

appearances in the assembly (multiplicity) is calculated. This number is then compared to the frequency of the k-mer in a high quality, usually short-read set from the same individual.

Accuracy is measured with the consensus quality value (QV), which is a log-scaled probability of error (QV=-10log10(error rate)). When a reference genome is available, the QV of an assembly can be computed by comparing the two assemblies and counting the mismatches to then compute the probability of error. When a reference is not available, the QV may be computed with a reference-free k-mer based approach. A QV of 20 corresponds to one error every 100 bp, while a QV of 50 to one error every 100,000 bp. For the 2022 Telomere-To-Telomere (T2T) assembly of the human genome, a staggering QV74 was achieved by using a combination of Merfin and other tools (Fang & Wang, 2022).

The presence of **contaminant** DNA in the sequenced samples can add exogenous sequences in the assembly or decrease the mapping rate, as a certain percentage of reads will belong to the contaminant and thus not map on the assembly. To avoid this issue, decontamination can be performed prior to assembly with tools like Decontaminer (Sangiovanni et al., 2019) and Kraken (Wood & Salzberg, 2014), or after assembly with tools like Blobtools (Challis et al., 2020).

3. Genome annotation

Genome assembly is usually followed by genome **annotation**, the identification of repetitive and functional elements on the genome, such as protein-coding genes, untranslated regions of transcripts, transcription start sites, start and stop codons, predicted CDSs, transposable elements, tRNAs, non-coding RNAs. Genome annotation is usually conducted by first performing repeat annotation and then “hiding” (masking) the repetitive sequences from the genome to perform gene annotation (Angel et al., 2018), as repetitive sequences complicate sequence alignment. After gene annotation is performed, a putative function is assigned to protein-coding genes; this step is called functional annotation.

3.1. Repeat annotation

Repetitive sequences can represent a large portion of a plant’s genome, varying by lineage from 15% in Algae to >80% in Gymnosperms (X. Luo et al., 2022). Studying the occurrence and sequence divergence of repeats can help uncover the dynamics of genome evolution and genome size variation. Repeats can be split in two categories according to their positioning: tandem and interspersed repeats (**Figure 5**). Tandem repeats are made up of a basic unit, generally made of short sequences, repeated multiple times consecutively, and they make up a large portion of centromeres and telomeres. (Liao et al., 2023). Interspersed repeats originate from the proliferation of transposable elements (TEs), which are DNA elements with the ability to move and copy themselves into different parts of a host genome. In plants, retrotransposons of the LTR family make up on average 21% of the genome, with a peak of 50% in Gymnosperms; thus, they seem to be the major drivers for the massive genome sizes in this clade (X. Luo et al., 2022). By contrast, in humans they only make

up 9% of the genome, and the LINE and SINE family of retrotransposons make up 34% of the genome. The identification and masking of repeat sequences from a genome is generally a two-step process in which first a custom set of repetitive sequences is modelled from the genomic sequence of interest (i.e. with RepeatModeler (Flynn et al., 2020)) and then the families are used to identify individual instances of said repeats in the genome (i.e. with RepeatMasker (Smit et al., n.d.)). The repeats are replaced either by N characters (hard masking) or lowercase characters (soft masking). Another approach is to use a pre-made library of repeats for the species or lineage of interest, for example, the RepBase (Bao et al., 2015) or Dfam (Hubley et al., 2016) databases. The repeats are then classified into families with sequence similarity to a reference database. The presence and integrity of LTR repeats, which have a very conserved structure, can be used to evaluate the completeness of a genome assembly by computing the LTR Annotation Index (LAI) (Ou et al., 2018).

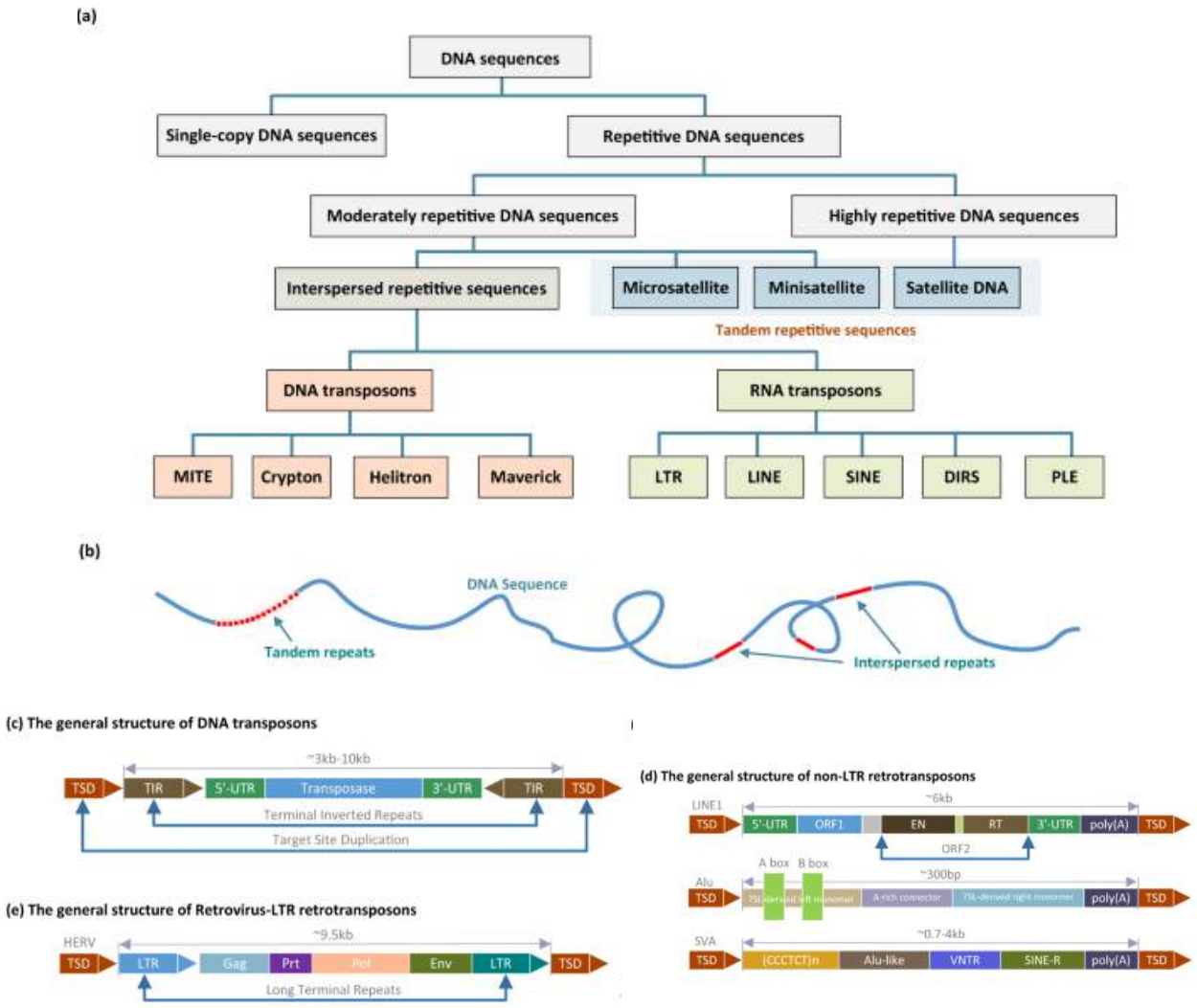


Figure 5. Classification of repetitive elements and structure of interspersed repeats (transposons). Different families of TEs are more prevalent in different clades.
Image credits: https://doi.org/10.1038/s42003-023-05322-y

3.2. Gene annotation

There are two families of approaches to annotation of protein-coding genes: *ab initio* and evidence based (Yandell & Ence, 2012). *Ab initio* gene annotation is based on developing a model that will predict the position of coding sequences exclusively based on the genomic sequence, generally employing Hidden Markov Models (HMM). Evidence-based annotation exploits various types of evidence of transcription, such as expressed sequence tags (ESTs), transcriptomic sequencing (RNAseq), gene models of a closely related species and alignments of protein sequences (Angel et al., 2018). As the expressed gene fraction varies depending on tissue type and developmental stage, it is best to get transcriptomic evidence from many different sample types. The best type of evidence would generally be full-length total RNA sequencing (ONT or IsoSeq), but short read sequencing is more common and cheaper. When the main objective is retrieval of protein-coding genes, selecting mRNA specifically can help in lowering the sequencing amount; moreover, reducing the amount of rRNA, which is a significant portion of the total RNA in a cell, is essential. Several library preparation protocols are available, such as poly-A enrichment. Still, lowly expressed or extremely cell-type specific transcripts ca be hard to catch; as such, the most popular pipelines for gene annotation combine one or more *ab initio* prediction models to evidence-based annotation. RNAseq evidence can be mapped on the genome with a splice-aware aligner or assembled *de novo* in transcript models and can be used as training data for an *ab initio* model.

Refinement tools can be used to compare gene models to the available evidence, compare and combine different annotation, and increase the quality of the predictions. The completeness and quality of a gene annotation can be evaluated with BUSCO, by comparison of basic statistics to closely related organisms (i.e. number of transcripts, gene length, number of exons), by comparing the gene models to the evidence (AED), and by evaluating the functional annotation results. See **Table 3** for a list of genome annotation software.

3.3. Functional annotation

Functional annotation pairs gene models to a putative function, generally based on homology to known proteins with an annotated function. EggNOGmapper (Cantalapiedra et al., 2021), Blast2GO (Conesa et al., 2005), AHRD (Boecker, 2021) can be used to perform homology-based functional annotation. However, the concept that function necessarily follows from sequence has been challenged multiple times. Protein structure is not always directly correlated to the protein sequence. Moreover, a significant portion of proteins are intrinsically disordered, meaning, they do not present a stable structure that we were able to observe experimentally or predict computationally. As a result, even in the widely studied model plant *Arabidopsis thaliana*, only 80% of the genes have a recorded putative function. Alternative methods which exploit deep learning or language models, such as Fantasia (Martínez-Redondo et al., 2024), in order to surpass the limits of methods based on sequence similarity have recently been proposed and are starting to be used.

SOFTWARE	REFERENCE	NOTES
Evidence alignment		
BLAST	(Altschul et al., 1990)	Find matches and align protein and transcript evidence
Miniprot	(H. Li, 2023)	Align protein evidence
StringTie	(M. Pertea et al., 2015; Kovaka et al., 2019)	Transcript assembly from RNAseq
Transdecoder	(B. Haas, n.d.)	CDS predition
Trinity	(Grabherr et al., 2011)	Transcript assembly from RNAseq
Exonerate	(Slater & Birney, 2005)	Aligns protein and transcript evidence to a genome
Gene predictors		
Augustus	(Stanke & Waack, 2003)	
SNAP	(Korf, 2004)	
FGENESH	(Salamov & Solovyev, 2000)	
GeneMark	(Lukashin, 1998; Brůna et al., 2020; Bruna et al., 2023)	
Combined pipelines		
PASA	(B. J. Haas, 2003)	
MAKER	(Cantarel et al., 2008; Campbell et al., 2014)	Computes Annotation Edit Distance (AED) for gene model quality evaluation
Ensembl	(Curwen et al., 2004; EMBL-EBI, 2024)	
BRAKER3	(Gabriel et al., 2023)	
GeSeq	(Tillich et al., 2017)	Annotates organellar genomes; combines several predictors and visualization tools
Refine/evaluate annotation		
EVidenceModeler	(B. J. Haas et al., 2008)	
PASA	(B. J. Haas, 2003)	
Ingenannot	(*BIOGER / Ingenannot · GitLab*, 2024)	Computes protein and transcript AED for gene model quality evaluation
TSEBRA	(Gabriel et al., 2021)	Transcript selector for BRAKER
Non-coding RNA		
tRNAscan-SE	(Chan et al., 2021)	tRNA prediction
RNAmmer	(Lagesen et al., 2007)	rRNA prediction
FEELnc	(Wucher et al., 2017)	Long non coding RNA

Table 3. List of some of the available tools for coding and non-coding gene annotation, divided by the step of the genome annotation pipeline they can be used in. Software: name of the tool. Reference: paper on its development or application. Notes: additional information on its use or the type of data it employs.

4. Reduced representation sequencing and population genomics

4.1. Reduced representation sequencing

Conservationists have made large use of genetic methods by investigating a small number of neutral markers, such as the number of repetitive units of one or more microsatellites, to investigate the phylogenetic relationships, genetic diversity and population structure of species of interest (Supple and Shapiro, 2018). The decrease in cost of Next and Third Generation Sequencing approaches has made genome sequencing more accessible. With DNA barcoding, short sequences from specific genes or intergenic sequences (mitochondrial markers for animals, plastidial and selected nuclear genes for plants) are sequenced; these sequences are conserved enough that they can be amplified from the template DNA with the same set of primers at large evolutionary scale, but divergent enough to allow species identification and phylogenetic inference (Kress & Erickson, 2008). When investigating individuals of the same species, more resolution is required. It is possible to genotype an individual by completely sequencing its genome and comparing it to a reference. This allows identification of the most complete set of genetic variants between individuals. The most used genetic markers are Single Nucleotide Polymorphisms (SNPs), which are a change in the sequence of a singular nucleotide. SNPs can generally be identified by sequencing with short reads of at least 30x coverage, mapping the reads on the reference genome, and identifying ("calling") variants between the sequenced individual and the reference. However, when the interest is studying different populations of the same or different species under a population genetics framework, it is generally necessary to genotype tens or even hundreds of individuals, making whole genome sequencing prohibitive because of the cost and computational resources required (Fuentes-Pardo & Ruzzante, 2017). Several methods were developed to reduce cost while still obtaining more information than what is possible with traditional genetic markers. These approaches are called Reduced Representation Sequencing (RRS), and their main characteristic is that they focus on sequencing a selected portion of the genome, so that high coverage can be obtained on a reduced number of genomic loci which can be accurately genotyped (Davey et al., 2011). The way the portion of the genome is selected can vary from use of restriction enzymes (RE) to predetermined probes designed to capture conserved genes. In RE-based methods such as RADseq and GBS the genomic DNA is digested with restriction enzymes and sequencing is performed only in proximity of the restriction sites (Baird et al., 2008; Elshire et al., 2011; Ziarsolo et al., 2021). RRS has limitations compared to WGS, but it has a better resolution and provides a more complete set of information than traditional genetic approaches (Scheben et al., 2021). Other methods to reduce cost may be used; a set of genotyping approaches for non-model species is summarised in **Table 4** (from Deglialberti, M., Paleni, C., Fainelli. F., Lambertini, C., Urpì, S., submitted).

Method	Pros	Cons
Reduced representation sequencing (RRS) (RADseq, Baird *et al.*, 2008; GBS, Elshire *et al.*, 2011; K-seq, Ziarsolo *et al*,. 2021)	+ Requires less sequencing data than WGS + Accurately identifies thousands of molecular markers + Allows genotyping of large populations + Reference genome is not strictly necessary	– Reduced set of variants – Missing data – Dependent on digestion efficiency (RE-based methods) – Data analysis is easier with a reference genome
Pool-seq Pooled sequencing of groups of individuals to detect genetic differences among groups (Schlötterer *et al*., 2014)	+ Accurately identifies variants in the whole genome + Allows genotyping large populations or organisms that are too small for DNA extraction or groups of individuals with different phenotypic traits + Allows accurate allele frequency estimations in sequenced pools	– Loss of information on the genotype of the single individual – Unequal DNA amount in the sequencing pool can bias results – Requires alignment to a reference genome – Can miss rare variants
Target enrichment Sequencing of conserved target regions across taxa detected with hybridization probes (Johnson *et al.*, 2019 for Angiosperms)	+ Requires low amount of sequencing data + Accurately identifies hundreds of molecular markers + Can be applied to highly divergent species + Does not require a reference genome + Can be applied to damaged DNA	– Can only capture highly conserved sequences – Lower amount of markers than RRS – Subject to selection bias
Transcriptome sequencing (or RNA-seq) Total mRNA is transcribed and sequenced (Theissinger *et al.*, 2023)	+ Accurately identifies variants on transcribed/coding regions + Requires lower amount of sequencing data + Does not require a reference genome	– Bias in variants detected – Potential challenging RNA isolation – Can only be applied to fresh samples (flash frozen, RNAlater) – Transcriptome is highly variable depending on tissue types and environmental conditions
Genome skimming Sequencing the genome at low coverage to recover high copy regions (mitochondria, plastids, rDNA…) (Straub *et al.*, 2012)	+ Cheap and low amount of data required + Recovers organellar genomes + Does not need a reference genome + Can be used to design primers for nuclear genes + Can be applied to damaged DNA (i.e. herbarium specimens)	– Cannot be used to genotype nuclear DNA markers – Highly repetitive regions are difficult to handle computationally

Table 4. Summary of some methods for reduced representation sequencing. Method: name of the method / family of methods, short explanation and references (either for its development or reviews of its applications). Pros: Best use cases and positive aspects of the methods. Cons: drawbacks or limitations of the methods.

4.2. Population genomics

The aim of population genetics is to investigate genetic differences within and among populations according to evolutionary theory. Several factors impact the genetic differences between species or populations of the same species: natural selection, mutation, genetic drift, gene flow or migration. Reproductive or geographical isolation, for example, will lead to non-random mating in a population and will cause differences in allelic frequencies in different groups of individuals. This is referred to as population structure. Knowing the structure of a population helps clarify the evolutionary history of populations or species and is a prerequisite for studying genotype-phenotype associations. Genetic information can be used to infer the demographic history of a population, such as the presence of bottlenecks at some point in their history.

As the number and density of genetic markers increased, thanks to high-throughput sequencing, population genetics is transitioning into population genomics. Different genes may have experienced different rates and modes of evolution, thus increasing the number of markers and computing a consensus of the different models increases the accuracy of our model; however, computational complexity also rises. Moreover, failing to sample copies of a same gene in some taxa may lead to incorrect tree topologies (incomplete lineage sorting, ILS). Linkage disequilibrium (LD) is the non-random association of alleles in different loci. This is either due to the loci being at a short distance from each other on the chromosome or having some functional link to a shared phenotype under selection. Some methods will obtain better results if the markers are not in linkage disequilibrium, as the information they provide is redundant, and so filtering ("pruning") based on the genetic distance or correlation may be applied beforehand. On the other hand, LD can also be used to detect admixture between populations, natural selection, or markers associated to certain phenotypes.

Introduced in the 1980s, the coalescent model described the genealogical history of DNA sequences and traces back their history in a phylogenetic tree (Tajima, 1983). With this model, the times of divergence between different branches of a phylogenetic tree and the length of the branches can be estimated, as well as other events in the demographic history. The coalescent was later expanded for a multi-species and multi-locus case (Multi Species Coalescent, MSC) (Jiao et al., 2021) and the MSC is at the basis of most of the models used today.

Species delimitation is a subset of population genetics aimed at recognizing whether different taxonomical units belong to the same species. The definition of species is a fundamental problem in biology, which is nowadays addressed under multiple points of view, including genetics (De Queiroz, 2007). Concordance on a split between two populations in multiple loci can be considered a sign of reduced gene flow, or complete genetic isolation between them; at that point, this can support the delimitation of two different species (Z. Yang & Rannala, 2010). Determining precisely what a species is has effects on conservation policies, which generally are not aimed at conserving subspecies or populations, but rather prioritize the survival of endangered species. Species delimitation generally consists in applying population genetics methods to

determine whether there is genetic separation or reduced gene flow between different populations and integrating this data with morphological and phenological observations or hybridization tests. Ultimately, as different definition of species exist, different interpretations may be given to the same set of observations; moreover, the putative mode of speciation, which is something that computational models cannot take into account, should be considered as a factor to explain the patterns of genetic diversity distribution (i.e.: are the populations in sympatry? Is there a geographic barrier? Is there a phenological barrier such as flowering time discordance?). Applying different and complementary analysis types can counteract their individual weaknesses and help elucidate complex situations, especially in situations with high degree of admixture (see **Table 6** for a list of some of the tools and approaches).

4.3. Summary statistics

A set of summary statistics can be helpful in inferring the demographic history of the population. One of the effects of non-random mating is a change in the percentage of heterozygosity. According to Hardy-Weinberg equilibrium (HWE), the rate of homozygous and heterozygous individuals under random mating can be determined from the allele frequencies, that can be different in different populations because of selection or genetic drift. If individuals tend to mate preferentially inside subpopulations, the frequency of heterozygosity in the subpopulation should be lower than what would be expected under HWE for the whole population. This departure from HWE can be a signature for ongoing evolution and it can be quantified with fixation indexes (F). Several estimators for various F statistics (Wright, 1949) have been developed. At the population level, F_{IS} is a measure of inbreeding, while F_{ST}, which is the most used metric, compares the expected heterozygosity of subpopulations to the expected heterozygosity of the total population. F_{ST} is a measure of the degree of differentiation between subpopulations, or alternatively, the proportion of genetic divergence between individuals due to the division in subpopulations. The lower the F_{ST}, the higher the gene flow between (sub)populations (**Table 5**). The F_{ST} can be estimated also for any genomic locus and can reveal whether the locus is subject to selection (Lewontin & Krakauer, 1973). Several estimators for computing F_{ST} take into account the presence of multiple loci, the effect of random sampling of individuals and loci, and other factors that may cause bias in the computation. However, the meaning and interpretation of the F_{ST} can be tricky and highly depends on the system under study.

Genetic structure level inferred from F_{ST}	
<0.05	Little structure
0.05-0.15	Moderate structure
0.15-0.25	Large structure
>0.25	Very large structure

Table 5. Guidelines for interpreting genetic structure inferred by F_{ST}. Values close to 1 can also be due to clonal populations with few mutations.

The effective population size or N_e is the number of individuals that a theoretical population would need to reach the same level of random genetic drift as the population under study (Nadachowska-Brzyska et al., 2022). N_e is small for populations with a high level of inbreeding and is high for populations with a high level of gene flow. Evaluating N_e is particularly important for small populations subject to extinction risk and is correlated with the level of genetic diversity. Additionally, demographic events in a population's history leave signatures in the genome-wide genetic variation that can be fit to a specific demographic model including events such as bottlenecks, population expansions and contractions.

The D-statistics or ABBA-BABA statistics (Durand et al., 2011) can be used to infer whether events not supported by tree topology (i.e. reticulate relationships such as gene flow, admixture, horizontal gene transfer) can explain the observed data.

4.5. Non-parametric methods

Non-parametric methods try to summarize the genetic data to infer evolution without fitting an explicit statistical model, by clustering individuals with similar genetic profiles (Alhusain & Hafez, 2018). The easiest method is computation of genetic distance between individuals and plotting a distance matrix, which can be used to derive clusters of individuals with higher similarity within clusters. Several distance metrics are available that can also adapt to polyploidy. Principal Component Analysis (PCA) is a dimensionality reduction method which can also be applied to genetic data. This allows plotting the individuals in a lower dimensionality space to visualize their genetic distance. The reduced data can also be employed for clustering; clusters of individuals will correspond to groups of individuals with a more similar genetic makeup and thus with a population structure. For example, DAPC (Discriminant Analysis of Principal Components) applies k-mean clustering and Discriminant Analysis to PCA-reduced genetic data to infer the best number of clusters that can effectively divide the individuals and performs a dimensionality reduction that maximizes between-group variation while minimizing within-group variation (Jombart et al., 2010).

4.6. Parametric methods

Parametric methods fit an explicit statistical model and try to estimate the parameters of the model. These methods are based on the concept of the coalescent theory, which is a way of modelling the ancestral state of the population starting from the present data. Parametric approaches rely on a series of assumptions about the data and are sensitive to sample size, and care is needed when interpreting results. For example, several methods rely on the assumption of random mating between subpopulations (Hardy-Weinberg Equilibrium or HWE), which may be violated in many real-world scenarios. Bayesian inference and Maximum Likelihood have been widely applied in this field, as well as simulation-based methods that rely on the Markov Chain Monte Carlo family of methods. Some methods aim at reconstructing a certain number of ancestral populations, or clusters, from which the sampled individuals derive (Pritchard et al., 2000; Alexander et al., 2009a). Each individual has a certain percentage of their genetic makeup coming from a certain ancestral

cluster, or a certain posterior probability of belonging to a certain cluster. As the number of clusters k is a hyperparameter that has to be defined by the user, it is standard practice to run the algorithms for several values of k and choose the best number according to several metrics (such as the Bayesian Information Criterion or BIC). The clusters may represent subpopulations, different populations or species, and indicate the history of admixture or gene flow between populations. Finally, the evolutionary history of individuals or population can be explicitly modelled via a phylogenetic tree. As different loci may have different evolutionary histories, each locus may result in a distinct phylogenetic tree. There are different methods to compute the individual trees, create a consensus when more trees are equally good, and optionally integrate events such as hybridization or migration in the modelling.

SOFTWARE	REFERENCE	NOTES
Nonparametric methods		
DAPC (implemented in *adegenet* R package)	(Jombart et al., 2010; Jombart & Ahmed, 2011)	Clustering and dimensionality reduction
EIGENSTRAT/SmartPCA	(Price et al., 2006)	
PLINK	(Purcell et al., 2007)	Multipurpose software for analysis of genotype/phenotype data
Parametric methods		
STRUCTURE	(Pritchard et al., 2000)	Infer population structure in a Bayesian framework with MCMC simulations
ADMIXTURE	(Alexander et al., 2009a; Alexander & Lange, 2011)	Infer population structure in a maximum likelihood framework
BEAST	(Bouckaert et al., 2019)	Infer a tree phylogeny in a Bayesian framework with MCMC simulations
TREEMIX	(Pickrell & Pritchard, 2012)	Infer a tree phylogeny with admixture and hybridization events in a maximum likelihood framework
Fastsimcoal	(Marchi et al., 2024)	Estimate Ne and infer demographic history with MCMC simulations in a maximum likelihood framework
ABCtoolbox	(Wegmann et al., 2010)	Approximate Bayesian Computation to estimate Ne and infer demographic history with MCMC simulations
BPP	(Flouri et al., 2018)	Bayesian MCMC to infer demographic history and species delimitation from genetic data under the MSC model
ASTRAL	(C. Zhang et al., 2018),	Infer species tree from gene trees under the MSC model
SNAPP	(Bryant et al., 2012)	Estimate species tree from genetic data in a Bayesian framework with MCMC simulations

Table 6. List of some approaches for population genetics and species delimitation inference divided in parametric and nonparametric methods. Software: tool name. Reference: paper on its development. Notes: notes on the use cases and methods employed by the tool with distinction between Bayesian and maximum likelihood frameworks.

5. The genus *Salvia*

5.1. Stamen architecture and taxonomy

The genus *Salvia* is at the center of one of the main collections in the Brera Botanical Garden collection, due to its long history of culinary, medicinal and cosmetic use. Being the largest genus in the mint family, *Salvia* comprises almost 1000 species worldwide. Historically, taxonomic studies in *Salvia* were based on the most prominent characteristic of the genus: the staminal lever mechanism (Claβen-Bockhoff et al., 2003). The mechanism, combined with gynodioecy, which is the presence in the population of plants developing exclusively female flowers in addition to plants developing hermaphrodite flowers, seems to have the effect of increasing cross-pollination frequency. Flowers of *Salvia* present two stamens (compared to the four in Lamiaceae), modified in a lever-like system; generally, the two thecae in each stamen are separated by an elongated connective tissue and one theca is sterile. The two filaments of the stamens are connected by sterile connective tissue at the base (the lever of the lever system), a structure that prevents access to nectar by insects. When the insects visit the flower, they need to push on this structure to reach the nectary, at the base of the flower. Under this weight, the lever is triggered and the fertile anthers descend from the upper lip of the corolla depositing the pollen on the back of the insect. The ovary presents a longer stylus to receive pollen from a visiting pollinator (Claßen-Bockhoff, 2017) (**Figure 6**).

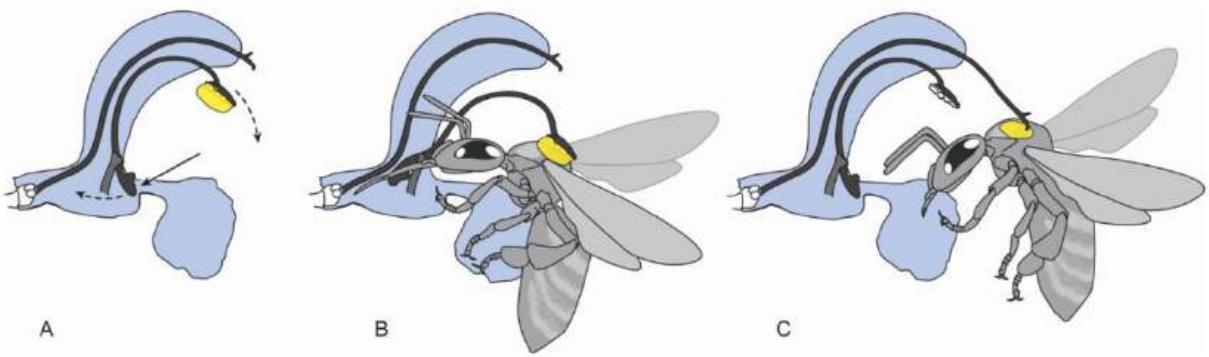


Figure 6. Pollination mechanism in *Salvia pratensis*. In B), the insect triggers the mechanism by accessing the nectary at the base of the flower. In C), a flower at a later stage of development present a longer stylus and can receive the pollen. Image credits: Claßen-Bockhoff, R. (2017). Stamen construction, development and evolution in Salvia s.l. *NVEO - NATURAL VOLATILES & ESSENTIAL OILS Journal | NVEO*, 28–48

This mechanism was long studied in *S. pratensis*, one of the most common European species. However, it was later found out that the staminal lever evolved multiple times independently in the genus and is not present in all species, being almost completely absent from some subgenera (Walker & Sytsma, 2007); moreover, given its independent evolution, the lever is not a homologous structure, and the stamen follow different modes of development in the genus, which have been studied extensively from a mechanistic point of view (Claßen-Bockhoff et al., 2004; Claßen-Bockhoff, 2017) (**Figure 7**).

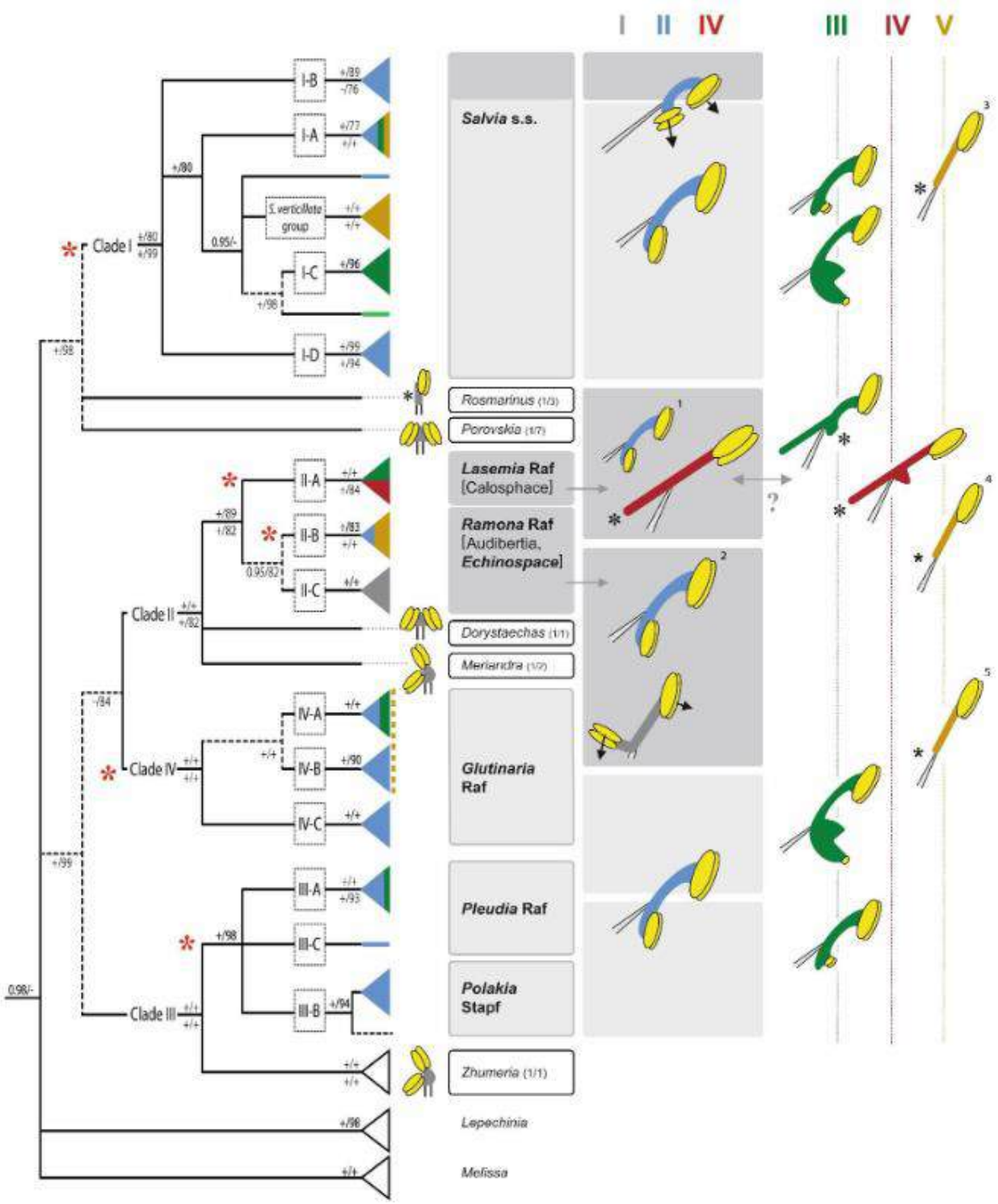


Figure 7. Evolution of stamen types in different *Salvia* lineages. *S. pratensis* has type III stamens (Monothecate stamens with inhibited thecae and curved mobile connectives).
Image credits: Claßen-Bockhoff, R. (2017). Stamen construction, development and evolution in Salvia s.l. *NVEO - NATURAL VOLATILES & ESSENTIAL OILS Journal | NVEO*, 28–48

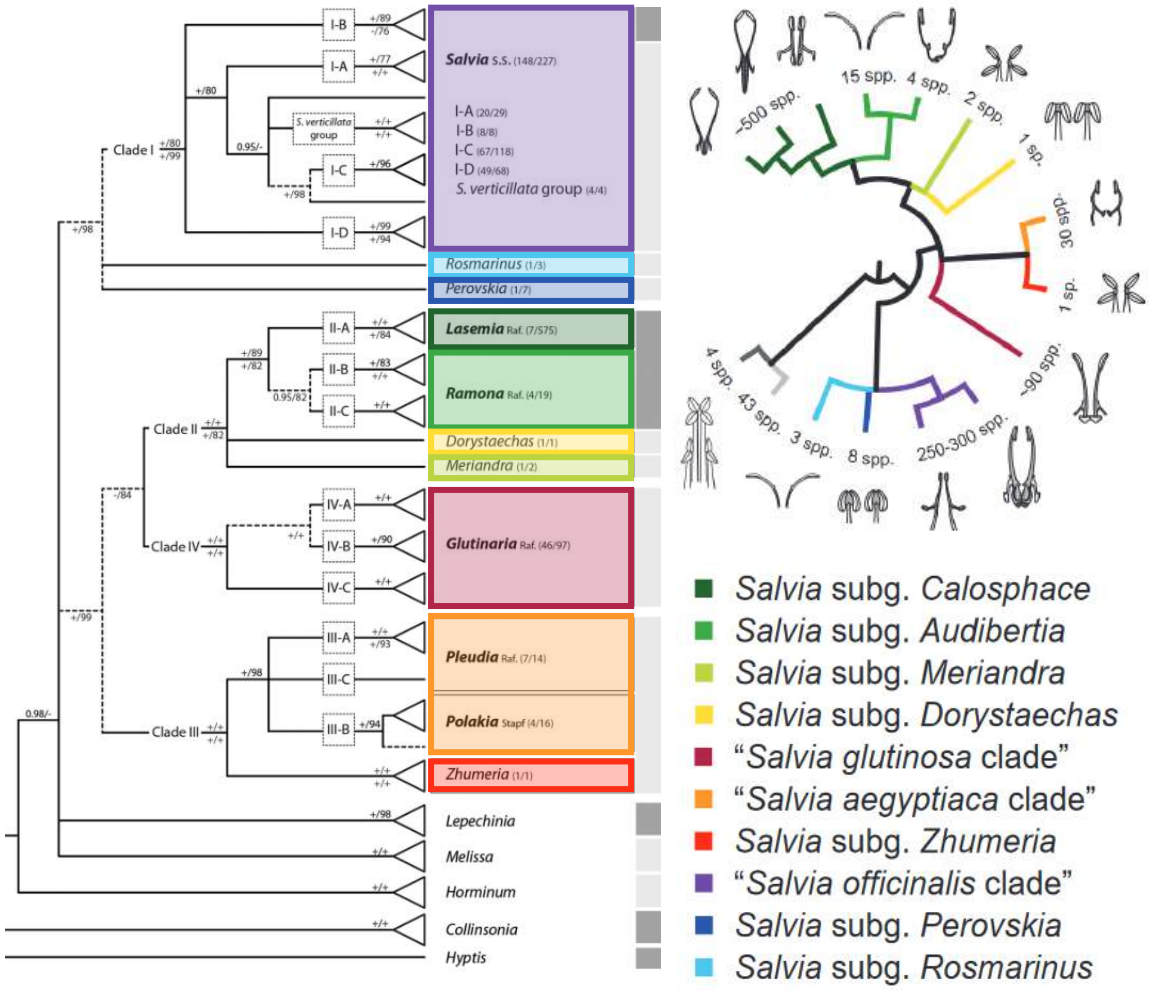


Figure 8. Correspondences between taxonomical classification of Salvia in Drew *et al*. (2017) (right) and Will & Claßen-Bockhoff (2017) (left). Kriebel et al. (2020) largely follows the Drew et al (2017) classification, with the following distinctions: "*Salvia officinalis* clade" is split in subgenera *Salvia*, *Sclarea* and *Heterosphace*; *Zhumeria* includes the "*Salvia aegyptiaca* clade" (*Pleudia* and *Polakia*); "*Salvia glutinosa* clade" = *Glutinaria*.

As genetic data started to contradict taxonomy based on stamen morphology, it became clear that the traditional *Salvia* genus was not monophyletic and is closely related to five other genera (*Rosmarinus, Zhumeria, Perovskia, Dorystaechas* and *Meriandra*) The work of Will & Claßen-Bockhoff (2017) and Kriebel et al. (2019) identified 4 main clades and 11 subgenera (**Figure 8**). The clades reflect the geographical distribution of the species of the subgenera: Clade I (*Salvia* s.s. with *Salvia officinalis* as type species) includes Old World species (Africa, Mediterranean and Southwest Asia), Clade II includes American species, Clade III goes from Southwest Asia to Northern Africa, and species in Clade IV can be found in East Asia with very few exceptions.

Salvia includes species of commercial interest. The common kitchen sage, *Salvia officinalis* and rosemary, *S. rosmarinus*, are commonly used in traditional cooking and herbal medicine. Traditional Chinese medicine uses the red roots of danshen (*S. miltiorrhiza*) and its importance is highlighted by the number of papers on

its genome. Chia, native to Mexico and Guatemala and sometimes referred to as an orphan crop, which are minor crops that tend to not be traded internationally but are important for local food security, and is obtained by seeds of *Salvia hispanica* or of a few other species. Several species and cultivars are grown as horticultivars, such as *S. splendens* and *S. microphylla*, or used in the essential oil industry. There is a wealth of literature on the bioactive effects of *Salvia* leaf extract, root extract, or essential oils.

5.2. *Salvia* in Italy

The Mediterranean area was for some time considered the center of origin of the genus. Later studies reassigned the origin of *Salvia* s.s. to Asia (Will & Claßen-Bockhoff 2017); still, West Asia, Northern Africa and Europe remain important centers of diversity. However, the diversity of *Salvia* is generally not known to the general public. In Italy, there are 19 native species according to the latest checklist (Bartolucci et al., 2024) and a few naturalized or casual alien species (Pignatti, 2018; Galasso et al., 2024). A summary of the species is available in **Table 7**. *S. pratensis*, belonging to subgenus Sclarea, is one of the most abundant *Salvia* species in Europe north of the Mediterranean (**Figure 9**). Of the *Salvia* species native to Italy, five are endemic or have a restricted distribution range, and three of them are morphologically similar to *S. pratensis*, to the point that their identification and species delimitation has been long debated. This similarity is due to the high phenotypic plasticity and morphological variation occurring within and among *S. pratensis* populations. The three species part of this *S. pratensis* group are *S. ceratophylloides* (localized to Reggio Calabria and critically endangered), *S. haematodes* (traditionally identified in the Central and Southern regions) and *S. saccardiana* (found in the North-Eastern regions of Veneto, Trentino Alto Adige and Friuli Venezia Giulia, as well as North-West Croatia and Slovenia).

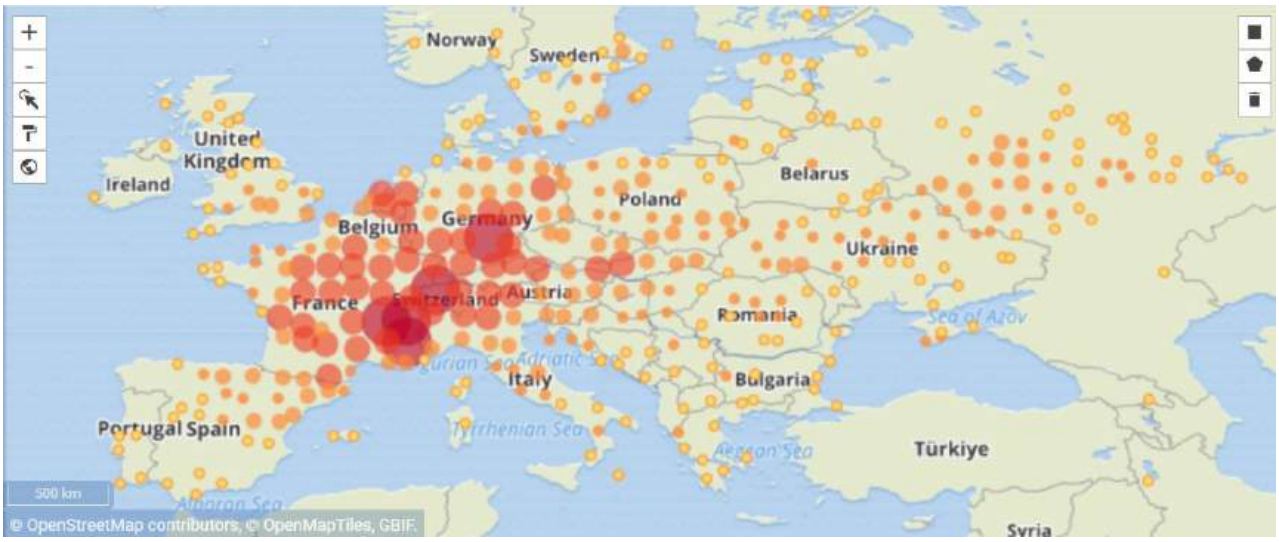


Figure 9. Distribution of *S. pratensis* in Europe, from GBIF. (Salvia pratensis L. in GBIF Secretariat (2023). GBIF Backbone Taxonomy. Checklist dataset https://doi.org/10.15468/39omei accessed via GBIF.org on 2024-11-11)

Species	Chorological group	Native in	Altitude
***S. aethiopis* L.**	Mediterranean southsiberian	VdA and Piemonte (Cryptogenic), Abruzzo	100-1000 m
***S. argentea* L.**	Stenomediterranean	Lazio, Abruzzo, Puglia, Basilicata, Calabria, Sicilia	0-1400 m
***S. ceratophylloides* Ard.**	Endemic Italian	Reggio Calabria, on Aspromonte hills (endemic)	250-400 m
***S. clandestina* L.**	SE European	ubiquitous except VdA, Lombardia, Trentino A.A., Veneto, Friuli V.G.	0-1400m
***S. desoleana* Atzei & V.Picci**	Endemic Sardinia	Sardinia (endemic)	100-500m
***S. fruticosa* Mill. subsp. *thomasii* (Lacaita)**	North Mediterranean-Montane	Lazio, Puglia, Basilicata, Calabria, Sicilia (endemic)	0-1000 m
***S. glutinosa* L.**	Orophilous eurasiatic	Alpes and Apennines from Northern Italy to Calabria	100-1600 m
***S. haematodes* L.**	Endemic Italian	Emilia-Romagna, Marche, Umbria, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria (endemic)	0-1600m
***S. officinalis* L.**	Stenomediterranean - west	Carsean region in Friuli V.G. and Southern Italy; often naturalized	0-300 m
***S. nemorosa* L.**	Sudsib. (Steppe)	Piemonte, Emilia-Romagna, Marche, Toscana, Lazio, Umbria, Abruzzo, Molise	0-600m
S. pratensis* L. / *S. pratensis* subsp. *bertolonii	Eurymediterranean	Ubiquitous except Puglia, Basilicata, Calabria	0-1600 m
***S. rosmarinus* Schleid.**	Stenomediterranean	Mediterranean coast	0-800m
***S. saccardiana* (Pamp.) Del Carr. & Garbari**	Endemic SE-Alpic, present in NW Croatia	Trentino A.A., Veneto, Friuli V.G.	1000-1400m
***S. sclarea* L.**	Eurymediterranean	Ubiquitous except Trentino A. A., Veneto, Marche, Calabria	0-900 m
***S. verbenaca* L.**	Mediterranean-Atlantic	Ubiquitous except VdA, Trentino A. A.	0-1400m
***S. verticillata* L.**	South European-Caucasian	Northern-Central Italy to Lazio, Basilicata	0-1800m
***S. virgata* Jacq.**	SE European (subpontic)	Emilia-Romagna to Basilicata, Sardinia	0-1000m
***S. viridis* L.**	Steno-Mediterranean	Toscana, Lazio, Campania, Puglia, Basilicata, Calabria, Sicilia	0-1000m

Table 7. Native Italian Salvia species with information on their chorological group, native locations and altitude where they can be found. Chorological and presence information from (Pignatti, 2018) and the Portal to the Flora of Italy (dryades.units.it/floritaly). VdA = Valle d'Aosta; Friuli V.G. = Friuli Venezia Giulia; Trentino A.A. = Trentino Alto Adige.

Salvia pratensis L. (**Figure 10**), (Del Carratore et al., 1998) is frequent in Central and Southern Europe, where it grows in dry meadows and pastures. Concerns about its conservation were raised in the Netherlands and United Kingdom due to extreme mowing and intensive agriculture (Hegland et al., 2001). However, it is generally an extremely frequent plant and not under threat. Microsatellite markers have been developed for this species (Krak et al., 2020). Several commercial varieties of *S. pratensis* with different colors and morphology of the flower (for example, varieties with pink or white flowers instead of purple) can be purchased. It is an important source of nectar for bees and other insects and blooms from spring to summer, though flowering can be protracted to autumn after meadow mowing. *Salvia pratensis* has been used as a model system to study genetic erosion in relation to population size (Ouborg et al., 1991; Treuren et al., 1991; Ouborg & Treuren, 1994, 1995) and gynodioecy (B. Zhang & Claßen-Bockhoff, 2019).

Salvia saccardiana (Pamp.) Del Carr & Garbari (**Figure 11**) was initially described as *S. pratensis* var *saccardiana* (Pamp., 1904) and classified as an endemic species of NE Italy by Carratore et al. (1999). The species was described from samples coming from the south-eastern Alps (Euganean Hills, Bellunese Prealpine district, Cadore and Carnia) and it was shown to be distinct from *S. pratensis* morphologically and chorologically, as well as having a different chromosome number. It was reported in 2013 to be present in Slovenia (Dakskobler et al., 2013). Despite this, it was often given a subspecies rank by several authors. (Balant et al., 2019) reviewed historical records of this species in Italy and performed morphological analyses on living plants from Croatia and Slovenia, concluding that *S. saccardiana* is clearly separated from *S. pratensis*. *S. saccardiana* is reported as having larger leaves and flowers and a more pronounced scent. The preferred habitat is also slightly different from that of *S. pratensis*, as *S. saccardiana* thrives on dolomitic bedrock in a thermophytic fringe forest of Trifolio-Geranietea class, while *S. pratensis* prefers dry meadows and grasslands (Festuco-Brometea) (Dakskobler et al., 2013).

Salvia haematodes L. (**Figure 12**), phenotypically similar to S. pratensis, has its distribution range in central and southern Italy. The name *haemaetodes* was assigned by Del Carratore et al. (1998) to describe plants with dark, red markings on the leaves, that look like droplets of blood. The reference to blood is captured also in the common Italian name "Salvia sanguigna". In Pignatti (2018) *S. pratensis* subsp. *haematodes* (L.) Briq. is described, with abundant glandular hairs on the flower, in contrast with subsp. *pratensis* which lacks them. Subsequently, Pignatti referred to the subsp. *haematodes* as most frequent in Italy and subsp. *pratensis* as not observed. The 2018 and subsequent checklists (Bartolucci et al., 2018; Galasso et al., 2024) assigned the rank of species to *S. haematodes* and Conti et al. (2005) reported its occurrence in the Southern regions of Italy, while other records were reported in Central Italy (*Portale Della Flora d'Italia - Portal to the Flora of Italy*, n.d.). *Salvia haematodes* has been used to treat male infertility in humans and mice.

Salvia pratensis subsp. *bertolonii* (Vis.) Soó was described by Visiani in 1847 and applied to populations along the East Adriatic coast. Morphological criteria for distinction between *S. bertolonii* and *S. pratensis* have been different between different authors. While several studies focused on the different essential oil profile of the two, following studies (Balant et al., 2019) were not able to distinguish between *S. bertolonii* and *S. pratensis* on morphometric data.

Salvia ceratophylloides (**Figure 13**) was described by Arduino in 1764 from plants in the Padua Botanical Garden and recognized by Lacaita in 1921 as from Calabria. *S. ceratophylloides* has always been rare and it was considered extinct between 1997 to 2008. compared to *S. pratensis*, the leaves of *S. ceratophylloides* have a very distinctive shape and scent. The plant only grows in an extremely restricted range, in a few localities around Reggio Calabria, and is threatened by habitat fragmentation and land use changes (Spampinato et al., 2019). For these reasons it is classified as "critically endangered" according to IUCN redlist criteria. Some authors report it as a variety of *S. pratensis*. Pignatti (2018) includes it in the *S. pratensis* group

but recognizes it as an endemic species. The demography and genetic structure of the populations have been recently studied (Laface et al., 2022, 2024), and indicate a severe bottleneck in the evolutionary history of this taxon corresponding to the period of presumed extinction, but despite that, the current populations are not suffering from inbreeding depression.

At the time of writing, few genetic studies have been performed on these species, and inter-species relationships within the *S. pratensis* group are still unknown. A summary of available cytogenetic and genome size data for some selected species is available in **Table 8**. *S. pratensis* and *S. ceratophylloides* populations have been separately studied with species specific markers, however, a study including all these species is not present in literature. Investigating gene flow, population structure, and evolutionary distance between these species is important to corroborate their taxonomic rank and shed light on the evolution of the Italian endemic species.

Species	Chromosome n° (2n)	Ref.	Genome size (2C, pg)	Ref.
***S. aethiopis* L.**	22	A	2.99	L
***S. argentea* L.**	22	B		
***S. ceratophylloides* Ard.**	54; inferred hexaploid	C	3.71	*
***S. clandestina* L.**	42	D		
***S. desoleana* Atzei & V. Picci**	44; inferred tetraploid	E	3.11	L
***S. fruticosa* Mill.**	14	D	1.15	L
***S. glutinosa* L.**	16	F	2.14	F
***S. haematodes* L.**	18	D	1.13	L
***S. officinalis* L.**	14	G	1.17	L
***S. nemorosa* L.**	12, 14	H	1.11-1.19	L
***S. pratensis* L.**	18	D	1.12	L
***S. rosmarinus* Schleid.**	24	I	2.45	M
***S. saccardiana* (Pamp.) Del Carr. & Garbari**	10	J		
***S. sclarea* L.**	22	K	1.32-1.37	L
***S. verbenaca* L.**	54, 56, 56 + 1B, 56 + 2B, many others in literature; inferred hexa- to octa-ploid	D, K	2.90-2.94	L
***S. verticillata* L.**	16	K	1.50	L
***S. virgata* Jacq.**	16, 32	D, K	2.31	*
***S. viridis* L.**	16	H	1.06	L

Table 8. Karyological and genome size data for native Italian species followed by respective references. Chromosome counts and genome size could vary greatly between different populations. Sources: A: (del Caldo, 1971); B: (Giordani & Peruzzi, 2012); C: (Salmeri, 2019); D: (Kamari et al., 1996); E: (Atzei & Picci, 1982; Corrias, 1983); F: (Temsch et al., 2010); G: (Siljak-Yakovlev et al., 2010); H: (Haque, 1981); I: (Romano et al., 1996); J: (Del Carratore et al., 1999); K: (Ranjbar et al., 2015); L: (Maynard & Ruter, 2022); M: (Pellicer et al., 2010); *: De Paola L., unpublished data. When possible, publications with samples coming from Italy have been prioritized.

Figure 10. *Salvia pratensis* L. From The New Plantfinder, Dryades Project, University of Trieste. (https://dryades.units.it/cercapiante/index.php) (accessed 13/11/2024)
Image credits: Andrea Moro, Dipartimento di Scienze della Vita, Università degli Studi di Trieste.

Figure 11. *Salvia saccardiana* (Pamp.) Del Carr & Garbari. From The New Plantfinder, Dryades Project, University of Trieste. (https://dryades.units.it/cercapiante/index.php) (accessed 13/11/2024)
Image credits: Sergio Ballestrin.

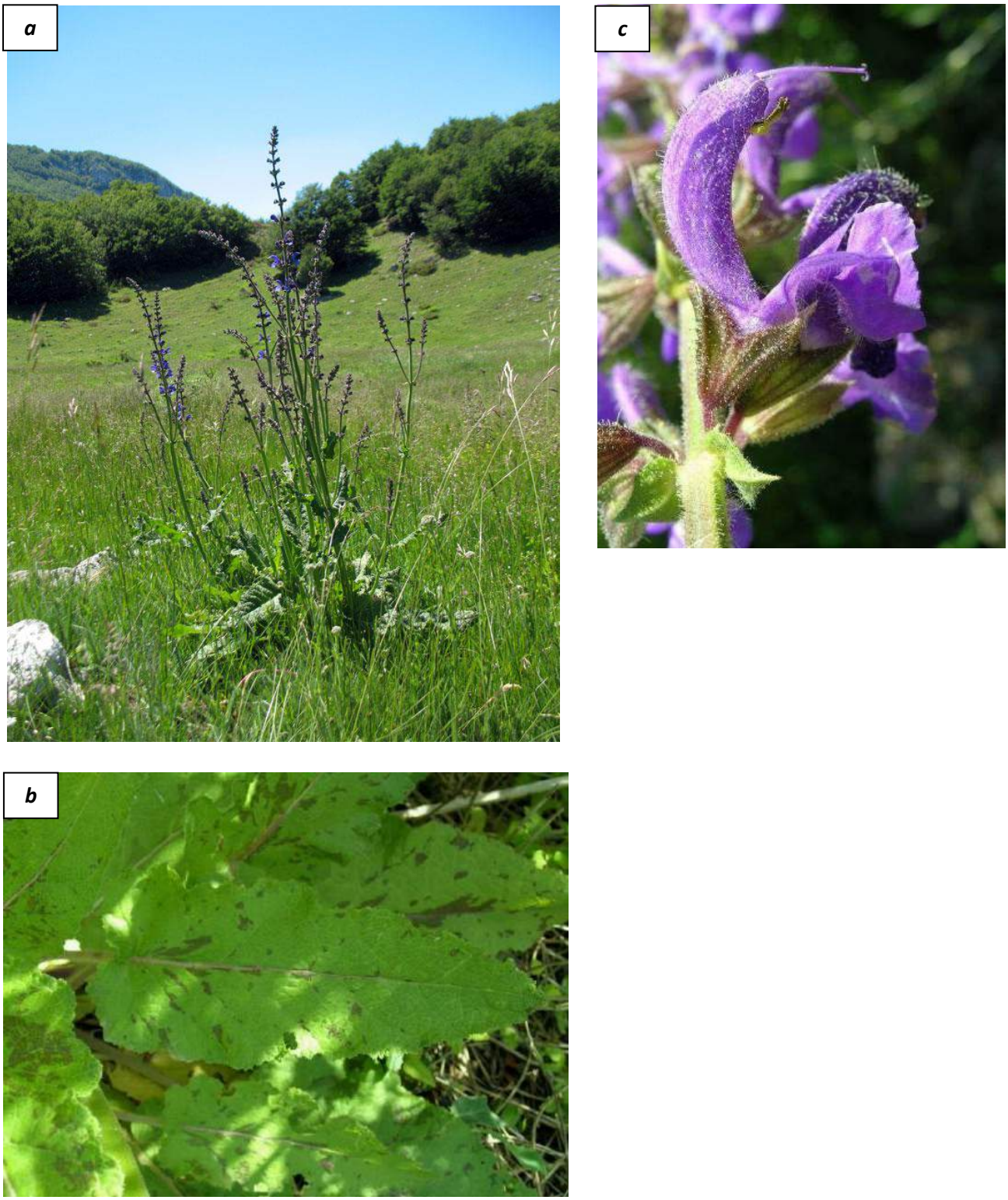


Figure 12. *Salvia haematodes* L. From Acta Plantarum.
a) Salvia sanguigna, Ugento (LE), 10 m, mag 2019. Credits: Quintino Giovanni Manni
b,c) *S. haematodes*, Rocca di Mezzo (AQ), June 2008. Credits: Gianluca Nicolella.
https://www.actaplantarum.org/galleria_flora/galleria1.php?id=3254 (accessed 13/11/2024).

Figure 13. *Salvia ceratophylloides* Ard.
Spampinato *et al.* https://doi.org/10.5772/intechopen.84905

6. The MADS-box family of transcription factors

The MADS domain transcription factor family is a family of transcription factor genes involved in flower development (Ng & Yanofsky, 2001; Shore & Sharrocks, 1995). The family is present in all eukaryotic organisms and takes its name from the four founding members of the family, on which it was first studied, MCM1 *Saccharomyces cerevisiae*, AGAMOUS *Arabidopsis thaliana*, DEFICIENS *Antirrhinum majus* and SRF human. These genes encode an extremely conserved domain of 56 amino acids, which is responsible for DNA binding, called the MADS-domain, located at the N-terminus in most proteins of this family (**Figure 14**). MADS-box transcription factors bind the DNA as homo- or heterodimers to modulate the expression of their target genes. MADS-box genes play a fundamental role in the ABCE model of floral organ determination (Theißen, 2001), based on initial studies of *Arabidopsis thaliana* and *Antirrhinum majus* homeotic flower mutants (Coen & Meyerowitz, 1991). In this model, floral organ identity is explained by the expression of class A, B and C homeotic genes; where expression of A alone specifies sepals, the combinatorial expression of both A and B specifies petal identity, expression of B and C specifies stamens, and C alone specifies the identity of carpels (**Figure 15**). The model was developed based on the observation of homeotic mutants carrying loss-of-function mutations in genes of each class, displaying homeotic conversions of floral organ identity. The model was expanded to include E genes, necessary for petal, stamen and carpel formation, and all ABCE genes were found to be encoding transcription factors, almost exclusively belonging to the MADS-domain family. MADS-box genes can generally be divided into Type I and Type II. In plants, Type II MADS, also called MIKC-MADS, present a highly conserved keratin-like (K) domain and have a complex intron-exon structure, with on average seven exons per gene. The name MIKC comes from the structure of the Type II proteins, composed of a MADS domain (M), an intervening domain (I), a K domain, and the carboxy-terminal (C). In contrast, Type I genes have only the MADS domain conserved and generally have one or two exons (Gramzow & Theissen, 2010). Type I genes in *Arabidopsis thaliana* include three subfamilies, Mα, Mβ and Mγ (Pařenicová et al., 2003), whereas the MIKC genes include MIKC* (or Mδ) and $MIKC^C$. Most of the genes which are involved in the ABCE model of flower development belong to the $MIKC^C$ family: in *A. thaliana*, APETALA1 and APETALA2 are controlling sepal and petal identity; APETALA3 and PISTILLATA carry out the B class function specifying petal and stamen identity, while AGAMOUS the C-class function determining stamen and carpel identity. The SEPALLATA 1 to 4 genes, the E-class genes, are necessary to facilitate the A, B and C functions in all four floral whorls (Jetha et al., 2014; Hugouvieux et al., 2024). MADS-box genes are involved in many other processes, from flowering time determination to fruit and ovule development (Hartmann et al., 2000; Samach et al., 2000; Ratcliffe et al., 2003; Liljegren et al., 2000; Colombo et al., 2010).

The MADS-box gene family has undergone several duplication events in flowering plants (Gramzow & Theissen, 2010), making functional redundancy, neofunctionalization, subfunctionalization and gene loss very common. The *A. thaliana* genome encodes 108 MADS-box genes (Pařenicová et al., 2003); the *Zea mays* genome 211 genes (Zhao et al., 2021) and the *O. sativa* genome 75 genes (Arora et al., 2007).

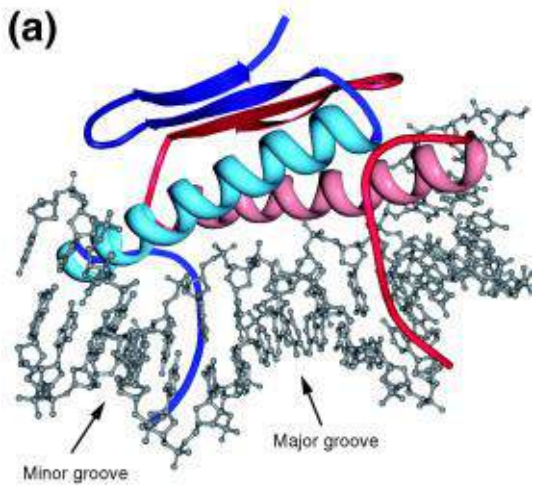


Figure 14. 3D structure of the MADS domain bound to DNA. From Gramzow & Theissen, 2010.

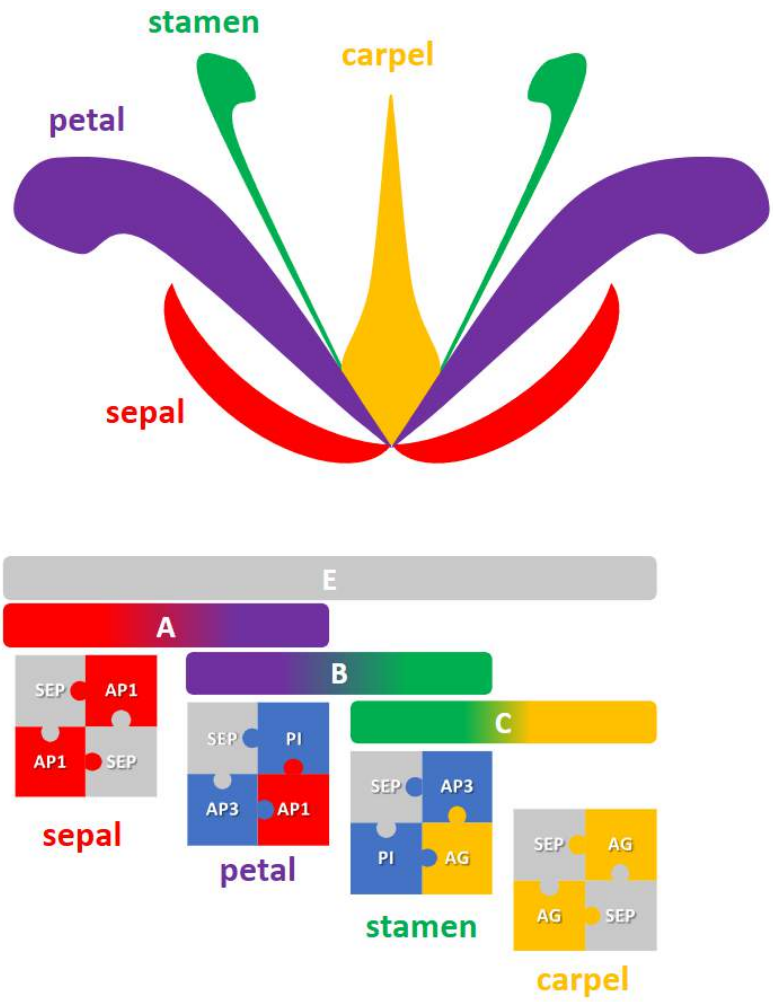


Figure 15. Simplified representation of the ABCE model of flower development. Combinatorial expression of genes of classes A, B and C is necessary for specification of reproductive organs (sepals, petals, stamens and carpels). Expression of genes with the E class function is necessary for all reproductive organs. Adapted from Gramzow & Theissen, 2010.

7. *Althaea officinalis*

Althaea officinalis is a native European plant belonging to the *Malvaceae* family. Its roots in various preparations have emollient and expectorating properties (Committee on Herbal Medicinal Products (HMPC) (EMA). Assessment report on Althaea officinalis L., radix. July 12, 2016). Several hectares of land are cultivated with *A. officinalis* by Aboca S.p.A Società Agricola for the adhesive properties of their root extracts to mucous membranes. Aboca is interested in improving the production of these molecules by selecting genotypes with higher adhesive properties. The species has not been domesticated so far. The main compounds of interest are polysaccharides and mucilage, whose production is highest in drought conditions. The adhesive activity of root extracts can be tested *in vitro*. However, given that the germplasm cultivated in Aboca does not have a single origin, there is variation in the production of secondary metabolites, and thus bioactivity, between different genotypes, even when exposed to the same environmental conditions.

Figure 16. *Althaea officinalis*. From The New Plantfinder, Dryades Project, University of Trieste. (https://dryades.units.it/cercapiante/index.php) (accessed 13/11/2024)
Image credits: Andrea Moro, Dipartimento di Scienze della Vita, Università degli Studi di Trieste.

The first issue for domestication is self-pollination, necessary to obtain “pure lines”. So far the plant has been propagated by seed via spontaneous open pollination and it is unclear whether there are self-incompatibility issues or if proterandry or protogyny (the earlier development of either of the reproductive organs, typical of the Malvaceae family) can hinder self-pollination. Moreover, Aboca processing methods are designed for large quantities of plant material and are not adequate to test the activity of individual plants. Plant clonal propagation to increase the quantity of root extract from the same genotype is not an option at the status

of the art. For genotype-phenotype association, the best approach is to carry out a Quantitative Trait Locus (QTL) analysis. In QTL, two strains or lines that differ in phenotype and genotype are crossed to obtain a heterozygous, hybrid progeny, which is then self-fertilized; the segregation of phenotypes in F2 is followed in order to identify genetic markers that occur more frequently together with the phenotype of interest (Miles & Wayne, 2008; Powder, 2020). The progeny is then self-crossed for at several generations to obtain a pure homozygous lineage that ensures phenotypic stability. Thanks to the advancement in genome sequencing, it is however possible to take a different approach that reduces the necessary time for the generation of the inbred lines. For example, approaches like pool sequencing (see **Table 4**) allow for the identification of alleles with different frequency in groups of individuals. When these individuals differ for a phenotype of interest, for example, resistance to a pathogen, it is possible that those alleles with different frequencies are also linked to that phenotype. This approach is similar to Bulk Segregant Analysis, where the population to analyze is derived from crossing of two inbred lines that display extreme phenotypes for a trait of interest and pools of individuals with extreme phenotypes are then defined and genotyped (Majeed et al., 2022) (**Figure 17**). The drawback of an approach that does not include inbred lines is that the populations to be analyzed require a large amount of present genetic and phenotypic variation and that population structuring needs to be carefully evaluated and accounted for in order to avoid spurious associations.

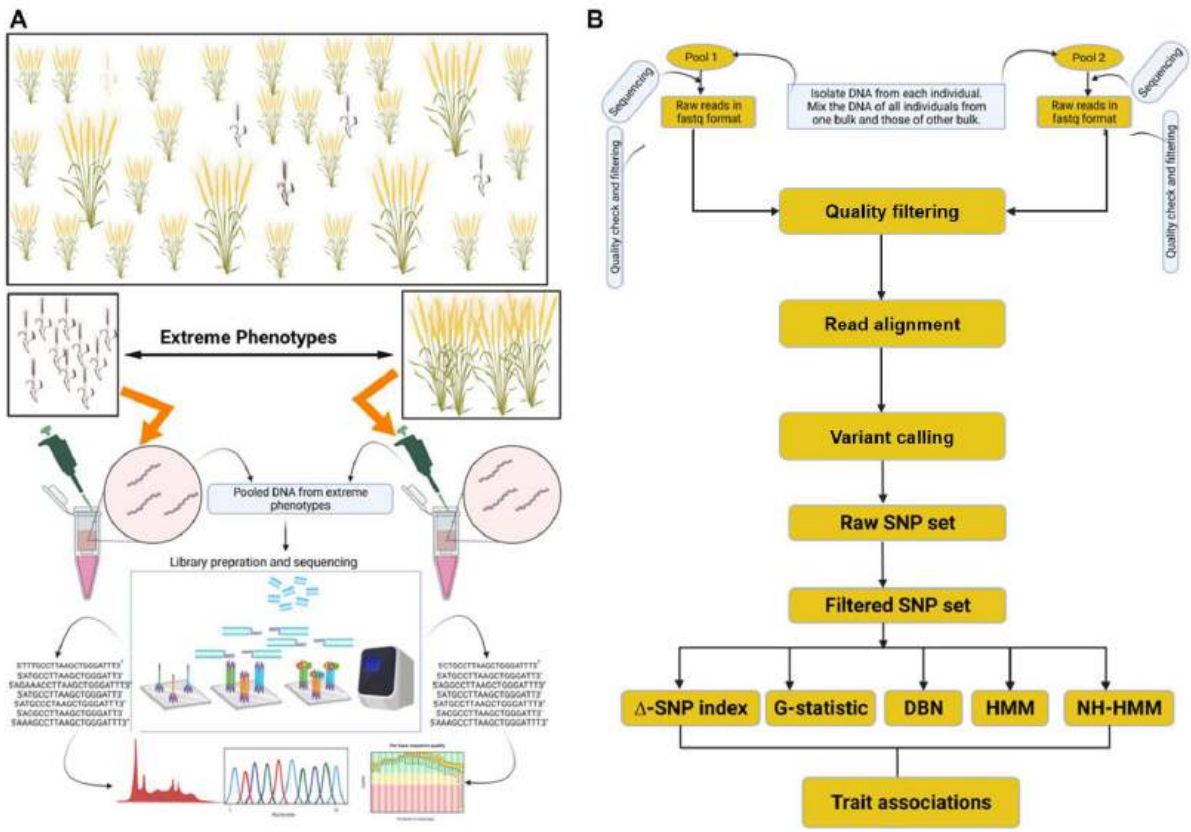


Figure 17. Library preparation and data analysis workflow for Bulk Segregant Analysis sequencing (BSA-seq). Adapted from: Majeed, A., Johar, P., *et al* (2022). Harnessing the potential of bulk segregant analysis sequencing and its related approaches in crop breeding. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.944501

A reference genome is a prerequisite for any Genotype-Phenotype association study. *Althaea officinalis* does not have a sequenced genome, and has 42 chromosomes (Zonneveld, 2019). There are no ploidy estimates for this species, but it could possibly be a hexaploid with a haploid number of 7, like related *Abutilon theophrasti* and *Malva sylvestris*. Moreover, genome annotation and functional annotation are necessary, to identify potential SNPs in the proximity or within genes related to specialized metabolites production. As no genetic studies have been carried out in Aboca on this plant, several aspects needed to be addressed, before tackling the challenge of selecting more productive genotypes:

- Is *A. officinalis* self-compatible?
- Are there significant differences in the adhesive properties of individual plants from the collection?
- Are these differences heritable?
- Can we obtain "pure" lines of plants with high and low activity?
- How big is the genome and can it be sequenced?

8. Other species – *Campanula thyrsoides* and *Solanum*

Campanula thyrsoides subsp. *carniolica* is an endemic subspecies from the Eastern Alps. In Italy, it is only found in Veneto and Friuli, and Monte Serva, within the Dolomiti Bellunesi National Park, is the western limit of the species. The populations living in Monte Serva are extremely small and have reduced in number due to habitat fragmentation and grazing. In collaboration with the Orsenigo group in the University of Pavia, the project aims at evaluating the genetic variability inside these populations to assess whether they are subject to adverse effect due to the small population size and, if that is the case, which genotypes from other populations of this subspecies are more similar to the original population and may be used for introduction. Genetic studies on this species have mostly focused on the other subspecies, *C. thyrsoides* subsp. *thyrsoides*, in Switzerland.

Two projects in the genus *Solanum* aim at genetically characterizing and placing within a background of known accessions and commercial species a set of putative heirloom varieties of tomato (*S. lycopersicum*) and potato (*S. tuberosum*) from the collection of the seed bank in the University of Pavia, which have been studied from the morphological point of view by Graziano Rossi, Silvano Lodetti and Michela Landoni. These studies fall under the project "ConseRvAzione, VAlorizzazione e caratterizzazione morFologicA e genetica di VArietà ortive "minori" locali: la RAVA e la FAVA".

Results and discussion

The Results section of this thesis presents results coming from five main research streams. Different methods employed during the three years are compared and evaluated, and results are discussed in the context of the state of the art. In chapter **1. The *Salvia pratensis* genome**, the results on sequencing, assembly and annotation of a reference genome for *Salvia pratensis* are discussed. The chapter illustrates and compares several different approaches for all steps of the assembly and annotation process and compares the *S. pratensis* genome to other sequenced genomes from the *Salvia* genus. Results from transcriptome sequencing are also briefly discussed. Chapter **2. Flower morphology and MADS-box genes**, presents morphological observation of sexual dimorphisms in *Salvia pratensis* and the results of identification and transcriptional characterization of the MADS-box family of transcription factors in *Salvia*. Chapter **3. Species interrelationships among Italian *Salvia*** illustrates results of the population genetics part of the PhD project. First, the results of the sampling campaign are described. Then, success of the genotyping method chosen (GBS) are critically evaluated. Finally, genetic data is analysed to investigate the phylogeny of Italian *Salvia* and species delimitation within the *S. pratensis* group. In chapter **4. *Althaea officinalis***, the experimental design of a study for improving the genetic resources of this plant in Aboca S.p.A. is proposed. Additionally, results of preliminary work aimed at evaluating the basic suitability of *A. officinalis* for a genomic study are described. Finally, chapter **5. Other species** illustrates results obtained with the GBS protocol for *Campanula thyrsoides*, *Solanum lycopersicum* and *Solanum tuberosum*. A list of publications and manuscripts in preparation based on the results of this PhD project, along with unrelated projects, is available in the **Appendix**.

1. The *Salvia pratensis* genome

We sequenced, assembled and annotated the genome of *S. pratensis*. We compared different long read technologies and several methods for assembly and polishing. Finally, we annotated the genome with coding and non-coding features and performed comparative analyses with other plant species.

1.1. Genome assembly

Before assembly, the genomic properties of *S. pratensis* were investigated with k-mer spectrum analysis of short reads (GenomeScope), which gave an estimate of the genome size, heterozygosity and repeated content. We performed this analysis on three accessions, P9_PEC_herm, P10_SN_herm and P11_PI_fe, which were selected from an initial collection of seeds obtained by the Botanical Garden of Brera via seed exchange for their Italian origin and the high germination rate (**Table 9**). The three accessions were sequenced with the DNBseq short read technology by BGI (**Table 10**,**Table 11**) and genomic properties were investigate with k-mer spectrum analysis performed with GenomeScope and FindGSE**Table 12** (**Figure 18**) as well as flow cytometry (**Table 11**). Genome size estimates for k-mer size analysis tend to be smaller than experimental evidence from flow cytometry and also vary greatly depending on model parameters. Still, we obtained an estimate for the genome size of *S. pratensis* of around 400Mb for the haploid genome and a ploidy estimation of 2. A more punctual analysis of this data and considerations on k-mer analysis are in (Paleni, 2021) (master's thesis). We then selected accession P11_PI_fe and performed genome assembly based on ONT (**Table 12**) and short read data. Later, the accession was also sequenced with the PacBio HiFi technology.

species	origin	sex	Code
***S. pratensis* L.**	Pecetto di Valenza, harvested from the wild (C. Pieri)	Hermaphrodite	P9_PEC_herm
***S. pratensis* L.**	Self-pollinated P9_PEC_herm	-	P9_PEC_f1
***S. pratensis* L.**	Seme Nostrum, harvested in Friuli V.G. /Veneto	Hermaphrodite	P10_SN_herm
***S. pratensis* L.**	Orto Botanico di Pisa (OBP-283/01)	Hermaphrodite	P11_PI_herm
***S. pratensis* L.**	Orto Botanico di Pisa (OBP-283/01)	Female	P11_PI_fe

Table 9. *Salvia pratensis* accessions sown and sequenced for the study. Sex of P9_PEC_f1 could not be determined as it was harvested while still in the cotyledon stage.

Sample	P9_PEC_herm	P10_SN_herm	P11_PI_fe
Total Processed Reads (M)	173.16	186.58	173.43
Total Processed Bases (Gb)	25.87	27.94	25.91
Mean duplication (%)	7.49	9.69	10.08
Estimated Genome Coverage (X)	60	60	60
GenomeScope analysis			
Genome size (Mb)	336.226	324.923	329.217
Heterozygosity rate (%)	1.31	1.2	1.7
Repeated content (%)	42.3	40.8	41.5

Table 10. Short read sequencing data from three *S. pratensis* accessions. Top: basic statistics (number of reads, number of bases, duplication, estimated coverage computed as sequenced bases / genome size). Bottom: Results of GenomeScope k-mer spectrum analysis performed with k=23. The three accessions have similar genome sizes and mid to high levels of heterozygosity for a diploid. Around 60% of the genome is repetitive.

	GenomeScope	FindGSE	Flow cytometry
Haploid genome size (Mb)	329	466	420
Heterozygosity rate (%)	1.7	1.5	-
Repeated content (%)	41.5	50.8	-

Table 11. K-mer spectrum analysis results on short read data from accession P11_PI_fe. Estimated genome size varies greatly between tools and between different values of k for the same tool. Results are reported for k=23.

	ONT	HIFI
Total Raw Reads (M)	2.36	3.16
Total Raw Bases (Gb)	74.96	28.76
Mean read quality	12.3	
Mean read length (b)	31,665.4	9,078
Read N50 (kb)	42	12
Estimated Genome Coverage (X)	**150**	**57**

Table 12. Summary of long read sequencing data (Oxford Nanopore Technologies, ONT; PacBioHiFi) from *S. pratensis* accession P11_PI_fe. Estimated coverage computed as sequenced bases / genome size

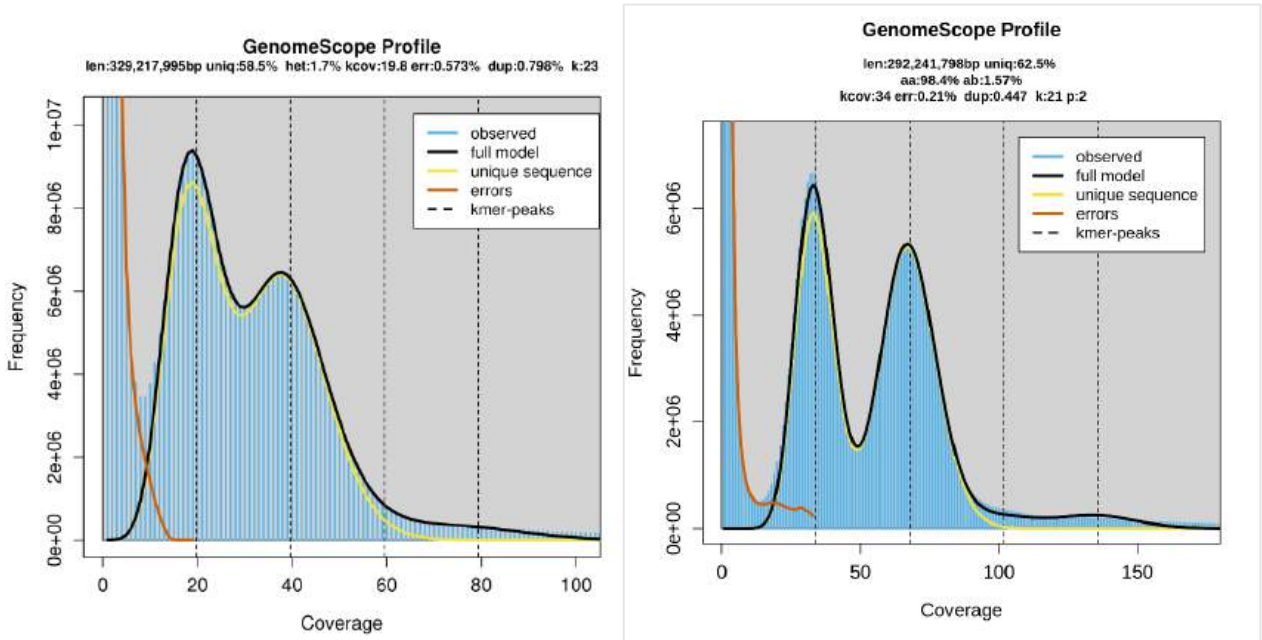


Figure 18. K-mer spectrum analysis of the short (left) and HiFi (right) read set on accession P11_PI_fe. The heterozygous peak is clearly visible and both read sets agree on the percentage of heyerozygosity (1.5-1.7%). Length estimates are quite discordant (329 Mb vs 292 Mb) and lower than other estimates (see **Table 11**).

Several assemblers were employed (Canu, Shasta, Flye, Raven) and the results from each tool were evaluated on contiguity, completeness and consensus quality (=agreement with the short-read set). As a first polishing step, all assemblies were polished by using short reads with the software Polca. The assembly results were mainly split in two categories: 800-Mb assemblies (Canu, Flye) and 400-Mb assemblies (Raven, Shasta) (**Table 13**). The best results in terms of contiguity were obtained by Canu and Shasta, with N50 over 1 Mb and largest contigs above 15 Mb. The best results in term of completeness were obtained by Canu and Flye, with BUSCO values above 96%, a k-mer completeness of 93% after polishing and a QV above 25 (**Table 14**, **Figure 19**)..

ASSEMBLY	# CONTIGS	LARGEST CONTIG	TOTAL LENGTH	N50	L50
Canu	1190	17,121,532	872,093,460	1,746,614	117
Flye	3372	7,085,168	806,597,457	827,432	235
Raven	3156	2,257,482	486,082,912	197,845	766
Shasta	957	15,045,789	487,598,328	1,293,629	90
Canu + Polca	1190	17,198,122	875,527,643	1,752,644	117
Flye + Polca	3372	7,081,282	806,422,700	826,986	235
Raven + Polca	3156	2,257,482	486,044,959	197,761	766
Shasta + Polca	957	15,059,929	488,122,758	1,294,024	90

Table 13. Assembly results on ONT reads before (top) and after (bottom) short read polishing with Polca. This table includes contiguity metrics. The haploid genome size, and thus the expected assembly size, is 400 Mb. The total length divides the assemblies in two major categories: duplicated (800Mb) and unduplicated (400Mb) assemblies. According to contiguity, Canu and Shasta are the best performers, with largest contigs above 10 Mb and a N50 above 1 Mb. Shasta has the least amount of contigs (below 1000).

	K-MER COMPLETENESS %		QV		ERROR RATE	
Assembly	Unpolished	Polished	Unpolished	Polished	Unpolished	Polished
Canu	87.72	94.40	21.63	25.43	0.0068	0.0028
Flye	87.46	93.60	21.97	26.33	0.0063	0.0023
Raven	69.33	72.99	19.78	21.04	0.0105	0.0078
Shasta	70.76	76.57	19.13	21.23	0.0122	0.0075

Table 14. Merqury reference-free evaluation of ONT assemblies before and after polishing. Polishing and evaluation were performed with the short read set. The k-mer completeness is higher in the duplicated assemblies (Canu and Flye). This is expected as the unduplicated assemblies should be leaving out half of the heterozygous k-mers. However, the QV is also higher in the duplicated assemblies. This is potentially due to a negative effect of using reads coming from both haplotypes for assembly correction, leading to the formation of chimeras in the genome sequence.

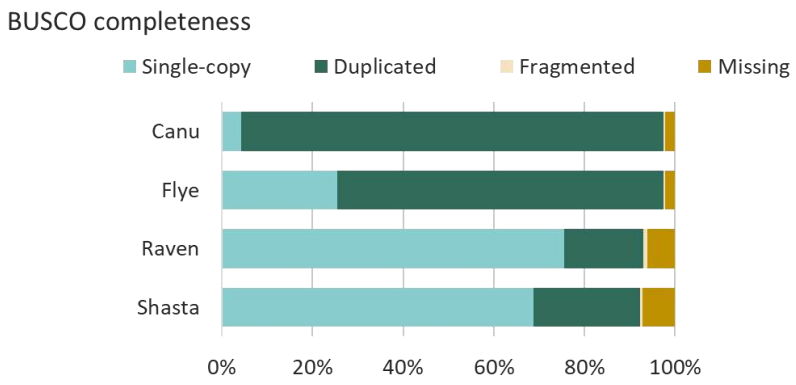


Figure 19. BUSCO completeness of ONT assemblies. The 800-Mb assemblies (Canu, Flye) have an extremely high level of duplication but are more complete.

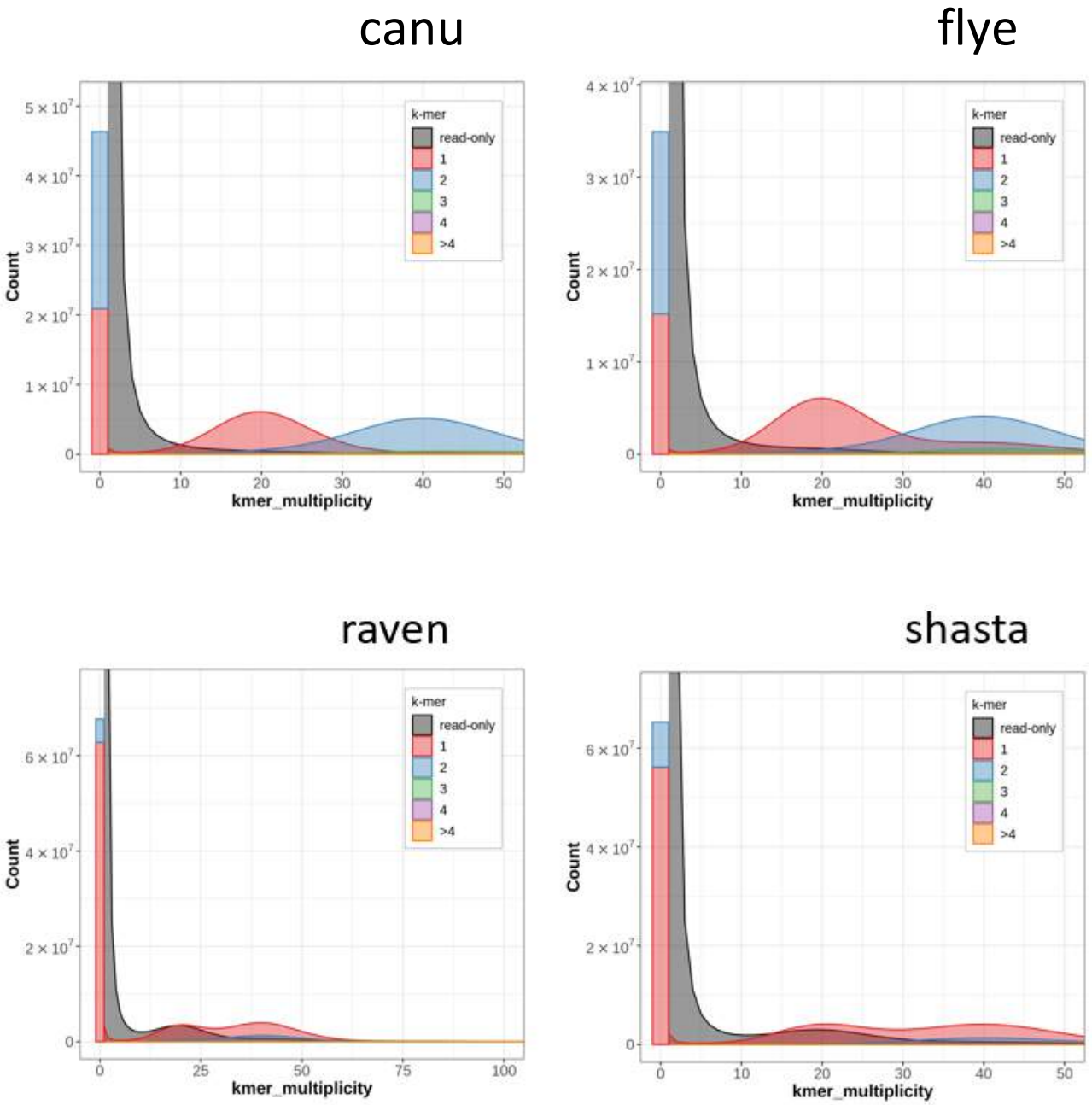


Figure 20. Merqury evaluation of ONT assemblies after Polca correction. The 800-Mb assemblies present a completely separated 2-copy peak (blue) at depth 40, which means the homozygous regions are present twice in the assembly, confirming duplication seen in BUSCO analysis. The 400-Mb assemblies present a peak of read-only k-mers (black) at depth 20, which represent the heterozygous kmers left out of the assembly. The number of assembly-only k-mers (barplot at the left) is extremely high.

The major issue encountered with ONT assembly was failure to collapse haplotypes. When the heterozygosity rate or error rate of reads is high, assemblers may have a hard time collapsing haplotypes, leading to duplicated sequences being left in the assembly. Being a diploid, highly heterozygous genome, some level of artificial duplication was expected. However, both Canu and Flye returned an almost completely duplicated assembly, as can be seen by the k-mer spectrum (**Figure 20**), the percentage of duplicated BUSCOs (**Figure 19**), and the inflated assembly size compared to the around 400 Mb expected from flow cytometry and k-

mer spectrum analysis (**Table 13**). By contrast, the assemblers which resulted in a less duplicated assembly (Raven and Shasta) had much lower quality values and a higher percentage of missing BUSCOs. Polishing increased the QV, with a more marked increase in the duplicated assemblies. This is possibly due to reads from the two haplotypes correctly mapping on the respective positions in the duplicated assembly and thus reducing the percentage of haplotype switching. However, quality was far from ideal in all assemblies. This is highlighted by the high number of assembly-only k-mers in Merqury evaluation (**Figure 20**), as well as the very low consensus quality; the double sequences are not phased haplotypes, but rather chimeras that mix the two copies of each chromosome.

To exclude the scenario of having unintentionally sequenced a tetraploid individual, we checked allele frequencies on a few highly conserved genes (BUSCOs). The idea behind this approach is that a true tetraploid would display allele frequencies with peaks along the ¼, ½, and ¾ frequency. The distribution of alternative allele frequencies had a peak around 0.5 (**Figure 21**). This confirmed to us we were in the presence of two haplotypes, and so, two sets of homologous chromosomes.

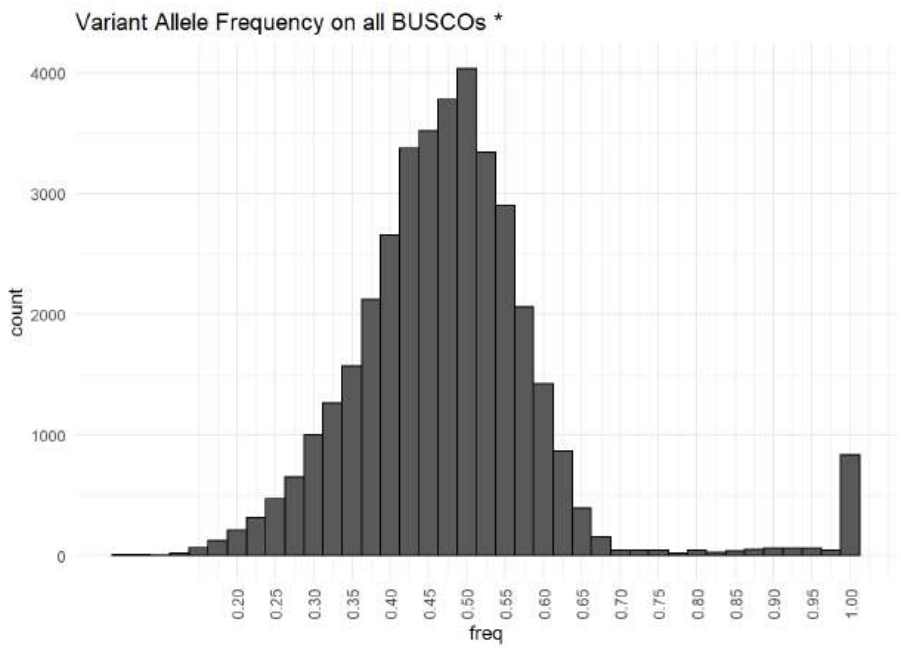


Figure 21. Frequency of alternative alleles on BUSCOs. The peak at 50% confirms this is a diploid organism.

Since Canu was best performing assembler in terms of contiguity, completeness and quality scores, we selected this assembly to improve it with post-processing, starting from the raw assembly results. The final post-processing pipeline is explained in **Materials and methods** (**ONT assembly** section). To remove duplication and increase the consensus quality, we tested different combinations of haplotype purging and polishing tools (**Table 15**, **Table 16**, **Figure 22**). Ultimately, the best results were obtained by NextPolish correction followed by haplotype purging with Purge_haplotigs and we were able to obtain a pseudohaploid assembly, with size 373 Mb, without compromising BUSCO scores (**Figure 23**). However, the assembly quality was still on the low end (QV 30). The combination of a high rate of heterozygosity of the individual and the

high read error rate of the ONT R9.4.1 chemistry did not allow for a good quality assembly even after multiple rounds of polishing. The final error rate of almost 0.1% corresponds to one erroneous base every 1,000 bp; this may cause issues in gene and coding sequence prediction. Moreover, Canu took almost one month to run, making it almost impossible to optimize assembly parameters, whereas all other assemblers, under the same conditions and the same machine, were much faster.

ASSEMBLY	LENGTH (BP)	N. CONTIGS	N50 (MB)	L50
1_canu_spra11.contigs	872,093,460	1,190	1,74	117
2_canu.filtered	869,472,723	1,159	1.74	117
3_canu.filtered.nextpolish	877,205,980	1,159	1.76	117
4_canu.filtered.purged	372,914,401	156	4.32	26
6_canu.filtered.purged.nextpolish	491,445,944	156	4.46	26
7_canu.filtered.purged.polca	393,725,887	156	4.34	26
8_canu.filtered.purged.3polca	374,657,576	156	4.34	26
9_canu.filtered.nextpolish.purged	374,741,681	159	4.36	26

Table 15. Contiguity metrics of polishing and purging steps on Canu ONT assembly. Analysis steps are sequentially listed in the row names. Contigs: initial Canu assembly. Filtered: Blobtools filtering. Purged: haplotype purging with Purge_haplotigs. Polca: short-read based Polca correction. Nextpolish: long and short read-based NextPolish correction.

ASSEMBLY	K-MER COMPLETENESS %	QV	ERROR RATE	MAP SHORT READS %
1_canu_spra11.contigs	87.72	21.69	0.0068	98.54
2_canu.filtered	87.71	21.69	0.0068	95.68
3_canu.filtered.nextpolish	95.19	28.80	0.0013	97.49
4_canu.filtered.purged	69.87	21.99	0.0063	87.96
6_canu.filtered.purged.nextpolish	76.29	28.66	0.0014	90.69
7_canu.filtered.purged.polca	74.68	27.81	0.0017	92.04
8_canu.filtered.purged.3polca	74.82	28.22	0.0015	90.41
9_canu.filtered.nextpolish.purged	74.76	30.11	0.0009	90.65

Table 16. Reference-free Merqury evaluation of polishing and purging steps on the Canu ONT assembly. Purging reduces k-mer completeness as it excludes heterozygous kmers from the assembly. Polishing decreases the error rate. Assembly 9 was chosen as the final assembly. As it had the best results in terms of QV.

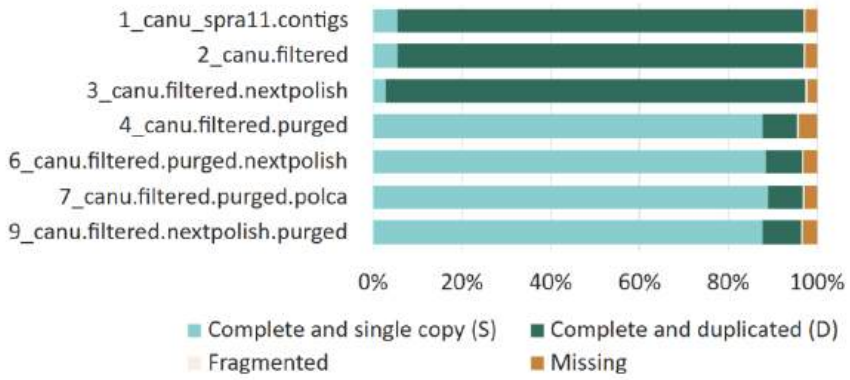


Figure 22. BUSCO evaluation of purging and polishing steps on the Canu ONT assembly. Purging greatly decreases duplications, but the order in which purging and polishing are performed does not have a big impact on the completeness.

assembly	Spra11_v9 (ONT)
Length (Mb)	373
n. contigs	159
N50 (Mb)	4.3
L50	26
N90 (bp)	326,971
L90	562
Mapping rate of short reads (%)	90.65
Mapping rate of ONT reads (%)	99.45
BUSCO score	C:96.4%[S:87.8%,D:8.6%], F:0.5%,M:3.1%,n:2326
Merqury completeness (%)	74.76
Merqury QV	30.1197
Error rate (%)	0.0972
LAI	21.56

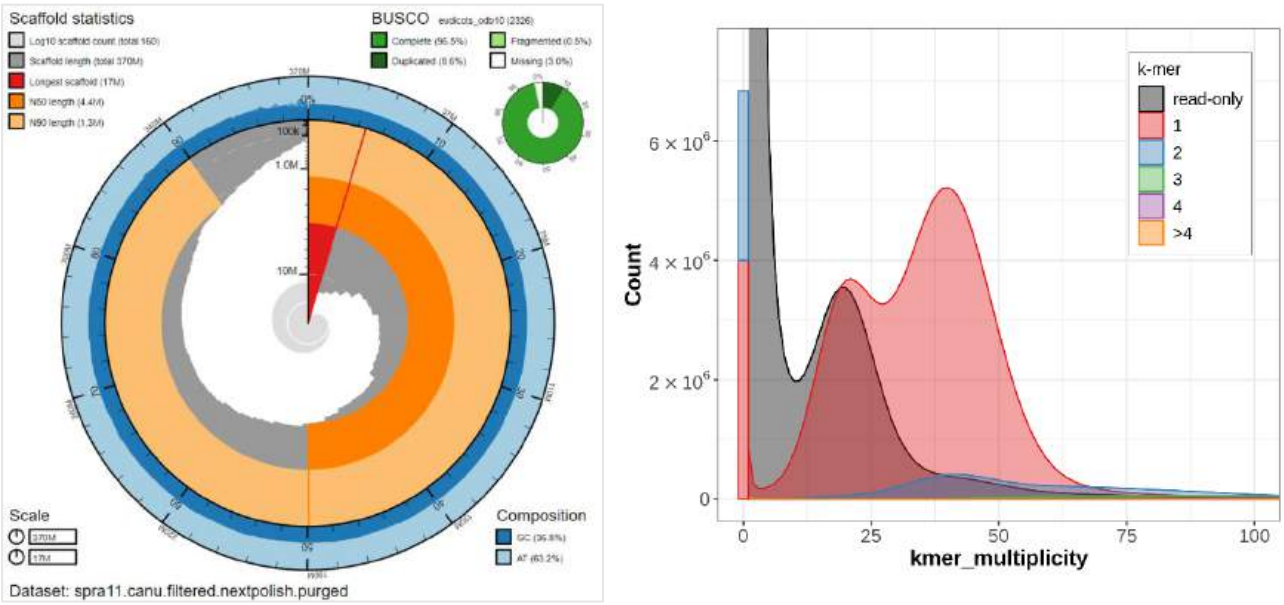


Figure 23. Table of basic statistics, snailplot and k-mer spectrum of the final ONT assembly for *Salvia pratensis* (Spra11_v9). The assembly has a length of 373 Mb and is made of 159 contigs, with the longest being 17 Mb. The N50 is megabase-scale indicating a quite good contiguity. The assembly has a good level of completeness and limited duplications, as can be seen by the mapping rates, k-mer completeness and k-mer spectrum; however, the error rate is quite high (QV 30). The snailplot summarizes contiguity statistics and BUSCO scores, while the k-mer spectrum highlights the low level of duplication, but also the high amount of assembly-only k-mers (as seen by the height of the barplot at kmer_multiplicity of 0).

During a later phase of the project, we were able to get HiFi reads for the same individual (**Table 12**). While the read length for these libraries was shorter than generally expected, with an average read length of 9kb, we still proceeded with assembly with Hifiasm and HiCanu, a Canu fork for HiFi reads. Hifiasm had the better performance, while HiCanu and Flye displayed the same issue of duplication as Canu and had a lower contiguity and accuracy (**Table 17**, **Table 18**, **Figure 24**). Compared to the ONT assembly, the HiFi assembly had highly contiguous, almost chromosome-level contigs, with the longest contig measuring 70 Mb and a N50 of 50 Mb, and a high agreement with the short-read set, despite no read correction being performed. Moreover, no contaminants were identified in the assembly. Most short contigs were made up of short repeat elements, such as rDNA clusters.

Assembly	# contigs	Largest contig	Total length	N50	L50
Hifiasm	189	73,767,105	465,905,149	50,850,824	4
HiCanu	2135	36,963,196	909,790,098	13,119,990	20
Flye	2034	7,415,628	830,282,647	1,451,669	155

Table 17. Contiguity metrics of HiFi assemblies. Both Flye and HiCanu produced a duplicated assembly. Hifiasm had the best results in terms of contiguity, with the longest contigs and N50.

Assembly	K-mer completeness %	QV	Error rate %
Hifiasm	77.5	50.65	8.60345e-04
HiCanu	95.22	42.27	5.92047e-03
Flye	95.02	50.03	9.92216e-04

Table 18. Merqury reference-free evaluation of the HiFi assemblies performed on the short-read set. The QV is much higher than that of ONT assemblies even after polishing. Hifiasm and Flye performed similarly, with Hifiasm correctly excluding some of the heterozygous k-mers.

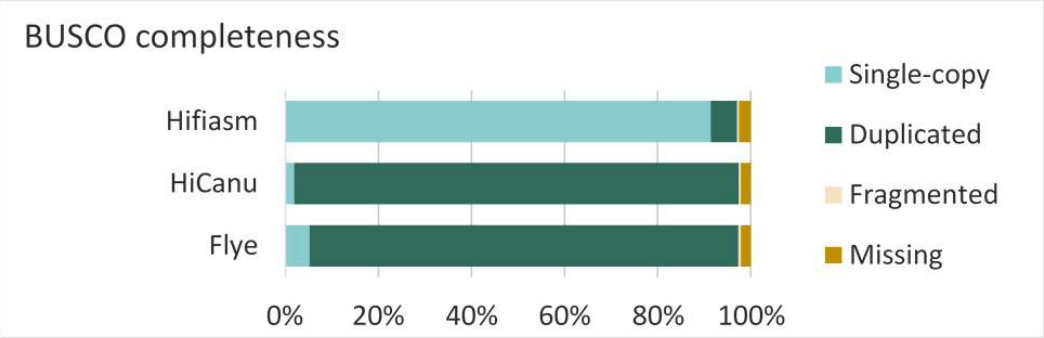


Figure 24. BUSCO completeness of HiFi assemblies. HiCanu and Flye produced a duplicated assembly, but the completeness is high in all assemblers.

We selected the Hifiasm assembly and we performed purging of haplotypes and artificial duplication (**Table 19**, **Table 20**). The final assembly has a consensus quality >50, with a 419.9 Mb assembly length over 19 contigs, 96% BUSCO completeness, minimal duplication levels, and a LAI of 27 (**Figure 25**). Due to the lack of HiC data, chromosomal scaffolds could not be achieved. However, four contigs present both telomeric and centromeric-like tandem repeat clusters, and probably represent almost complete chromosomes (see next section). The error rate of 0.0006% corresponds to one error every 166,000 bp. The k-mer completeness is

around 76%, which is similar to the completeness of the ONT assembly. However, the level of assembly-only k-mers is much lower, which is reflected in the higher QV. This indicates that a much higher percentage of heterozygous positions are correctly phased on the small scale, reducing the number of chimeras and misassemblies and thus increasing the agreement with the short-read set.

ASSEMBLY	LENGTH (BP)	N. CONTIGS	N50	L50
Spra_11.hifiasm (Spra1c)	465,905,149	189	50,850,824	4
Spra_11.hifiasm.filtered	464,381,876	159	50,850,824	4
Spra_11.hifiasm.filtered.purged	421,080,003	37	50,850,824	4
Spra_11.hifiasm.filtered.purged.clip **(Spra2c)**	419,973,825	19	50,850,824	4
Spra2c_plus_organelles	420,428,973	24	50,85,0824	4

Table 19. Contiguity metrics of polishing and purging steps on Hifiasm HiFi assembly. Analysis steps are sequentially listed in the row names. Filtered: Blobtools filtering. Purged: haplotype purging with Purge_haplotigs. Clip: clipping to discard overlapping contig ends with purge_haplotigs. The last row reports statistics of the genome sequence plus the organellar genomes (chloroplast and mitochondrion). Purging and clipping greatly reduced the number of sequences by excluding shorter, probably repetitive contigs.

assembly	K-mer completeness %	QV	error rate	Map short reads %	Map HiFi %
Spra_11.hifiasm (Spra1c)	77.49	50.6533	8.6034E-06	96.33	99.88
Spra_11.hifiasm.filtered	77.40	50.9398	8.05416E-06	91.42	99.66
Spra_11.hifiasm.filtered.purged	76.22	51.9198	6.42717E-06		
Spra_11.hifiasm.filtered.purged.clip **(Spra2c)**	76.21	52.0335	6.26109E-06	90.93	99.62
Spra2c_plus_organelles				95.91	99.85

Table 20. Reference-free Merqury evaluation of polishing and purging steps on the Hifiasm assembly of HiFi data. The QV is above 50 and the mapping rate of both short and HiFi reads is above 95%.

Despite the shorter average read length, HiFi reads performed better in assembly compared to ONT reads. This highlights the impact of high base-level accuracy on the assembly of highly heterozygous genomes. Even though the final assembly is still pseudo-haploid, sequences coming from homologous copies can be correctly separated reducing the number of chimeras and thus increasing the accuracy. This means the stretches of sequence coming from the same haplotype are longer and only one haplotype is selected. Polishing with short reads was not necessary as it only gave very minor improvements in QV.

The 9 chromosomes of *S. pratensis* are arranged in 19 contigs in the Spra2c reference, with several contigs displaying almost chromosome-level contiguity, consistent with Salvia chromosome scaffold sizes (8-70 Mbp for *S. miltiorrhiza* and *S. officinalis*).

Organellar genomes could also be retrieved. Chloroplast assembly is a relatively easy task, as evidenced by the large number of chloroplast genome generally available. Indeed, chloroplast genomes could be assembled starting from even short reads with pipeline GetOrganelle and their structure was later confirmed in the HiFi assembly (**Figure 26**). Interestingly, the chloroplast of P11_PI_fe presents a large inversion in the LSC region when compared to other accessions (**Figure 27**).

Assembly	Spra2c (HiFi)
Length (Mb)	419
n. contigs	19
N50 (Mb)	50.9
L50	4
N90 (Mb)	29.62
L90	8
Mapping rate of short reads (%)	90.93
Mapping rate of HiFi reads (%)	99.85
BUSCO score	C:96.9%[S:91.4%,D:5.20%], F:0.4%,M:2.7%,N:2326
Merqury completeness (%)	76.22
Merqury QV	52.03
Error rate (%)	0.00062611
LAI	27.32

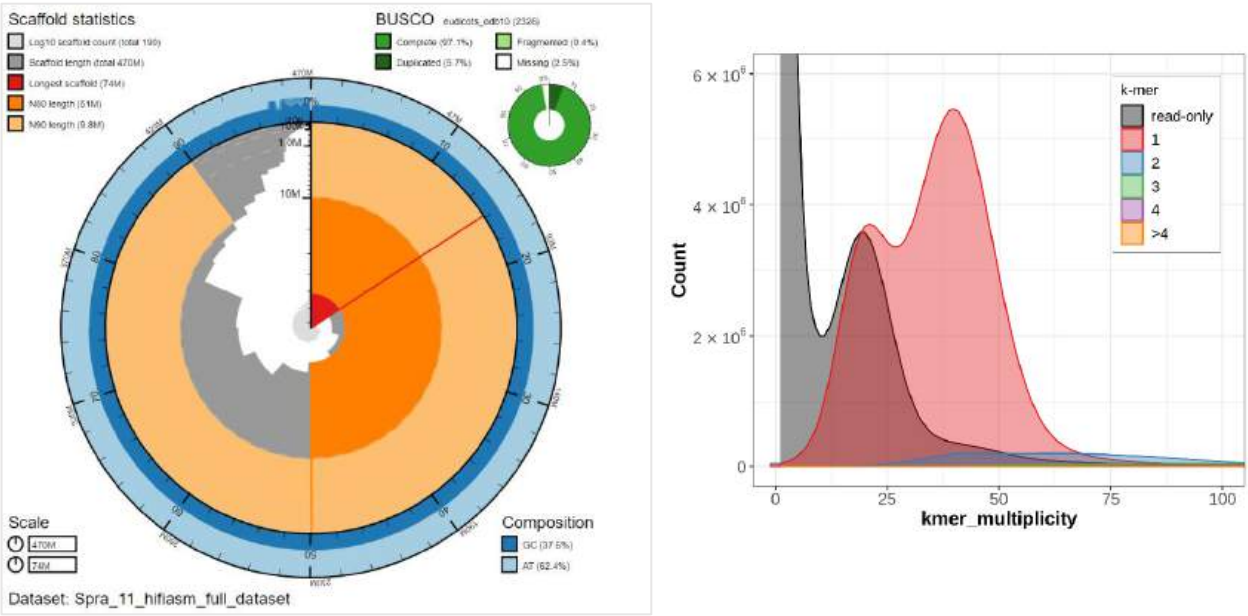


Figure 25. Statistics, snailplot and Merqury evaluation of the final HiFi assembly for *Salvia pratensis* (Spra2c). The assembly length is bigger than the ONT assembly (419 Mb vs 373 Mb), but the number of contigs is much lower (19 vs 159). The contiguity is much better, with the longest contig spanning 70 Mb and a N50 of 50 Mb. Several contigs are indeed chromosome-scale, and 50% of the assembly is represented by only 4 sequences. The level of completeness is quite high and the consensus quality also indicates a high quality assembly (QV50). The k-mer spectrum highlights minimal levels of duplication and minimal assembly-only sequences. Half of the heterozygous k-mers are excluded, as highlighted by the black peak at kmer_multiplicity ca. 20.

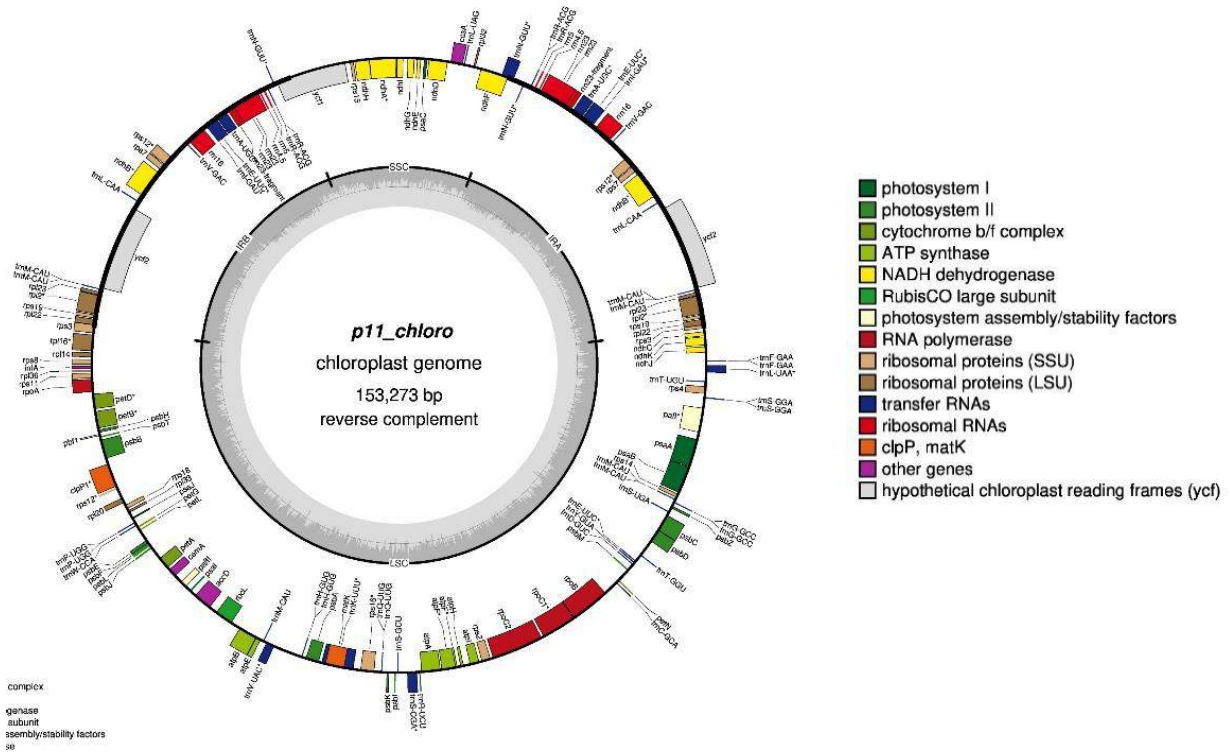


Figure 26. Annotated chloroplast sequence of P11_PI_fe. The LSC, SSC and inverted repeats are highlighted in the inner circle, while the outer circle reports gene sequences colored by functional categories (right).

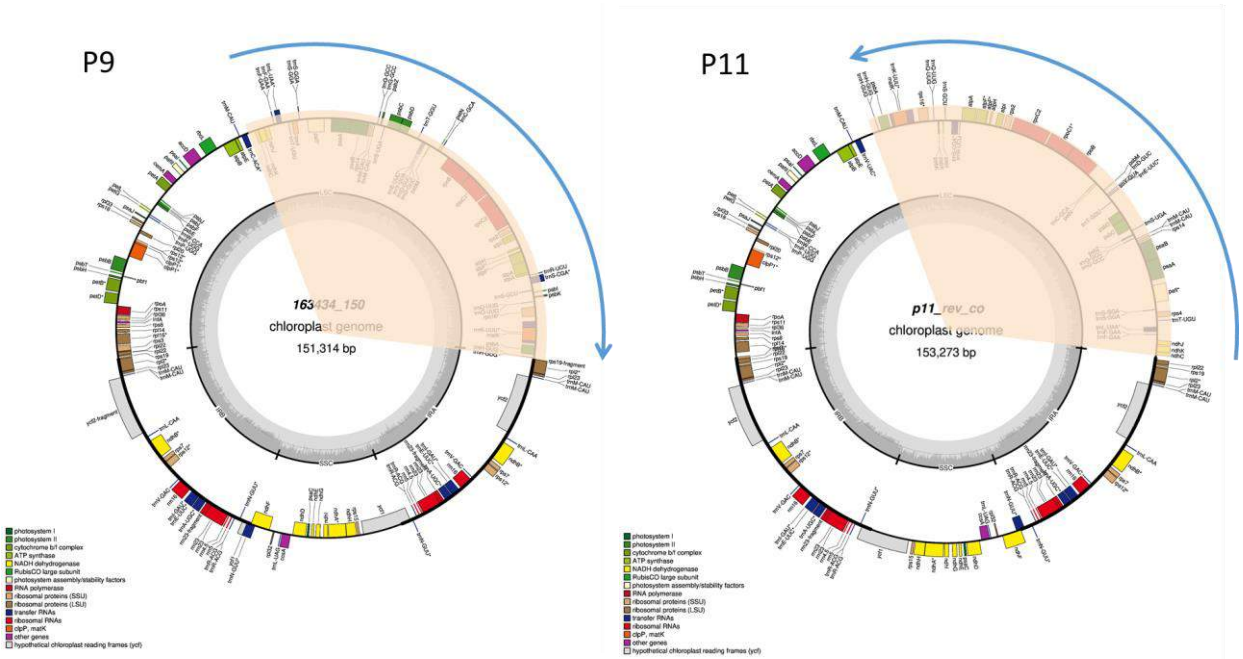


Figure 27. Annotated chloroplast sequences of P9_PC_herm (left) and P11_PI_fe (right). Highlighted in orange is the inverted segment between the two, as can be seen by the relative positioning of several genes.

Mitochondrial plant genomes are still an unsolved problem. For animals, mitogenome assembly is generally easy as mitogenomes are short and several pipelines are available (Uliano-Silva et al., 2023). In plants, the mitochondrion is not a singular, circular sequence, as it is often represented, but rather can exist in more complex structures, with a mix of circular and linear branched sequences that can recombine inside the cell (Kozik et al., 2019; Møller et al., 2021). For this reason, there is no "officially recommended" way to achieve plant mitogenomes (Štorchová & Krüger, 2024). Additionally, transfer of mitochondrial DNA to the nuclear genome (Nuclear insertions of mitochondrial origin, NUMTs) (Ramos et al., 2011) complicate the assembly process. Several approaches were tested in the project, starting from both HiFi and ONT data. Ultimately, the the OATK pipeline (Zhou, 2023/2024) was selected. We identified four sequences with an average coverage of 250X and BLAST matching to other species' mitochondria (**Figure 28**). One of the sequences can be represented as a "lasso", with the circular part composed of three single-copy sequences of 55, 13 and 16 kb and an inverted repeat of 3kb, and a 49 kb linear "tail". The largest circle has two single-copy regions of 42 and 44 kb and a repeat of 6kb. The smaller circle is 46kb in length. Finally, the shortest sequence is a linear sequence of length 18kb. The total length of mitochondrial sequences is 301.875 bp, placing it below the other *Salvia.*

	Query name	Type	Length	Hits	Query cover
	NC_037304.1	nucl	367,808	170	21.38%
	NC_023209.1	nucl	499,236	358	41.89%
	OQ001564.1	nucl	268,341	182	70.98%
	OQ001565.1	nucl	39,827	37	57.98%
	CM031494.1	nucl	347,308	303	54.56%

Figure 28. Assembly graph of selected assembly segments from OATK mitochondrial assembly of HiFi reads. Matches to other mitochondrial sequences are colored.

Approximately 5% of the short-read set comes from organellar genomes, as can be seen by the mapping rate of short reads on the assembly with and without the organellar sequences. By contrast, a very minor percentage of long reads maps on the chloroplast or mitochondria. This difference could possibly be due to the differences in DNA extraction protocols, as many methods for high molecular weight (HMW) DNA extractions that are performed for long-read sequences include as a first step nuclei isolation, which naturally excludes most of the organellar DNA. An additional 4% could not be mapped and is potentially coming from contaminants or unassembled regions.

The genome sequence was mainly developed as a tool for population genomics and gene annotation. As such, reaching a fully complete, telomere-to-telomere, haplotype-phased genome assembly was outside the main goal of the project. The upgrade from ONT to HiFi reads was deemed necessary as we feared that the high error rate in the assembly and the high percentage of chimeric sequences would negatively impact read mapping and coding sequence prediction. The change in sequencing technology also had the positive effect of highly increasing the contiguity of the assembly, other than the quality. While for several genomes, scaffolding could be performed by using a reference genome of a different individual from the same or a very closely related species, this was not possible for us; synteny analysis revealed that even the closest relative with an available reference genome, which is *S. officinalis*, has high levels of genomic rearrangement compared to *S. pratensis* (see **Figure 33**). As such, reference-based scaffolding would not have been helpful to achieve an accurate representation of the *S. pratensis* genome. Additional experiments, for example, HiC sequencing, could be performed in the future to scaffold the contigs into chromosomes and achieve a chromosome-scale assembly for *S. pratensis*. This would also allow for more punctual synteny and genome evolution analysis and would be a requirement in case we are going to develop future projects on sequencing other Italian *Salvia*, as an example, the endemic taxa strictly related to *S. pratensis*.

1.2. Transcriptome sequencing

To improve genome annotation, we carried out transcriptome sequencing on five types of plant tissue. This was done to maximize the captured gene space, as different genes are expressed at each developmental stage. We employed short read RNA-seq in paired end mode and obtained around 50M reads per sample (**Table 21**) with a mean mapping rate of 82%. Generally, for alternative isoform annotation and gene annotation, full-length transcriptome sequencing is recommended. However, given this was not our primary focus, we decided to go for short reads as they can still add a lot of value during genome annotation, focusing instead on increasing the developmental stages included. The mapping rate of transcripts to the genome was above 80% for most samples. This is a somewhat low number for the standards of model organisms, however, several factors impact mapping in novel non-model organisms. As such, the number of reads was still deemed sufficient for further analyses.

We modelled transcripts from the short-read set using StringTie. This resulted on 38,000 to 42,000 transcripts in each set, for a combined total of 60,604 transcripts over 32,691 genes. Transdecoder was used to identify putative CDSs and resulted in a set of 56,702 CDSs over 24,832 genes, which was evaluated with BUSCO in protein mode at 96.1% completeness (**Table 21**).

Tissue type	Plant code	Raw reads (M)	Clean reads (M)	Uniquely mapped	Multi-mapping	Unmapped	Assembled transcripts
Seedling	P9_PEC_f1	50.44	49.06	74.89%	4.03%	21.08%	38,771
Leaf	P11_PI_fe	47.33	45.60	80.27%	3.91%	15.82%	40,581
Inflorescence	P10_SN_herm	59.38	59.11	83.65%	3.25%	13.10%	43,470
Inflorescence0	P11_PI_fe	53.07	52.13	84.55%	3.04%	12.41%	43,198
Inflorescence1	P10_SN_herm	49.52	48.08	53.81%	2.60%	43.59%	40,867
Inflorescence1	P11_PI_fe	64.42	63.96	82.79%	2.95%	14.26%	44,879
Inflorescence2	P11_PI_fe	42.88	41.08	83.49%	2.63%	13.88%	41,290
Open Flower	P10_SN_herm	59.09	58.64	83.88%	3.48%	12.64%	42,360
Open Flower	P11_PI_fe	114.55	108.08	83.46%	3.20%	13.34%	47,486
Immature fruit	P11_PI_herm	41.69	39.27	82.78%	5.09%	12.13%	42,487

Table 21. RNA-seq statistics. Plant code: accession of the sequenced plant for that specific tissue sample. Raw reads: number of reads before quality trimming. Clean reads: number of reads after quality trimming. Uniquely mapped: number of reads that map in only one position in the Spra2c genome. These are the reads that are generally used for most analyses. Multi-mapping: reads mapping to multiple genomic locations, possibly due to assembly or gene duplications. Unmapped: reads that did not map to the Spra2c genome. Possibly due to contaminants, or missing sequences in the assembly. Assembled transcripts: Number of transcripts that were modelled by StringTie from the mapped RNA-seq reads.

1.3. Genome annotation

Repeat annotation was performed with the RepeatModeler - RepeatMasker pipeline, classifying 61% of the genome as repetitive (**Figure 29**). Retroelements make up 27% of the total genome sequence, among which 15% belong to the Gypsy and 7% to the Copia families. A very low amount of LINE sequence was retrieved (1.94%). The genome is made up of 1.16% of simple repeats and 0.17% of low complexity repeats.

Repeat families corresponding to the telomeres and centromeres could be identified. Telomeres seem to be composed of simple repeats (AAACCCT)n followed by longer repeated motifs (identified as rnd-5_family-3307) (**Figure 30**, **Figure 32**, **Figure 31**). Centromeres are composed of alternating and partially overlapping 91 and 66 bp motifs, generally interspersed with transposons (repeat families rnd-1_family-39 and rnd-6_family-901) (**Figure 30**, **Figure 31**). Centromeres and telomeres mostly correspond to peaks in the “Other repeats” track on **Figure 29**. The centromere of contig1 is flanked on one side by a cluster of rRNA repeats between positions 27,000,000 and 28,500,000.

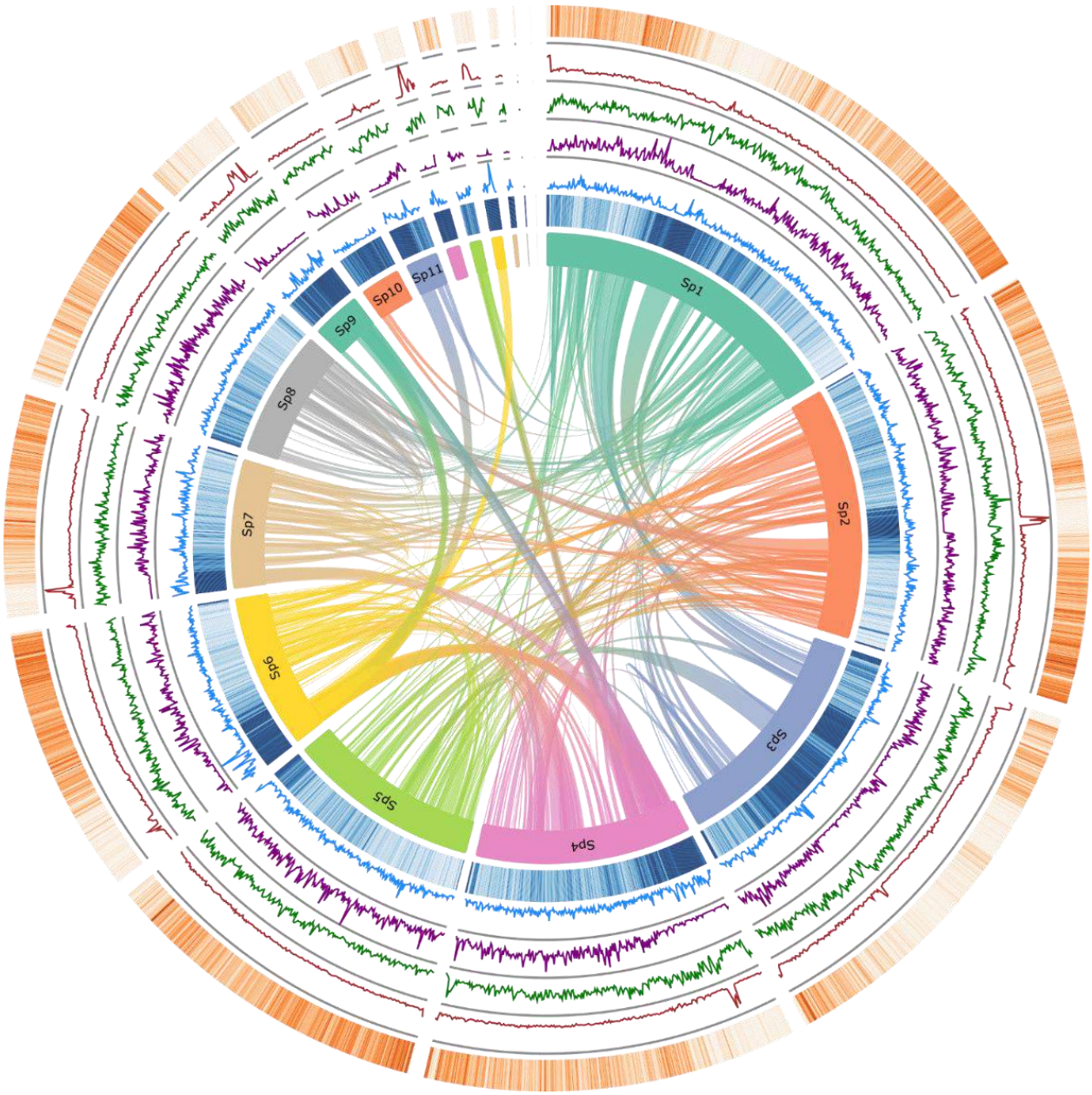


Figure 29. Summary plot of *S. pratensis* genome annotation. The outer heatmap (orange) is gene density. The inner line plots represent density of different types of repeats, in order: other repeats; LTR retrotransposons; DNA transposons; LINEs/SINEs. The inner heatmap (blue) is total repeat density. Ribbons connect duplicated segments of the genome, based on gene sequences (MCScanX analysis).

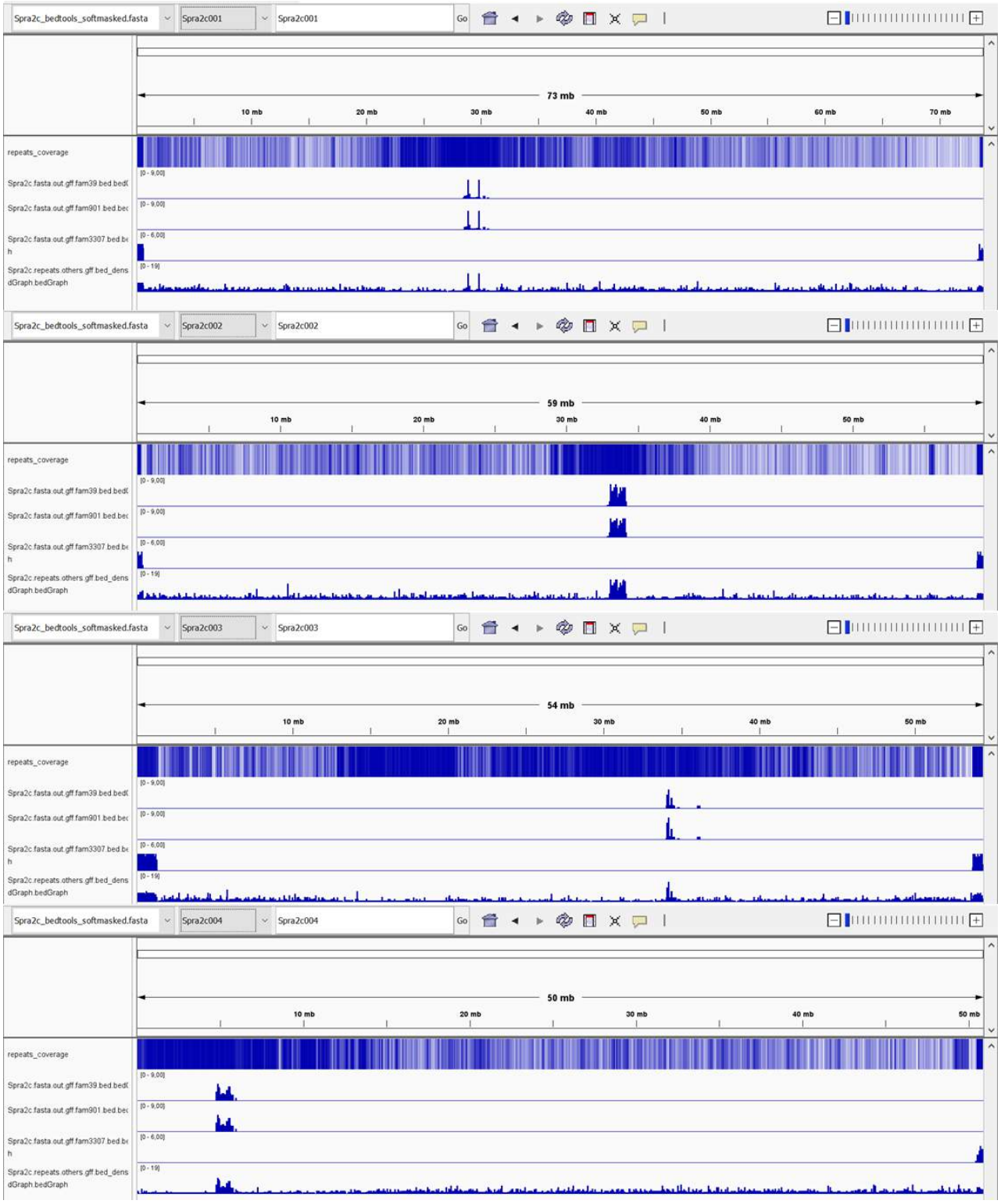


Figure 30. Telomeric and centromeric tandem repeat clusters in contigs 1-4. The heatmap under each of the contig sequences is the density of all repetitive elements, while densities of specific repeat families are represented as histograms. Fam39 and fam 901 were identified as some of the centromere-associated tandem repeats while 3307 is associated to telomeres. The «other repeats» tracks includes all repeats that were not assigned to a family (=generally transposons) and as such most of the tandem repeats, and as such can be a proxy for telomere and centromere positions.

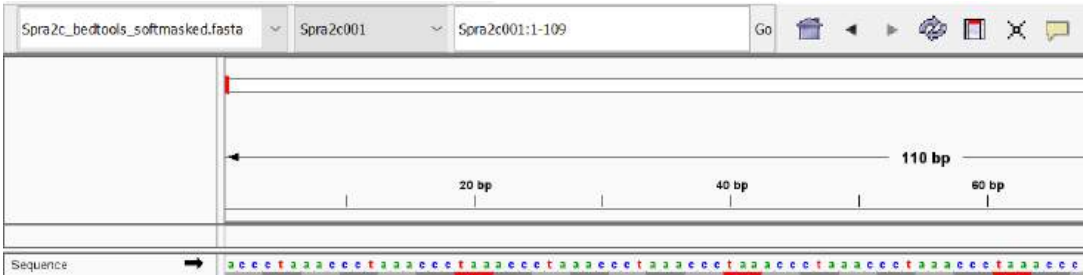


Figure 32. Example of telomeric tandem repeat sequence.

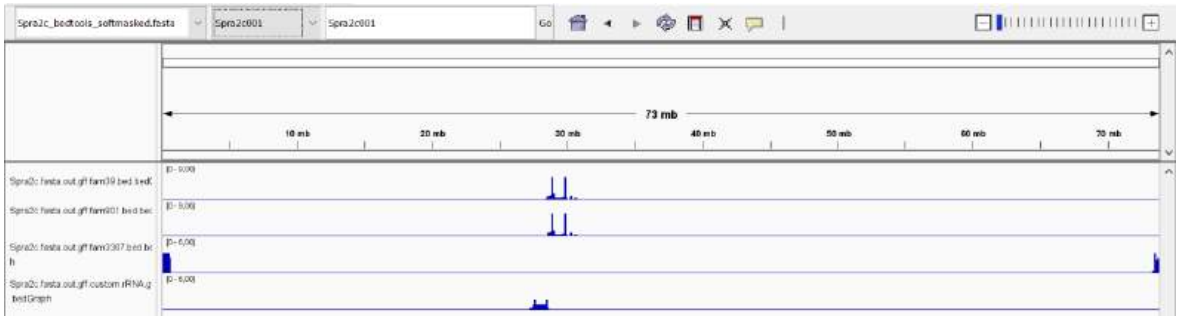


Figure 31. Centromeric (rows 1-2), telomeric (row 3) and rRNA repeats (row 4) in contig Spra2c001.

Gene annotation was carried out via several pipelines (**Table 22**). For some of those, either MAKER or Ingenannot were employed to evaluate the gene models with the Annotation Edit Distance (AED) score or add untranslated regions (UTRs) to the gene models. First, Gemoma was used to perform reference-guided annotation using several different combinations of *Salvia* species and *A. thaliana* as a guide. The results from Gemoma gene modelling depended heavily on the reference used. Adding RNAseq data greatly increased the number of alternative transcripts per gene annotated from 1.4 to 1.9. However, it seemed that the annotated gene models relied too heavily on the gene models from the other species; thus, the several isoforms annotated for each gene were mainly coming from different references rather than actual RNAseq alignments. Then, the BRAKER2 pipeline was employed, with or without RNAseq data, resulting in around 40,000 genes. Of these, around 31,000 have an AED < 0.5. The BUSCO slightly decreases with RNAseq data, but the percentage of single exon genes was 21%, closer to the expectation of 20% for plants. A very minimal number of alternative isoforms is annotated, which is somewhat expected with short read data. A third annotation was performed with the newest version of BRAKER (BRAKER3). BRAKER3 has stricter filters on the final pipeline results, thus outputting only around 28,000 gene models. Intermediate results from Augustus modelling were retrieved, which consisted of around 37,000 gene models. The MAKER pipeline was also tested and gave a very high number of gene models, above 50,000, which were then filtered based on AED to around 29,000. However, the BUSCO score of the MAKER annotation was much lower than the other annotations and even the assembled transcripts, and upon inspection in a genome browser many predicted CDSs appeared truncated. Additionally, we attempted to merge results from the several gene predictors trained via the BRAKER or MAKER pipelines with EVidenceModeler, which resulted in a high confidence set of 29,770 gene models. However, EVidenceModeler does not model alternative transcript isoforms.

	Species/	n°genes	n° mRNAs	rnas per gene	genes per Mb	n° single exon genes	%s.e.g	longest gene	mean length	BUSCO c	BUSCO f	BUSCO m
	Reference species											
	A. thaliana	27,562	52,177	1.89	229.68	5,291	19	367,996	2,418			
	S. splendens	55,405	79,251	1.43	68.73	8,577	15	794,448	3,426			
	S. hispanica	32,922	46,508	1.41	102.41	5,348	16	163,900	3,118			
	S. miltiorrhiza	32,434	49,187	1.51	61.01	5,978	18	1,184,262	4,219			
	S. rosmarinus	50,388	57,652	1.14	45.18	15,545	31	216,532	2,755			
	S. officinalis	31,713	38,063	1.20	66.54	5,874	19	273,924	4,031			
	***S. pratensis* annotations**											
	Assembled transcripts	32,691	60,604	1.85	77.84	8,560	26	1,103,053	4,394	96.1	0.7	3.2
GEMOMA	ath shis sspl	33,448	48,048	1.44	79.64	8,757	26	58,819	2,341	97.8	0.2	2.0
	ath shis sspl sros	40,080	58,101	1.45	95.43	13,923	35	58,819	2,370			
	ath_shis_sros_sspl_smil	44,154	66,188	1.50	105.14	15,745	36	58,819	2,341			
	ath_shis_sspl_smil	38,660	57,231	1.48	92.05	11,203	29	58,819	2,528			
	ath_shis_sros_sspl_smil_soff	49,233	72,481	1.47	117.23	20,413	41	58,819	2,127			
	ath_shis_ssp_smil, with rna	38,598	73,008	1.89	91.91	11,280	29	58,819	2,536			
BRAKER	braker2	39,418	42,058	1.07	93.86	11,586	29	47,374	2,183	97.2	0.5	2.3
	braker2 with rna	40,018	41,653	1.04	95.29	8,352	21	47,516	2,499	95.8	0.5	3.7
	braker2 with rna, aed < 0.5	31,358	32,872	1.05	74.67	5,594	18	47,516	2,770	95.7	0.5	3.8
	braker3	28,451	29,064	1.02	67.74	6,516	23	286,743	2,896	96.7	0.3	3.0
	braker3, aed < 0.5	27,752	28,342	1.02	66.08	6,097	22	286,743	2,945	95.7	0.5	3.8
	braker3, maker utr	25,745	26,332	1.02	61.30	6,058	24	69,111	3,109	96.5	0.4	3.1
	braker3 augustus	37,527	39,108	1.04	89.36	10,019	27	47,516	2,413	96.8	0.4	2.8
	braker3 augustus, maker utr	32,045	33,339	1.04	76.30	8,491	26	47,885	2,744	96.6	0.4	3.0
MAKER	maker	57,291	57,291	1.00	136.42	6,223	11	101,017	3,255	89.8	0.7	9.5
	maker, aed < 0.25	29,060	29,060	1.00	69.19	2,131	7	101,017	4,244	82.5	0.6	16.9
	maker_mult_tr	55,279	68,184	1.23	131.62	6,036	11	101,017	3,452	91.9	0.8	7.3
	evidencemodeler	29,770	29,770	1.00	70.89	7,699	26	81,222	2,671	96.7	0.4	2.9

Table 22. Gene annotations for reference species and results of *S. pratensis* gene annotation using different pipelines. GeMoMa: the species used as reference are listed on the left. BRAKER: the BRAKER version as well as inclusion of RNAseq data and quality filters are listed on the left. MAKER: the filter and the multiple-transcript mode were applied. EVidenceModeler was run to aggregate and select annotations performed by Augustus, SNAP and GeneMark. Columns: total n° of genes, total n° of mRNAs (=n° of transcripts), average n° of transcript isoforms per gene (n° mRNAs/n°genes), average n° of genes per Mb (n°genes/genome size), n° of single exon genes (s.e.g.), rate of s.e.g., length of the longest gene, mean gene length, BUSCO completeness statistics (c = complete BUSCOs; BUSCO f = fragmented BUSCOs; BUSCO m = missing BUSCOs).

The best results in terms of BUSCO score and gene model accuracy were obtained by BRAKER. However, the high confidence gene set selected by BRAKER3 was too restrictive and only resulted in 25,757 gene models once UTRs were added with MAKER, as MAKER applies some additional filters that eliminate overlapping or problematic gene models. Therefore, to recover the expected approximate number of 30,000 gene models, the Augustus intermediate result was selected (as recommended by the BRAKER3 authors) to add UTRs and perform following analyses. This resulted in a gene set of 32,045 gene models with a BUSCO score of 96.6%. However, the annotation of alternative isoforms is lacklustre and could be improved by long-read transcriptomic data, as well as the percentage of single-exon genes, which is slightly higher than expected.

The gene models were additionally filtered for potentially TE-related protein domains and poorly supported gene models, meaning, gene models that either present sequence similarity to TE-associated protein domains or have no protein and transcript evidence associated to them. The final protein-coding gene count of the annotation was 31,337 gene models and 32,814 coding sequences.

The assembled transcripts were analysed to identify potential long noncoding RNAs with the FEELnc pipeline. This pipeline first excludes transcripts that overlap a gene model, then classifies the coding potential of the remaining transcripts, then assigns a relative location to each lncRNA to the closest gene model. 4,984 transcripts were thus classified as candidates for lncRNA, of which 3,094 "genic" (located inside a gene) and 1,806 "intergenic". tRNA and rRNA annotation identified 885 and 3,794 loci respectively with almost all the rRNA genes in contigs 1 and 17, while tRNAs were spread more equally between contigs 1, 2, 5, 6, 7 and at a lower number in contigs 3 and 4 (**Table 23**).

contig	length	protein-coding genes	tRNAs	rRNAs
Spra2c001	73,767,105	5,921	123	2,649
Spra2c002	59,079,562	5,405	121	
Spra2c003	54,364,029	2,402	69	3
Spra2c004	50,850,824	3,509	80	
Spra2c005	44,911,161	4,751	140	4
Spra2c006	36,361,082	3,268	140	4
Spra2c007	30,512,240	2,586	128	1
Spra2c008	29,618,631	2,621	71	1
Spra2c009	11,692,271	228	6	1
Spra2c010	9,752,273	210	2	14
Spra2c011	7,735,308	193	5	
Spra2c012	3,477,718	22		30
Spra2c013	2,990,901	178		1
Spra2c014	2,869,586	31		5
Spra2c015	1,282,195	7		
Spra2c016	267,529	2		75
Spra2c017	413,248	0		1,002
Spra2c018	19,942	2		
Spra2c019	8,220	0		4
total	**419,973,825**	**31,336**	**885**	**3,794**

Table 23. Annotation statistics split by sequence. The majority of rRNAs cluster in the longest contig Spra2c001 around the centrosome. The number of protein coding genes is correlated to contig length.

1.4. Functional annotation

Sequence homology-based functional annotation was performed by comparing the predicted protein set to the EMBL SwissProt, the UniProt, and the Arabidopsis reference proteome using DIAMOND for fast protein alignment. The annotations of relative matches were propagated to the *S. pratensis* genes using AHRD and resulted in 87% of the proteins having at least one annotation. This relatively low percentage could be due to the choice of a lower confidence gene model set from the previous step. Indeed, functional annotation on the high-confidence gene set resulted in a function being annotated to 95% of the gene models. However, in this case, a higher number of gene models was prioritized. We then used OrthoFinder to identify orthologues in *Salvia pratensis* and seven other species: *A. thaliana, E. guttata, O. sativa, Solanum lycopersicum, Salvia hispanica, Salvia miltiorrhiza and Salvia officinalis*. These species were chosen as representatives of the major plant taxa that *Salvia pratensis* belongs to: *A. thaliana* for eudicots, *S. lycopersicum* for Asterids, *E. guttata*

for Lamiales. The *Salvia* genomes were included for a comparison with a close relative. Finally, *O. sativa* was added as a monocot representative and to act as an "outgroup". Overall, 95% of *S. pratensis* genes were assigned to an orthogroup (**Table 24**). Of these, 8.6% were assigned to orthogroups exclusive to *S. pratensis*, compared to 10.9 and 5.8% of *S. miltiorrhiza* and *S. officinalis* respectively. This data could also explain the number of genes that could not be annotated with sequence homology.

	Atha	***Mgut***	***Osat***	***Shis***	***Slyc***	***Smil***	***Soff***	***Spra***
Number of genes	27,562	28,140	35,804	32,922	34,074	32,434	31,713	31,336
Number of genes in orthogroups	24,794	26,135	27,812	31,842	28,938	31,595	30,351	29,852
Number of unassigned genes	2,768	2,005	7,992	1,080	5,136	839	1,362	1,484
Percentage of genes in orthogroups	90	92.9	77.7	96.7	84.9	97.4	95.7	95.3
Percentage of unassigned genes	10	7.1	22.3	3.3	15.1	2.6	4.3	4.7
Number of orthogroups containing species	13,729	15,262	13,948	15,413	15,214	15,511	16,788	16,746
Percentage of orthogroups containing species	56.5	62.8	57.4	63.4	62.6	63.8	69.1	68.9
Number of species-specific orthogroups	952	470	2,173	335	957	442	503	569
Number of genes in species-specific orthogroups	4,377	2,128	8,563	2529	4,894	3,549	1,845	2,692
Percentage of genes in species-specific orthogroups	15.9	7.6	23.9	7.7	14.4	10.9	5.8	8.6

Table 24. Results of Orthofinder analysis. *Atha=Arabidopsis thaliana, Mgut=Mimulus guttatus, Osat=Oryza sativa, Shis=Salvia hispanica, Slyc=Solanum lycopersicum, Smil=Salvia miltiorrhiza, Soff=Salvia officinalis, Spra=Salvia pratensis*

1.5. Whole genome duplications and gene family analysis

Genome duplication analysis was carried out with MCScanX and Ksrates. The analyses highlighted several duplicated segments in the *S. pratensis* genome (**Figure 29**). Comparing the genome to other *Salvias* in a synteny analysis, we can see the level of collinearity is higher when *S. pratensis* is compared to *S. officinalis*, but drastically lower in *S. miltiorrhiza* (**Figure 33**). This is expected as *Salvia* is a large genus with high divergence times between Mediterranean and Asian species.

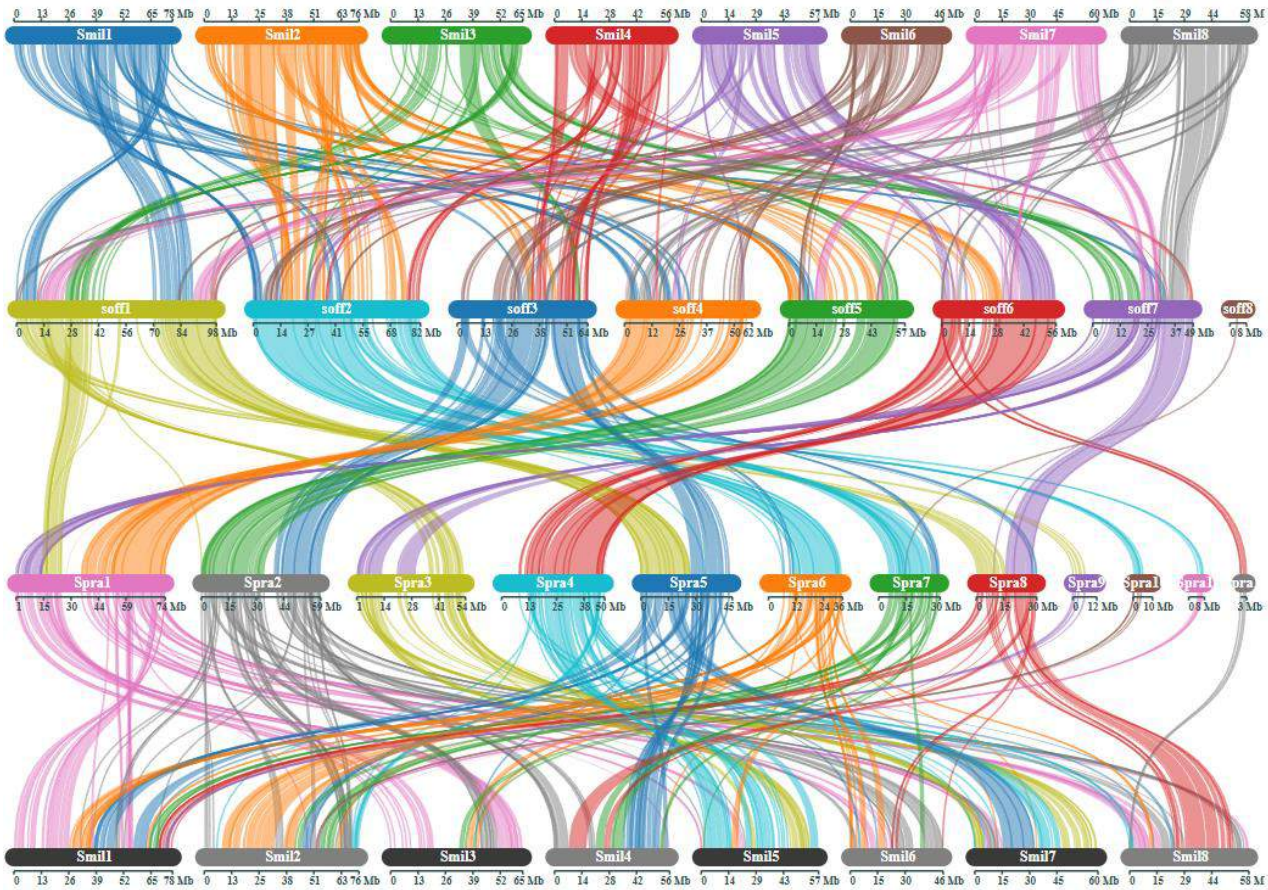


Figure 33. Collinearity between the genomes of *S. miltiorrhiza (Smil)* and *S. officinalis (Soff), S. officinalis* and *S. pratensis (Spra), S. pratensis* and *S. miltiorrhiza.*

WGD analysis revealed a whole genome duplication shared by all three *Salvia* species and *Mimulus guttatus* (or *Erythrante guttata*) which belongs to the Lamiales (**Figure 34**). Finally, gene family contraction/expansion analysis was performed with OrthoVenn (**Figure 35**). *Salvia* shows expansions in gene families related to GO terms such as oxidoreductase activity, translation, and several biosynthetic or catabolic processes (phenylpropanoid, secologanin, xyloglucan). To plot the time-calibrated tree, divergence times for *S. officinalis*, *S. miltiorrhiza*, *A. thaliana, S. lycopersicum*, *M. guttatus* e *O. sativa* were retrieved from TimeTree (http://timetree.temple.edu/), but no data was available for *S. pratensis*; the divergence time between *S. pratensis* and *S. officinalis* was estimated to around 15MYA from this analysis. This is within the confidence interval for the estimates of the divergence times between clade I-D (*S. officinalis*) and I-C (*S. sclarea*) from (Will & Claßen-Bockhoff, 2017).

Overall, the results from the annotation show that the genome obtained from *Salvia pratensis*, despite not being chromosome level, is adequate to perform several different types of comparative genomic analysis, and gives similar metrics and statistics to other sequenced *Salvia* genomes, which were used as a comparison.

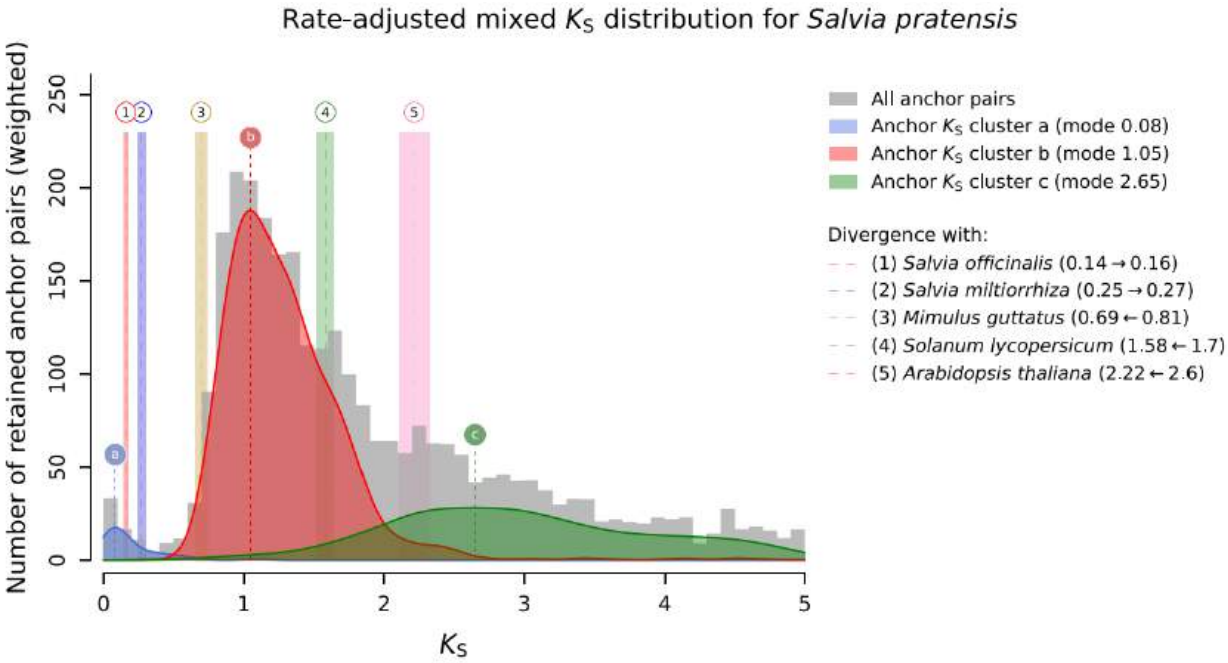


Figure 34. Whole genome duplication analysis for *Salvia pratensis* and five other species.

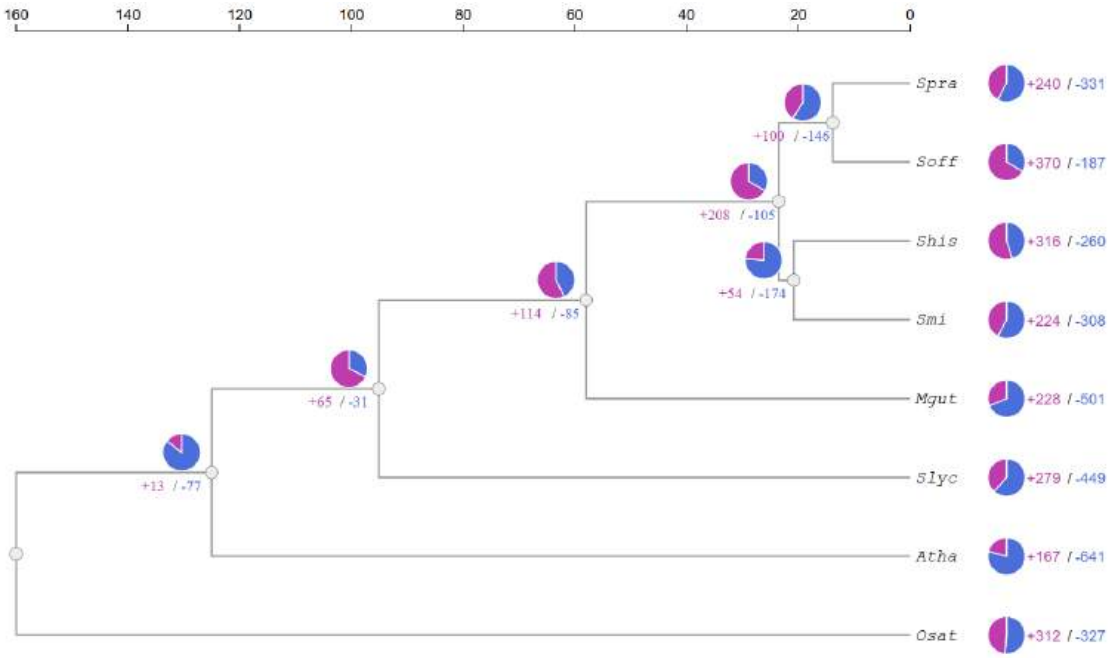


Figure 35. Gene family expansions and contractions analysis. Divergence times were retrieved from the TimeTree database (http://timetree.temple.edu/) for most of the species (data for *S. pratensis* was not available).

2. Flower morphology and MADS-box genes

Thanks to our production of plants from seed in green house to obtain suitable plant material for WGS (including selfing treatments to obtain homozygous lineages) and the field collections that we made for the GBS study, we observed large variation in flower traits in *Salvia pratensis* among the plants that we handled that led us to investigate the MADS-box gene family involved in flower development.

2.1. Preliminary observations of *Salvia pratensis* flowers

Salvia pratensis is a perennial species that has underground buds that resprout every year. Seeds can be set several times a year and germinate on the same year they are produced; however, the seedlings are unlikely to flower during the first year. In greenhouse, our seed germination rate was no higher than 10% and flower set was very variable from population to population and in general more abundant in the second year after sowing in all populations.

We were able to observe gynodioecy in several populations in the field and in botanical garden, in pot plants obtained from seeds from wild populations, and in one plant in the permanent collection at Brera Botanical garden . Generally, a plant produces only one flower type – either hermaphrodite or female – and the female plant in Brera Botanical garden produced only female flowers every year. However, we observed considerable phenotypic variation in stamen traits among flowers of different plants (**Figures 35 to 38**) from a simple reduction of the anther and filament length (unfunctional stamens) to a completely mutated, petaloid stamen phenotype, similar to that of homeotic *Arabidopsis thaliana* mutants. We also identified plants with a visibly shortened stamen, in which the anthers appeared well formed and the lever mechanism was functional. This observation suggested gradual variation in stamen phenotype. We termed this phenotype “short stamen”, a new stamen phenotype in Salvia pratensis not described in the literature.

We started a study of stamen phenotypic variation in *S. pratensis* both in wild populations and in common garden in the Città Studi Botanical Garden with plants obtained from seeds collected in wild populations. We counted the number of hermaphrodite, female and short stamen plants, classified stamen morphotype, and measured stamen lengths in three flowers of each plant (**Figure 40**). The ratio of female to hermaphrodite flowers changed among different populations (table) with no clear pattern of variation dependent on population environment conditions or on mendelian segregation ratios. The PCA of **Figure 41** shows the variation pattern of stamen morphotypes in the studied populations. “short stamen” appears as a distinct phenotype between female and hermaphrodite flowers.

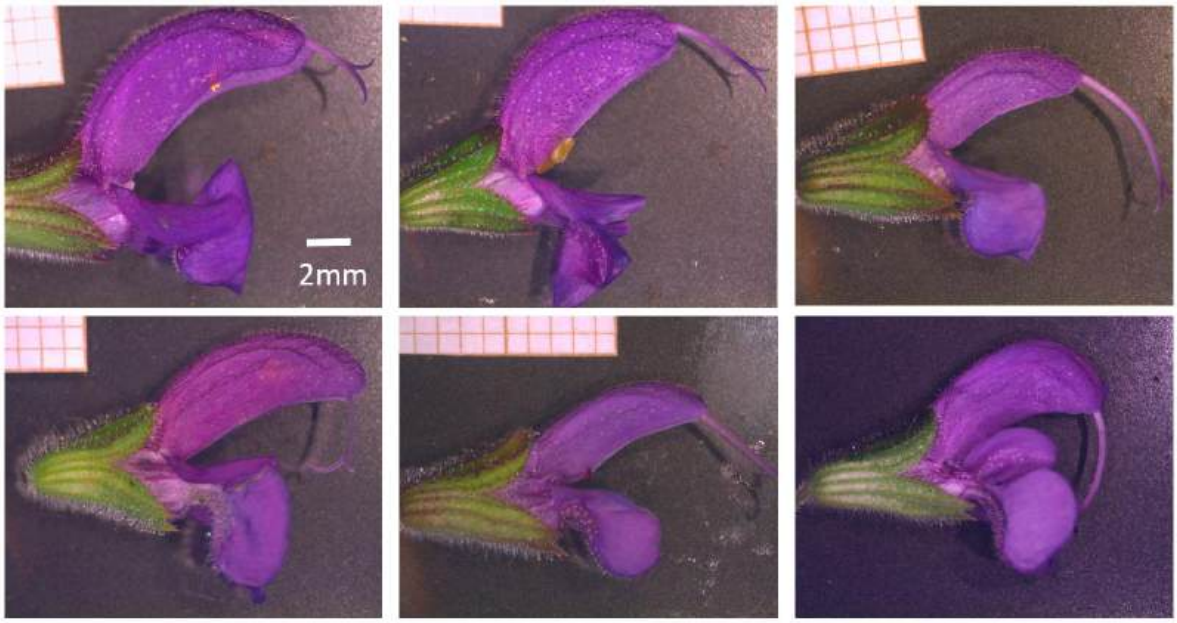


Figure 36. Flowers of *Salvia pratensis*. From top left: hermaphrodite, «short stamen» and female flower from a wild population in Inveruno (MI). From the bottom left: hermaphrodite flower of accession P9 grown in Città Studi; female flower of accession P11 grown in Città Studi; female flower with petaloid stamens grown in Brera.

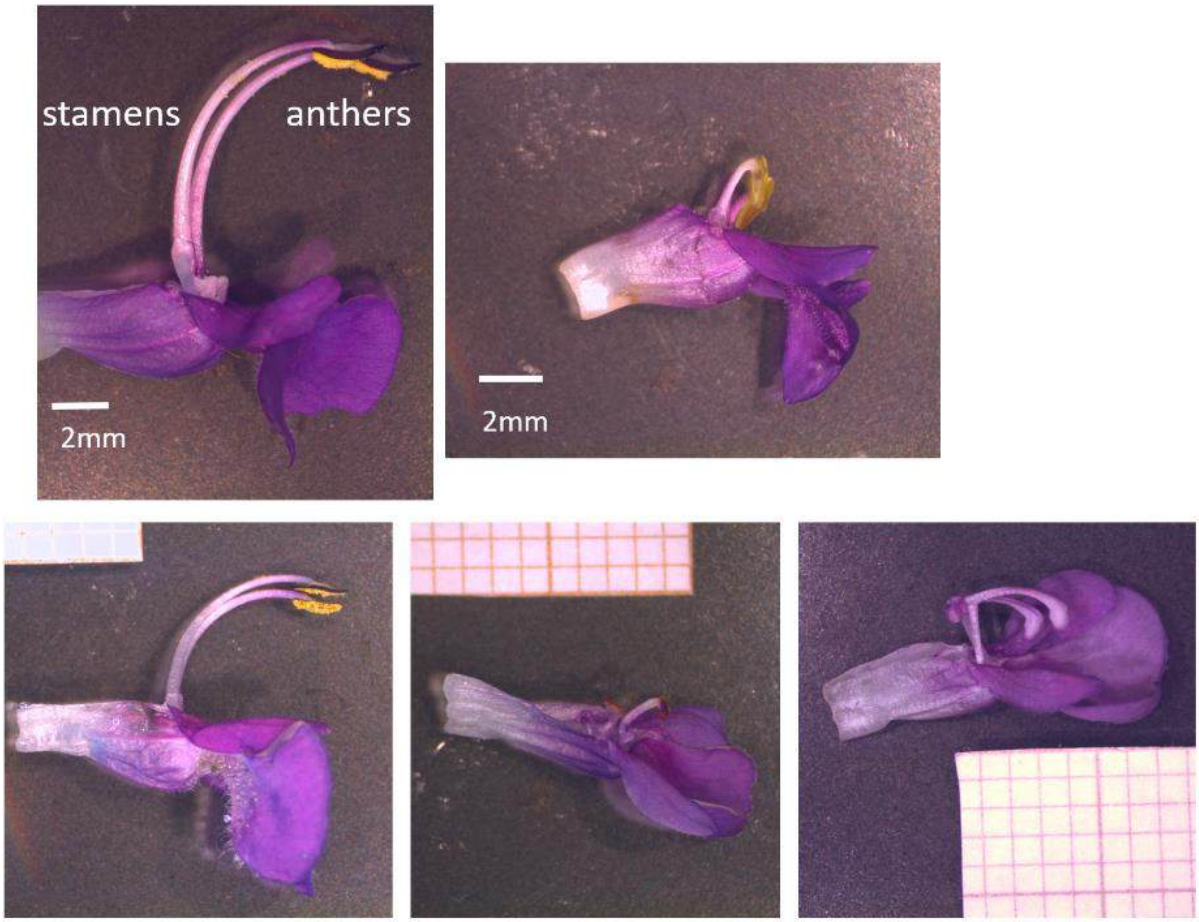


Figure 37. Flowers of *Salvia pratensis* with the upper petals removed to expose the stamens. Top: hermaphrodite and "short stamen" flower from a wild population (Inveruno, MI). Note the difference in length of the filament. Bottom: hermaphrodite flower of accession P9 grown in Città Studi; female flower of accession P11 grown in Città Studi; female flower with "petaloid" stamens grown in Brera.

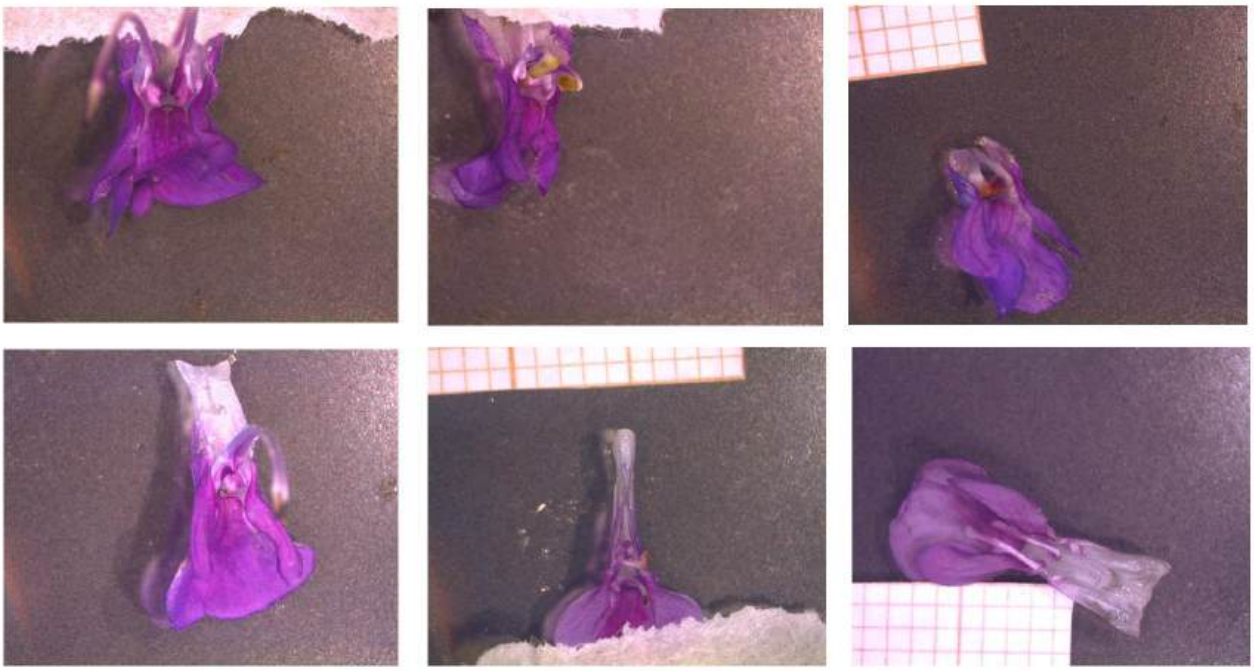


Figure 38. *Flowers of* Salvia pratensis *with upper petals removed, view from above. Top left, hermaphrodite, short stamen and female from Inveruno; hermaphorodite P9; female P11; female "petaloid" Brera.*

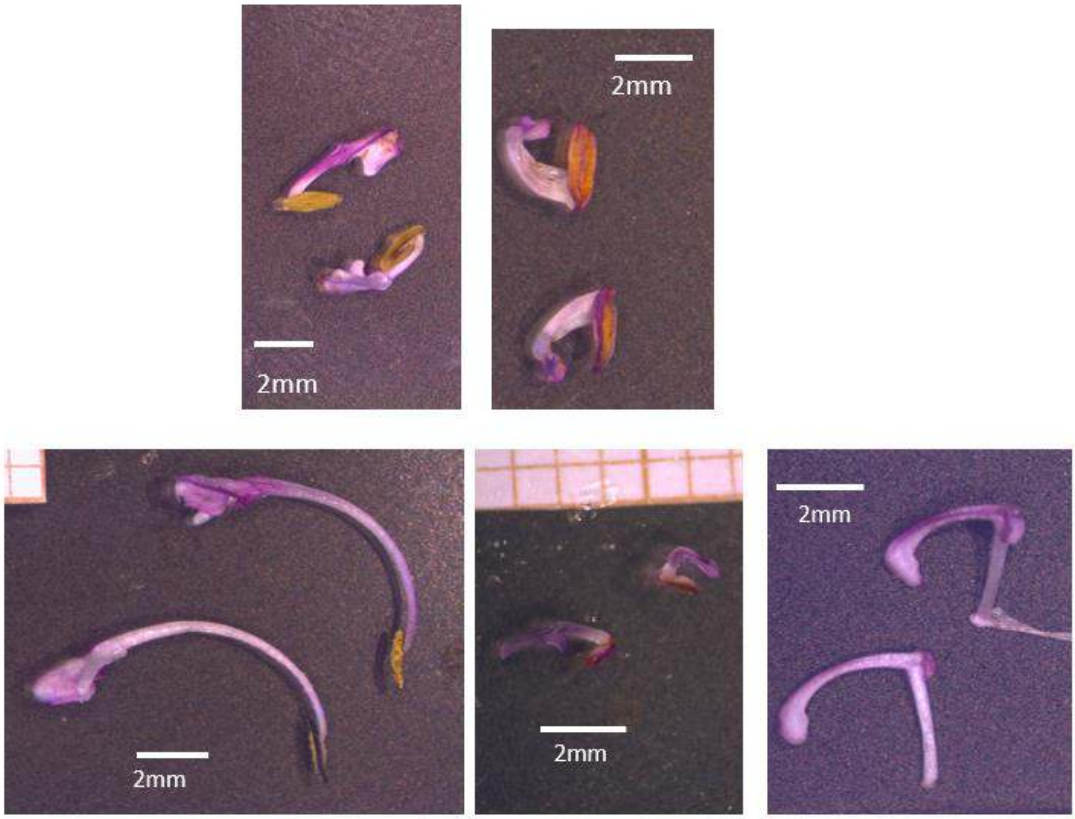


Figure 39. Stamens of *Salvia pratensis*. Top: "short stamen" and female from Inveruno (MI). Bottom: hermaphrodite P9, Female P11, "Petaloid" female P11.

		Number of plants			Percentage of plants		
Population	**Date**	**Hermaphr.**	**Short stamen**	**Female**	**Hermaphr.**	**Short stamen**	**Female**
Parco di Monza (MB), grown in CS	10/04/2024	7	6	0	53.85%	46.15%	0.00%
Suzzara (MN), grown in CS	12/04/2024	2	0	6	25.00%	0.00%	75.00%
Sassonia, grown in CS	12/04/2024	0	2	0	0.00%	100.00%	0.00%
Inveruno (MI), patch 1	13/04/2024	4	8	8	20.00%	40.00%	40.00%
Inveruno (MI), patch 2	13/04/2024	2	1	8	18.18%	9.09%	72.73%
Rasiglio, Monte San Pietro (BO)	29/04/2024	15	0	0	100.00%	0.00%	0.00%
Valle S. Felicita 1 (VI)	03/07/2024	20	8	2	66.67%	26.67%	6.67%
Tezze 5 (TN)	04/07/2024	31	1	2	91.18%	2.94%	5.88%
S Boldo 1 (TV)	04/07/2024	11	0	5	68.75%	0.00%	31.25%

Table 25. Sex ratios in different *S. pratensis* populations either grown in Città Studi (CS) from seeds collected in the wild, or from wild populations.

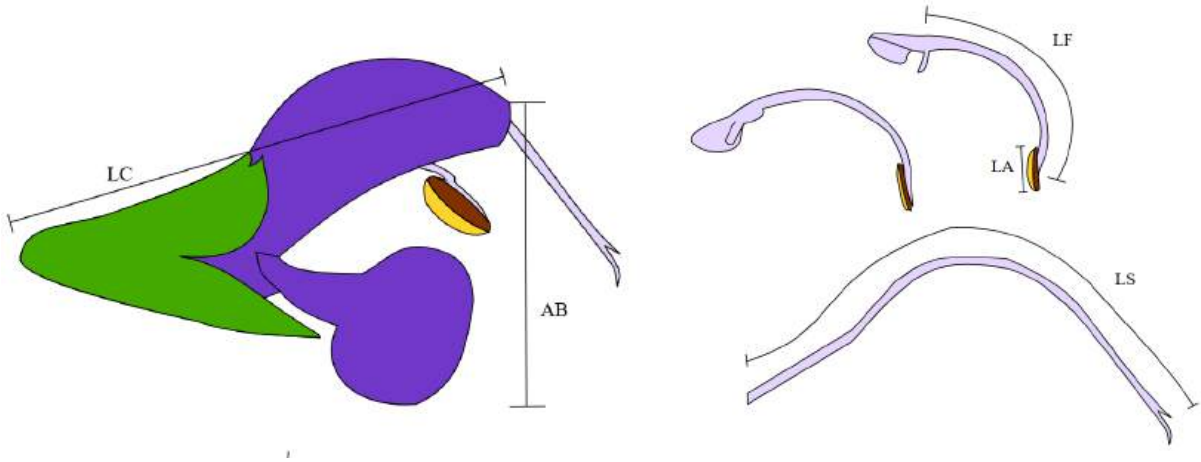


Figure 40. Measurements of flower morphology.

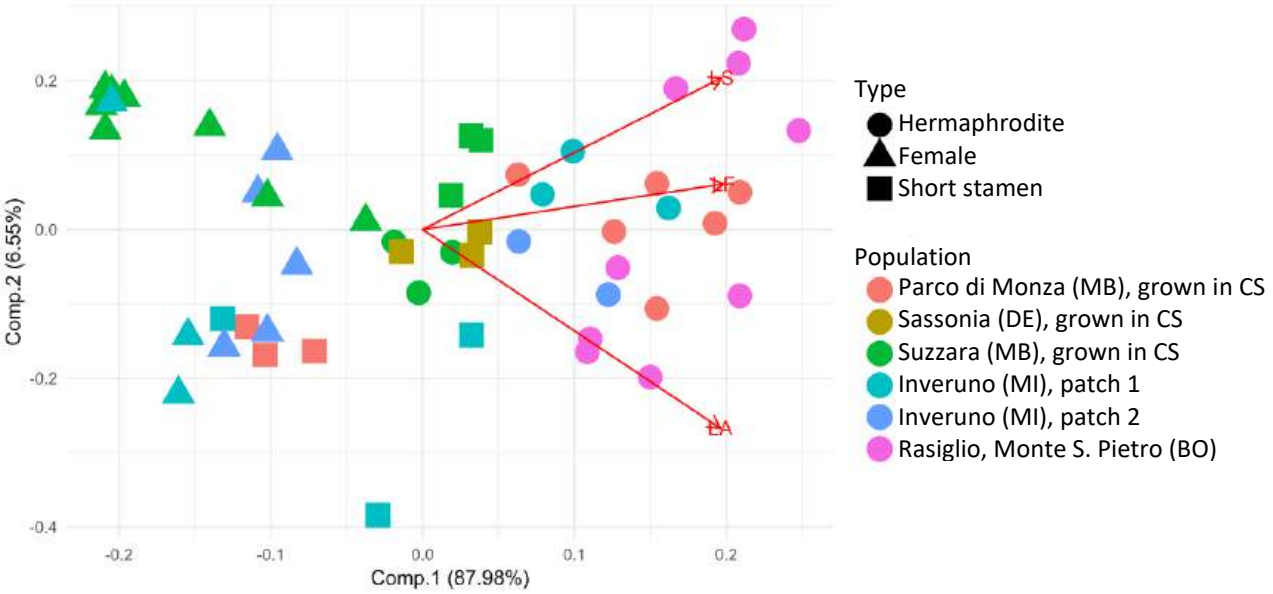


Figure 41. Principal component analysis (PCA) of *S. pratensis* flowers from different wild and cultivated populations. LF (filament length) is the variable that most contributes to the first principal component. Flowers appear to form a gradient, from the upper left to the bottom right, from female flowers (triangle) to hermaphrodites (circle).

For genome assembly, we initially chose plant P11_PI_fe from our collection based on higher leaf production. This is because a large amount of leaf material is needed for high molecular weight (HMW) extraction for long-read sequencing. During the first year, the plant did not produce flowers. In the second and following years, the plant produced exclusively female flowers, occasionally displaying a petaloid phenotype (**Figure 37**, **Figure 38**). This sparked our interest, and we started wondering whether this phenotype could be reconducted to the regular gynodioecious condition of *S. pratensis* or to a mutation in this specific individual, and in particular if this could be a homeotic mutant. While we were not able to answer this question completely, this also put the MADS-box genes on our radar and we decided to investigate their expression in more detail, thanks to the annotated genome assembly we achieved for *S. pratensis*. Ultimately, sequencing more individuals could also help investigate whether there are large-scale differences in the genome between the sequenced plant and other *S. pratensis*.

Figure 42. Sequenced *Salvia pratensis* individual (P11_PI_fe) with detail of female flowers.

2.2. MADS-box family annotation and expression

Given our interest in flower development and male sterility, we focused our attention on the most well-known players in the ABCE genetic model of flower development, the MADS-box family of transcription factors. We used the HMM model of the SRF-TF DNA binding site to scan the set of predicted proteins from the BRAKER3-Augustus annotation. We then confirmed the presence and conservation of protein domains with MEME (**Figure 44**) and CDD (**Figure 45**), which respectively identify shared motifs in the set of *S. pratensis* MADS candidates and identify conserved domains compared to known protein domains. With this, the initial set of 74 candidates was narrowed down to 68 (**Figure 43**, **Figure 47**).

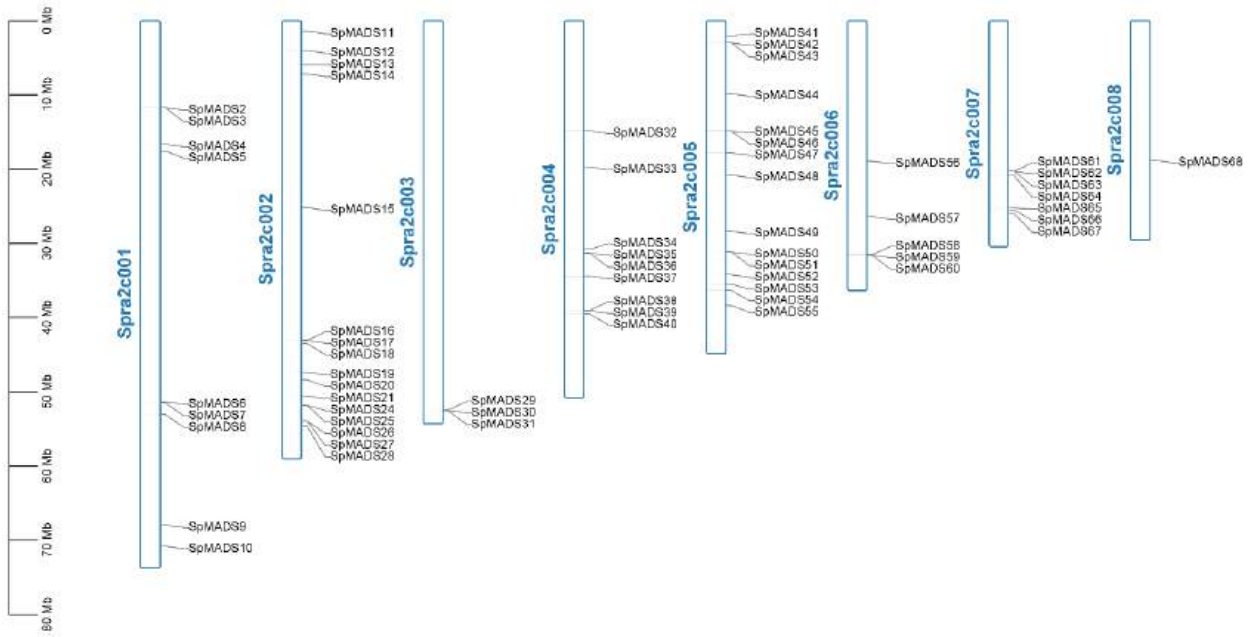


Figure 43. Position of the *SpMADS* genes on the Spra2c reference. Only contigs 1-8 are represented as no *SpMADS* are found on contigs 9-19.

We performed whole-protein and MADS-box domain only alignments between the *S. pratensis* candidates and the *Arabidopsis thaliana* known MADS-box protein sequences (**Figure 46**). In *A. thaliana*, MADS-box genes are generally split into two groups: Type I MADS, including the Mα, Mβ and Mγ subfamilies, and Type II MADS, including the Mδ (or MIKC*) and MIKC (or MIKC^{C}) subfamilies. The Type II MADS have been reported to generally evolve via WGD events and are more conserved between different species, while Type I evolve via segmental duplication (Nam et al., 2004) and have much higher birth/death rates. This can also be seen in the phylogenetic analysis of the *S. pratensis* and *A. thaliana* MADS-domain genes; in the MIKC and Mδ well-defined clades can be identified that support orthology between specific *A. thaliana* and *S. pratensis* genes, whereas clades in the Type I subclasses disproportionately include genes from a singular species, indicative of gene duplication after speciation. Additionally, Type I MADS are mostly mono-exonic, and UTR annotation could not be achieved, possibly due to the generally lower levels of expression in the RNAseq data, while Type II MADS show a mean of 10 exons per gene.

Subfamily Mβ was polyphyletic in the *S. pratensis* – *A. thaliana* MADS phylogeny, but with poor bootstrap support. The *S. pratensis* Mβ proteins all form a clade with AGL47 and AGL82, while the other 20 *A. thaliana* Mβ proteins form a clade with the Mγ subfamily. However, this part of the tree is poorly resolved.

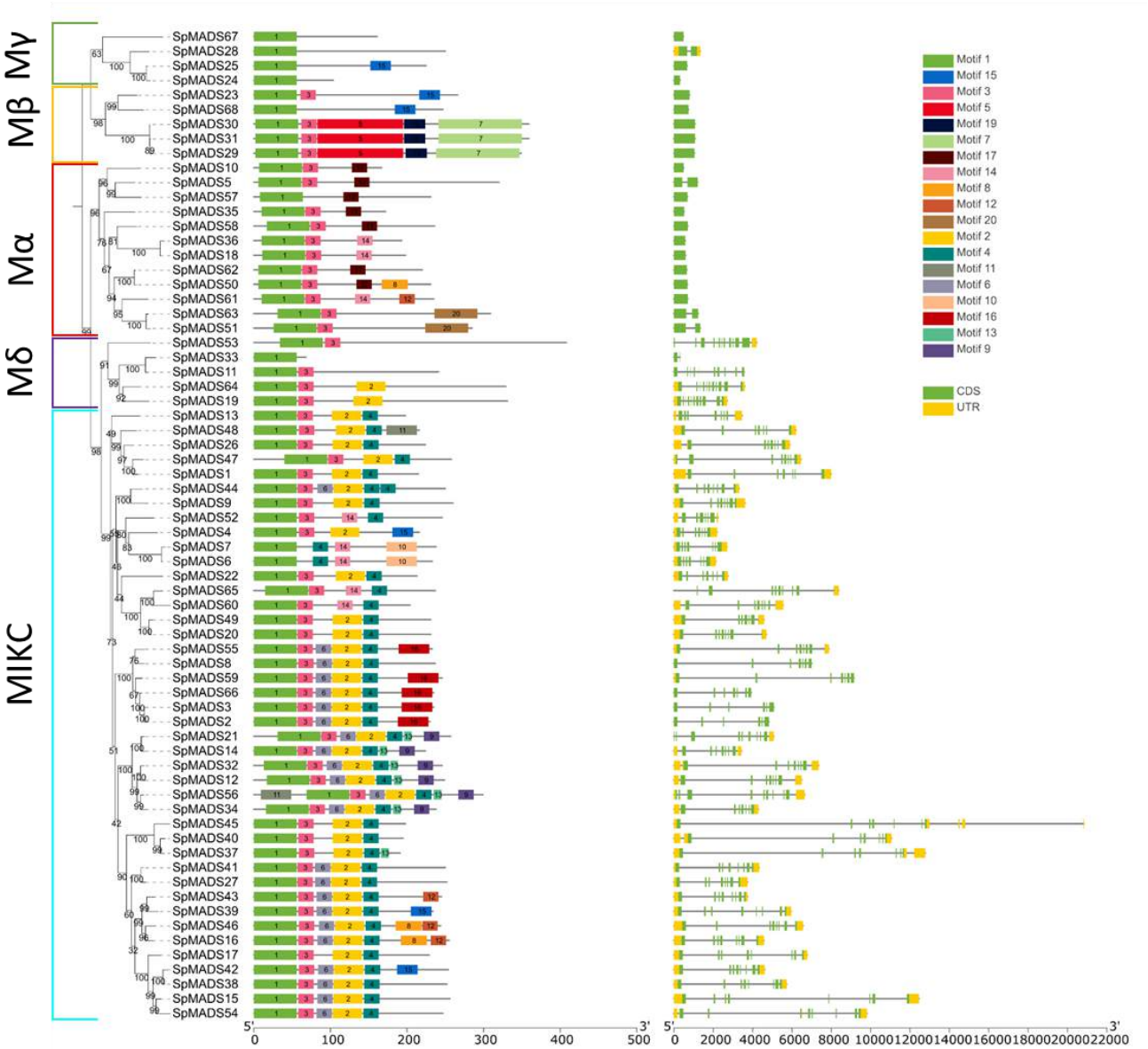


Figure 44. Conserved domain analysis performed with MEME on the predicted protein sequences (left) and intron-exon structure (right) of the 68 *SpMADS* genes. The differences in the gene structure of Type I MADS (Mα, Mβ, Mγ), mostly monoexonic, and Type II MADS (MIKC and Mδ), with several introns and exons, is evident, as is the presence of the K-box (domain 2, in yellow).

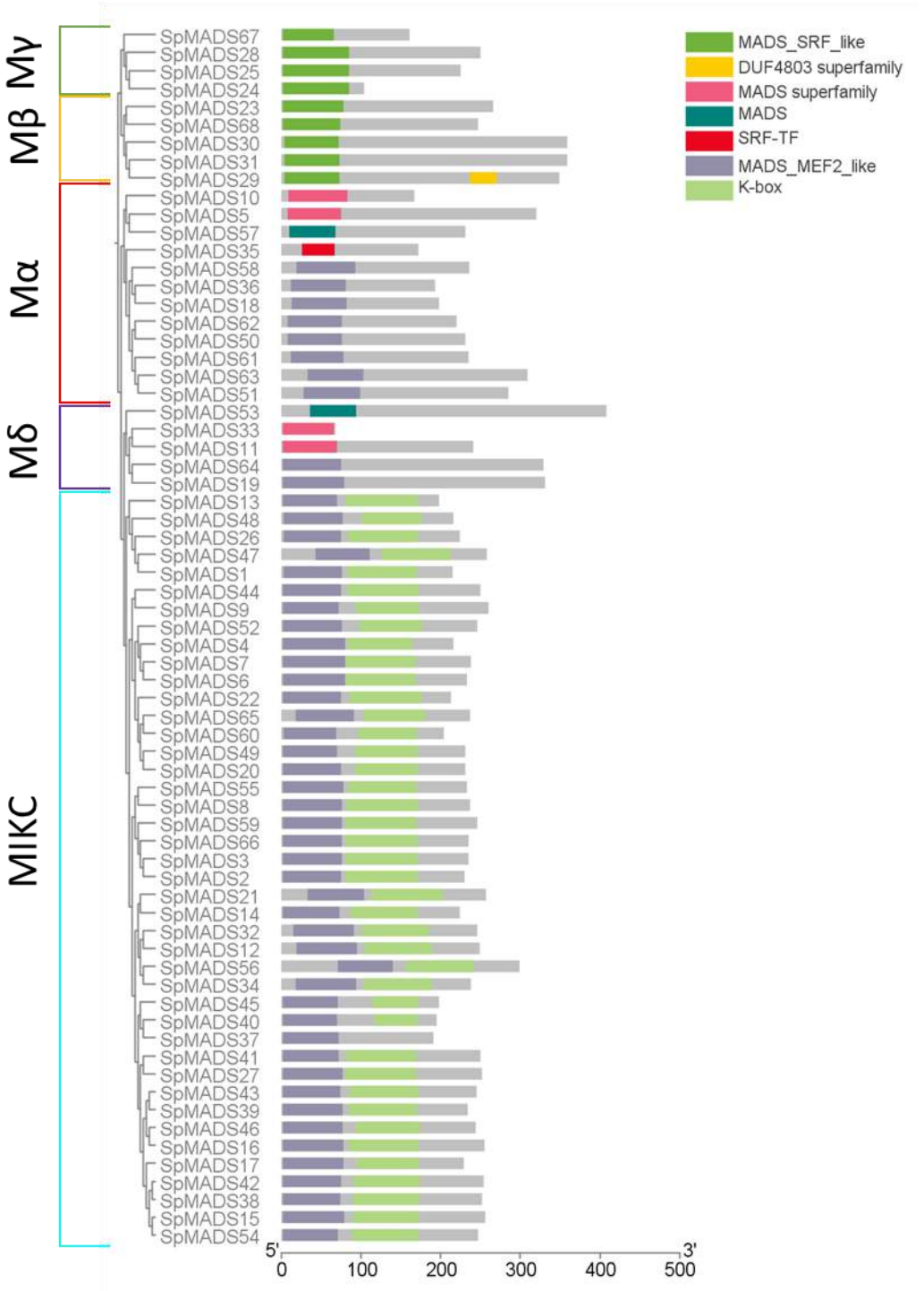


Figure 45. NCBI Conserved Domain Database (CDD) search of the predicted protein sequences for the 68 *SpMADS* genes. The MADS_MEF2_like, MADS_SRF_like and K-box domains are prevalent.

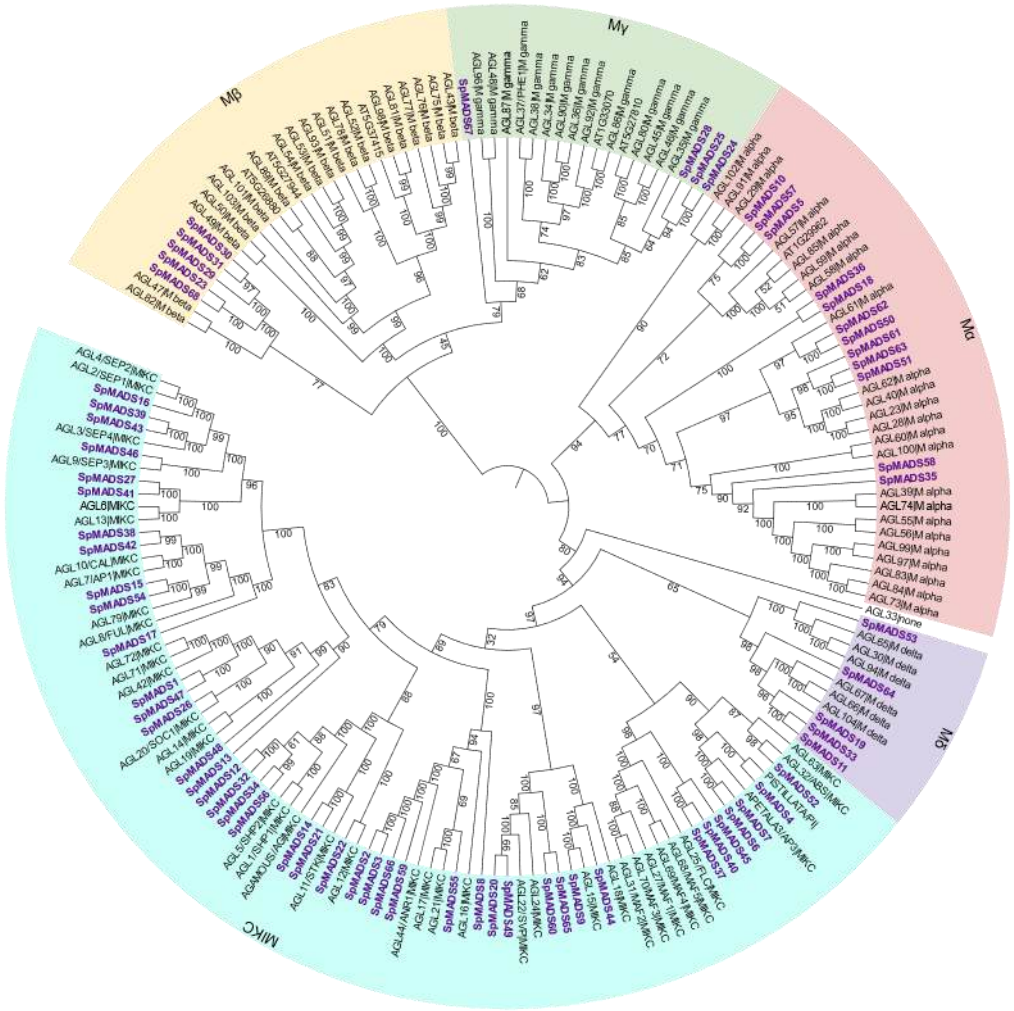


Figure 46. Cladogram of the SpMADS proteins with the *A. thaliana* MADS-box transcription factors. The 5 subfamilies of Parenicova et al, 2003 are highlighted. The tree was derived from full length protein alignments and bootstrap results are shown. *SpMADS* genes are highlighted with a bold and violet font.

To give further support to the identity of these genes, we analysed their presence in the transcriptomic data available. While the data is not suitable for a quantitative gene expression analysis or a comparative analysis, given the lack of replicates, we chose a qualitative approach and classified the transcripts in 5 discrete categories according to their abundance in the different tissue samples (from “Not found” to “Highly abundant”; **Figure 47**). We found abundant expression of most of the MIKC genes in all flowering stages and low expression of the Type I genes as expected. We also identified some genes with high expression in the root and seedling samples, which could have other roles during development. Ultimately, it will be interesting to perform a differential expression analysis on several female and hermaphrodite plants at early stages of flower development to check whether large changes in the transcriptomic profiles are linked to the different flower morphotypes. Alternatively, quantitative PCRs on specifically the MADS-box could also provide some answers.

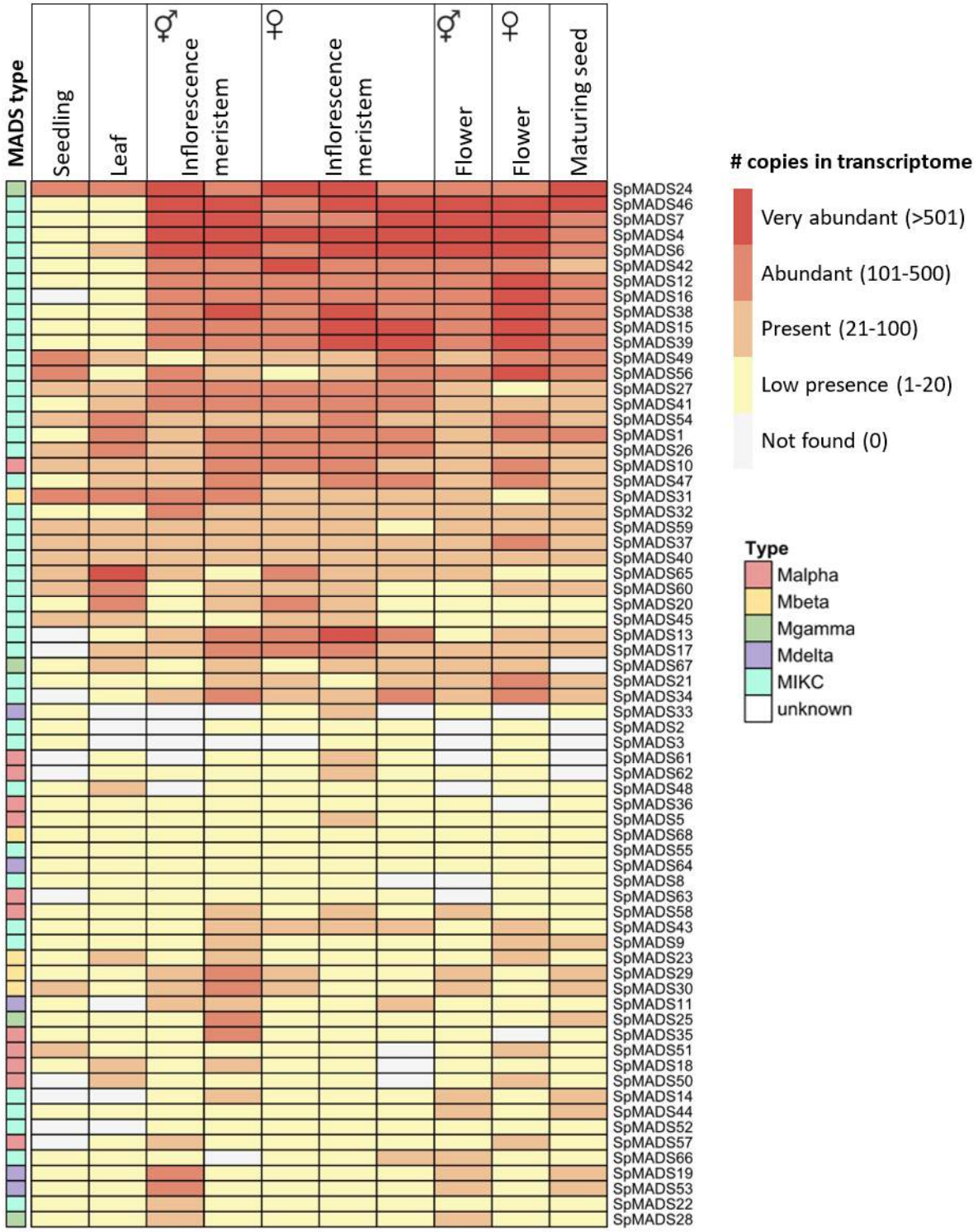


Figure 47. Qualitative transcription profiling of the 68 *SpMADS* in RNA-seq data from 10 samples of *Salvia pratensis* from 5 different tissue types. The transcripts were quantified, normalized by library size and transcript length, and represented as a qualitative variable with 5 different levels (see legend on the right). The subfamily of each gene is reported on the left and the sex of the plant used for sequencing (female or hermaphrodite) at the top.

To perform a comparative analysis, we analysed the genomes of *S. officinalis* and *S. miltiorrhiza* and found respectively 72 and 68 MADS-box genes respectively. Consistently with other plants, the number of MIKC genes is conserved, while the number of Type I genes is much more variable and is strongly reduced in *Salvia.* We then further focused on specific genes of the ABCE model and identified orthologs of the following *A. thaliana* genes: APETALA3, PISTILLATA, AGAMOUS and SEEDSTICK. We reason that given our anomalous, petaloid phenotype in some flowers, it is possible that AGAMOUS could be involved. This is because mutations in the AG have been known for a long time to generate homeotic mutants in *A. thaliana* with extra petals and no stamens. For several of these, we identified two copies in *Salvia* for each *A. thaliana* genes, and often one of the copies is more diverging while the other is more conserved. This is possibly caused by either functional redundancy, meaning, one of the copies is no longer needed, or subfunctionalization, such that one of the copies assumes a slightly different role in development.

A detailed description of the identified orthologs is available in: Paleni et al. "Mining the *Salvia pratensis* genome to study the MADS-box gene family and its role in flower development" (attached).

3. Species interrelationships among Italian *Salvia*

We investigated the phylogeny of Italian *Salvia* by characterizing a collection of leaf samples, coming from wild populations, with the Genotyping-By-Sequencing (GBS) method. We evaluated the suitability of the method on a test dataset and we then proceeded with analysis of a larger dataset. This data allowed us to perform phylogenetic and species delimitation analysis to investigate separation between *S. pratensis* and endemic cryptic species and to evaluate the level and distribution of genetic diversity within and among populations of *S. pratensis*.

3.1. Sampling campaign

The sampling campaign, led by the Botanical Garden of Brera, was highly successful and resulted in a collection of 1,227 samples over 18 Italian regions and 16 taxa in 220 populations (**Figure 48** and **Materials & Methods**). Additionally, 4 samples from Germany were added. Contributors (**Table 26**) received paper or plastic bags with silica gel; they were instructed to sample plants with at least a 2m distance between them, to avoid collecting leaves from the same genotype, and report in an Excel file a series of mandatory (species, date, coordinates) and optional metadata associated to the sample. To organize the collection, all data was aggregated in excel files and analysed in R. Each population and sample were given a unique alphanumeric code to allow identification of the sample throughout all library preparation steps. The code combines a species code, a location code and two sequential numbers for population and sample (see **Materials & Methods** section). To integrate the collection with endemic species *S. saccardiana*, an additional campaign was conducted in July 2024 (see **Materials & Methods**). Finally, we reached out to Boštjan Surina from the Natural History Museum Rijeka (Croatia), who has worked on the *S. pratensis* group (Balant et al., 2019), and he provided samples of *S. saccardiana* and *S. bertolonii* from Croatia.

Of all collected samples, 155 have already been analysed with GBS (**Figure 51**) and an additional 96 are being sequenced at the time of writing. The vast majority of samples come from the most common species, *S. pratensis*, *S. glutinosa* and *S. verbenaca* (**Figure 49**). The overrepresentation of *S. pratensis* comes from both its frequency and our specific focus on this group; particularly *S. saccardiana* which was subject to a more intensive sampling campaign. The overrepresentation of northern Italian regions, and in particular Lombardy, in our sampling (**Figure 50**) is likely a result of sampling bias due to our location, as it was more likely to end up contacting and obtaining samples from people in our general area. Importantly, the sequenced samples of *S. saccardiana* were identified as *S. pratensis* subsp. *saccardiana* by the collector and as such their code was PR.VI.2 and they appear within *S. pratensis* for some of the plots.

3.2. GBS library preparation

DNA extraction and GBS library preparation required optimization. Two major factors impacted the protocol success: a) the nature of silica-dried material, which is less optimally conserved than other methods (i.e. LN2); b) the aromaticity of Salvia, which translates to presence of specialized metabolites in the leaves, which

complicate both extraction and digestion. Details on the optimization of the protocols are available in the **Materials & Methods** section (**Optimization of DNA extraction from silica-dried leaf for GBS, page 111**; **GBS library preparation, page 114**). We then assessed the success of the protocol by evaluating the sequencing data: number of reads, mapping rate, number and coverage of sequenced GBS sites, and finally, number and quality of markers (single nucleotide polymorphisms or SNPs).

Figure 48. Summary map of retrieved samples from different *Salvia* species. One circle is plotted for each combination of species and province. The size of the dot corresponds to the number of samples from that species in the corresponding province.

Alberto Colatore, Andrea Mologni, Andrea Pompa, Andrea Stucchi, Andrea Truzzi, Antonio Pica, Beppe di Gregorio, Boštjan Surina, Camilla Pieri, Cancellieri Laura, Carla Lambertini, Carolina Elena Girometta, Chiara Paleni, Colleoni Gianmaria, D. Saulle, Domenico Gargano, Donato Palermo, Elisa Pelagani, Emanuela Carli, Fabrizio Bonali, Fabrizio Buldrini, Filippo Prosser, Francesco Santi, G. Pazienza, Giacomo Cangelmi, Giorgio Perazza, Giulia Tomasi, Goffredo Filibeck, I. Bonini, L. Forte, L. Montoni, Larissa de Paola, Leopoldo de Simone, Lisa Brancaleoni, Mar. Ord. Antonio Zoccola, Marco Pieri, Marco Valecic, Maria Speranza, Mariacristina Villani, Mats Gustafsson, Mattia Pallanza, Mauro Manca, P. Castagnini, Riccardo Copiz, S. Cannucci, Silvio Scortegagna, Simon Pierce, Ten. Vito Fazio, Thomas Abeli, Željka Modrić Surina	
Dipartimento di Scienze della Terra e dell'Ambiente, Università degli Studi di Pavia, Fondazione Museo Civico di Rovereto, Museo Orto Botanico Università degli Studi di Bari "Aldo Moro", Orto Botanico di Bologna, Orto Botanico di Padova, Orto Botanico di Siena, Orto Botanico di Urbino, Parco Pineta di Appiano Gentile e Tradate, Parco Regionale Campo dei Fiori, Parco Regionale dei Colli di Bergamo, Reparti Carabinieri Biodiversità di Vallombrosa e Pratovecchio , Servizio Biodiversità della Regione Friuli-Venezia Giulia, Giardino Botanico Carsiana, Università degli Studi della Tuscia, Università degli Studi di Ferrara, Università degli Studi di Milano, Università degli Studi di Siena, Università degli Studi Roma Tre, Università della Calabria, Università di Bologna, Prirodoslovni Muzej Rijeka	

Table 26. List of individuals and institutions that contributed to the collection.

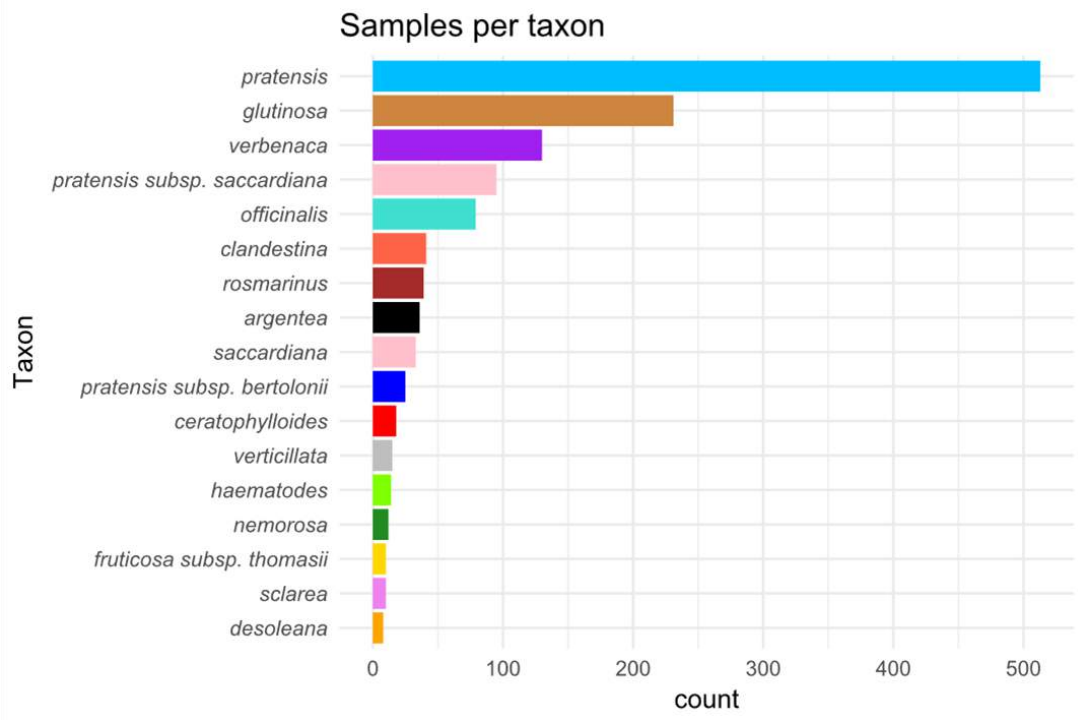


Figure 49. Statistics of the *Salvia* collection. Number of collected samples for each taxonomic unit.

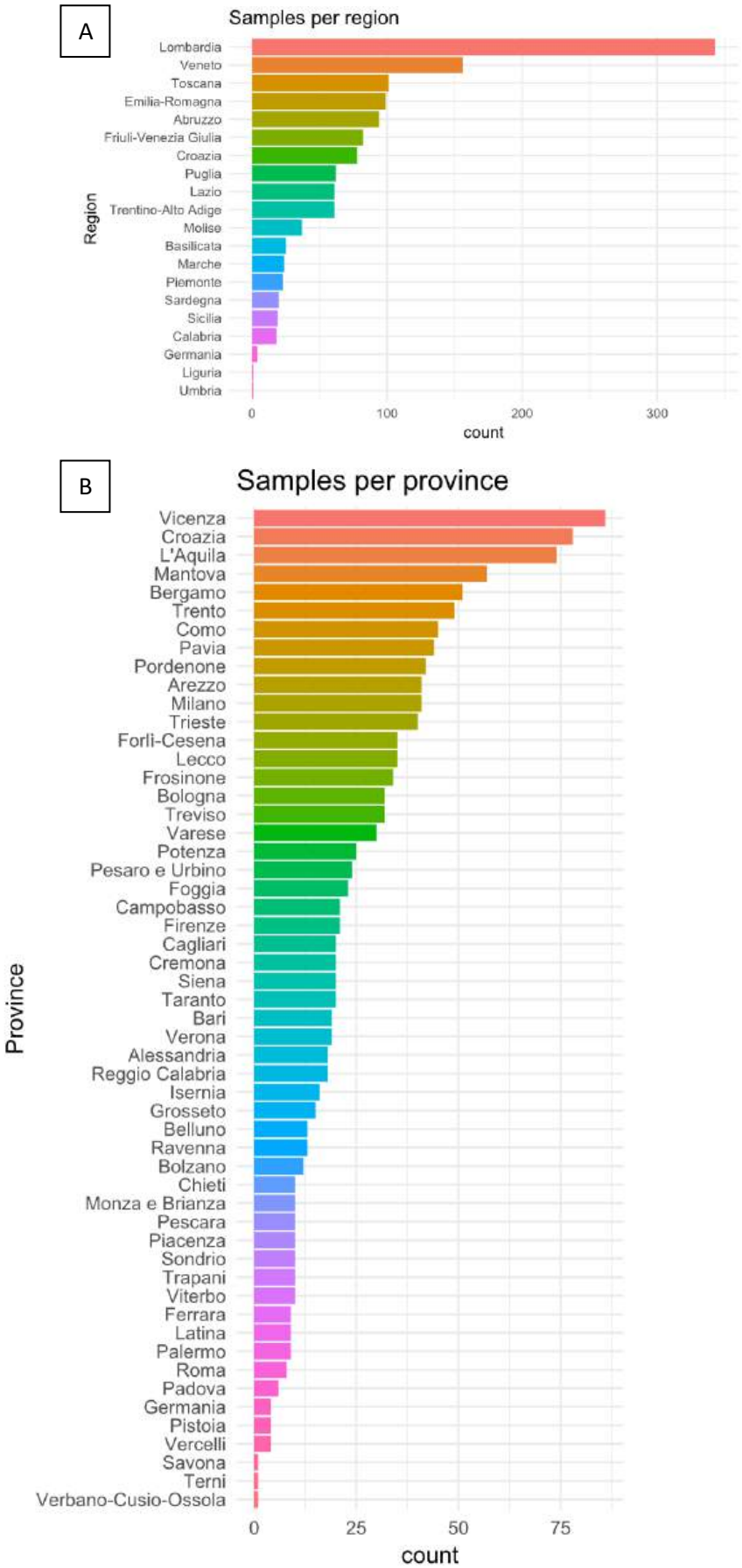


Figure 50. Statistics of the *Salvia* collection. Number of collected samples from different administrative regions of Italy. Region and province correspond to Italy's first and second level administrative regions respectively, plus Germany and Croatia.

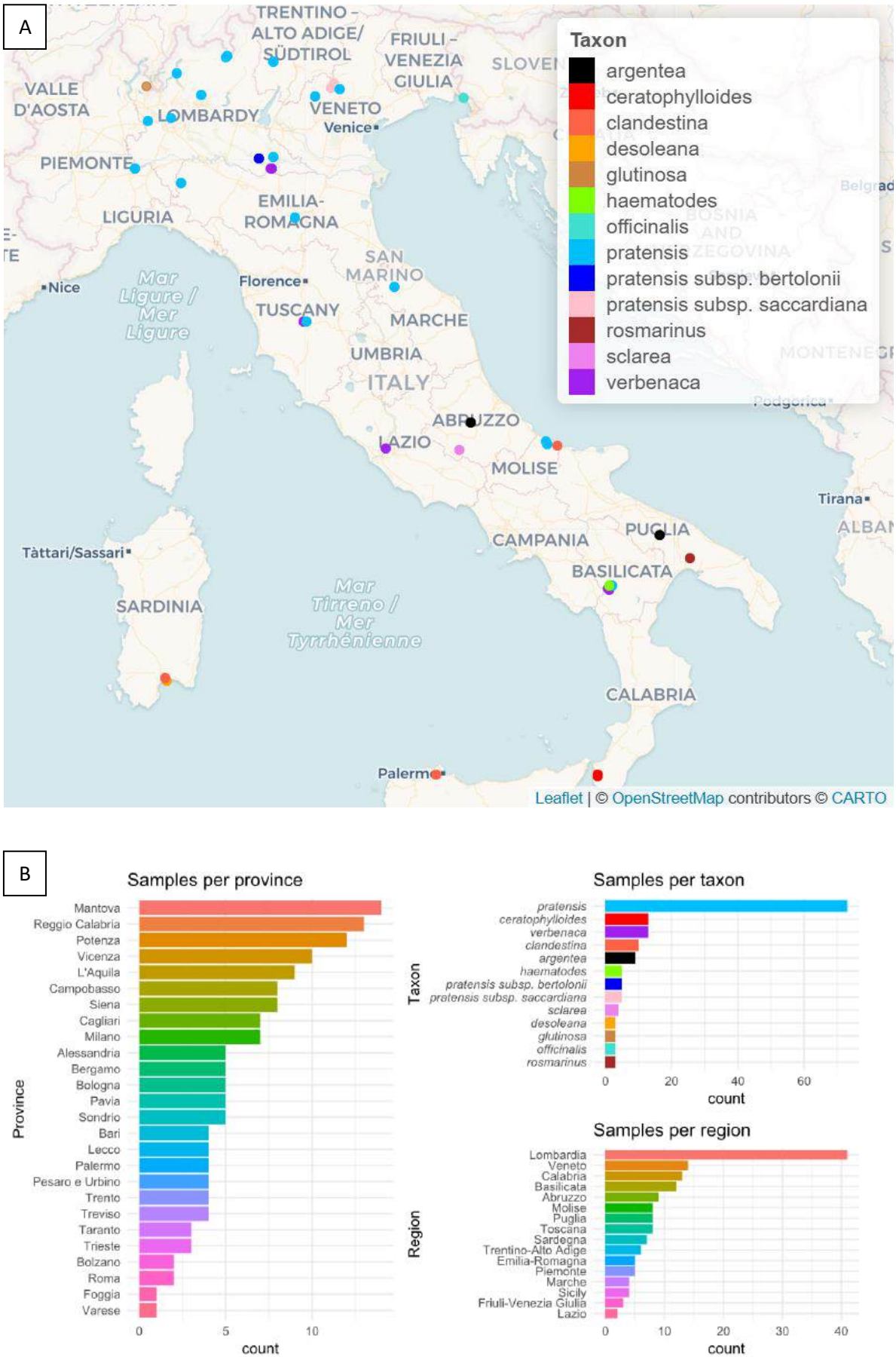


Figure 51. Map (***a***) and statistics (***b***) of sequenced samples from the *Salvia* collection.

3.3. Sequencing data evaluation

From the test library for the GBS protocol, 1.2M reads and 378 Gb of data were obtained, with a mean quality score of 36.6. From the second round of sequencing, 1.3M reads and 396 Gb of data with a mean quality of 36.55 were obtained. After receiving the raw sequencing data from the test library, we first used the inline barcode for demultiplexing, meaning, divide the sequencing reads in different files corresponding to the different samples. We then optimized the read trimming and filtering steps to adapt to this specific library structure (details in **Materials & Methods; Sequencing, read quality evaluation and trimming**). The reads were then subject to a filtering step to remove low quality, unpaired, or adapter-contaminated reads. Around 25% of the reads were lost in this process (**Figure 52**, **Figure 53**): this may be due to our impossibility to perform size selection on the sequencing library due to lack of instrumentation which was initially required for the GBS protocol, causing the presence of very short fragments which translate to short reads after adapter trimming.

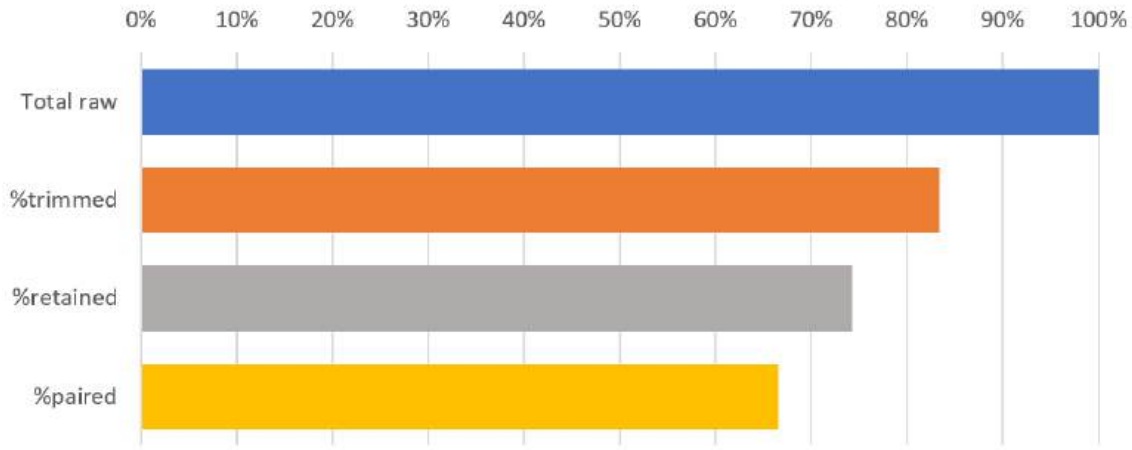


Figure 52. Overview of number of reads throughout GBS trimming and filtering steps on the test library (POOL 1).

- Total raw: demultiplexed reads.
- % trimmed: reads after quality trimming (only paired reads are kept for trimming)
- % retained: reads after quality filtering
- % paired: percentage of retained reads which are also paired. These reads will be kept for the following analyses (mapping).

The read mapping rate was highly dependent on the evolutionary distance of each *Salvia* species to the reference *S. pratensis* (**Figure 54**), with the closest species reaching above 80% average mapping rate and the most distant, *Salvia glutinosa*, reaching a little above 40%. The quantity of reads per sample and the mean sequencing depth (figure) was sufficient for accurate genotyping at a good number of GBS sites and ultimately only two samples were dropped out of the analysis due to low sequencing depth. With the second round of sequencing we reached a total of 155 samples from 12 taxonomic units. This round performed similarly to the first sequencing round in read number and quality. Overall, only 9 samples in total were removed from downstream analysis due to low sequencing depth. A table of all sequenced samples and their metadata, along with information of all the bioinformatics steps from demultiplexing to mapping is available online: https://unimibox.unimi.it/index.php/s/QE7ixdrimpoZGZd.

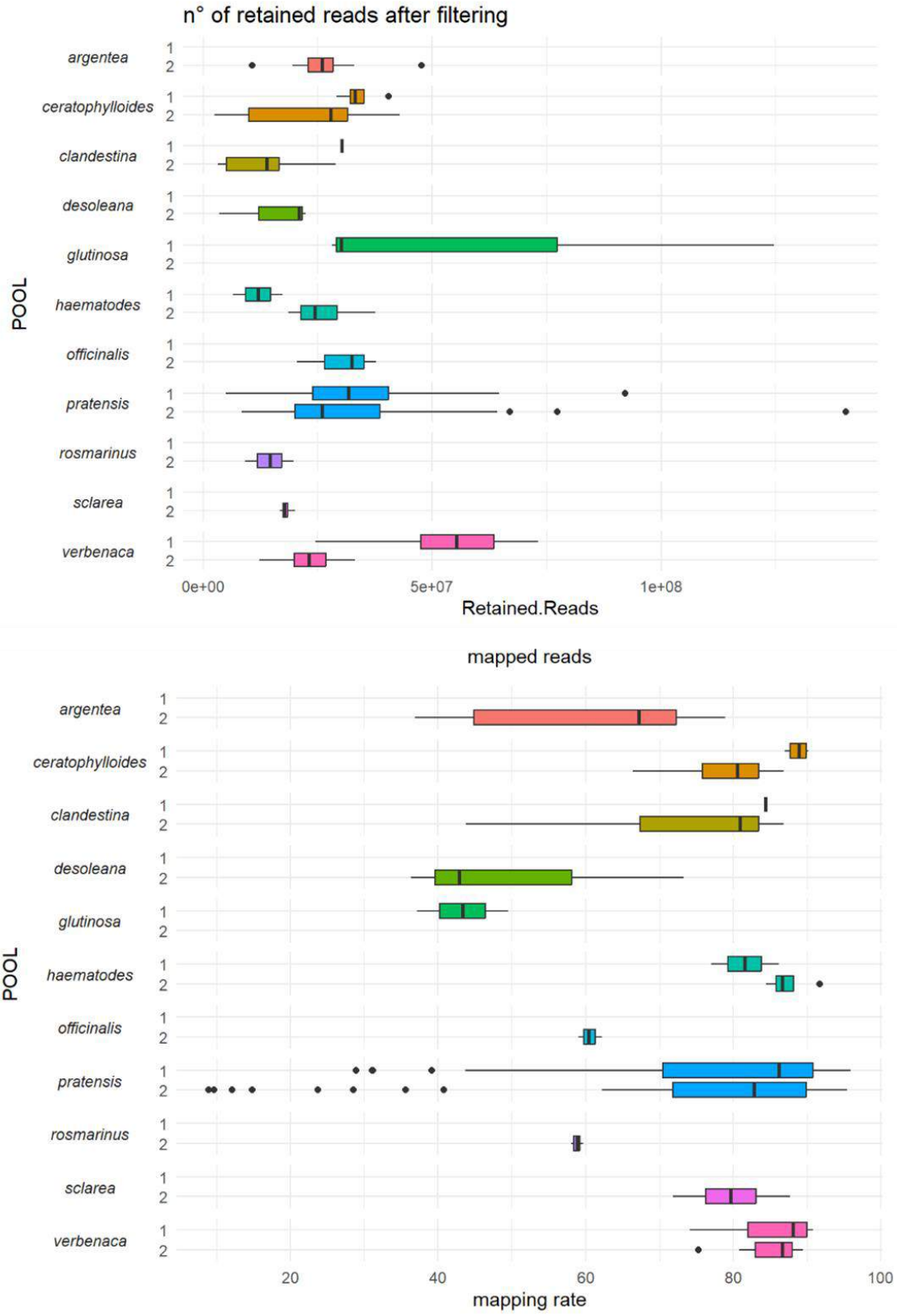


Figure 53. Boxplot of reads per sample (top) and mapping rate (bottom) of different *Salvia* species throughout the two rounds of sequencing GBS libraries. The samples are split by sample and sequencing pool to highlight the effects of evolutionary distance from the reference genome and check whether there are batch effects.

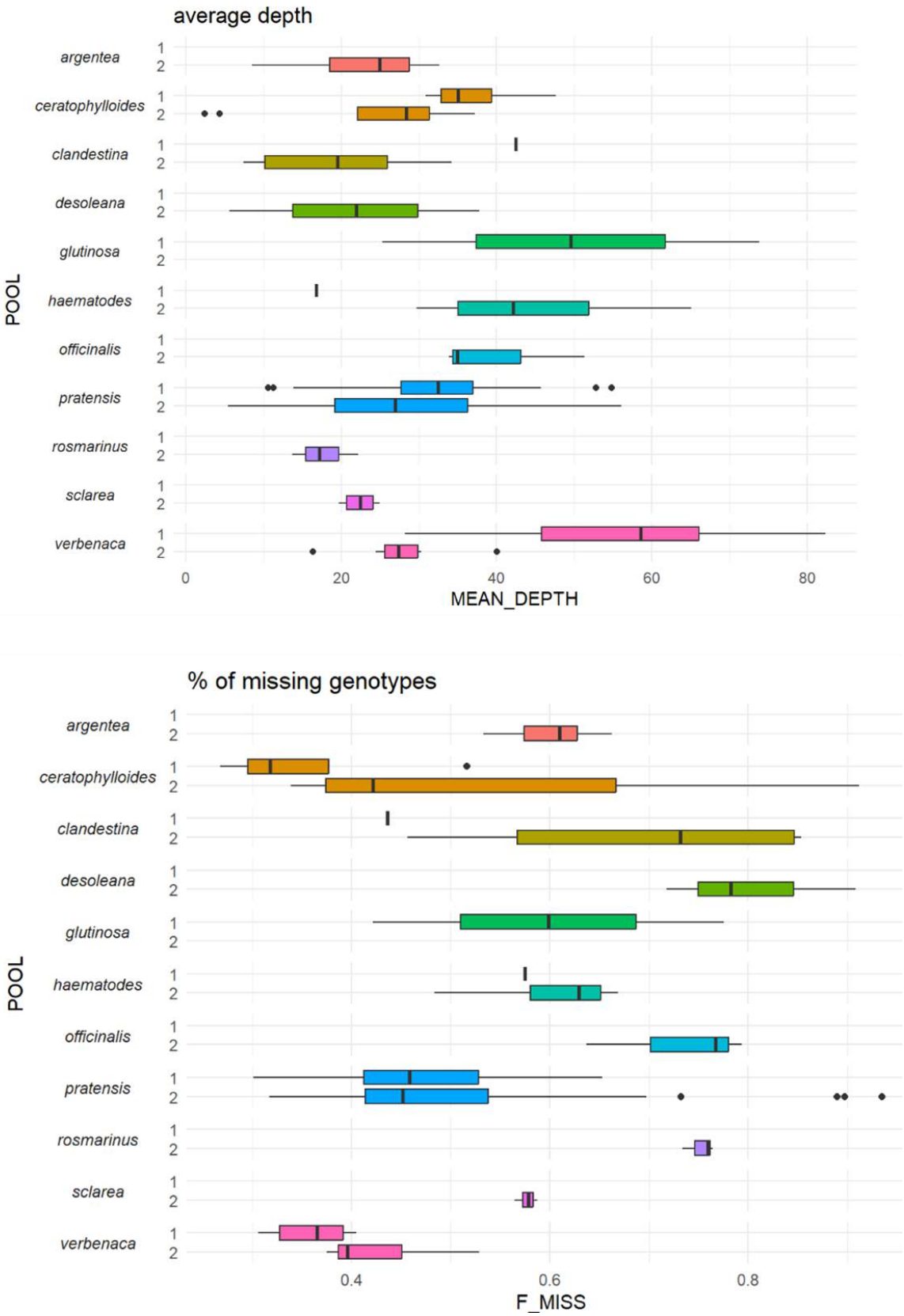


Figure 54. Average sequencing depth (top) and frequency of missing genotypes (bottom) of different *Salvia* species throughout the two rounds of sequencing the GBS libraries.

Variant calling resulted in more than 10M SNPs calls over 1.6M GBS sites (**Figure 55**). After removing missing data, removing polyallelic sites, and applying quality filtering, we obtained 28,558 variant sites. The sites were filtered based on distance to only keep SNPs with a minimum distance of 1kb and resulted in a set of 2,429 variant sites (**Figure 56**).

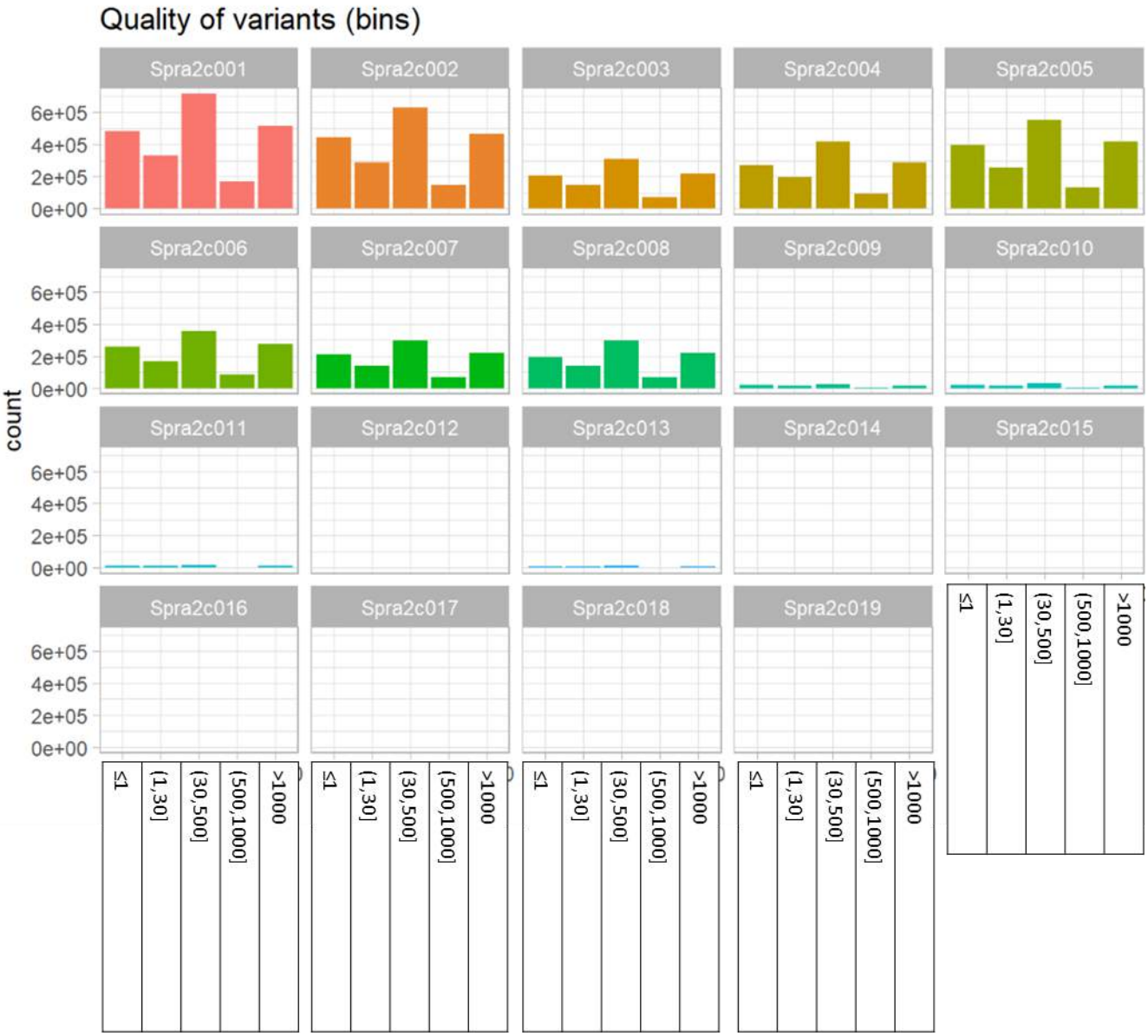
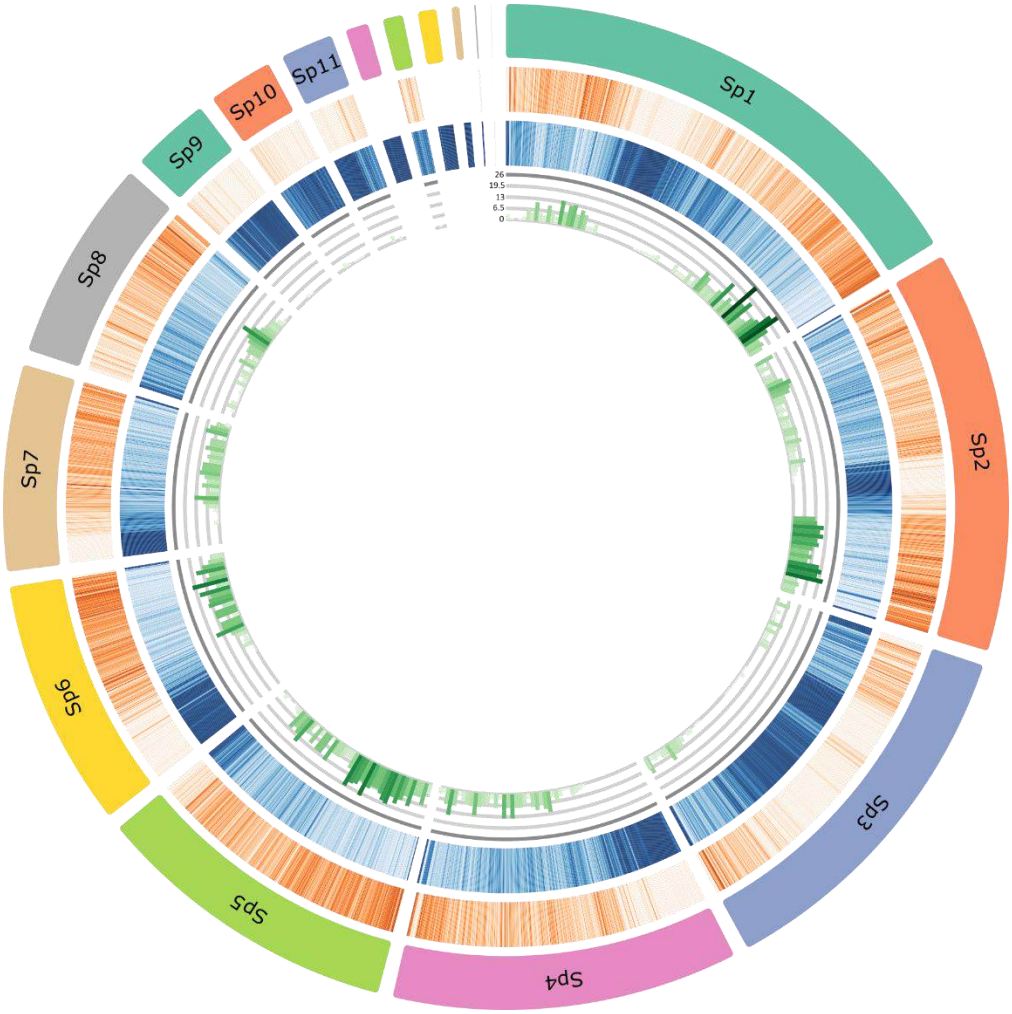


Figure 55. Binned quality of called variants, grouped by contig. Quality distribution seems to not be dependent on the contig. Most of the variants were identified in contigs 1-8 (the longest ones).



Contig	Length	N° variants	Variants/Mb
Spra2c001	73,767,105	479	6.49
Spra2c002	59,079,562	463	7.84
Spra2c003	54,364,029	136	2.50
Spra2c004	50,850,824	250	4.92
Spra2c005	44,911,161	456	10.15
Spra2c006	36,361,082	263	7.23
Spra2c007	30,512,240	196	6.42
Spra2c008	29,618,631	168	5.67
Spra2c009	11,692,271	3	0.26
Spra2c010	9,752,273	8	0.82
Spra2c011	7,735,308	6	0.78
Spra2c013	2,990,901	1	0.33
Total		2,429	

Figure 56. Position of the 2,429 SNPs selected for analysis along the Spra2c reference. Outer layer: contigs. Outer heatmap (orange): gene density. Inner heatmap (blue): repeat density. Inner histogram (green): variant density in 1Mb bins. The table lists contigs, their length, the number of variants placed on a contig out of the selected 2,429, and the average amount of variants per Mb. Only contigs with at least one selected variant are reported.
SNP calling and filtering tends to filter out regions with exceedingly high coverage: thus, selected SNPs are scarce in highly repetitive regions. Contigs 1-8, which are the longest, have an average variant density of 6.40. Spra2c003 has a much lower amount of contigs, possibly due to the high fraction of it covered by repetitive content. By contrast Spra2c005 has the highest variant density (10 variants per Mb).

3.4. Phylogenetic relationships and ploidy of Italian *Salvia*

We first analysed SNP data to infer ploidy based on the frequency of alternative alleles in the read set (**Figure 57**). In *S. pratensis*, the major peak is at 50%. *S. ceratophylloides* has a clear peak at frequency 25%, which is a strong indication for tetraploidy or even octoploidy. *S. verbenaca*, instead, shows a tall peak at 33%. This suggest that the variant is present in 1/3rd of the chromosomes, indicating a condition of triploidy or hexaploidy. All other species have a peak at 50%.

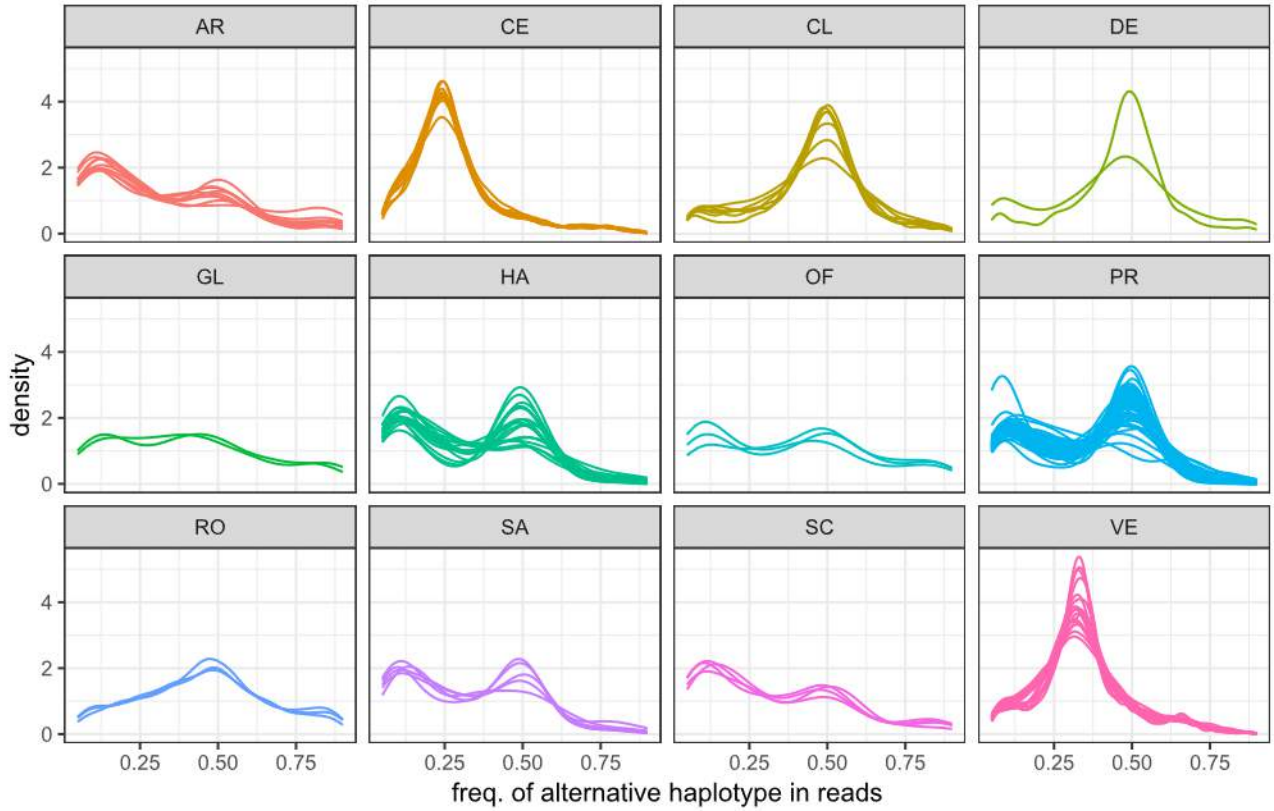


Figure 57. Alternative allele frequency in the read set of GBS *Salvia* samples, divided by taxon. The leftmost peaks at around 0.12 are the "tail" of the distribution centered around 0 (=positions with two reference alleles), which was cut.

We analysed the genetic data with Principal Component Analysis (PCA) to check whether samples from the same species or population would cluster together according to their genetic makeup. As we are comparing species with different ploidies, variant calling was performed assuming diploidy for all samples. The PCA shows that the samples were split in three major groups on the PC1-PC2 plot (**Figure 58a**). The first group includes *S. glutinosa*, *S. rosmarinus*, *S. argentea* , *S. sclarea* , *S. desoleana* and *S. officinalis*. The second includes *S. verbenaca*, *S. clandestina* and *S. ceratophylloides*. The third includes *S. pratensis*, *S. saccardiana* and *S. haematodes.* PC1 explains 23% of the variance, while PC2 explains 18%. Overall, the different species can be clearly distinguished as separate clusters in the plot. The exceptions are three samples of *S. pratensis*, which cluster with *S. verbenaca*, and were observed on further inspection to have possibly been misidentified during sampling, as their leaf shape was much more similar to *S. verbenaca*. Curiously, *S. ceratophylloides* is much closer to *S. verbenaca* and *S. clandestina* than *S. pratensis*, despite being traditionally associated to the *S. pratensis* group. On the PC4 axis, which explains 6.5% of the variance (**Figure 58b**), a split is evident

between three different groups inside the *S. pratensis* group: *S. pratensis* samples coming from Nothern and Central Italy, *S. saccardiana* samples, and *S. haematodes* together with samples of *S. pratensis* coming from Southern Italy (PR.CB and PR.PZ, which will also be referred to as *S. haematodes* from now on). Samples identified as *S. pratensis* subsp. *bertolonii* (PR.MN.2) are not separated from *S. pratensis* on any axis.

We performed Discriminant Analysis of Principal Components (DAPC) as an unsupervised clustering method on the PCA-reduced data (**Figure 59**). The Bayesian Information Criterion, which is generally used to choose the best number of clusters for the algorithm, did not reach a global minimum. The mean BIC has the sharpest decline between consecutive values of K between K=14 and 16, after which the rate of change is much smaller; at this point most taxa are separated, with the exception of *S. desoleana* and *S. sclarea*, for a total of 11 taxa. Even for values of K at which *S. sclarea* and *S. argentea* are assigned to the same cluster (K<13), the *S. pratensis* group shows the same separation in three groups as in the PCA. We also performed ADMIXTURE analysis (**Figure 60**). At K=4 the clusters correspond to the *S. pratensis* group, the *S. verbenaca* group (*S. verbenaca*, *S. clandestina* and *S. ceratophylloides*), the *S. argentea* group (*S. argentea*, *S. desoleana* and *S. sclarea*) and the *S. glutinosa* – *S. rosmarinus* – *S. officinalis* group, with the latter presenting components from the *pratensis* group. At K=9, which is a local minimum of the cross-validation error, *S. pratensis* group is separated in the three components (*pratensis – haematodes – saccardiana*), *S. clandestina* belongs to the same ancestral group as *S. verbenaca*, and the *S. argentea* group is also not separated. At higher values of K, admixture between different groups is identified within *S. pratensis*. This is possibly due to the higher number of *S. pratensis* samples compared to the other species. These components do not show a particular geographical structure within *S. pratensis*.

We computed a neighbour-joining tree with 1,000 bootstrap replicates rooted on *S. glutinosa*, which belongs to Clade IV *sensu* Will & Clasen-Bockhoff (subgenus *Glutinaria*) (**Figure 61a**). The tree is in agreement with the *Salvia* phylogenies from literature, with *S. rosmarinus* being a sister group to all other *Salvia* and *S. officinalis* being a sister group to the *Sclarea* subgenus. The tree also shows a clade formed by *S. ceratophylloides*, *S. clandestina* and *S. verbenaca*. Samples of the *S. pratensis* group show the same separation as in the PCA, with *S. saccardiana* and *S. haematodes* forming two separate clades and *S. pratensis* from Northern and Central Italy and *S. bertolonii* forming a third clade. Which of the three clades is sister to the other two is not well supported, with 90% bootstrap support for *S. saccardiana* being the sister group, however, with a short branch length. Within *S. pratensis*, samples from the same population have a slight tendency to form clades, but they are not well supported, showing possibly a low level of genetic isolation, and very short branch lengths. *S. clandestina*, which is considered a synonym of *S. verbenaca* by the Plants Of the World Online Database (POWO) (*Salvia Clandestina L. | Plants of the World Online | Kew Science*, n.d.) forms a separate and well supported clade to *S. verbenaca*. We then performed Maximum Likelihood phylogeny estimation (**Figure 61b**). The tree shows high 97bootstrap support for *S. saccardiana* as sister group to *S. haematodes* and *S. pratensis* but has poor resolution in the *S. verbenaca* group.

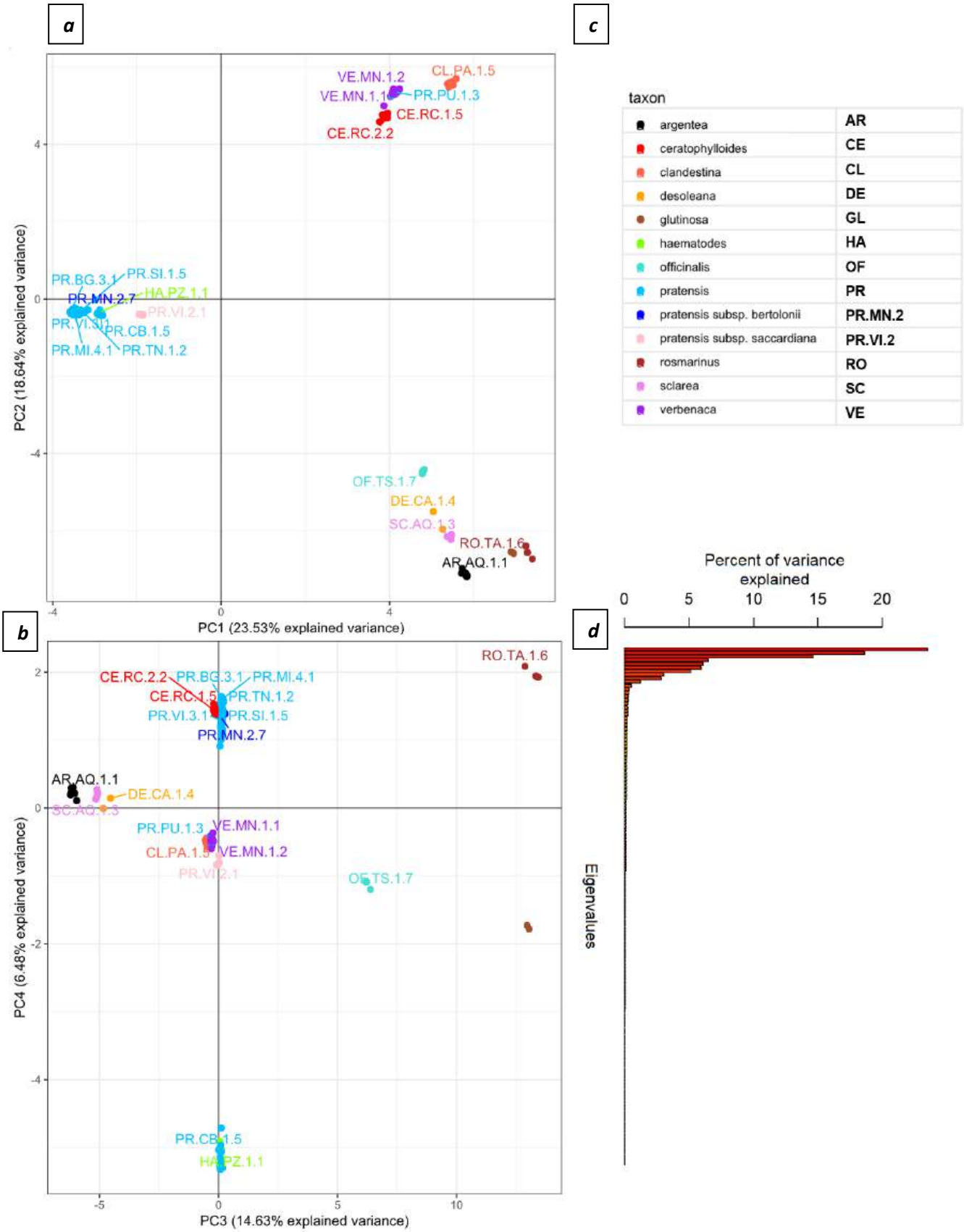


Figure 58. Principal component analysis (PCA) of GBS data on *Salvia* samples. Each point in the PC space is a sample, color-coded by taxon. Some sample names have been plotted for ease of interpretation. ***a)*** PC1-PC2 axis***. b)*** PC3-PC4 axis. ***c)*** Color and abbreviation legend. ***d)*** Percentage of variance explained by each component.

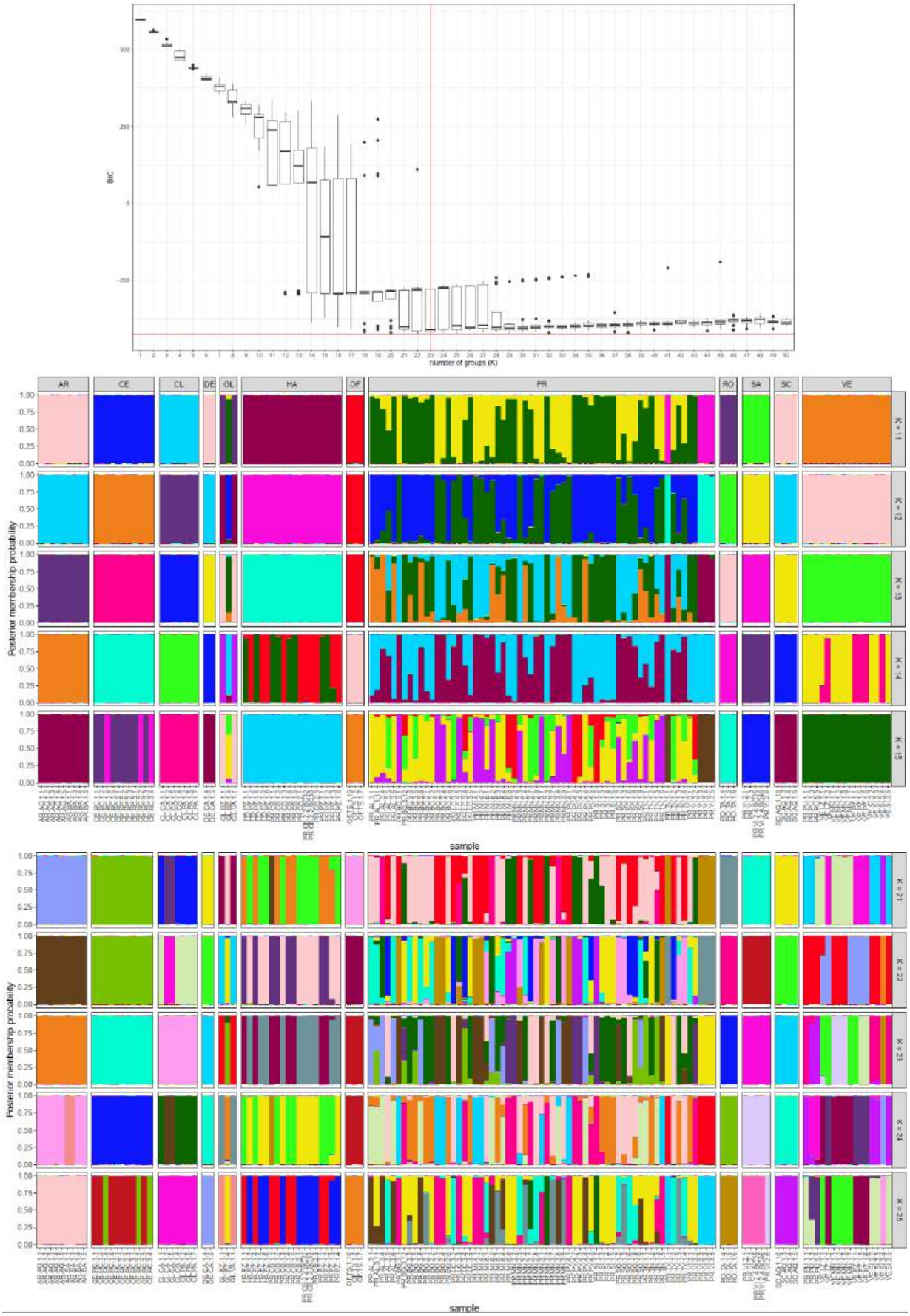


Figure 59. DAPC clustering results on *Salvia* GBS data. Top: BIC for the model at different values of K (=number of clusters) over 20 replicates. Bottom: posterior probability for each sample to belong to each cluster at selected values of K. Samples are grouped by taxon.

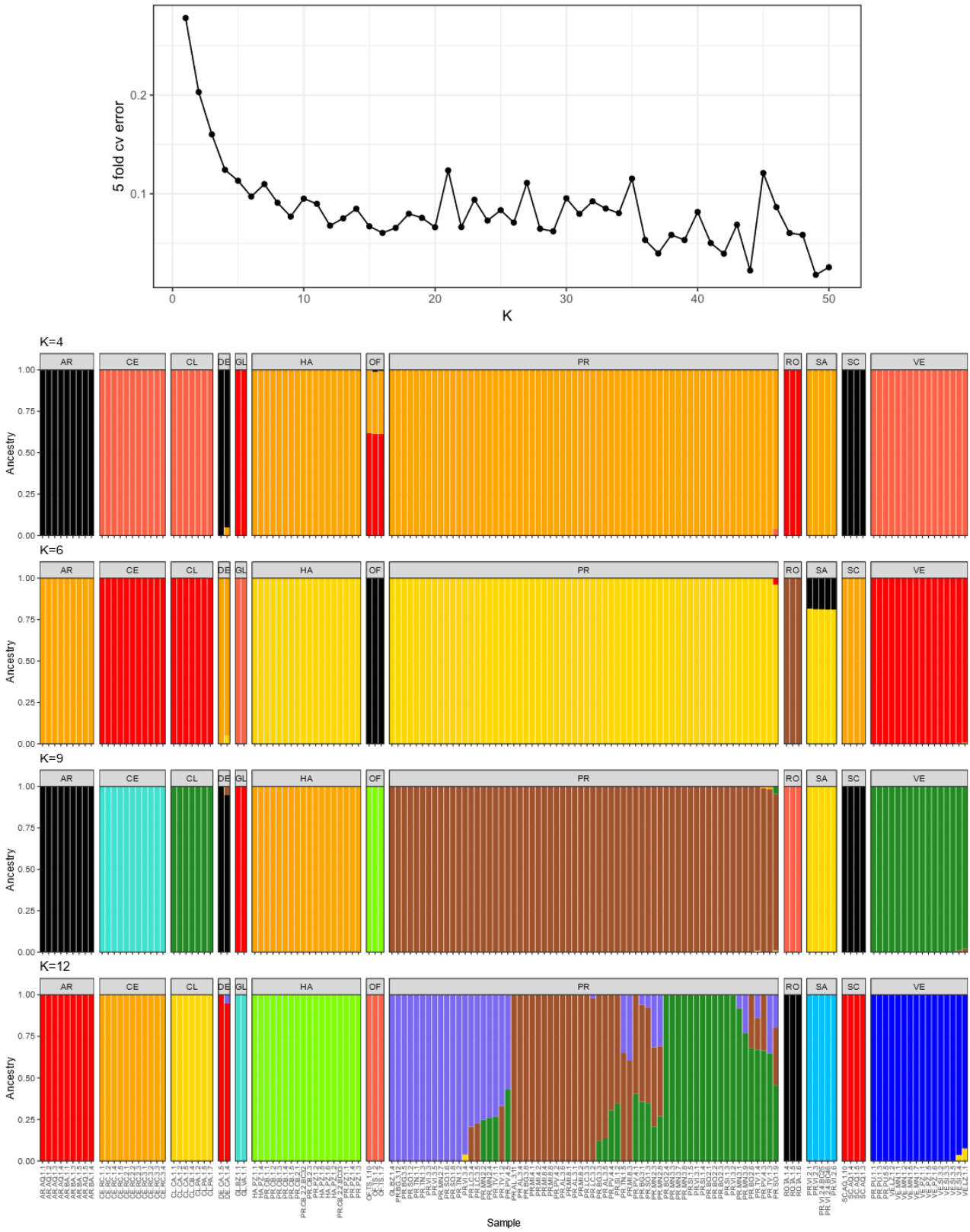


Figure 60. ADMIXTURE analysis of *Salvia* samples. Top: 5-fold cross validation error. Bottom: ADMIXTURE results plotted at selected values of K (local minima). Samples are split by taxon.

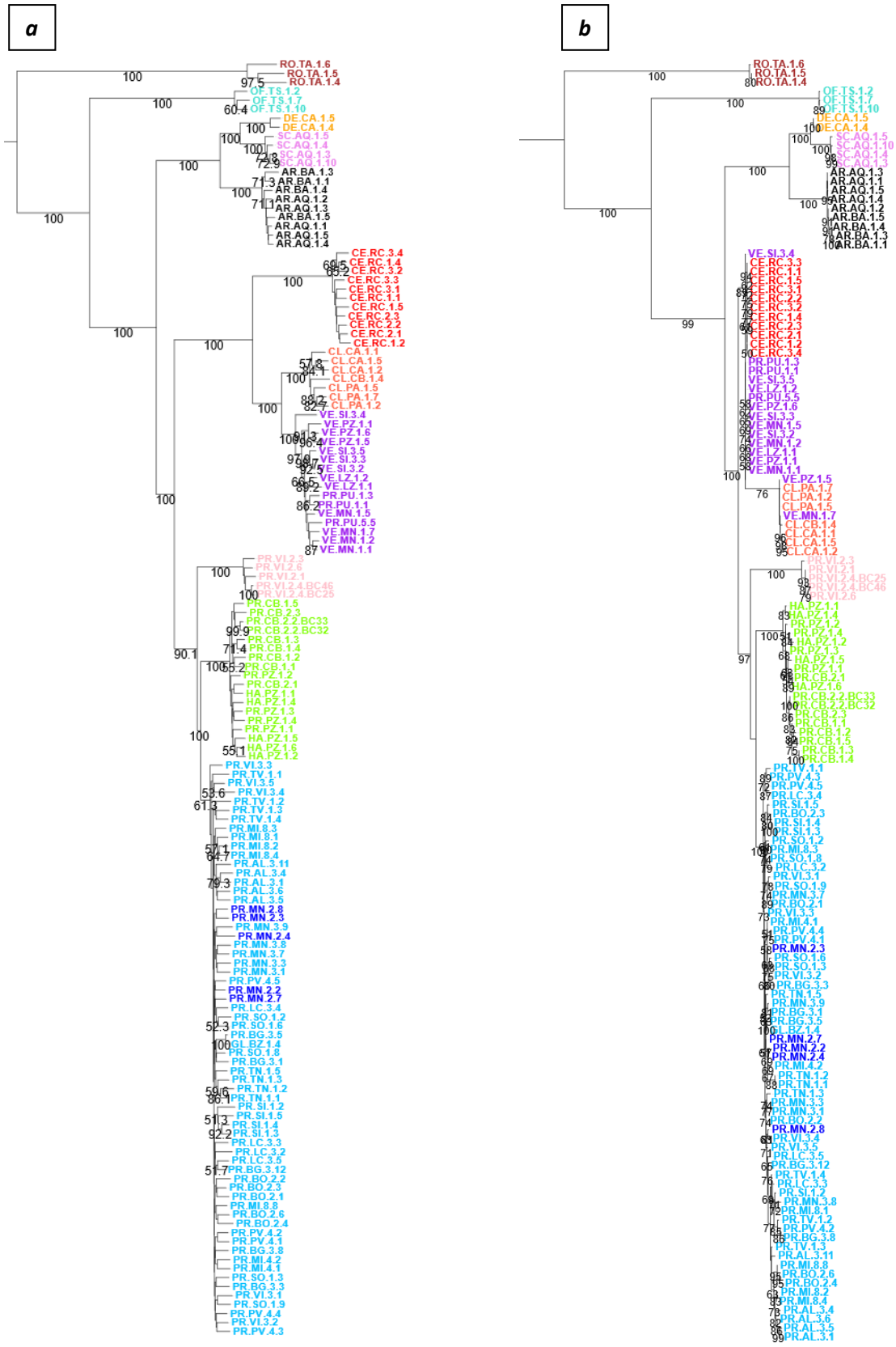


Figure 61. ***a)*** Neighbor-joining tree of *Salvia* samples. ***b)*** Maximum likelihood phylogeny of *Salvia* samples. See Figure 55 for color and abbreviation legend.

Placement of *S. ceratophylloides* in a clade with *S. verbenaca* and *S. clandestina* (*S. verbenaca* group) was unexpected, as *S. ceratophylloides* has historically been considered part of the *S. pratensis* group, and their flower morphology is much more similar. The *S. verbenaca* group shows different levels of ploidy which complicate the phylogeny. Variant calling methods that consider different ploidy levels could help in disentangling this phylogeny. Admixture analysis does not seem to suggest a hybrid origin for *S. ceratophylloides*, at least within the analysed species. Karyological data (Salmeri, 2019) showed a chromosome count of 2n=54 for *S. ceratophylloides* and inferred a hexaploidy arrangement with reference to n=9 for *S. pratensis*. This is in contrast with the sequencing data which suggest a tetraploid or possibly octoploid; additionally, *S. ceratophylloides* may be more related to *S. verbenaca*, and thus not share a base chromosome number with *S. pratensis*. *S. clandestina* has a different ploidy level than *S. verbenaca*, to which however it is very closely related. Chromosome counts for the two species are of 2n=42 and 2n=56 respectively (Kamari & Garbari, 2000). The count of 56 chromosomes for *S. verbenaca* is again in contrast with ploidy of 3 or 6 inferred from sequencing data. Differently from *S. clandestina*, neither *S. verbenaca* nor *S. ceratophylloides* display a major peak in allele frequencies at 50%, which would be expected in a process or re-diploidization or in recent autopolyploids.

These species were shown to have extremely small chromosomes, complicating chromosome counting and karyotype reconstruction. The discordance between ploidy inferred by chromosome counting and bioinformatic analysis could be explained by different haploid sets for species of the *S. pratensis* and *S. verbenaca* groups, which would negatively impact reference-based ploidy estimation approaches. This is supported by the high rate of chromosomal rearrangement between *Salvia* species (as seen in **Figure 33**). K-mer based methods for ploidy estimation cannot be applied to GBS data but could be applied to WGS data in the future and provide additional information. Obtaining reference genomes of species belonging to the *S. verbenaca* group would also allow identification of syntenic regions between *S. pratensis* and these species and shed more light on the genome evolution of the genus; however, the extremely high number of chromosomes and varying ploidy levels make this more complex.

3.5. *Salvia pratensis* and endemics

We then investigated genetic variation inside the *S. pratensis* group (*S. pratensis*, *S. saccardiana* and *S. haematodes*). We repeated SNP selection and kept 6,188 SNPs with MAF > 0.05 for analysis. Exclusion of less closely related species increased the amount of shared GBS sites and the number of variants that could be accurately genotyped. From the principal component analysis, the split in three separate lineages is evident (**Figure 62**). There is no gradient of genetic variation from north to southern Italy, but rather, three genetically isolated groups. Samples of *S. pratensis* from Trentino to Tuscany cluster together, while samples from Molise cluster with *S. haematodes* despite Tuscany approximately equidistant from both Trentino and Molise. *S. saccardiana* is genetically isolated from population of *S. pratensis* from the same geographical area. We

performed clustering with DAPC and computed the posterior probability for each sample to belong to different clusters (**Figure 63**). The BIC does not reach a minimum in the values tested but has a sharp decline for k=3. Accordingly, the samples split unambiguously in clusters corresponding to the three taxa and increasing the number of clusters only adds components inside the *S. pratensis* samples. We performed ADMIXTURE analysis, which selected K=3 as the best number of groups according to cross-validation error, and obtained the same results (**Figure 64**).

We computed nucleotide diversity statistics (**Table 27**) and pairwise F_{ST} between populations of the *S. pratensis* group (**Table 28**). The two endemic taxa *S. saccardiana* and *S. haematodes* show lower levels of nucleotide diversity and heterozygosity compared to *S. pratensis*, possibly due to smaller population sizes. The pairwise F_{ST} between populations of the three taxa clearly shows the same split that was observed in the other analyses. The F_{ST} values between populations of *S. pratensis* from northern and central Italy (including subsp. *bertolonii*) range between 0.07 and 0.15. There does not seem to be a strong correlation between geographical distance and genetic isolation. By contrast, the F_{ST} between *S. saccardiana* and *S. pratensis* and *S. haematodes* and *S. pratensis* range between 0.4 and 0.5, which is a very high range of values. *S. saccardiana* and *S. haematodes* display even higher F_{ST} values of 0.68 between each other. Populations of *S. haematodes* have pairwise F_{ST} values between them in the same range as among populations of *S. pratensis*.

	Pop.	Priv. all.	Indv	Ho	He	π	Fis
PR	PR.AL.3	0	4	0.15561±0.00348	0.12906±0.00256	0.1475±0.00292	-0.01490
	PR.BG.3	0	5	0.17639±0.00362	0.14339±0.00256	0.15932±0.00285	-0.03171
	PR.BO.2	0	5	0.17838±0.00368	0.14521±0.00261	0.16134±0.0029	-0.03257
	PR.LC.3	0	4	0.17253±0.00387	0.13257±0.00261	0.15151±0.00299	-0.03869
	PR.MI.4	0	2	0.17248±0.00429	0.11697±0.00262	0.15596±0.0035	-0.02479
	PR.MI.8	0	5	0.15614±0.00366	0.12226±0.00253	0.13584±0.00281	-0.03747
	PR.MN.2	0	5	0.17574±0.00350	0.14477±0.00255	0.16086±0.00283	-0.02890
	PR.MN.3	0	5	0.16987±0.00359	0.14217±0.00262	0.15797±0.00292	-0.02112
	PR.PV.4	0	5	0.18169±0.00376	0.14529±0.00264	0.16143±0.00294	-0.03872
	PR.SI.1	0	4	0.16424±0.00391	0.12057±0.00257	0.1378±0.00294	-0.04864
	PR.SO.1	0	5	0.16861±0.00367	0.13201±0.00256	0.14668±0.00285	-0.04296
	PR.TN.1	0	4	0.15532±0.00361	0.12444±0.00254	0.14222±0.00291	-0.02363
	PR.TV.1	0	4	0.16514±0.00367	0.13001±0.00255	0.14858±0.00292	-0.03081
	PR.VI.3	0	5	0.16407±0.00338	0.13547±0.00249	0.15052±0.00277	-0.02727
HA	HA.PZ.1	0	5	0.09336±0.00303	0.07352±0.00214	0.08169±0.00238	-0.02212
	PR.CB.1	0	5	0.10969±0.00334	0.08110±0.00224	0.09011±0.00249	-0.03774
	PR.CB.2	0	4	0.11489±0.00369	0.07681±0.00226	0.08779±0.00258	-0.04930
	PR.PZ.1	0	4	0.10841±0.00353	0.07659±0.00225	0.08753±0.00257	-0.03760
SA	PR.VI.2	751	5	0.06438±0.00275	0.04403±0.00171	0.04892±0.0019	-0.02943

Table 27. Population genetics statistics of taxa within the *S. pratensis* group computed at the population level. Pop: population. Priv. all.: private alleles in the population (=only present in that one population). Indv: n° of samples belonging to the population. Ho: observed heterozygosity, He: expected heterozygosity, π: genetic diversity, Fis: inbreeding coefficient. PR, HA, SA refer to the main three *S. pratensis* lineages in the text (PR=Northern Italy; HA=*haematodes* and Southern Italy; SA=*saccardiana*, identified by the collector as *S. pratensis* subsp. *saccardiana*).

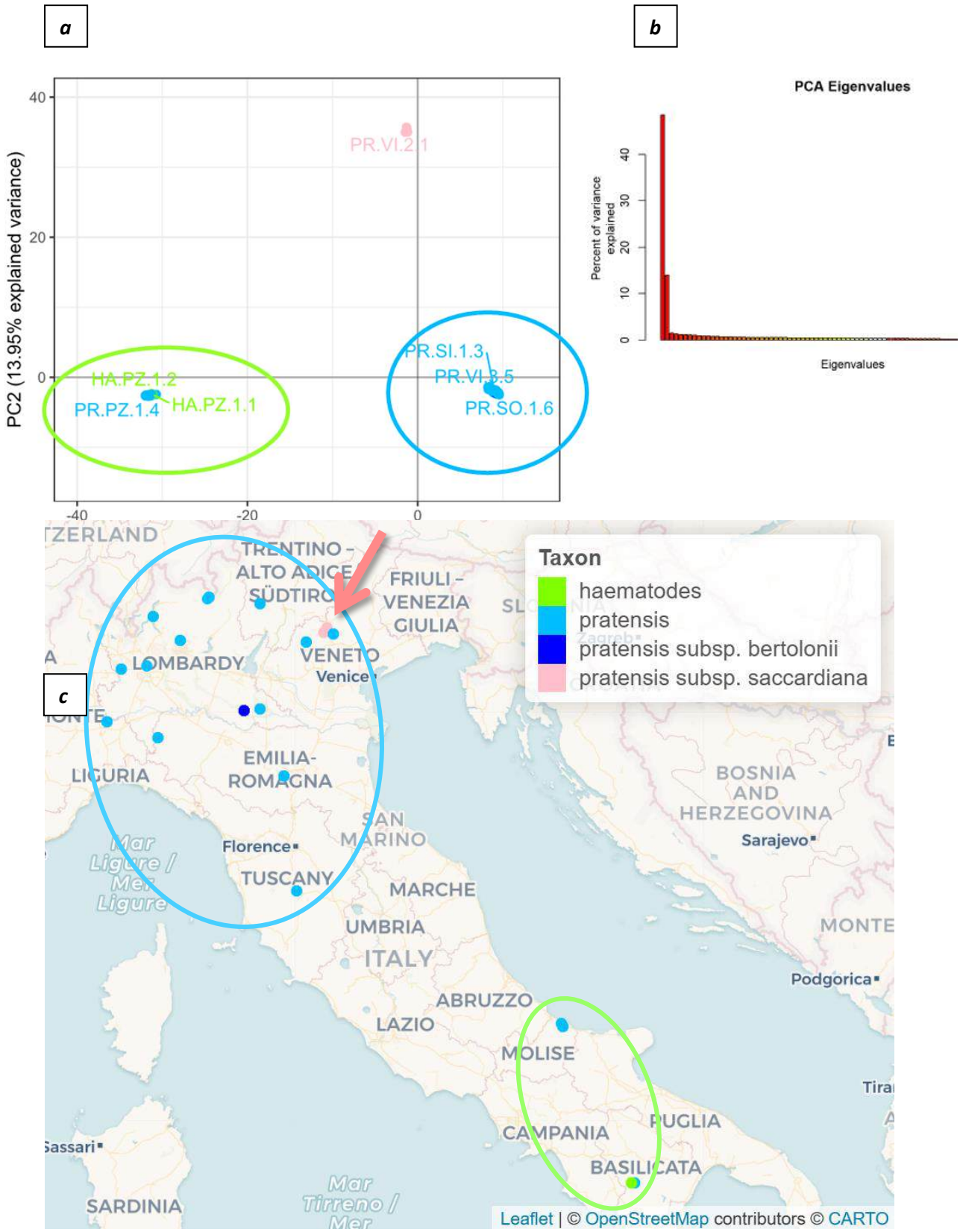


Figure 62. ***a)*** PCA of samples within the Salvia pratensis group. ***b)*** Percentage of variance explained. ***c)*** Origin of samples from a) indicated by circles or a pink arrow for population of S. saccardiana PR.VI.2.

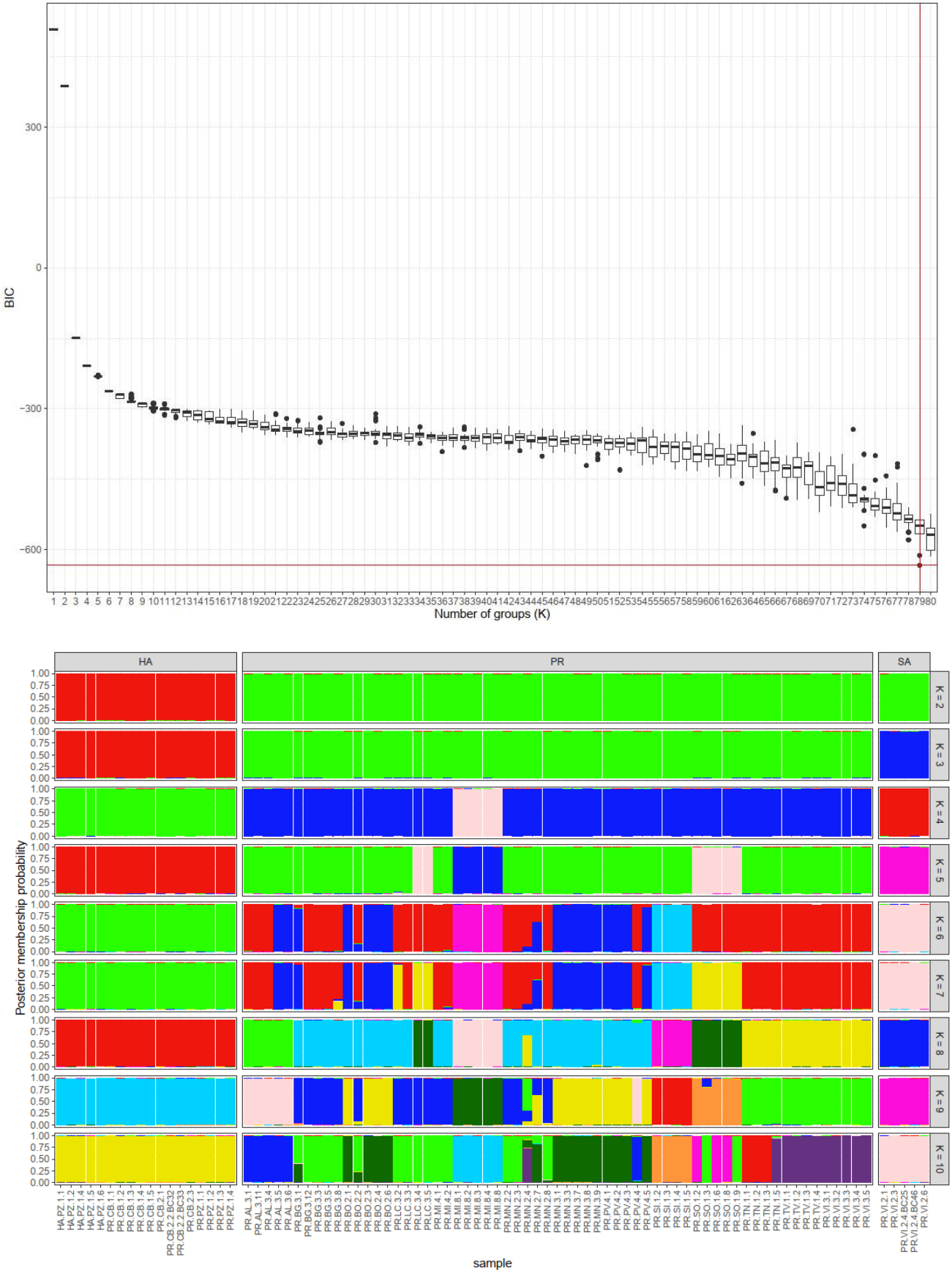


Figure 63. DAPC posterior probability analysis on samples from the S. pratensis group. Top, Bayesian information criterion. Bottom, results for K from 2 to 10.

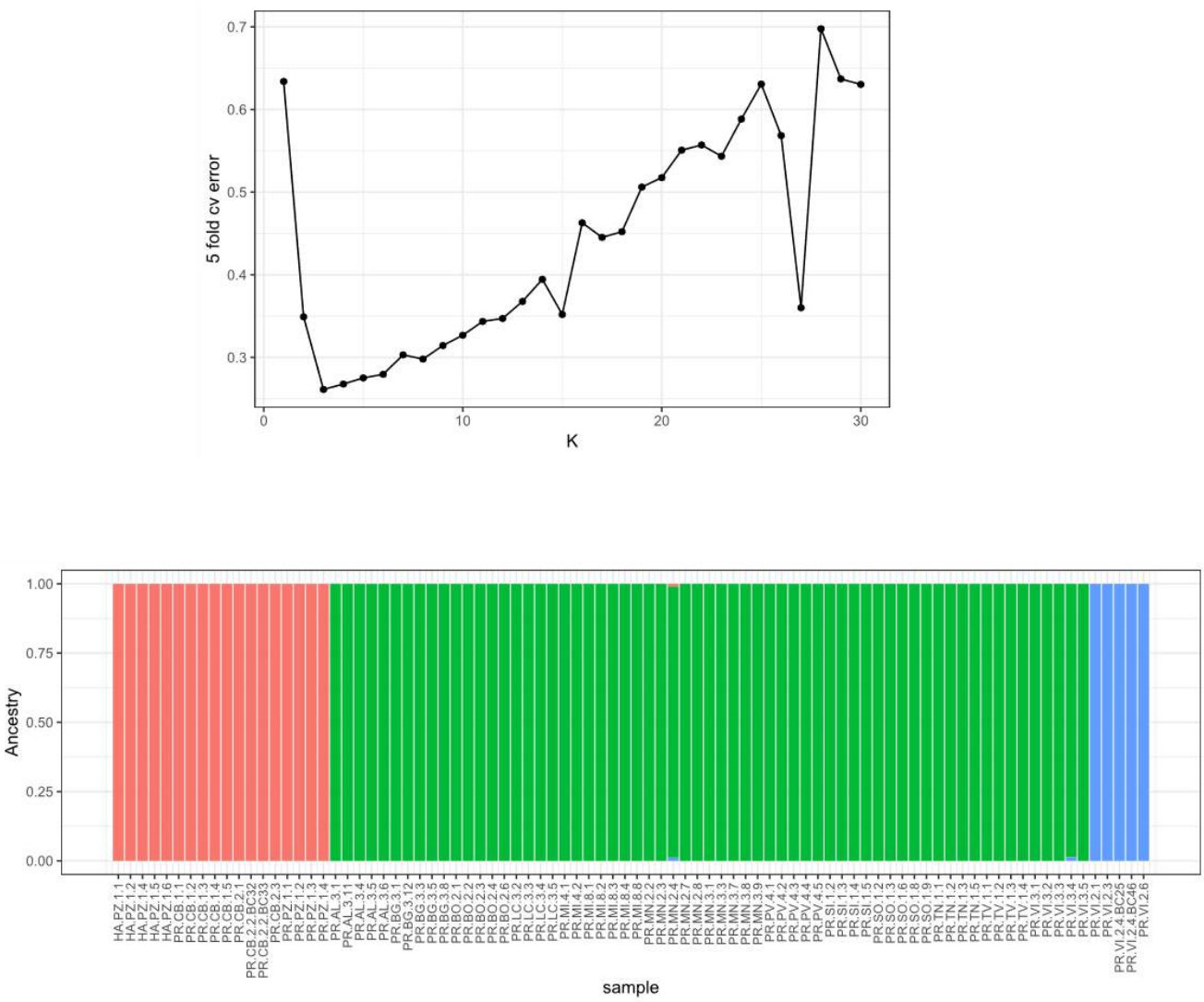


Figure 64. ADMIXTURE results on samples from the S. pratensis group. Top, 5-fold cross validation error. Bottom, ADMIXTURE results for K=3.

		PR														HA			
		PR.AL.3	PR.BG.3	PR.BO.2	PR.LC.3	PR.MI.4	PR.MI.8	PR.MN.2	PR.MN.3	PR.PV.4	PR.SI.1	PR.SO.1	PR.TN.1	PR.TV.1	PR.VI.3	HA.PZ.1	PR.CB.1	PR.CB.2	PR.PZ.1
SA	PR.VI.2	0.471	0.423	0.428	0.467	0.537	0.475	0.417	0.437	0.431	0.496	0.452	0.475	0.454	0.426	0.670	0.668	0.686	0.680
HA	PR.PZ.1	0.451	0.409	0.413	0.445	0.504	0.453	0.404	0.423	0.416	0.469	0.434	0.454	0.432	0.410	0.069	0.097	0.100	
	PR.CB.2	0.458	0.417	0.422	0.453	0.512	0.462	0.411	0.431	0.424	0.477	0.442	0.462	0.440	0.418	0.102	0.065		
	PR.CB.1	0.448	0.411	0.416	0.445	0.495	0.453	0.406	0.425	0.419	0.467	0.435	0.453	0.432	0.411	0.095			
	HA.PZ.1	0.449	0.411	0.415	0.446	0.499	0.454	0.405	0.425	0.418	0.470	0.435	0.454	0.432	0.409				
PR	PR.VI.3	0.093	0.071	0.080	0.091	0.101	0.103	0.069	0.082	0.080	0.107	0.092	0.092	0.077					
	PR.TV.1	0.108	0.081	0.089	0.104	0.126	0.118	0.081	0.095	0.088	0.119	0.099	0.106						
	PR.TN.1	0.116	0.087	0.097	0.108	0.136	0.121	0.084	0.096	0.096	0.135	0.108							
	PR.SO.1	0.100	0.077	0.087	0.095	0.112	0.112	0.082	0.093	0.087	0.119								
	PR.SI.1	0.124	0.100	0.105	0.122	0.150	0.139	0.100	0.113	0.105									
	PR.PV.4	0.079	0.069	0.071	0.088	0.099	0.095	0.071	0.080										
	PR.MN.3	0.099	0.076	0.077	0.095	0.105	0.105	0.069											
	PR.MN.2	0.085	0.063	0.069	0.081	0.092	0.090												
	PR.MI.8	0.108	0.093	0.099	0.112	0.131													
	PR.MI.4	0.124	0.078	0.100	0.122														
	PR.LC.3	0.110	0.077	0.087															
	PR.BO.2	0.088	0.069																
	PR.BG.3	0.084																	

Table 28. Pairwise F_{ST} of populations from the *S. pratensis* group. Cells are colored in a scale blue-white-red from low to higher values of F_{ST}. PR, HA, SA refer to the main three *S. pratensis* lineages in the text (PR=Northern Italy; HA=*haematodes* and Southern Italy; SA=*saccardiana*, identified by the collector as *S. pratensis* subsp. *saccardiana*).

GBS data shows that *S. pratensis* and *S. saccardiana* are genetically isolated. Moreover, it shows that all samples from either *S. haematodes* or *S. pratensis* coming from Southern Italy belong to the same genetic background, which is also isolated from *S. pratensis* coming from Northern and Central Italy. We assign these southern samples to *S. haematodes* and hypothesize that they were mis-classified during sampling. Indeed, morphological differences between *S. haematodes* and *S. pratensis* are quite subtle and possibly require the use of a stereomicroscope to assess presence of glandular hairs. Unfortunately, our rather sparse sampling of central Italy did not allow us to identify a potential boundary separating *S. pratensis* in the north and *S. haematodes* in the south. Additional samples from Abruzzo were obtained and are already being sequenced, and the genetic information from those samples will be compared to morphological information from herbarium samples of the same populations. It is unclear whether this separation corresponds to a geographic barrier or if *S. haematodes* and *S. pratensis* live in sympatry. Recently, *S. haematodes* was reported as present as far north as Modena (Emilia-Romagna) (Buldrini et al., 2023). Perhaps, the two taxa found different refugia in the past and evolved separately, causing their different genetic backgrounds. In the future, this lack of information can easily be amended with a targeted sampling campaign in southern and Central Italy, accompanied by harvest of seeds and cross-pollination experiments. By contrast, *S. saccardiana* and *S. pratensis* live in close proximity, and thus their separation is more easily compatible with a reproductive isolation mechanism. From the literature, the two species have different chromosome counts (n=9 for *S. pratensis*, n=10 for *S. saccardiana*) (Carratore et al., 1999), but not different ploidy levels, which was confirmed by sequencing data. Adding samples from more populations, and especially from Croatia, could help understand better the evolutionary history and separation mechanisms between *S. pratensis* and *S. saccardiana*. Additionally, this could be integrated with some low coverage sequencing to retrieve plastidial sequences: the GBS protocol did not result in any reads coming from the chloroplast, which could tell another part of the phylogenetic history of these taxa.

In summary, our genetic data provides robust support for the species rank of *S. saccardiana*, *S. haematodes, S. ceratophylloides* and *S. clandestina*. However, the maximum likelihood phylogeny does not properly resolve a complete separation between *S. verbenaca* and *S. ceratophylloides* and *S. verbenaca* and *S. clandestina*. This could be amended with the inclusion of more samples of these species to obtain a more even coverage of genetic diversity throughout the peninsula. In any case, all methods place *S. ceratophylloides* as more closely related to *S. verbenaca* than *S. pratensis* in contrast with existing literature.

We already identified some homozygous variants (SNPs) exclusive to *S. haematodes*, which are candidates for development of markers. However, we are waiting for sequencing results of additional samples from Southern Italy to refine the set of markers and confirm their specificity. The same holds for *S. saccardiana*, as it would not be useful to develop markers from the single population that we analysed so far.

Overall, this study has offered novel insights into the phylogeny of Italian *Salvias* by sequencing species previously not included in most genus-wide phylogenies. Additionally, we corroborate the species rank of several taxa such as *S. haematodes*, *S. saccardiana* and *S. clandestina* by showing their genetic isolation. Genetic data shows promise to develop markers for identification of endemics. We also investigated patterns of genetic diversity and showed that *S. pratensis* only shows moderate structuring and a good level of genetic diversity.

4. *Althaea officinalis*

The project, in collaboration with Aboca S.p.A. Società Agricola, is aimed at identifying genotypes of *Althaea officinalis* with root extracts displaying a high adhesive activity in an *in vitro* test on cell cultures. The plants currently cultivated in Aboca originate from seed bought from a supplier in 2018 and were self-propagated by Aboca in the following years. There are no genomic resources for the species, and viability of self-pollinated seeds is unknown. The genome size is expected to be quite large from the literature. To attempt identification of variants associated to an interesting phenotype, an experimental design in multiple steps was proposed.

1. Evaluation of phenotypic and genetic variability.
2. Evaluation of self-pollination efficiency.
3. Genome sequencing.
4. Transcriptomic analysis of high and low activity plants.
5. Pooled sequencing of high and low activity sets of plants.

The design of the different steps will now be illustrated.

4.1. Evaluation of phenotypic and genetic variability

Two experimental fields with 100 plants each were set up in Aboca to obtain plant material for the study (**Figure 66**). The number was selected to hopefully capture as much of the diversity as possible. It was very relevant to keep a sufficient distance between individual plants to avoid confusion of genotype roots during the sampling stages. The plants were otherwise grown in the usual conditions, which include drought to maximise polysaccharide production. Young leaves with as little damage as possible were sampled during July of 2023 and stored to be used for DNA extraction down the line. At the end of the vegetative period, the plants were harvested, and the fresh and dry weight of 108 plants was measured (**Figure 65**). There was no significant difference between plants grown in the two plots, but there was some level of variability in root weight between plants, which could already possibly indicate some candidates for selection. After the roots were thoroughly dried, several extraction protocols were tested and adapted for use on the extremely little material coming from each individual plant. While the protocol was set up, the results on all the sampled plants will only be ready in early spring 2025. The tests will be focused on measuring whether different individuals present significant differences in activity.

Another small cultivation was set up in Città Studi. This set of plants was analysed to assess the genome size of *Althaea officinalis* and whether there would be variation in the ploidy of different individuals. The flow cytometry results indicate a genome size of 3.5 Gb with very little differences between the 25 tested individuals (**Table 29**).

Figure 66. Experimental field of *Althaea officinalis* in Aboca.

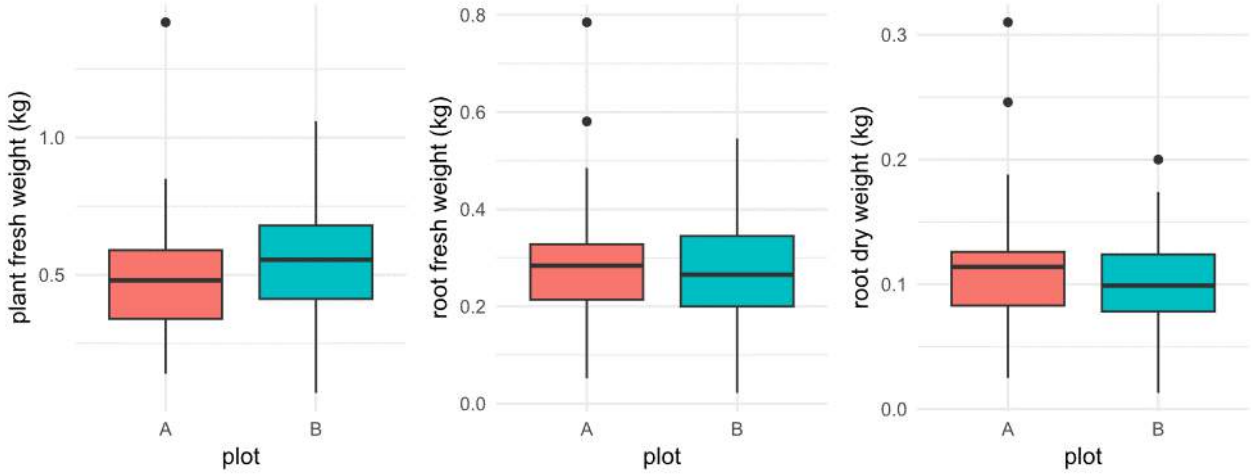


Figure 65. Plant and root weight from 108 plants of the experimental *A. officinalis* field.

plant	2c althaea (pg)	2c althaea (Gb)
30	3.73	3.65
65	3.50	3.42
66	3.52	3.44
68	3.66	3.58
68	3.63	3.55
75	3.64	3.56
78	3.62	3.54
88	3.59	3.51
99	3.52	3.45
100	3.64	3.56
104	3.70	3.61
107	3.74	3.66
108	3.59	3.51
113	3.59	3.51
130	3.74	3.66
153	3.57	3.49
159	3.56	3.48
160	3.63	3.55
161	3.29	3.22
162	3.57	3.49
163	3.47	3.39
189	3.55	3.47
191	3.59	3.51
193	3.55	3.47
300	3.62	3.54
MEAN	**3.59**	**3.51**

Table 29. Genome size estimation for *A. officinalis* with flow cytometry.

4.2. Evaluation of self-pollinating efficiency

Obtaining self-pollinated seeds from the plant under study is essential for genotype-phenotype association. It is crucial to assess whether any interesting difference in phenotype can be inherited: this implies measuring the trait on the progeny of the selected plant to check whether, in this case, a higher or lower level of root extract activity is preserved. To test viability of this approach, plants in the experimental field set up in Aboca were self-pollinated. *A. officinalis* seems to have protandric flowers which first develop mature anthers with pollen, and afterwards, a change in the morphology of the stylus can be observed, where the stigmas "open" and seem to be more receptive to pollen (**Figure 67**). Of the 108 tested plants, self-pollination was performed on 88 and seeds could be obtained from 49 plants. Additionally, seeds from cross-pollinated flowers (=not protected) of 102 plants were collected. Germination tests of these seeds are in progress and they should highlight whether self-pollinated seeds are viable.

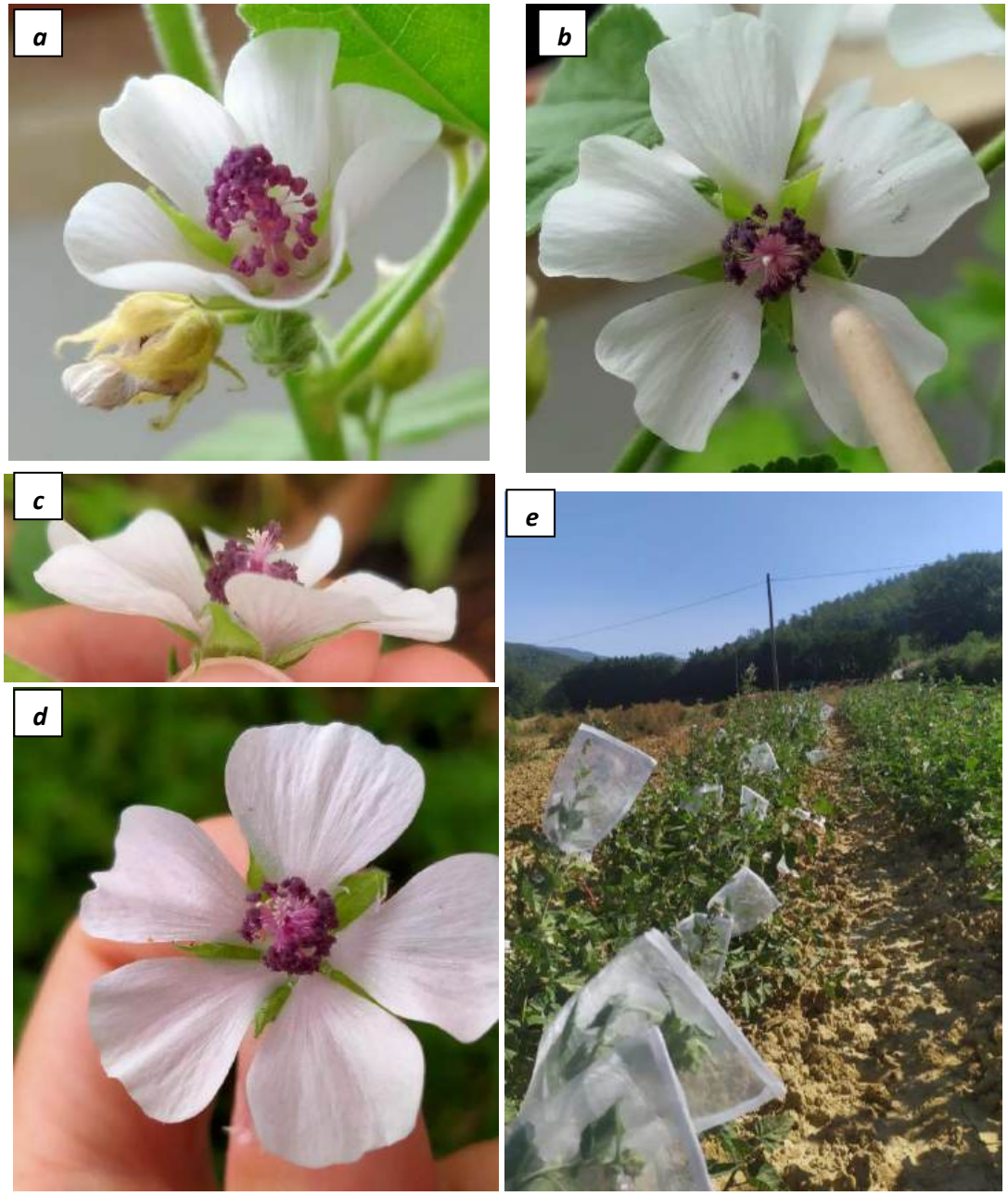


Figure 67. *Althaea officinalis* flowers. ***a)*** Immature flower. Pollen is not released by the anthers in this stage**.** ***b)*** Flower at a later stage of development. This flower produces pollen, which is a dark purple color. ***c, d)*** Flower with mature pistils. The stigmas appear open and receptive. The flower still produces pollen. ***e)*** Setup for self-pollination in Aboca. Mesh bags are closed around the flower before the flower can open to prevent cross-pollination, then the bags are opened and manual pollination is performed on flowers at stage ***d)***.

4.3. Genome sequencing

Steps 3-5 of the project are highly dependent on the results from steps 1 and 2, which are not yet available. However, the design includes genome sequencing of *Althaea officinalis*, in order to obtain a reference genome for mapping and variant identification. The genome assembly will be performed, if possible, on a self-pollinated F1 to reduce heterozygosity. The genome will be sequenced with PacBio HiFi and additionally, HiC, which will be extremely relevant, given the proposed polyploidy of the genus. Moreover, long read transcriptome sequencing will be used to annotate protein-coding genes. This will be crucial to perform RNAseq and to identify potential candidate genes related to differences in phenotype.

4.4. Transcriptome sequencing

To identify potential candidate genes for marker-assisted selection, generally Quantitative Trait Locus (QTL) analysis is performed on the progeny of "pure" lines for the phenotype of interest. Given we have no availability of pure lines, and have a long generation time for *A. officinalis*, two alternative analyses were designed to try and identify genes related to the phenotypic differences potentially observed during step 1. The first analysis is differential gene expression (DEG) analysis on root samples of a set of replicate plants with high root extract activity and a set with low root extract activity. Additionally, this could be integrated with DEG analysis on plants grown in different drought conditions. This analysis should identify genes involved in the production of metabolites that affect the adhesive activity and provide putative targets for selection.

4.5. Pool sequencing

The second genetic analysis is aimed at trying to identify variants involved with the phenotype of interesting while avoiding WGS of several tens or even hundreds of individuals, which would cause a very high cost. Briefly mentioned in the introduction, pool sequencing consists in pooling the genomic DNA of a set ("pool") of individuals and sequencing different pools separately. The pools should be genetically different and differ in a trait of interest: in this case, we would identify ideally 50-100 plants with a "high" adhesive activity and the same number with a "low" activity. Then, differences in allele frequency between the two pools can be identified. If there are variants with significant frequency differences between the two pools of individuals, they could be linked to genes involved in the phenotype of interest and used as markers of genotypes with high adhesive activity. While information on the genotype of the single individual is lost, this method can drastically reduce the cost of sequencing, as the coverage per individual needed is much lower. The process and analysis is similar to Bulk Segregant Analysis sequencing (BSA-Seq) (Majeed et al., 2022). The genotypes with high adhesive activity can later be retrieved with Sanger sequencing of selected marker regions. The analysis of the pools is complicated by several different factors. Several traits are controlled by multiple different genes, and a variant in a single gene may have negligible effect on the phenotype of the individual. Additionally, a high degree of genetic variation should be present in the starting collection, considering that we do not start from the hybrid progeny of pure lines as in use in plant domestication. Moreover, differences in DNA amount between different individuals of the pool might negatively impact allele frequency estimations.

Overall, the *Althaea* project is a proof of concept for Aboca, and a way to start thinking about their collection from a genetic point of view. The activities that were started during this PhD project will be continued in the next year and will serve as an example for potential future genomic research on other undomesticated species.

5. Other species

The GBS protocol was applied to *Campanula thyrsoides* subsp. *carniolica*, and endemic subspecies of *C. thyrsoides*, to study the genetic variation patterns, with the final aim of addressing conservation efforts by evaluating the levels of genetic diversity within the population living in the Dolomiti Bellunesi National Park and potentially select co-specific populations for breeding. Two other GBS studies concerned *Solanum lycopersicum* and *Solanum tuberosum* with the aim of genetically characterizing putative landraces from Lombardy conserved in the Pavia seed bank.

Campanula thyrsoides* subsp. *carniolica

We prepared a GBS sequencing library on a subset of 7 samples from *Campanula thyrsoides* subsp. *carniolica* from a larger sample collection. The aim of the test was to check suitability of the GBS protocol for this species. During library preparation, we could not obtain complete digestion of genomic DNA, possibly due to polysaccharides in the lattice of this species. The reads were analysed with STACKS in *de novo* mode, as this species does not have a reference genome. After exclusion of one failed sample, 14,000 variable sites were identified, and after distance thinning and exclusion of fixed sites, a high confidence set of 2,692 variants for 7 samples was obtained (**Figure 68**). A simple PCA analysis on the genetic data (**Figure 69**) showed separation between the different populations. Since the viability of GBS has been proved, sequencing on the full set of 150 samples is in progress, which will provide actual information on the genetic structure and genetic diversity of the populations. Additionally, a second RRS protocol, Kseq (Ziarsolo et al., 2021), was tested, but gave poor results, as sequencing failed completely on the read2.

Horticultural landraces

The GBS protocol was also applied to cultivars of potato (*Solanum tuberosum*) and tomato (*Solanum lycopersicum*) from the seed bank of the University of Pavia. These seeds were collected from local gardeners and farmers in Lombardy with the aim of characterizing putative traditional varieties, or landraces, that might have been preserved and passed down in farmer families. Within tomato, a set of 5,865 SNPs was selected and PC1 identified a separation between cultivars Perino, Riccio di Parma, Costoluto fiorentino and all the others (Figure **71**). Analysis of this data together with fruit morphometry is in progress. Analysis of 14 potato varieties highlighted a separation between red potatoes and white potatoes and shows high similarity between putative landraces and modern cultivars (such as several red potatoes and Desirée; Comasca biancona anc Kennebeck) suggesting that several landraces have unfortunately been lost (F**igure 70**).

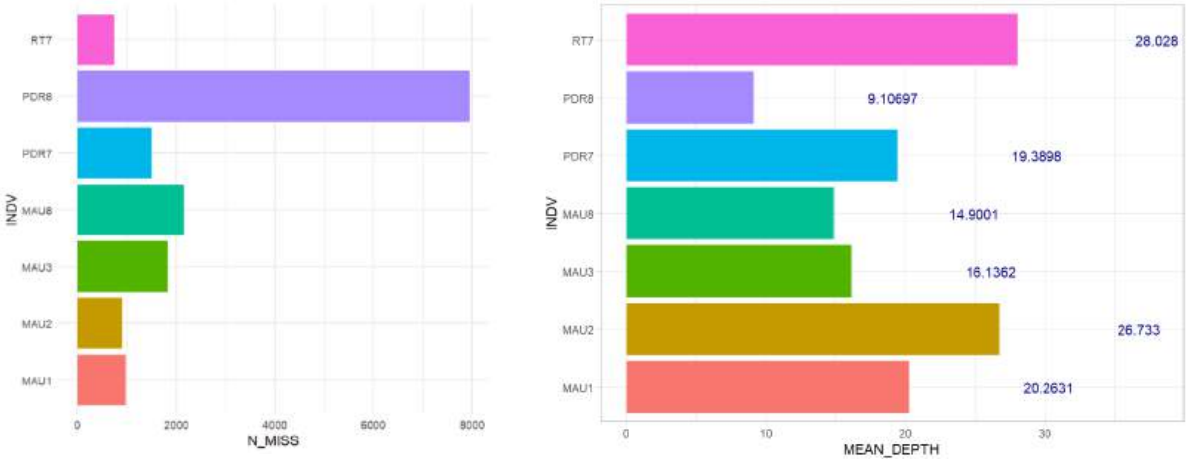


Figure 68. Missing sites per sample (left) and average sequencing depth (right) on GBS sites of 7 samples of *Campanula thyrsoides* susbp. *carniolica.*

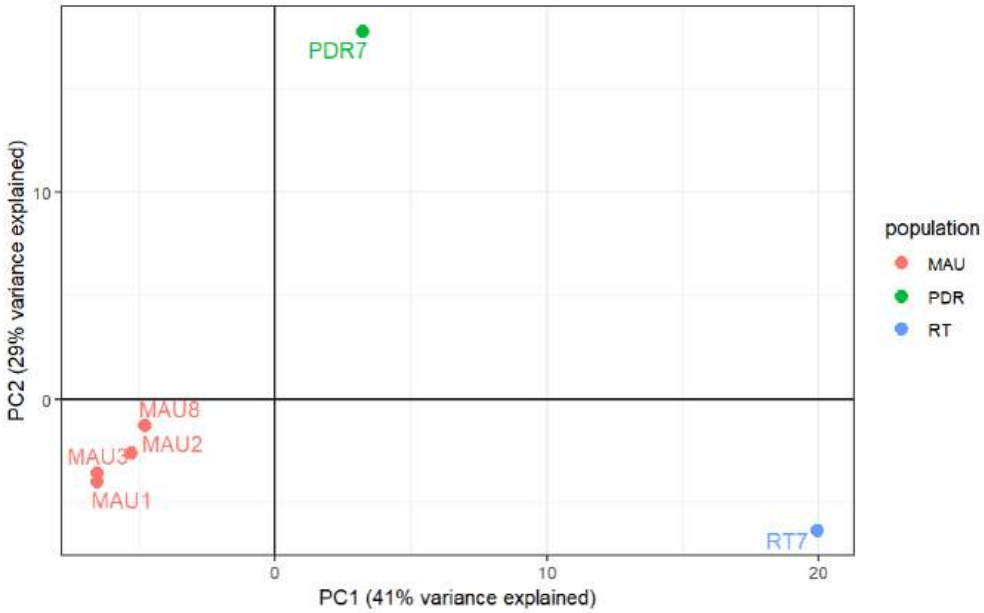


Figure 69. PCA on genetic data of 7 samples of *Campanula thyrsoides* subsp. *carniolica*. Good separation between the three populations is present.

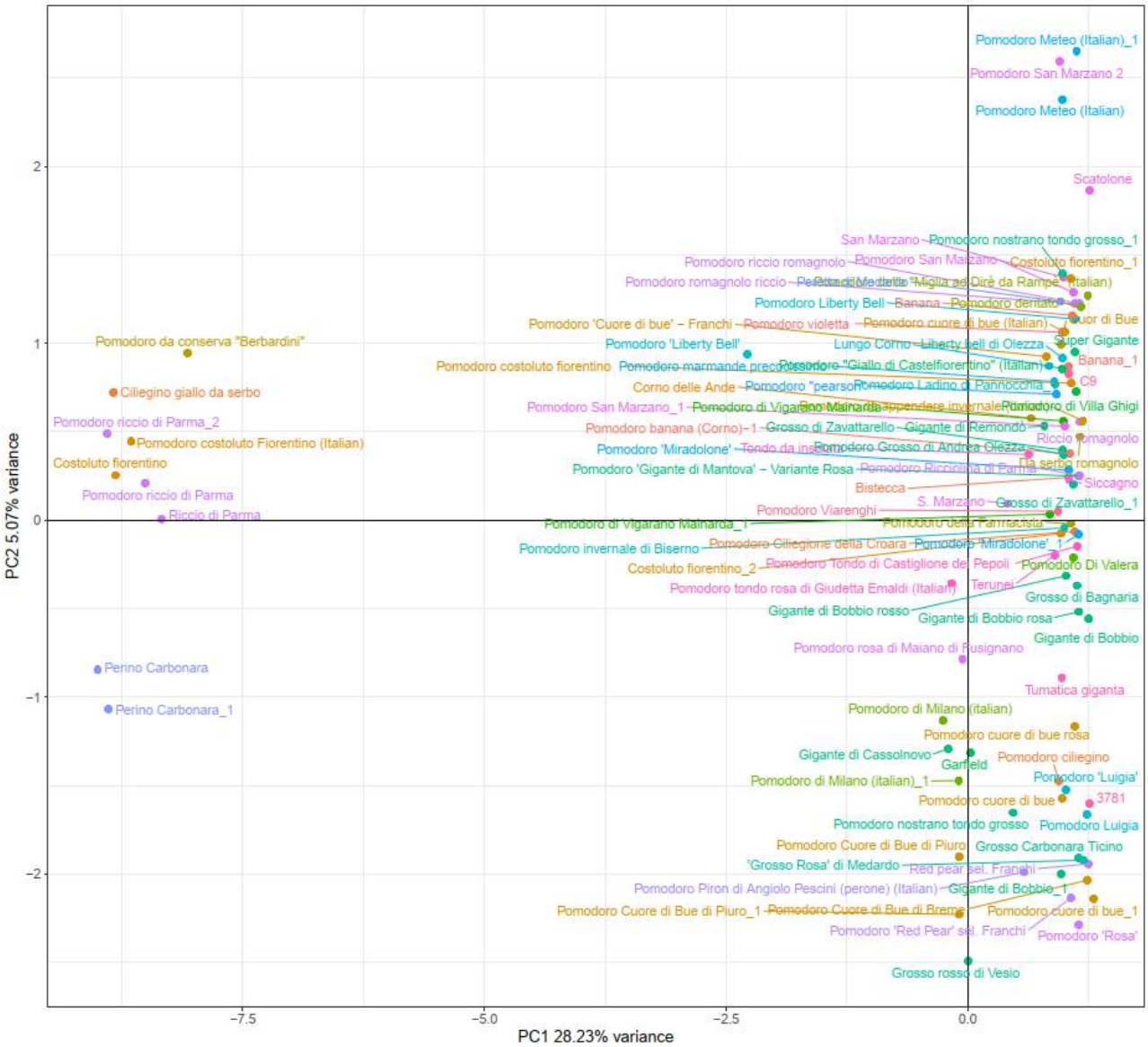


Figure 71. PCA on genetic data of 112 cultivars of tomato (*Solanum lycopersicum*). Two groups are identified on PC1.

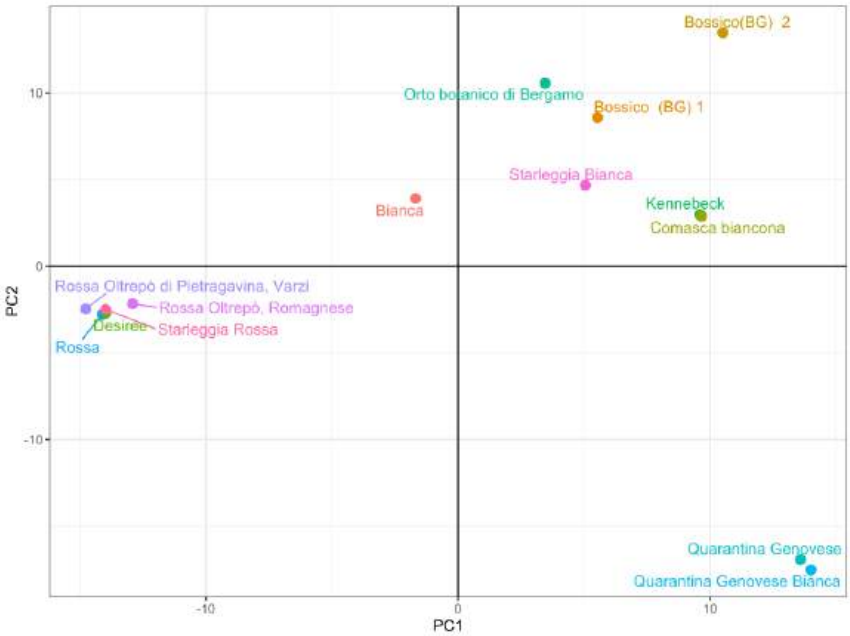


Figure 70. PCA on genetic data of 14 cultivars of potato (*Solanum tuberosum*). Several putative landraces are instead overlapping with commercial varieties such as red potatoes ("Rossa") overlapping with Desirée and "Comasca biancona" overlapping with Kennebeck.

Conclusions and future perspectives

The work carried out during this PhD project was aimed at developing genomic resources in a context where they had previously not been utilized, both for the collection of the Botanical Garden of Brera and within Aboca S.p.A.'s medicinal plant collection, to aid in conservation and genetic improvement.

With the *Salvia pratensis* genome sequencing project, we compared different long-read sequencing methods and ultimately were able to assemble a reference of enough quality for several applications. The genome was annotated with coding and non-coding genes and used for a gene family study of the MADS-box transcription factors including other *Salvia* genome; candidates for the abnormal flower genotypes have been identified. Future research in this area will be focused on further developing probes for RNA fluorescent *in situ* hybridization and analysing a large number of flower samples from female and hermaphrodite plants in order to pinpoint the molecular mechanisms responsible for differences in stamen phenotypes within *Salvia pratensis*, as well as potentially improving the reference genome to chromosome level with HiC data.

Thanks to a network of contributors, we were able to obtain a large collection of leaf samples from 15 *Salvia* species all throughout Italy. The collection was used during this project to develop a phylogeny for Italian *Salvias* and investigate species delimitation of endemic taxa within the *Salvia pratensis* group. The results from this part of the project corroborate species status for the investigated taxa (*S. saccardiana*, *S. haematodes*, *S. ceratophylloides*, *S. clandestina*) and potentially require a revision of the taxonomy of *S. haematodes*. However, resolution is limited by the low number of samples sequenced so far for *S. saccardiana* and the poor coverage of Southern and Central Italy for *S. pratensis* and *S. haematodes*. This could be explained by the fact that *S. pratensis* is less common in Southern Italy than in Northern Italy, however more samples can be obtained with a targeted sampling campaign during the next flowering season. For *S. saccardiana*, the sequences of additional samples should be ready in the first months of 2025. These results will aid in identification of species to integrate the Botanical Garden's collection with a special focus on endemics. On the other hand, no differentiation or diversity gradient linked to geographical origin was identified within *Salvia pratensis* populations.

The collaboration with Aboca was carried out with setup of experimental fields, self-pollination tests, genome size estimation and definition of a protocol to obtain root extracts from the single plant, in order to characterize the phenotype of interest. All of this is a necessary prerequisite for our genomic study. In the next year, tests on the samples obtained from the experimental filed will be performed to investigate whether the phenotypic variation in the adhesive properties of root extracts justifies the investigation of a genotype-phenotype association within the sample set of *Althaea officinalis*. In the meantime, self-pollinated plants will be employed for reference genome generation.

Acknowledgements

I would like to thank my PhD supervisor, Professor Martin Kater, for giving me the opportunity to work on these projects and offering invaluable leadership, and his Plant Functional Genomics lab for the support and collaboration. Professor Carla Lambertini has contributed greatly to the project and we shared many joys and frustrations while optimizing DNA protocols and sampling on site. Professor Aureliano Bombarely has been a wonderful mentor on the technical and personal side. Silvia Manrique has been a great help and support both on the project and during my time in Spain. From the Botanical Garden of Brera, Camilla Pieri and Cristina Puricelli handled most of the networking for the GBS sampling and seed exchange and storage, and Antonella Testa and Elisabetta Caporali gave us the opportunity to hold a webinar for the community. I would like to acknowledge Nicola Babolin for the *in situ* experiments, professor Lucia Colombo for her guidance, and MS student Maria Giacalone for her instrumental role in the lab in these past few months. From the PhD thesis committee, I want to thank Professors David Horner and Giorgio Binelli for their encouragement and Professors Matteo Dell'Acqua and Francesco Dal Grande for their suggestions. From Aboca S.p.A. Società Agricola, I wish to thank Fabrizio Zara for believing in the *Althaea* project and putting things in motion in the company. From Pavia, I would like to thank the Orsenigo group for their collaboration and understanding on the *Campanula* project. It has been great to share part of the journey with Federico Fainelli and Martina Deglialberti. Larissa de Paola has been a wonderful friend and colleague. Finally, I am deeply grateful for my parents, my partner Oscar, and my friends, who have supported me and believed in my academic journey for many years.

Materials and methods

Plant samples

Salvia pratensis genome sequencing

For genome sequencing, seeds of *S. pratensis* were obtained by the Botanical Garden of Brera through exchange with other gardens or collected in the wild, sown on universal soil and grown in the Città Studi Botanical Garden (Milan). Large, undamaged leaves were collected and used fresh for DNA extraction or flash frozen in liquid nitrogen (for HMW DNA and RNA extraction). For transcriptome sequencing Inflorescences were collected at several stages of development from the three of the planted plants and immediately frozen in liquid nitrogen. For the bud stage, flowers were collected just before opening. For the seedling stage, seeds obtained were sown on agar plates, kept for 2 days in the dark at 4°C, then germinated at 22°C. Seedlings with cotyledons, roots but no true leaves were collected and frozen in liquid nitrogen. All flash-frozen samples were conserved at -80°C before use.

species	origin	code
S. pratensis L.	Pecetto di Valenza harvested from the wild (Camilla Pieri)	p9
S. pratensis L.	Seme Nostrum, harvested in Friuli V.G. /Veneto	p10
S. pratensis L.	Orto Botanico di Pisa (OBP-283/01)	p11

Table 30. Salvia accessions sown for the study.

Genotyping-by-Sequencing

For Genotyping-By-Sequencing, silica dried leaf samples of different *Salvia* species were obtained thanks to a network of collaborators coordinated by the Botanical Garden of Brera including more than 40 contributors throughout 2022 and 2023. An additional sampling trip aimed at recovering different populations of *S. saccardiana* in Veneto, Trentino Alto Adige and Friuli-Venezia Giulia was conducted in 2024 (**Table 33**) and samples from Croatia were received in September 2024. For each contributor, data related to the samples was collected with an Excel file and R scripts were used to aggregate the data with R v4.3.1. Package Sf v1.0-17 was used to handle geographical and coordinate data and Leaflet v2.2.2 for plotting maps. Incorrect or missing coordinates were adjusted. A unique progressive code which includes species and province was assigned to each population and each sample. The code is composed as follows:

- Two letters to represent the species (PR for *pratensis*, OF for *officinalis*, and so on) (**Table 31Table 31**);
- Two letters to represent the Italian province or the country the population comes from (province codes are used, i.e. MI for Milan, VI for Vicenza, and so on; DE for Germany; HR for Croatia) (**Table 32**);

- One progressive number to identify a population of that species in that region (for example, the first population of *pratensis* sampled in the Bergamo province receives code PR.BG.1, the second PR.BG.2, the first population of *pratensis* sampled in the Vicenza province receives code PR.VI.1, and the first population of *glutinosa* from Bergamo receives code GL.BG.1);
- One progressive number to identify the sample inside the population it belongs to (PR.MI.1.1 is the first sample from the population PR.MI.1; PR.MI.1.2 is the second sample from the same population; and so on).

The total list of samples is available online (https://unimibox.unimi.it/index.php/s/FsxwwiPW6i5EH6j). The analyses of these samples are still in progress.

Species	Species code	N° samples	N° populations
S. argentea	AR	36	6
S. ceratophylloides	CE	18	3
S. clandestina	CL	41	6
S. desoleana	DE	8	2
S. fruticosa	FR	10	1
S. glutinosa	GL	231	42
S. haematodes	HA	14	3
S. nemorosa	NE	12	3
S. officinalis	OF	79	9
S. pratensis	PR	513	96
S. pratensis subsp. *bertolonii*	PR	25	3
S. rosmarinus	RO	39	8
S. saccardiana / *S. pratensis* subsp. *saccardiana*	SA / PR	128	15
S. sclarea	SC	10	1
S. verbenaca	VE	130	25
S. verticillata	VT	15	3

Table 31. List of species abbreviations and statistics for the *Salvia* collection.

Province name	Province code	N° samples	N° populations
Alessandria	AL	18	4
Arezzo	AR	41	7
Bari	BA	19	3
Belluno	BL	13	3
Bergamo	BG	51	6
Bologna	BO	32	12
Bolzano	BZ	12	1
Cagliari	CA	20	4
Campobasso	CB	21	7
Chieti	CH	10	1
Como	CO	45	14
Cremona	CR	20	2
Ferrara	FE	9	3
Firenze	FI	21	3
Foggia	FG	23	4
Forlì-Cesena	FC	35	7
Frosinone	FR	34	4

Grosseto	GR	15	3
Isernia	IS	16	4
L'Aquila	AQ	74	10
Latina	LT	9	2
Lecco	LC	35	4
Mantova	MN	57	6
Milano	MI	41	10
Monza e Brianza	MB	10	1
Padova	PD	6	2
Palermo	PA	9	1
Pavia	PV	44	8
Pesaro e Urbino	PU	24	5
Pescara	PE	10	1
Piacenza	PC	10	2
Pistoia	PT	4	1
Pordenone	PN	42	4
Potenza	PZ	25	5
Ravenna	RA	13	3
Reggio Calabria	RC	18	3
Roma	RM	8	3
Savona	SV	1	1
Siena	SI	20	4
Sondrio	SO	10	1
Taranto	TA	20	2
Terni	TR	1	1
Trapani	TP	10	4
Trento	TN	49	7
Treviso	TV	32	6
Trieste	TS	40	4
Varese	VA	30	12
Verbano-Cusio-Ossola	VB	1	1
Vercelli	VC	4	1
Verona	VR	19	3
Vicenza	VI	86	12
Viterbo	VT	10	1

Table 32. List of location abbreviations and statistics for the *Salvia* collection.

Location	Source
M. Grappa, Valle di Santa Felicita (VI)	Portal to the Flora of Italy, *Salvia saccardiana* (Pamp.) Del Carr. & Garbari, observation by Sergio Ballestrin (*Salvia Saccardiana (Pamp.) Del Carr. & Garbari - Portale Della Flora d'Italia / Portal to the Flora of Italy*, n.d.); (Carratore et al., 1999)
Val Tramontina (PN)	Acta Plantarum, *Salvia saccardiana* (Pamp.) Del Carr. & Garbari, observation by Giovanni Dose, 2010 (*Salvia Saccardiana (Pamp.) Del Carr. & Garbari, Acta Plantarum*, 2010)
Valstagna (VI)	Previous sampling by Silvio Scortegagna (2022); (Carratore et al., 1999)
Costa (VI)	Previous sampling by Silvio Scortegagna (2022)
Tezze (TN)	(Carratore et al., 1999)
Meduno (PN)	Observation on INaturalist by manuelgrifalconi 21/04/2024 (Grifalconi, 2024)

Table 33. List of locations visited for *Salvia saccardiana*.

Althaea officinalis

For **flow cytometry analysis** and self-pollinations tests of *Althaea officinalis*, two sets of plants were cultivated at the Città Studi Botanical Garden and in Aboca fields. A small set of seedlings from Aboca crops was repotted in Milan in universal soil. Young, undamaged leaves were collected for flow cytometry analysis and the plants where then used for self-pollination in summer 2023, by bagging flowers to avoid outcrossing. Selfed seeds were harvested in 2024 and are ready to be sown again in 2025 to increase homozygosity and facilitate genome assembly. An experimental field was set up in Aboca with 200 plants. 100 of these were collected at the end of their vegetative period and their fresh weight was measured. Then, fresh and dry weight of roots was measured. Young leaf samples were collected from 100 plants for genetic analyses. The leaves were immediately placed in a test tube and put in dry ice for transport from the field to the lab and then flash-frozen in liquid nitrogen and preserved at -80°C.

Flower morphology and measurements

Limited to *S. pratensis* we quantified sex ratios (hermaphrodite: female flowers) in wild populations by checking stamen morphology in 20 plants per patch when possible, three flowers in each plant, keeping a distance of at least 2m between different sampled plants. We assigned the flowers to three flower phenotypes (hermaphrodite, female or short stamen). In almost all cases, plants produced only one flower phenotype, and so we assigned that phenotype category to the plant.

For flower measurements, we picked three flowers per plant, one from the top, one from the bottom, and one from the middle of the inflorescence. The flowers were kept moist in water and measured with a digital calliper. We took measurements of the flower's "mouth" opening (AB), the length of the corolla (LC), the length of the stylus once extracted from the corolla (LS), and the length of the stamen dividing filament (LF) from anthers (LA) once carefully removed from the flower, using tape to straighten them to ensure measurement accuracy.We took pictures of different flowers under a stereomicroscope. A PCA of measured traits was conducted in R with the Stats package v4.3.1.

Flow cytometry

Flow cytometry was performed on 25 individual *A. officinalis* plants in order to estimate genome size. Nuclei extraction and staining were performed in Aru buffer following (L. Yang et al., 2019) using as a standard *Nicotiana benthamiana*. Samples of *A. officinalis*, *N. benthamiana* and the mixed samples (for which leaves of the two plants were chopped at the same time in the same plate) were analysed with a BD Accuri™ C6 Personal Flow Cytometer using filters 585/40, 670 LP to visualize propidium iodide-stained nuclei. Data was analysed with BD Accuri™ C6 Plus Software. To eliminate debris, a PE-A lower threshold of 10,000 was set. Additionally, gates were set on the PE-A/PE-H and FITC-A/PE-A density plots. 5,000 events inside the filtering gate were recorded aiming for at least 1,000 events inside each 2C peak. Afterwards, gates were set manually

on each fluorescence histogram to separate peaks for the sample and standard. For each peak, the number of events, mean fluorescence (mean PE-A) and CV of PE-A were evaluated. Finally, the 2C value for the sample (*Althaea*) was computed with the formula: 2C (sample) = 2C (standard) * mean PE-A(sample)/mean PE-A(standard), where **standard** is *Nicotiana benthamiana* with a 2C value of 6.4 pg.

Nucleic acid extraction

Whole genome sequencing

For short-read sequencing, genomic DNA was extracted from fresh leaf material of three *Salvia pratensis* accessions with a CTAB-based protocol. Around 200 mg of fresh leaves were frozen with liquid nitrogen and quickly ground into a fine powder with a mortar and pestle. 800 µl of CTAB buffer (CTAB 2%, Tris-HCl pH 8 100 mM, EDTA pH 8 20 mM, NaCl 1.4 M), pre-heated to 60°C, and 1.6 µl of β-mercaptoethanol 0.2% were added to each sample in centrifuge tubes and mixed. Tubes were incubated at 60°C for 40 minutes mixing periodically. The same volume of phenol-chloroform-isoamyl alcohol 24:24:1 was added (800 µl) and the samples were centrifuged at 13,000 rpm for 10 minutes. Surnatant was carefully extracted and mixed with 550 µl of isopropanol. Samples were incubated at -20°C for 30 minutes and centrifuged at 13,000 rpm for 30 minutes. A washing in ice-cold 70% ethanol was performed followed by another centrifuge at 13,000 rpm for 5 minutes. Surnatant was carefully removed without disturbing the pellet and washing was repeated and the pellet was dried at 37°C for 5 minutes to ensure evaporation of the ethanol. The genomic DNA was resuspended in 50 µl of MilliQ water and conserved at -20°C. DNA was extracted from four samples taken from the same individual for each accession and merged in the same sample after extraction. A purification step was performed using a Monarch Genomic DNA Cleanup kit by New England BioLabs (NEB #T3010) following the Enzymatic Cleanup Protocol according to the manufacturer's instructions. DNA quality was assessed with electrophoresis on agarose gel and DNA purity was analyzed using a Thermo Scientific Nanodrop One. Finally, DNA concentration was determined with a QuBit 4 fluorometer by Thermo Fisher Scientific.

For HiFi sequencing, high molecular weight (HMW) DNA extraction was performed by the Next Generation Sequencing Platform (University of Bern, Bern, Switzerland) starting from flash-frozen leaves of accession P11.

Transcriptome sequencing

RNA was extracted with the MN NucleoSpin spin column kit (Macherey-Nagel) or with a modified LiCl protocol followed by Qiagen RNeasy spin column purification. For the LiCl protocol, frozen samples were ground with a small pestle, then 500 µl of extraction buffer (100 mM Tris-HCl pH 7.5, 1% SDS, 100 mM LiCl, 10 mM EDTA) and 400 µl of phenol were added before vortexing. The samples were centrifuged at 3,000 rpm for 5 minutes and the surnatant was carefully extracted. LiCl 8 M was added, 1/3rd of the volume of the surnatant. Samples were kept overnight at 4°C, then centrifuged the next day at 4°C, 13,000 rpm for 30

minutes. After removing the surnatant, the pellet was washed twice with 500 µl EtOh 70%. The RNA pellet was dried and suspended in 20 µl of water. Quality was assessed with electrophoresis on agarose gel and DNA purity and concentration were analyzed using a Thermo Scientific Nanodrop One.

Optimization of DNA extraction from silica-dried leaf for GBS

Starting from dried leaf, our first question was whether we would be able to obtain genomic DNA of sufficient quality for NGS approaches. Most samples arrived in Milan in good condition. Usually, silica gel was replaced upon arrival to the lab. Silica gel was also dried in an oven and recycled. Several DNA extraction methods were tested, from a CTAB buffer-based method to different commercial kits (**Figure 72** - **Figure 74**).

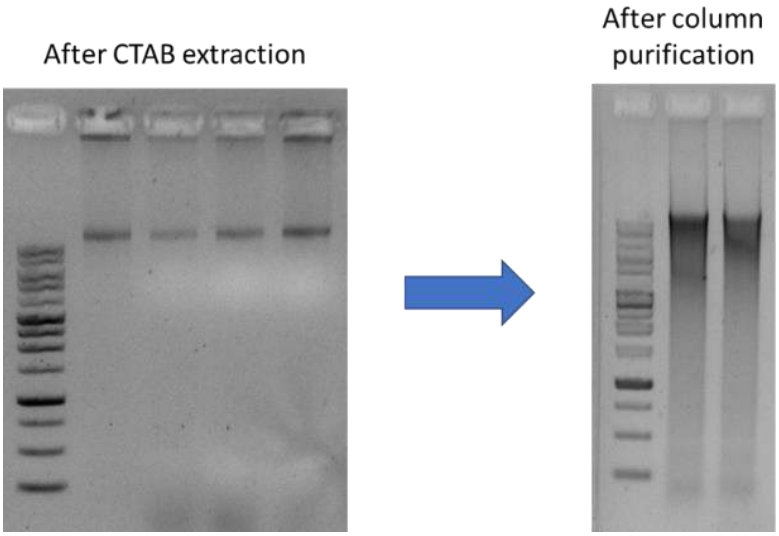


Figure 72. Negative effect of column purification on CTAB-extracted samples.

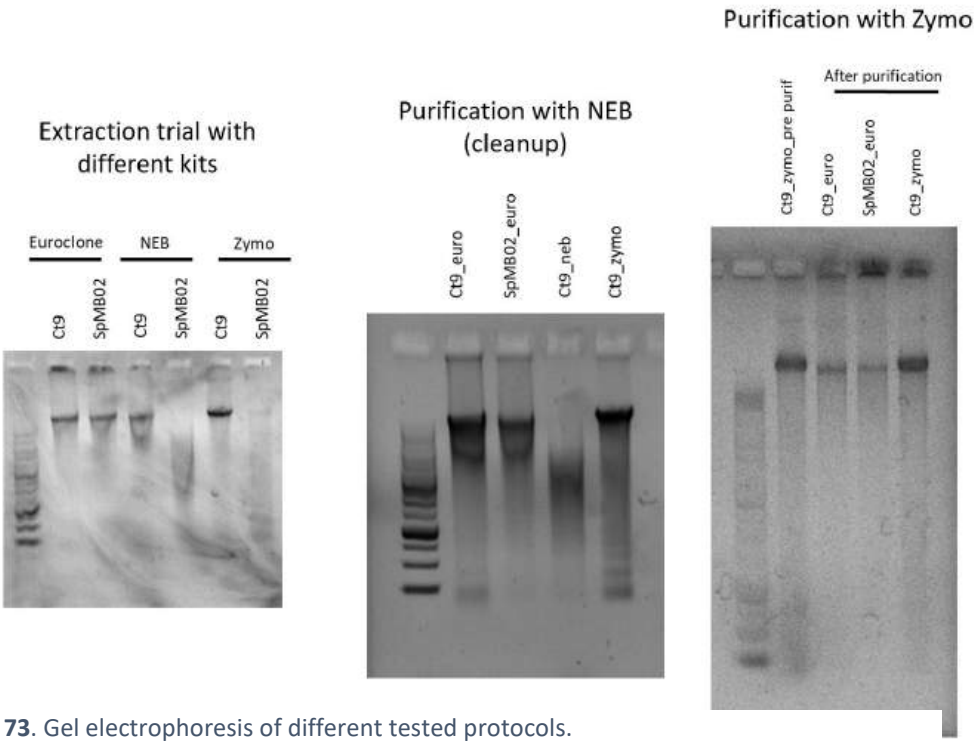


Figure 73. Gel electrophoresis of different tested protocols.

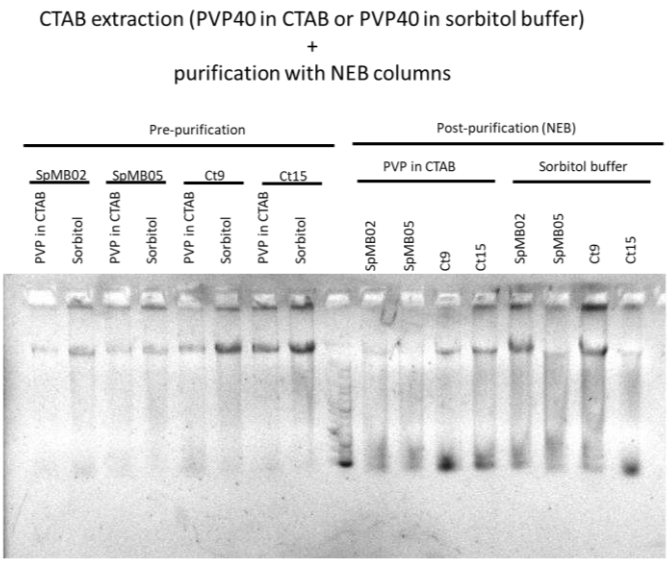
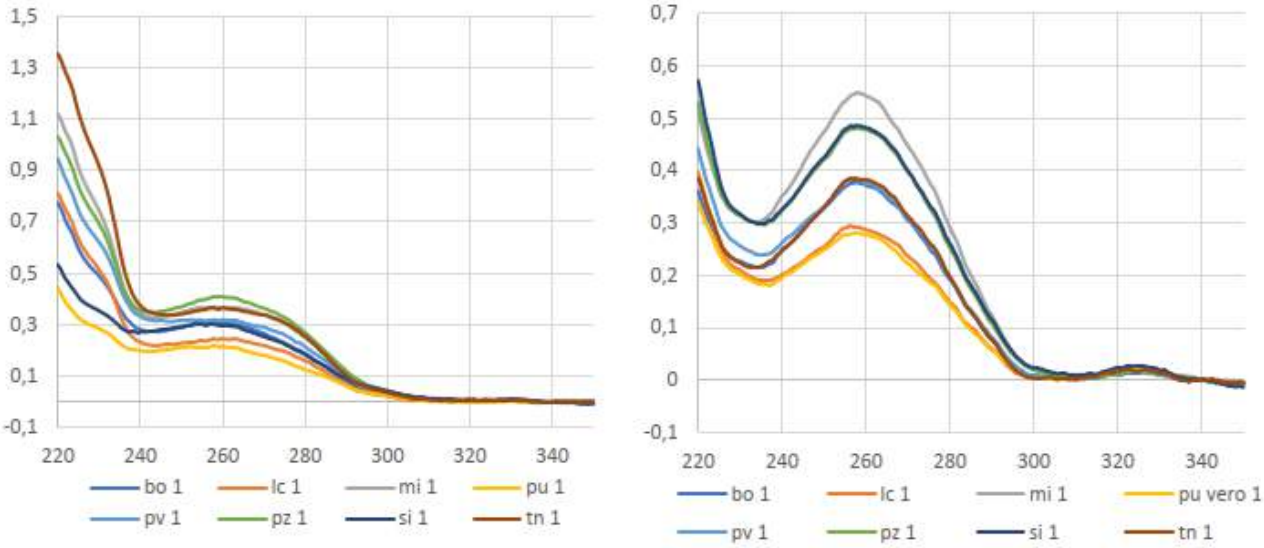


Figure 74. Different tested kits. Euroclone followed by Zymo purification gives the best result for *Salvia* in terms of DNA integrity, at the expense of DNA quantity.

The final chosen protocol is as follows: 20 mg of silica-dried leaves of each sample were flash-frozen in liquid nitrogen on the day of the extraction and ground into a fine powder with the use of a Qiagen TissueLyser and stainless steel beads, for one minute at 30 Hz two times, and immersed into LN2 again after the first grinding round. Afterwards, we performed DNA extraction with the Euroclone SpinNAKER Universal Genomic DNA kit according to the manufacturer's instruction with the following adjustments: 1 hour of proteinase K digestion; first centrifuge 21,000g for 2'; following centrifuges at 14,000g; warming the elution buffer at 80°C; incubation of the column with the elution buffer for 10' at each elution step; double elution in elution buffer, each with 50 ul of buffer, for a final volume of 100 ul of DNA. Following extraction, most samples still had high levels of impurity, so we followed up with a DNA purification using the Zymo Genomic DNA Clean & Concentrator kit with elution buffer at 65°C and eluting twice with 15 ul for a final volume of 30 ul. DNA concentration and quality were assessed with a Nanodrop spectrophotometer and a Qubit florometer (5-20 ng/ul). DNA integrity was assessed with 0.8% agarose gel electrophoresis (**Figure 75**). The Zymo kit results often in a loss and degradation of DNA. Integrity of genomic DNA is important to make sure that the digestion efficiency is as high as possible and that the highest number of selected fragments is recovered. For this reason, for the latest collected samples, the purification step was skipped, after assessing that digestion efficiency would not be negatively impacted, to prioritize saving as many samples as possible at an acceptable concentration. While the yield with these kits was relatively low, never higher than 30 ng/ul and more often around 10 ng/ul or even lower, the DNA retrieved with more efficient methods such as CTAB tended to have too many impurities and fail in the cleanup step.

		EXTRACTION			PURIFICATION					
Sample Name	q.ta (mg)	Nucleic Acid(ng/uL)	A260/ A280	A260/ A230	Nucleic Acid(ng/uL)	A260/ A280	A260/ A230	QUBIT		
								1	2	Mean
bo 1	21.9	15.479	1.641	0.632	18.628	1.941	1.638	15.5	15.6	15.55
lc 1	19.8	12.226	1.527	0.475	14.337	1.935	1.368	12.5	12.6	12.55
mi 1	21.9	18.375	1.468	0.494	27.163	1.864	1.73	25.2	25.4	25.3
pu 1	22.7	10.734	1.7	0.752	13.874	1.935	1.376	11.4	11.5	11.45
pv 1	23.2	16.002	1.488	0.515	18.807	1.902	1.468	16	16	16
pz 1	21.6	20.364	1.513	0.587	23.881	1.887	1.521	20	20.2	20.1
si 1	19.4	14.941	1.616	0.842	24.114	1.851	1.521	16	16	16
tn 1	19.5	18.232	1.422	0.397	19.111	1.937	1.709	16	15.8	15.9



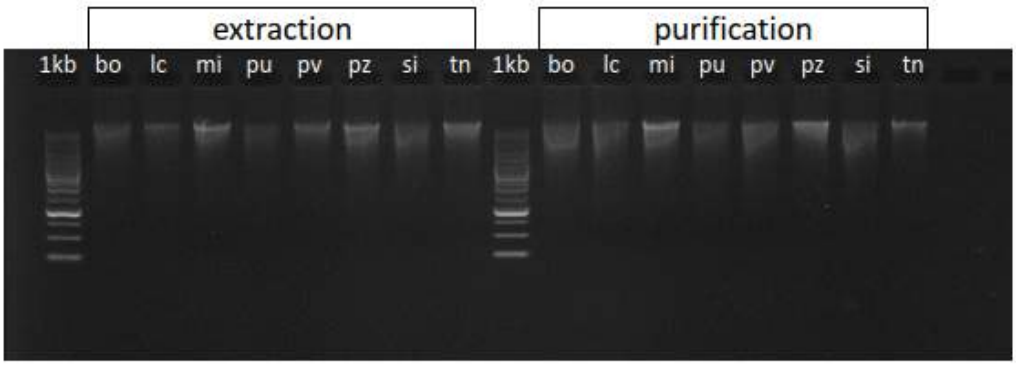


Figure 75. Example of results for the final selected protocol for DNA extraction and purification. Extraction was performed with the Euroclone spinNAker kit. Purification was performed with the Zymo Genomic DNA Clean & Concentrator kit. The protocol chosen allows for purification of relatively clean DNA that is not significantly degraded. This set of samples comes from different populations of Salvia pratensis.

GBS library preparation

GBS library preparation largely followed Elshire *et al.*, 2012 (adapted by S. Manrique, 2022) with a few modifications. Tables after various steps report the original protocol compared to the other tested protocols.

0. **Preparation of adapters**. 96 barcoded adapters and 1 common adapter Table 34were prepared as follows. The lyophilized adapter was resuspended in sterile TE for a concentration of 200 pmol. 25 ul of forward and 25 ul of reverse adapter were added to a tube along with 50 ul of TE. Then, annealing of each adapter with its complementary (upper and lower adapter) to form a double stranded adapter was performed in a thermocycler as follows:
    - 2' at 95°C
    - Ramp from 95°C to 25°C, lowering the temperature by 0.1°C per second
    - 30' at 25°C

    To prepare the first dilution (DIL1), 5.6 ul of annealed barcoded adapter was mixed with 995 ul of TE; while 100 ul of common adapter was mixed with 900 ul of TE. The amount of double stranded DNA was measured with a Qubit instrument using a High sensitivity dsDNA kit (Invitrogen). Then, adapter stocks (STOCK) were prepared mixing 300 ng of each annealed barcoded adapter with 300 ng of common adapter and with TE for a final volume of 200 ul. Then, a working adapter solution (WORK) was prepared diluting the STOCK 1:4 with sterile milliQ water. The concentration of this mix is 0.6 ng/ul.
1. **Dilution of DNA**. Genomic DNA was diluted to a concentration of 10 ng/ul with sterile water or TE.
2. **Digestion**. 50 ng of DNA for each sample were digested with ApeKI (see table for the reagents). Digestion was performed in a thermocycler at 75°C for 4h with lid on at 80°C. To check success of the digestion 5 ul per sample were run on a 2% agarose gel electrophoresis and compared to a negative control without ApeKI and a negative control of a DNA sample without the reaction mix. The following tests were performed to obtain the most efficient digestion conditions for our samples (**Figure 76, Figure 77**).

	DIGESTION			
Reagent	**Original (ul)**	**Set 2 (ul)**	**Set 3 (ul)**	**Final (ul)**
H2O	7	47	12	15
NEB buffer 3.1 10X	2	2	2	2.5
ApeKI	1	1	1	2.5
DNA	10	10	10	5
Total reaction volume	20	50	25	25
DNA amount in digestion	100 ng	100 ng	50 ng	50 ng
Digestion time at 75°C	2h	2h	2h	4h

3. **Ligation**. 6 ul of 0.6 ng/ul mix of the barcoded and common adapters were added to each sample (20 ul). Then, 5 ul of NEB T4 ligase 10x reaction buffer and 1.6 ul of NEB T4 DNA ligase were added, along with 17.4 ul water for a final reaction volume of 50 ul. Ligation was performed with 2h at 22°C, lid off, in a thermocycler, followed by 30' at 65°C to inactivate the enzyme.
4. **Pooling and purification of barcoded fragments**. 10 ul for each sample were mixed and the resulting DNA was split in two halves and purified with the use of two NEB Monarch ® PCR & DNA cleanup kit (#T1030L) columns according to the manufacturer's instructions. Each column was eluted in a final volume of 24 ul for a final amount of 48 ul of template pooled DNA.
5. **PCR amplification**. The clean pooled template DNA was amplified in a PCR reaction with the aim of performing a rough fragment size selection (100-500 bp) and adding the Illumina primers to the oligonucleotide sequences of the adapter sequences The reaction mixes tested to obtain the most efficient amplification conditions were the following (**Figure 78**).

AMPLIFICATION				
Reagent	**Original**	**Alt 1 (final)**	**Alt 2**	**Alt 3**
H2O	36	34	34	32
NEB 5x Taq Master Mix	10	10	10	10
Primer 1 12.5 uM	1	2	1	2
Primer 2 12.5 uM	1	2	1	2
Total	48	48	46	46
Clean pooled ligated DNA	2	2	4	4
Total	50	50	50	50

The cycle conditions were as follows:

- 60'' at 95°C
- 18 cycles of:
  - 30'' at 95°C
  - 20'' at 62°C
  - 30'' at 68°C
- 5' at 68°C

6. **Cleanup**. 8 independent PCR reactions for each 96x library were combined in 2 tubes and each of the tubes was purified with the NEB Monarch ® PCR & DNA cleanup kit and eluted in a final volume of 20 ul. The two elutions were combined and the fragment size profile of the library iwas analysed with a Tapestation using the High sensitivity D1000 ScreenTape (Agilent) and the concentration measured with a Qubit Fluorometer with the dsDNA High sensitivity kit (Invitrogen) (about 40 ng/ul). The library was then sent for Illumina sequencing (Genewiz/Azenta).

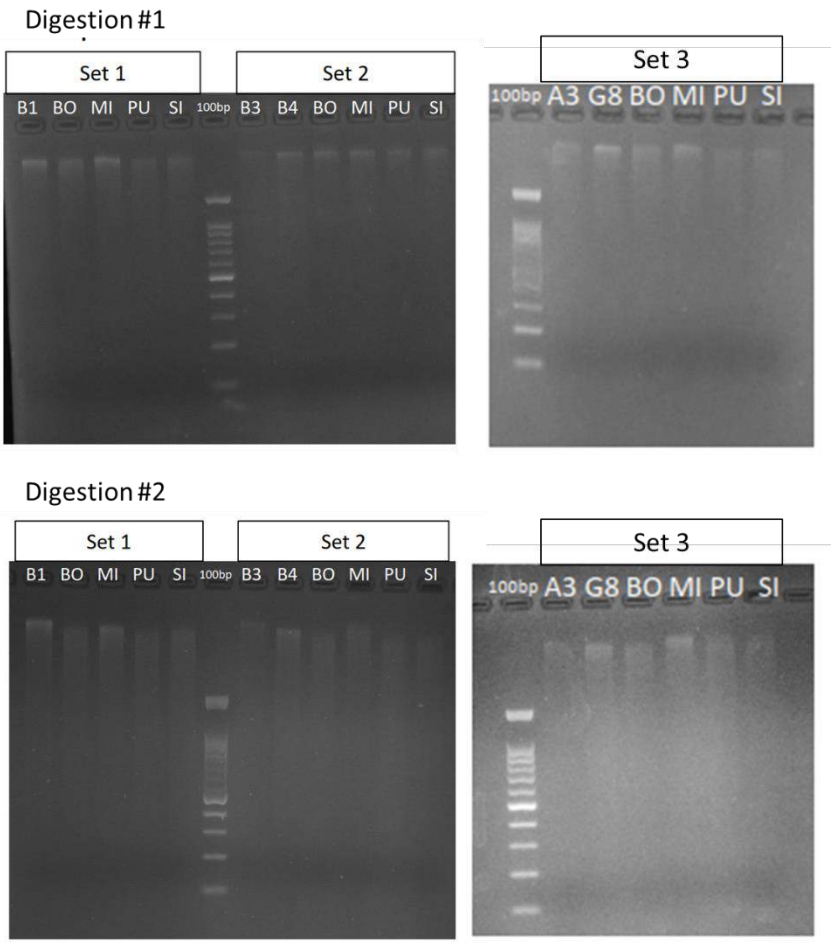


Figure 76. Optimization of the ApeKI digestion step. The different digestion protcols were compared after ligation on the efficiency of the amplification reaction. Given the uniformity of the PCR product, the conditions of Set 3 were selected for further improvement. Afterwards, the conditions of Set 3 (50 ng of DNA, 2.5 ul of ApeKI, 2.5 ul of r3.1 10X buffer, final reaction volume of 25 ul) gave the best and most consistent results for digestion and amplification.

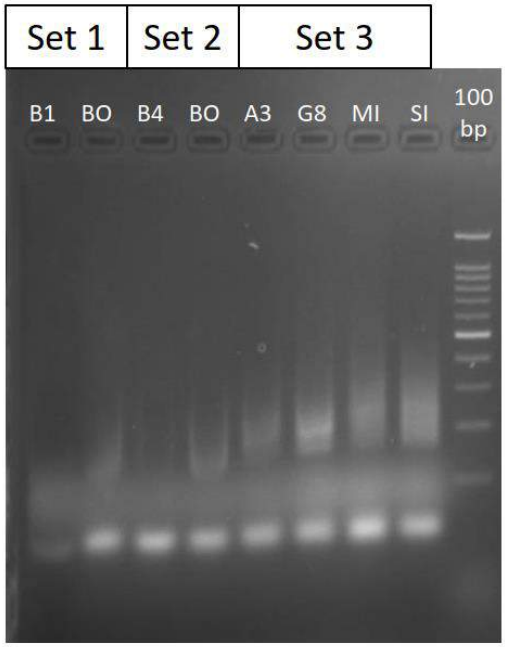


Figure 77. Optimization of the ApeKI digestion step. The different digestion protcols were compared after ligation on the efficiency of the amplification reaction. Given the uniformity of the PCR product, the conditions of Set 3 were selected for further improvement. Afterwards, the conditions of Set 3 (50 ng of DNA, 2.5 ul of ApeKI, 2.5 ul of r3.1 10X buffer, final reaction volume of 25 ul) gave the best and most consistent results for digestion and amplification.

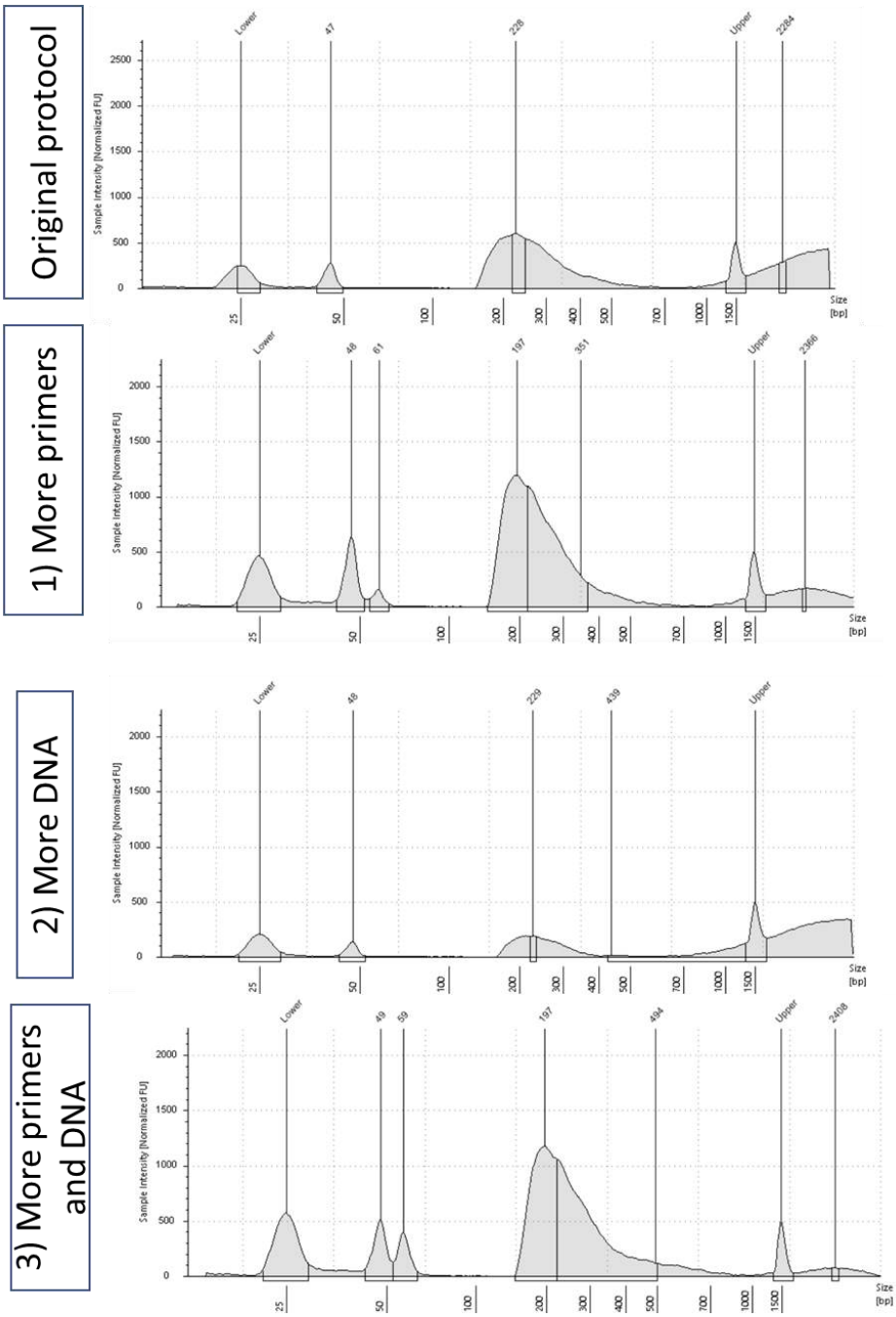


Figure 78. Optimization of the PCR step. Shown is capillary electrophoresis of a GBS sequencing library on a pool of 48 Salvia samples, performed with a Tapestation instrument. While the length of amplified fragments is similar between different protocols, the efficiency of amplification (height of the peak) changes depending mostly on the amount of primers used. Top: original protocol. 1) Amount of forward and reverse primer doubled. 2) Amount of template pooled DNA doubled. 3) Both template and primers doubled.

Table 34. List of adapters and barcodes for GBS library preparation.

Primers used for the *Salvia* library	
Common adapter positive strand	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Common adapter negative strand	CWG-AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Indexed adapter positive strand	ACACTCTTTCCCTACACGACGCTCTTCCGATCT-barcode
Indexed adapter negative strand	CWG-barcode-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
PCR primer A	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PCR primer B	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Newly developed primers (in testing)	
Common adapter positive strand	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Common adapter negative strand	CWG-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
New PCR primer with i5 index	AATGATACGGCGACCACCGAGATCTACGATCAGACACTCTTTCCCTACACGACGCTCTTCCGATCT
New PCR primer with i7 index	CAAGCAGAAGACGGCATACGAGATATATGCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

BC #	Barcode sequence (+ strand)
1	CTCC
2	TGCA
3	ACTA
4	CAGA
5	AACT
6	GCGT
7	CGAT
8	GTAA
9	AGGC
10	GATC
11	TCAC
12	TGCGA
13	CGCTT
14	TCACC
15	CTAGC
16	ACAAA
17	TTCTC
18	AGCCC
19	GTATT
20	CTGTA
21	ACCGT
22	GCTTA
23	GGTGT
24	AGGAT
25	ATTGA
26	CATCT
27	CCTAC
28	GAGGA

BC #	Barcode sequence (+ strand)
49	CTATTA
50	GCCAGT
51	GGAAGA
52	GTACTT
53	GTTGAA
54	TAACGA
55	TGGCTA
56	TATTTTT
57	CTTGCTT
58	ATGAAAC
59	AAAAGTT
60	GAATTCA
61	GAACTTC
62	GGACCTA
63	GTCGATT
64	AACGCCT
65	AATATGC
66	ACGTGTT
67	ATTAATT
68	ATTGGAT
69	CATAAGT
70	CGCTGAT
71	CGGTAGA
72	CTACGGA
73	GCGGAAT
74	TAGCGGA
75	TCGAAGA
76	TCTGTGA

29	GGAAC
30	GTCAA
31	TAATA
32	TACAT
33	TCGTT
34	GGTTGT
35	CCAGCT
36	TTCAGA
37	TAGGAA
38	GCTCTA
39	CCACAA
40	CTTCCA
41	GAGATA
42	ATGCCT
43	AGTGGA
44	ACCTAA
45	ATATGT
46	ATCGTA
47	CATCGT
48	CGCGGT

77	TGCTGGA
78	ACGACTAC
79	TAGCATGC
80	TAGGCCAT
81	TGCAAGGA
82	TGGTACGT
83	TCTCAGTC
84	CCGGATAT
85	CGCCTTAT
86	AACCGAGA
87	ACAGGGAA
88	ACGTGGTA
89	CCATGGGT
90	CGCGGAGA
91	CGTGTGGT
92	GCTGTGGA
93	GGATTGGT
94	GTGAGGGT
95	TATCGGGA
96	TTCCTGGA

Sequencing, read quality evaluation and trimming

Short read whole genome sequencing was performed by BGI Genomics on three DNA samples from three *Salvia pratensis* accessions on a DNBSEQ platform and produced three 150-bp paired-end libraries. Reads underwent a first cleaning procedure by BGI with SOAPnuke to eliminate reads with more than 28% match to the adapter sequence, more than 50% bases <Q20, or more than 3% N. Once this data was delivered, we performed additional quality control and trimming with Fastq-mcf v1.05 (Aronesty, 2011) with parameters -q 30 -l 50. Base read quality statistics were computed with both Fastq-stats v1.01 (Aronesty, 2011) and FastQC v0.11.5 (Andrews, n.d.) and aggregated with MultiQC v1.11 (Ewels et al., 2016).

ONT libraries were prepared and sequenced by KeyGene N. V. (Wageningen, The Netherlands) on a PromethION instrument with one R9.4.1 cell using the 1D Genomic DNA by ligation SQK_LSK110 library prep kit on 2 ug of Ampure purified genomic DNA. The library was loaded on one PromethION FLO PRO002 (R9.4.1 pore) flow cell. After ~24hrs a nuclease flush was performed and another ~375 ng library was loaded to increase the yield. Basecalling was performed (PromethION version: 21.05.20/Guppy5.0.13/MinKNOW 4.3.11) using the Super accuracy model.

HiFi libraries were prepared by the Next Generation Sequencing Platform (University of Bern, Bern, Switzerland) on the P11 *S. pratensis* accession and sequenced on two 8M SMRT cells. After HMW gDNA extraction, the DNA was sheared with Megaruptor and libraries were prepared with the SPK3.0 SMRTbell library prep kit and a bead-based size selection to remove fragments shorter than 5kb. A second library was

size selected to remove fragments below 10kb. Reads from both cells were merged and read length and quality were inspected with LongQC (Fukasawa et al., 2020).

Stranded **mRNA** libraries were prepared by IGATech (Udine, Italy) and sequenced in paired-end 150bp mode on a NovaSeq 6000 instrument (Illumina, San Diego, CA). Read trimming and filtering was performed as described previously for DNA sequencing. Reads were mapped on the *S. pratensis* genome with STAR (Dobin et al., 2013). StringTie (M. Pertea et al., 2015) was ran on the alignments to generate transcript models and TransDeCoder (B. Haas, n.d.), Diamond (Buchfink et al., 2021) and HMMscan (*HMMER*, n.d.) were employed to identify the coding sequences.

GBS libraries were sequenced with a NovaSeq 6000 instrument (Illumina) in paired-end 150bp mode by Azenta Life Sciences (Leipzig, Germany). Demultiplexing was performed with STACKS 2.0 with parameters -e apeKI -r –inline-null. While reads coming from WGS start at random positions throughout the genome, RRS reads are extremely homogeneous, as they all start with the restriction start site and the barcode. Thus, regular read trimming approaches can remove these parts of the read, which are important to keep for demultiplexing. As such, we disabled the "skew" option from the read trimming software, which would have removed these. Moreover, RRS reads show a higher percentage of "PCR duplicates" in QC analysis, for the same reason. We changed the -w parameter, which is the window size for quality and homopolymer trimming, to 2 to remove the runs of G at the tail of reads coming from particularly short fragments. Since we were not able to perform size selection, raw reads have an extremely high percentage of adapter content, due to the presence of short fragments. Quality and adapter trimming was performed with Fastq-mcf v1.05 with parameters -q 30 -l 50 -w 2 -k 0 supplying the adapter sequences to remove low quality bases or reads shorter than 50bp after trimming. Afterwards, an additional filter by STACKS was applied with parameters -e apeKI -r -c -q and the adapter sequences.

k-mer analysis

Jellyfish 2.3.0 (Marçais & Kingsford, 2011) was used for k-mer counting and for producing the k-mer count histogram with the canonical k-mer counting option (-C) on trimmed *S. pratensis* short WGS reads. The histogram file was then uploaded to the GenomeScope web tool (Vurture et al., 2017) with the following parameters: read length 150, max k-mer coverage 2000 or -1.

Salvia pratensis genome assembly and annotation

ONT assembly

ONT reads, filtered for a minimum length of 5000 bp, were assembled with several assemblers. Canu v2.2 (Koren et al., 2017) was run with parameters genomeSize=430m minReadLength=5000. Raven v1.5.0 (Vaser & Šikić, 2021) was run with default parameters. Shasta v0.7.0 (Shafin et al., 2020) was tested with different parameter presets: Nanopore-Oct2021, Plants-Apr2021, and Plants-Apr2021 with a minimum read length of

10,000 bp (the latter option was included in the results). Flye v2.9-b1768 (Kolmogorov et al., 2019) was run with option –nano-hq.

Polishing of ONT assemblies was performed with Polca, downloaded with MaSuRCA-4.0.5, using the short read set (Zimin et al., 2013). NextPolish v1.4.0 (Hu et al., 2020) was employed with the following options: task = best, rerun = 3, genome_size = auto, sgs_options = -max_depth 100 -bwa, lgs_options = -min_read_len 5k -max_depth 100, lgs_minimap2_options = -x map-ont.

HiFi assembly

HiFi reads were assembled using Hifiasm v0.19.0-r534 (H. Cheng et al., 2021), Canu and Flye. For Canu, parameter -pacbio-hifi and genomeSize=430m were set. For Flye default parameters were used. For Hifiasm, the –primary flag was set and all following analyses were run on the primary assembly (p_ctg.gfa) Blobtools2 v2.6.4 (Challis et al., 2020) was employed to screen for potential contamination in the assembled sequences through BLAST (Camacho et al., 2009) searches and read alignments. According to Blobtools2 recommendations, BLAST hits were obtained with BLAST command line tools, version 2.12.0+,against the NCBI nucleotide (nt) dataset, downloaded on Feb 14, 2022, with the following parameters: -outfmt "6 qseqid staxids bitscore std salltitles" -max_target_seqs 10 -max_hsps 1 -evalue 1e-25. Genome indexing and short read alignments were performed with Bowtie-2 v2.3.4.1 (Langmead & Salzberg, 2012), while long read alignments with Minimap v2.1 (H. Li, 2018) with flag map-hifi. Contigs were then excluded from the assembly when the coverage from short reads was higher than 300x to remove short, highly repetitive or organellar sequences.

Purge_haplotigs v1.1.2 (Roach et al., 2018) was employed for the haplotype purging step through a 3-step process: generating a coverage histogram of HiFi reads (purge_haplotigs hist); analyzing coverage per contig and flagging suspicious contigs with coverage outside a specific range (purge_haplotigs cov with parameters -l 5 -m 50 -h 150); running the purging pipeline to discard redundant contigs (purge_haplotigs purge with parameter -a 80). Afterwards, purge_haplotigs clip was employed to trim overlapping contig ends.

Organelle genomes

The Conda GetOrganelle v1.7.5.0 release (Jin et al., 2020) was used to assemble chloroplast sequences from read files by using the get_organelle_from_reads.py script with the following parameters: max rounds (-R) 15; target genome size 150,000; organelle type embplant_pt. Then, GeSeq (Tillich et al., 2017) was used to annotate and draw the complete chloroplast sequence. Geseq configuration was as follows: circular; plastid (land plants); annotate IR; annotate rps12; support annotation by Chloë (v0.1.0); keep best annotation; BLAT search; HMMER profile search; ARAGORN v1.2.38; tRNAscan-SE v2.0.7. Other parameters were kept as default. Chloroplast assembly was confirmed through comparison with the contigs assembled from HiFi reads.

To assemble mitochondrial sequences several approaches were employed. First, HiFi reads were mapped on four mitochondrial sequences (*A. thaliana* NC_037304.1, *S. miltiorrhiza* NC_023209.1, *S. officinalis* OQ001564.1, *S. officinalis* OQ001565.1, *S. splendens* CM031494.1) and the mapped reads were assembled with Flye. Contigs matching with BLAST the mitochondrion were selected, thus excluding NUMTs, and long ONT reads were mapped on the assembly graph with GraphAligner. The reads mapping on each edge of the graph were counted with Bash scripts and the sum plotted manually. For manual curation, nodes with double the average coverage were duplicated, and the assembly graph was then solved manually trying to follow the most represented path according to ONT read mapping. Then, the Oatk pipeline (Zhou, 2023/2024), followed by manual curation, was used to assemble mitochondrial sequences from the entire set of HiFi reads. The angiosperm_mito HMM profile from OatkDB v20230210 was downloaded. Oatk was run with parameters: -k 1001 -c 150 -m angiosperm_mito.fam. Bandage (Wick et al. 2015) was employed to visualize the assembled sequences from the assembly graph output by Oatk and visualize the results of BLAST searches. Four of the assembled sequences with mostly mitochondrial matches and coverage around 200x were selected to keep as the *S. pratensis* mitochondrial sequences and the repeat structures were solved manually to obtain a linear fasta file. In particular, unitigs with coverage around 400x were classified as repetitive and duplicated and uniquely linked to unitigs with coverage around 200x (unique).

To evaluate gene space completeness, BUSCO v5.0.0 (Manni et al., 2021) was employed in genome mode with lineage eudicots_odb10. To evaluate k-mer completeness and duplication level, the assembly was compared to the short-read set with Meryl v1.3 and Merqury v1.3 (Rhie et al., 2020) with a k-mer size of 19.

Genome annotation

Repeats

Repetitive sequence annotation was performed with RepeatModeler v2.0.3 (Flynn et al., 2020) and RepeatMasker v4.1.2-p1, accessed through the DFAM-TETools container TETools-1.5 with Singularity v3.11.0-bionic. LAI (LTR Assembly Index) was computed for the assembly with LTR_retriever v2.8 (Ou & Jiang, 2018).

Genes

Gene annotation was performed with several different tools. First, GeMoMa v1.9 (Keilwagen et al., 2019) was employed, using different sets of reference genomes as input data. Then, first BRAKER2 v2.15, then BRAKER3 pipeline v3.0.3 via Singularity container, built 19/10/2023, were used employing as evidence both assembled *S. pratensis* transcripts and protein sequences from *Arabidopsis thaliana*, *S. miltiorrhiza, S. hispanica*, *S. splendens* and *S. rosmarinus* (citations for BRAKER3: Barnett et al., 2011; Bruna et al., 2023, 2023; Buchfink et al., 2015; Gabriel et al., 2021, 2023; Gotoh et al., 2014; Iwata & Gotoh, 2012; Kovaka et al., 2019; H. Li et al., 2009; Lomsadze, 2005; G. Pertea & Pertea, 2020; Quinlan, 2014; Stanke et al., 2006, 2008).

The annotations were evaluated with BUSCO v5.0.0 in protein mode to assess gene space completeness. Ingenannot v0.0.11 (*BIOGER / Ingenannot · GitLab*, 2024) was ran for AED calculation. Statistics and file format conversions were performed with AGAT v1.4.0 (Jacques Dainat, 2024). The best annotation was chosen in terms of number of genes being consistent with other *Salvia*, BUSCO completeness and AED.

UTR refinement was performed with MAKER v2.31.10 (Cantarel et al., 2008; Campbell et al., 2014). MAKER was also employed for genome annotation with Augustus and SNAP gene predictors. Functional annotation of the predicted proteins was performed by sequence homology search with the EMBL SwissProt database and the Arabidopsis protein set performed in DIAMOND (Buchfink et al., 2015). AHRD (Schoof, 2012/2024) was used to integrate the functional annotations, which were added to the gene annotation file with custom Perl scripts. Long noncoding RNA were annotated with the FEELnc pipeline v0.2 (Wucher et al., 2017) employing mode *shuffle*. Annotation of tRNAs was performed with tRNAscan-SE v2.0.5 (Chan et al., 2021) and rRNAs were annotated with Barrnap (Seemann, n.d.); both annotations were converted to gff with Perl scripts (*Jorvis/Biocode: Bioinformatics Code Libraries and Scripts*, n.d.).

Whole genome duplication, synteny and orthology analysis

WGD analysis was performed with ksrates v1.1.4 (Sensalari et al., 2022) via Singularity container. The coding sequences of *S. pratensis* (one isoform per gene) were compared to *A. thaliana*, *Erythrante guttata*, *Oryza sativa, Solanum lycopersicum, Salvia miltiorrhiza, Salvia officinalis*.

Whole genome synteny analysis was also performed with MCScanX (Y. Wang, 2011/2024) and visualized on AccuSyn (Siri, 2019/2024) (accusyn.usask.ca).

The protein sequences of *S. pratensis* were clustered in orthogroups with proteins of *A. thaliana*, *E. guttata*, *O. sativa*, *Solanum lycopersicum*, *Salvia hispanica*, *Salvia miltiorrhiza* and *Salvia officinalis* with Orthofinder v2.5.5 (Emms & Kelly, 2019) and OrthoVenn3 (J. Sun et al., 2023) (orthovenn3.bioinfotoolkits.net).

MADS-box factors

Gene family analysis

The MADS-BOX domain was used as a query to search the predicted protein set with HMMER. To confirm presence of the MADS-box, the candidate genes were analyzed with the Conserved Domain Database (J. Wang et al., 2023) Batch search and with MEME (Bailey & Elkan, 1994), (Bailey et al., 2015) with the following parameters: -protein -mod zoops -nmotifs 20 -minw 6 -maxw 200 -objfun classic -markov_order 0 -minsites 3 -p 10. The domains identified were compared with known ones with Tomtom (MEME Suite). Proteins that lacked a specific hit for the MADS-box domain were excluded. The longest predicted translated CDS of genes were aligned with Clustal-Omega (Sievers et al., 2011) and a phylogenetic tree was obtained with IQ-TREE (Minh et al., 2020), using MODELFINDER (Kalyaanamoorthy et al., 2017) to select the best substitution model; a consensus tree was built from 1000 replicates with UltraFast Bootstrap (Hoang et al., 2018). The midpoint-

rooted tree was visualized with ITOL online (itol.embl.de) (Letunic & Bork, 2024) and TBTools (C. Chen et al., 2020). The Canprot R package v2.0.0 (Dick, 2017) was used for computation of the protein physical properties.

For comparative analysis, a list of MADS-box TFs and their phylogenetic placement for *Arabidopsis thaliana* was downloaded from https://www.arabidopsis.org/browse/gene_family/MADSlike, sourced from (Pařenicová et al., 2003).

To search MADS-box genes in *S. miltiorrhiza* and *S. officinalis*, the reference genomes and annotations were downloaded, then the search was performed as specified for *S. pratensis* only including the longest isoform for each gene. The MADS-box domain of candidates of all three *Salvia* and *A. thaliana* were aligned to classify the proteins in different subfamilies. For each subfamily, full length protein alignments were performed between *Salvia*. Alignment and tree construction were performed as previously described.

Qualitative transcriptional profiling

To investigate expression of the MADS-box genes in the RNAseq performed for annotation, we followed a custom workflow, given the absence of replicates for each condition. The mean coverage per transcript was determined with Bedtools from the STAR mapping files. It was then multiplied by two scaling factors: one for library size, computed for each sample as the average mapped reads across all samples divided by the number of mapped reads of that sample; one for transcript length, computed for each gene as the average transcript length divided by the length of the transcript. The scaled coverage values so derived were split in 5 discrete categories to assign a presence value to each transcript: no presence (0), low presence (1-20), present (21-100), abundant (101-500), highly abundant (>501). Analyses and plots were performed in R.

Population genomics of *Salvia*

Reads were mapped on the *S. pratensis* genome previously developed, using BWA v0.7.17-r1188. Then, bamaddrg (Garrison, 2011/2024) was used to aggregate the files and assign read group names corresponding to each sample; the mapped reads were used for variant calling with FreeBayes (Garrison & Marth, 2012), called through a custom script for multithreading (MultiThreadFreeBayes; (Bombarely, n.d.)) with the following options: -m=30; -q=30; --limit-coverage=10000; --min-coverage=5. Reads for *Campanula thyrsoides* were instead analysed with the denovo map option on STACKS. The variants were then analysed and filtered. The following filters were applied: minimum quality 30; minimum depth 5; no missing genotypes. When specified, a minor allele frequency (MAF) filter of 0.5 was applied. The following samples were excluded from analysis of *Salvia* samples due to an excessive fraction of missing genotypes or a low number of reads: PR.VI.1.1; CE.RC.1.3; DE.CA.1.3; PR.PU.5.1; PR.VI.1.7; CE.RC.3.5; CL.PA.1.6; CL.CA.1.4; CL.CB.1.2; HA.PZ.1.5; PR.LC.3.4.

The vcf file of variant sites was filtered with Vcftools v0.1.15 (Danecek et al., 2011) imported in R with package VcfR v1.1.5.0 and analysed with functions from Ape v5.8 and Poppr v2.9.6 R packages. PCA was performed with function glPca. The variants were additionally analysed with STACKS population module for computation of population genetics statistics. Admixture analysis was performed with ADMIXTURE v1.3.0 (Alexander et al., 2009b). File type conversion from vcf was performed with Plink v1.90b5.2 64-bit. A phylogenetic tree was obtained with IQ-TREE (Minh et al., 2020), using MODELFINDER (Kalyaanamoorthy et al., 2017) to select the best substitution model according to the Bayesian Information Criterion; a consensus tree was built from 1000 replicates with UltraFast Bootstrap (Hoang et al., 2018).

Other reference genomes used

Genome versions used for analyses: *A. thaliana* Araport11 (Reiser et al., 2024) (phytozome-next.jgi.doe.gov/info/Athaliana_Araport11)*; M. guttatus* v2.0 (Hellsten et al., 2013) (phytozome-next.jgi.doe.gov/info/Mguttatus_v2_0)*; O. sativa* v7.0 (Ouyang et al., 2007) (phytozome-next.jgi.doe.gov/info/Osativa_v7_0)*; Salvia miltiorrhiza* GCF_028751815.1 (X. Pan et al., 2023) (ncbi.nlm.nih.gov/datasets/genome/GCF_028751815.1); *Salvia hispanica* GCF_023119035.1 (Zare et al., 2024) (ncbi.nlm.nih.gov/datasets/genome/GCF_023119035.1); *Salvia splendens* GCF_004379255.2 (Jia et al., 2021) (ncbi.nlm.nih.gov/datasets/genome/GCF_004379255.2); *Salvia rosmarinus* (D. Han et al., 2023) (J. Han, 2022); *Salvia officinalis* GWHBJVP00000000 (C.-Y. Li et al., 2022) (ngdc.cncb.ac.cn/gwh/Assembly/25978/show); *Solanum lycopersicum* ITAG4.0 (Hosmani et al., 2019) (phytozome-next.jgi.doe.gov/info/Slycopersicum_ITAG4_0). Retrieved from NCBI Genome (O'Leary et al., 2024), Phytozome (Goodstein et al., 2012), The National Genomics Data Center Genome Warehouse (M. Chen et al., 2021).

Server details

Computations for this work were performed on a Supermicro server with 160 threads, 3 Tb of RAM and 20 Tb of disk space configured in RAID 6, with Ubuntu server OS 4.15.0-147-generic.

References

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246. https://doi.org/10.1186/1471-2105-12-246

Alexander, D. H., Novembre, J., & Lange, K. (2009a). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alexander, D. H., Novembre, J., & Lange, K. (2009b). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alhusain, L., & Hafez, A. M. (2018). Nonparametric approaches for population structure analysis. *Human Genomics*, *12*(1), 25. https://doi.org/10.1186/s40246-018-0156-4

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. https://doi.org/10.1186/s13059-022-02823-7

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews, S. (n.d.). *FastQC, A quality control tool for high throughput sequence data*. Retrieved 18 August 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angel, V. D. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., & Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, *7*, ELIXIR. https://doi.org/10.12688/f1000research.13598.1

Aronesty, E. (2011). *ea-utils: 'Command-line tools for processing biological sequencing data'*. https://github.com/ExpressionAnalysis/ea-utils

Arora, R., Agarwal, P., Ray, S., Singh, A. K., Singh, V. P., Tyagi, A. K., & Kapoor, S. (2007). MADS-box gene family in rice: Genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics*, *8*(1), 242. https://doi.org/10.1186/1471-2164-8-242

Atzei, A. D., & Picci, V. (1982). « Salvia desoleana » Atzei et Picci, specie nuova dell'isola di Sardegna. *Webbia*, *36*(1), 71–78. https://doi.org/10.1080/00837792.1982.10670241

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36.

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), W39–W49. https://doi.org/10.1093/nar/gkv416

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Balant, M., Glasnović, P., Pečnikar, Ž. F., Clementi, M., & Surina, B. (2019). In search of an identity for *Salvia bertolonii* (Lamiaceae). *Phytotaxa*, *413*(2), Article 2. https://doi.org/10.11646/phytotaxa.413.2.2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, *27*(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Bartolucci, F., Peruzzi, L., Galasso, G., Albano, A., Alessandrini, A., Ardenghi, N. M. G., Astuti, G., Bacchetta, G., Ballelli, S., Banfi, E., Barberis, G., Bernardo, L., Bouvet, D., Bovio, M., Cecchi, L., Di Pietro, R., Domina, G., Fascetti, S., Fenu, G., … Conti, F. (2018). An updated checklist of the vascular flora native to Italy. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *152*(2), 179–303. https://doi.org/10.1080/11263504.2017.1419996

Bartolucci, F., Peruzzi, L., Galasso, G., Alessandrini, A., Ardenghi, N. M. G., Bacchetta, G., Banfi, E., Barberis, G., Bernardo, L., Bouvet, D., Bovio, M., Calvia, G., Castello, M., Cecchi, L., Del Guacchio, E., Domina, G., Fascetti, S., Gallo, L., Gottschlich, G., … Conti, F. (2024). A second update to the checklist of the vascular flora native to Italy. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *158*(2), 219–296. https://doi.org/10.1080/11263504.2024.2320126

Bernatchez, L., Ferchaud, A.-L., Berger, C. S., Venney, C. J., & Xuereb, A. (2024). Genomics for monitoring and understanding species responses to global climate change. *Nature Reviews Genetics*, *25*(3), 165–183. https://doi.org/10.1038/s41576-023-00657-y

BIOGER / ingenannot · GitLab. (2024, January 17). GitLab. https://forgemia.inra.fr/bioger/ingenannot

Blanca, J., Pons, C., Montero-Pau, J., Sanchez-Matarredona, D., Ziarsolo, P., Fontanet, L., Fisher, J., Plazas, M., Casals, J., Rambla, J. L., Riccini, A., Pombarelli, S., Ruggiero, A., Sulli, M., Grillo, S., Kanellis, A., Giuliano, G., Finkers, R., Cammareri, M., … Granell, A. (2021). European Vintage tomatoes galore: A result of farmers combinatorial assorting/swapping of a few diversity rich loci. *bioRxiv*, 2021.10.26.465840. https://doi.org/10.1101/2021.10.26.465840

Blaxter, M., Archibald, J. M., Childers, A. K., Coddington, J. A., Crandall, K. A., Di Palma, F., Durbin, R., Edwards, S. V., Graves, J. A. M., Hackett, K. J., Hall, N., Jarvis, E. D., Johnson, R. N., Karlsson, E. K., Kress, W. J., Kuraku, S., Lawniczak, M. K. N., Lindblad-Toh, K., Lopez, J. V., … Lewin, H. A. (2022). Why sequence all eukaryotes? *Proceedings of the National Academy of Sciences*, *119*(4), e2115636118. https://doi.org/10.1073/pnas.2115636118

Boecker, F. (2021). *AHRD: Automatically Annotate Proteins with Human Readable Descriptions and Gene Ontology Terms* [PhD dissertation, Universitat Bonn]. https://bonndoc.ulb.uni-bonn.de/xmlui/bitstream/handle/20.500.11811/9344/6314.pdf?sequence=1&isAllowed=y

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics (Oxford, England)*, *27*(4), 578–579. https://doi.org/10.1093/bioinformatics/btq683

Bombarely, A. (n.d.). *GenoToolBox GitHub page*. Retrieved 18 August 2021, from https://github.com/aubombarely/GenoToolBox

Botanic Gardens Conservation International. (n.d.). *Botanic Gardens and Plant Conservation | Botanic Gardens Conservation International*. Retrieved 6 November 2024, from https://www.bgci.org/about/botanic-gardens-and-plant-conservation/

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2), lqaa026. https://doi.org/10.1093/nargab/lqaa026

Bruna, T., Lomsadze, A., & Borodovsky, M. (2023). *A new gene finding tool GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes*. https://doi.org/10.1101/2023.01.13.524024

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Buldrini, F., Gianaroli, I., Bosi, G., Alessandrini, A., & Santini, C. (2023). Spontaneous vascular flora of the historical monumental cemetery of Modena (N-Italy). *Italian Botanist*, *15*, 111–136. https://doi.org/10.3897/italianbotanist.15.102589

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, *48*(1). https://doi.org/10.1002/0471250953.bi0411s48

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Carratore, F. D., Garbari, F., & Jarvis, C. (1999). Salvia saccardiana (Lamiaceae), an endemic species of NE Italy. *Plant Biosystems*, *133*(2). https://doi.org/10.1080/11263509909381544

Cartney, A. M. M., Formenti, G., Mouton, A., Panis, D. D., Marins, L. S., Leitao, H. G., Diedericks, G., Kirangwa, J., Morselli, M., Salces, J., Escudero, N., Iannucci, A., Natali, C., Svardal, H., Fernandez, R., Pooter, T. D., Joris, G., Strazisar, M., Wood, J., … Mazzoni, C. J. (2024). *The European Reference Genome Atlas: Piloting a decentralised approach to equitable biodiversity genomics* (p. 2023.09.25.559365). bioRxiv. https://doi.org/10.1101/2023.09.25.559365

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics*, *10*(4), 1361–1374. https://doi.org/10.1534/G3.119.400908

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096. https://doi.org/10.1093/nar/gkab688

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Molecular Plant*, *13*(8), 1194–1202. https://doi.org/10.1016/j.molp.2020.06.009

Chen, M., Ma, Y., Wu, S., Zheng, X., Kang, H., Sang, J., Xu, X., Hao, L., Li, Z., Gong, Z., Xiao, J., Zhang, Z., Zhao, W., & Bao, Y. (2021). Genome Warehouse: A Public Repository Housing Genome-Scale Data. *Genomics, Proteomics & Bioinformatics*, *19*(4), 584–589. https://doi.org/10.1016/j.gpb.2021.04.001

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., & Wong, G. K.-S. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*, *7*(3), giy013. https://doi.org/10.1093/gigascience/giy013

Claßen-Bockhoff, R. (2017). Stamen construction, development and evolution in Salvia s.l. *NVEO - NATURAL VOLATILES & ESSENTIAL OILS Journal | NVEO*, 28–48.

Claßen-Bockhoff, R., Speck, T., Tweraser, E., Wester, P., & et al. (2004). The staminal lever mechanism in Salvia L. (Lamiaceae): A key innovation for adaptive radiation? *Organisms Diversity & Evolution*, *4*(3). https://doi.org/10.1016/j.ode.2004.01.004

Colombo, M., Brambilla, V., Marcheselli, R., Caporali, E., Kater, M. M., & Colombo, L. (2010). A new role for the *SHATTERPROOF* genes during *Arabidopsis* gynoecium development. *Developmental Biology*, *337*(2), 294–302. https://doi.org/10.1016/j.ydbio.2009.10.043

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Conti, F., Abbate, G., & Alessandrini, A. (Eds.). (2005). *An annotated checklist of the Italian vascular flora*. Palombi Editori.

Corrias, D. S. (1983). Numeri cromosomici per la flora italiana:948-955. *Informatore Botanico Italiano*, *15*(1), 44–48.

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., & Clamp, M. (2004). The Ensembl Automatic Gene Annotation System. *Genome Research*, *14*(5), 942–950. https://doi.org/10.1101/gr.1858004

Dakskobler, I., Anderle, B., & Vreš, B. (2013). *SALVIA SACCARDIANA, OROBANCHE ALSATICA IN NEKATERE DRUGE NOVOSTI V FLORI SLOVENIJE SALVIA SACCARDIANA, OROBANCHE ALSATICA AND SOME OTHER NOVELITIES IN THE FLORA OF SLOVENIA*. https://www.semanticscholar.org/paper/SALVIA-SACCARDIANA%2C-OROBANCHE-ALSATICA-IN-NEKATERE-Dakskobler-Anderle/c84c49e94c8825809d07a2ef164c4e2ba25f3930

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

De Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology*, *56*(6), 879–886. https://doi.org/10.1080/10635150701701083

del Caldo, L. (1971). Numeri cromosomici per la flora italiana:35-38. *Informatore Botanico Italiano*, *3*, 71–73.

Del Carratore, F., Garbari, F., & Jarvis, C. (1998). The application of the Linnaean names Salvia pratensis, S. Agrestis, S. haematodes, S. verbenaca and S. clandestina (Lamiaceae). *Plant Biosystems*, *132*(2). https://doi.org/10.1080/11263504.1998.10654201

Del Carratore, F., Garbari, F., & Jarvis, C. (1999). *Salvia saccardiana (Lamiaceae)* , an endemic species of NE Italy. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *133*(2), 157–162. https://doi.org/10.1080/11263509909381544

Dick, J. (2017). *canprot: Chemical Analysis of Proteins* (p. 2.0.0) [Dataset]. https://doi.org/10.32614/CRAN.package.canprot

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

EMBL-EBI. (2024, October). *Gene annotation in Ensembl*. https://www.ensembl.org/info/genome/genebuild/index.html

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fang, L., & Wang, K. (2022). Polishing high-quality genome assemblies. *Nature Methods*, *19*(6), 649–650. https://doi.org/10.1038/s41592-022-01515-1

Fernández, P., Amice, R., Bruy, D., Christenhusz, M. J. M., Leitch, I. J., Leitch, A. L., Pokorny, L., Hidalgo, O., & Pellicer, J. (2024). A 160 Gbp fork fern genome shatters size record for eukaryotes. *iScience*, *27*(6), 109889. https://doi.org/10.1016/j.isci.2024.109889

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2018). Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution*, *35*(10), 2585–2593. https://doi.org/10.1093/molbev/msy147

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Formenti, G., Rhie, A., Walenz, B. P., Thibaud-Nissen, F., Shafin, K., Koren, S., Myers, E. W., Jarvis, E. D., & Phillippy, A. M. (2022). Merfin: Improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nature Methods*, *19*(6), 696–704. https://doi.org/10.1038/s41592-022-01445-y

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, *37*(3), 197–202. https://doi.org/10.1016/j.tree.2021.11.008

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, *26*(20), 5369–5406. https://doi.org/10.1111/mec.14264

Fukasawa, Y., Ermini, L., Wang, H., Carty, K., & Cheung, M.-S. (2020). LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 (Bethesda, Md.)*, *10*(4), 1193–1196. https://doi.org/10.1534/g3.119.400864

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2023). *BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA*. https://doi.org/10.1101/2023.06.10.544449

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: Transcript selector for BRAKER. *BMC Bioinformatics*, *22*(1), 566. https://doi.org/10.1186/s12859-021-04482-0

Galasso, G., Conti, F., Peruzzi, L., Alessandrini, A., Ardenghi, N. M. G., Bacchetta, G., Banfi, E., Barberis, G., Bernardo, L., Bouvet, D., Bovio, M., Castello, M., Cecchi, L., Del Guacchio, E., Domina, G., Fascetti, S., Gallo, L., Guarino, R., Gubellini, L., … Bartolucci, F. (2024). A second update to the checklist of the vascular flora alien to Italy. *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *158*(2), 297–340. https://doi.org/10.1080/11263504.2024.2320129

Garrison, E. (2024). *Ekg/bamaddrg* [C++]. https://github.com/ekg/bamaddrg (Original work published 2011)

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing* (arXiv:1207.3907). arXiv. https://doi.org/10.48550/arXiv.1207.3907

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, *15*(8), e1007273. https://doi.org/10.1371/journal.pcbi.1007273

Giordani, A., & Peruzzi, L. (2012). Numeri cromosomici per la flora italiana:1481-1486. *Informatore Botanico Italiano*, *44*(1), 171–174.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Gotoh, O., Morita, M., & Nelson, D. R. (2014). Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics*, *15*(1), 189. https://doi.org/10.1186/1471-2105-15-189

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Gramzow, L., & Theissen, G. (2010). A hitchhiker's guide to the MADS world of plants. *Genome Biology*, *11*(6), 214. https://doi.org/10.1186/gb-2010-11-6-214

Grifalconi, M. (2024, April 21). *Salvia dei prati (Salvia pratensis)*. iNaturalist; iNaturalist. https://www.inaturalist.org/observations/208469620

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K., & Flot, J.-F. (2021). Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics*, *22*(1), 303. https://doi.org/10.1186/s12859-021-04118-3

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas, B. (n.d.). *GitHub—TransDecoder/TransDecoder: TransDecoder source*. Retrieved 23 April 2024, from https://github.com/TransDecoder/TransDecoder

Haas, B. J. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Han, D., Li, W., Hou, Z., Lin, C., Xie, Y., Zhou, X., Gao, Y., Huang, J., Lai, J., Wang, L., Zhang, L., & Yang, C. (2023). The chromosome-scale assembly of the Salvia rosmarinus genome provides insight into carnosic acid biosynthesis. *The Plant Journal*, *113*(4), 819–832. https://doi.org/10.1111/tpj.16087

Han, J. (2022). *The chromosome-scale assembly of the Salvia rosmarinus genome reveals key genes in carnosic acid biosynthesis* (p. 356262331 Bytes) [Dataset]. figshare. https://doi.org/10.6084/M9.FIGSHARE.21443223.V1

Haque, M. S. (1981). Chromosome numbers in the genus Salvia Linn. *Proceedings, Indian Academy of Sciences. Section B, Biological Sciences*, *47*, 419–426.

Hartmann, U., Höhmann, S., Nettesheim, K., Wisman, E., Saedler, H., & Huijser, P. (2000). Molecular cloning of SVP: A negative regulator of the floral transition in Arabidopsis. *The Plant Journal*, *21*(4), 351–360. https://doi.org/10.1046/j.1365-313x.2000.00682.x

Hegland, S. J., Leeuwen, M. V., & Oostermeijer, J. G. B. (2001). Population structure of Salvia pratensis in relation to vegetation and management of Dutch dry floodplain grasslands. *Journal of Applied Ecology*, *38*(6). https://doi.org/10.1046/j.0021-8901.2001.00679.x

Hellsten, U., Wright, K. M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S. R., Schmutz, J., Willis, J. H., & Rokhsar, D. S. (2013). Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences*, *110*(48), 19478–19482. https://doi.org/10.1073/pnas.1319032110

Heslop-Harrison, J. S. (Pat), Schwarzacher, T., & Liu, Q. (2022). Polyploidy: Its consequences and enabling role in plant diversification and evolution. *Annals of Botany*, *131*(1), 1–10. https://doi.org/10.1093/aob/mcac132

HMMER. (n.d.). Retrieved 25 September 2023, from http://hmmer.org/

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hogg, C. J. (2023). Translating genomic advances into biodiversity conservation. *Nature Reviews Genetics*, 1–12. https://doi.org/10.1038/s41576-023-00671-0

Hosmani, P. S., Flores-Gonzalez, M., Geest, H. van de, Maumus, F., Bakker, L. V., Schijlen, E., Haarst, J. van, Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). *An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps* (p. 767764). bioRxiv. https://doi.org/10.1101/767764

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, *36*(7), 2253–2255. https://doi.org/10.1093/bioinformatics/btz891

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Hugouvieux, V., Blanc-Mathieu, R., Janeau, A., Paul, M., Lucas, J., Xu, X., Ye, H., Lai, X., Le Hir, S., Guillotin, A., Galien, A., Yan, W., Nanao, M., Kaufmann, K., Parcy, F., & Zubieta, C. (2024). SEPALLATA-driven MADS transcription factor tetramerization is required for inner whorl floral organ development. *The Plant Cell*, *36*(9), 3435–3450. https://doi.org/10.1093/plcell/koae151

Iwata, H., & Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research*, *40*(20), e161–e161. https://doi.org/10.1093/nar/gks708

Jacques Dainat. (2024). *NBISweden/AGAT: AGAT-v1.4.0* (Version v0.6.3) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3552717

Jenike, K. M., Campos-Domínguez, L., Boddé, M., Cerca, J., Hodson, C. N., Schatz, M. C., & Jaron, K. S. (2024). *Guide to k-mer approaches for genomics across the tree of life* (arXiv:2404.01519). arXiv. http://arxiv.org/abs/2404.01519

Jetha, K., Theißen, G., & Melzer, R. (2014). Arabidopsis SEPALLATA proteins differ in cooperative DNA-binding during the formation of floral quartet-like complexes. *Nucleic Acids Research*, *42*(17), 10927. https://doi.org/10.1093/nar/gku755

Jia, K.-H., Liu, H., Zhang, R.-G., Xu, J., Zhou, S.-S., Jiao, S.-Q., Yan, X.-M., Tian, X.-C., Shi, T.-L., Luo, H., Li, Z.-C., Bao, Y.-T., Nie, S., Guo, J.-F., Porth, I., El-Kassaby, Y. A., Wang, X.-R., Chen, C., Peer, Y. V. de, … Mao, J.-F. (2021). Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome. *Horticulture Research*, *8*, 177. https://doi.org/10.1038/s41438-021-00614-y

Jiao, X., Flouri, T., & Yang, Z. (2021). Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *National Science Review*, *8*(12), nwab127. https://doi.org/10.1093/nsr/nwab127

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02154-5

Jombart, T., & Ahmed, I. (2011). *adegenet 1.3-1*: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

jorvis/biocode: Bioinformatics code libraries and scripts. (n.d.). Retrieved 25 September 2024, from https://github.com/jorvis/biocode/tree/master

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kamari, G., & Garbari, F. (2000). Mediterranean chromosome number reports—10. *Flora Mediterranea*, *10*, 381–430.

Kamari, G., Garbari, F., & Felber, F. (1996). Mediterranean chromosome number reports—6. *Flora Mediterranea*, *6*, 223–337.

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 161–177). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_9

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P. L., & Phillippy, A. M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, *36*(12), 1174–1182. https://doi.org/10.1038/nbt.4277

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59. https://doi.org/10.1186/1471-2105-5-59

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michelmore, R. W., & Christensen, A. C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genetics*, *15*(8), e1008373. https://doi.org/10.1371/journal.pgen.1008373

Krak, K., Vít, P., Douda, J., & Mandák, B. (2020). Development of 18 microsatellite markers for Salvia pratensis. *Applications in Plant Sciences*, *8*(1), e11316. https://doi.org/10.1002/aps3.11316

Kress, W. J., & Erickson, D. L. (2008). DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences*, *105*(8), 2761–2762. https://doi.org/10.1073/pnas.0800476105

Laface, V. L. A., Cavallini, M., Di Iorio, A., Lombardo, G., Binelli, G., Sorgonà, A., Musarella, C. M., & Spampinato, G. (2024). Genetic structure of populations of *Salvia ceratophylloides* endemic to southern Calabria (southern Italy). *Heliyon*, *10*(16), e35875. https://doi.org/10.1016/j.heliyon.2024.e35875

Laface, V. L. A., Musarella, C. M., Sorgonà, A., & Spampinato, G. (2022). Analysis of the Population Structure and Dynamic of Endemic Salvia ceratophylloides Ard. (Lamiaceae). *Sustainability*, *14*(16), Article 16. https://doi.org/10.3390/su141610295

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175–195. https://doi.org/10.1093/genetics/74.1.175

Li, C.-Y., Yang, L., Liu, Y., Xu, Z.-G., Gao, J., Huang, Y.-B., Xu, J.-J., Fan, H., Kong, Y., Wei, Y.-K., Hu, W.-L., Wang, L.-J., Zhao, Q., Hu, Y.-H., Zhang, Y.-J., Martin, C., & Chen, X.-Y. (2022). The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Reports*, *40*(7), 111236. https://doi.org/10.1016/j.celrep.2022.111236

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics*, *39*(1), btad014. https://doi.org/10.1093/bioinformatics/btad014

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, M., Tang, L., Wu, F.-X., Pan, Y., & Wang, J. (2019). SCOP: A novel scaffolding algorithm based on contig classification and optimization. *Bioinformatics*, *35*(7), 1142–1150. https://doi.org/10.1093/bioinformatics/bty773

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., & Fan, W. (2012). Comparison of the two major classes of assembly algorithms: Overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37. https://doi.org/10.1093/bfgp/elr035

Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., & Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome. *Communications Biology*, *6*(1), 954. https://doi.org/10.1038/s42003-023-05322-y

Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L., & Yanofsky, M. F. (2000). SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. *Nature*, *404*(6779), 766–770. https://doi.org/10.1038/35008089

Lomsadze, A. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, *33*(20), 6494–6506. https://doi.org/10.1093/nar/gki937

Lukashin, A. (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, *26*(4), 1107–1115. https://doi.org/10.1093/nar/26.4.1107

Luo, J., Lyu, M., Chen, R., Zhang, X., Luo, H., & Yan, C. (2019). SLR: A scaffolding algorithm based on long reads and contig classification. *BMC Bioinformatics*, *20*(1), 539. https://doi.org/10.1186/s12859-019-3114-9

Luo, J., Wei, Y., Lyu, M., Wu, Z., Liu, X., Luo, H., & Yan, C. (2021). A comprehensive review of scaffolding methods in genome assembly. *Briefings in Bioinformatics*, *22*(5), bbab033. https://doi.org/10.1093/bib/bbab033

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Luo, X., Chen, S., & Zhang, Y. (2022). PlantRep: A database of plant repetitive elements. *Plant Cell Reports*, *41*(4), 1163–1166. https://doi.org/10.1007/s00299-021-02817-y

Majeed, A., Johar, P., Raina, A., Salgotra, R. K., Feng, X., & Bhat, J. A. (2022). Harnessing the potential of bulk segregant analysis sequencing and its related approaches in crop breeding. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.944501

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msab199

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Marchi, N., Kapopoulou, A., & Excoffier, L. (2024). Demogenomic inference from spatially and temporally heterogeneous samples. *Molecular Ecology Resources*, *24*(1), e13877. https://doi.org/10.1111/1755-0998.13877

Martínez-Redondo, G. I., Barrios-Núñez, I., Vázquez-Valls, M., Rojas, A. M., & Fernández, R. (2024). *Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models*. Evolutionary Biology. https://doi.org/10.1101/2024.02.28.582465

Maynard, R. C. I., & Ruter, J. M. (2022). DNA Content Estimation in the Genus Salvia. *Journal of the American Society for Horticultural Science*, *147*(3), 123–134. https://doi.org/10.21273/JASHS05175-21

Miles, C., & Wayne, M. (2008). Quantitative trait locus (QTL) analysis. *Nature Education*, *1*, 208.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Møller, I. M., Rasmusson, A. G., & Van Aken, O. (2021). Plant mitochondria – past, present and future. *The Plant Journal*, *108*(4), 912–959. https://doi.org/10.1111/tpj.15495

Nadachowska-Brzyska, K., Konczal, M., & Babik, W. (2022). Navigating the temporal continuum of effective population size. *Methods in Ecology and Evolution*, *13*(1), 22–41. https://doi.org/10.1111/2041-210X.13740

Nam, J., Kim, J., Lee, S., An, G., Ma, H., & Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(7), 1910–1915. https://doi.org/10.1073/pnas.0308430100

Nevers, Y., Warwick Vesztrocy, A., Rossier, V., Train, C.-M., Altenhoff, A., Dessimoz, C., & Glover, N. M. (2024). Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-024-02147-w

Ng, M., & Yanofsky, M. F. (2001). Function and evolution of the plant MADS-box gene family. *Nature Reviews Genetics*, *2*(3), 186–195. https://doi.org/10.1038/35056041

O'Leary, N. A., Cox, E., Holmes, J. B., Anderson, W. R., Falk, R., Hem, V., Tsuchiya, M. T. N., Schuler, G. D., Zhang, X., Torcivia, J., Ketter, A., Breen, L., Cothran, J., Bajwa, H., Tinne, J., Meric, P. A., Hlavina, W., & Schneider, V. A. (2024). Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Scientific Data*, *11*(1), 732. https://doi.org/10.1038/s41597-024-03571-y

Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky730

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons1[OPEN]. *Plant Physiology*, *176*(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Ouborg, N. J., & Treuren, R. V. (1994). The significance of genetic erosion in the process of extinction. IV.Inbreeding load and heterosis in relation to population size in the mint Salvia pratensis. *Evolution*, *48*(4). https://doi.org/10.1111/j.1558-5646.1994.tb05288.x

Ouborg, N. J., & Treuren, R. V. (1995). Variation in Fitness-Related Characters Among Small and Large Populations of Salvia Pratensis. *The Journal of Ecology*, *83*(3). https://doi.org/10.2307/2261591

Ouborg, N. J., van Treuren, R., & van Damme, J. M. M. (1991). The significance of genetic erosion in the process of extinction: II. Morphological variation and fitness components in populations of varying size of Salvia pratensis L. and Scabiosa columbaria L. *Oecologia*, *86*(3), 359–367. https://doi.org/10.1007/BF00317601

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., & Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Research*, *35*(suppl_1), D883–D887. https://doi.org/10.1093/nar/gkl976

Oxford Nanopore Technologies Ltd. (2018). *Nanoporetech/medaka* [Python]. Oxford Nanopore Technologies. https://github.com/nanoporetech/medaka (Original work published 2017)

Paleni, C. (2021). *Genomics for biodiversity conservation: The botanical garden of Brera and the Salvia collection*. Università degli Studi di Milano, Politecnico di Milano.

Pan, W., Jiang, T., & Lonardi, S. (2020). OMGS: Optical Map-Based Genome Scaffolding. *Journal of Computational Biology*, *27*(4), 519–533. https://doi.org/10.1089/cmb.2019.0310

Pan, X., Chang, Y., Li, C., Qiu, X., Cui, X., Meng, F., Zhang, S., Li, X., & Lu, S. (2023). Chromosome-level genome assembly of Salvia miltiorrhiza with orange roots uncovers the role of Sm2OGD3 in catalyzing 15,16-dehydrogenation of tanshinones. *Horticulture Research*, *10*(6), uhad069. https://doi.org/10.1093/hr/uhad069

Pařenicová, L., de Folter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., Cook, H. E., Ingram, R. M., Kater, M. M., Davies, B., Angenent, G. C., & Colombo, L. (2003). Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in *Arabidopsis*. *The Plant Cell*, *15*(7), 1538–1551. https://doi.org/10.1105/tpc.011544

Pellicer, J., Estiarte, M., Garcia, S., Garnatje, T., Peñuelas, J., Sardans, J., & Vallès, J. (2010). Genome size unaffected by moderate changes in climate and phosphorus availability in Mediterranean plants. *African Journal of Biotechnology*, *9*(37), Article 37.

Pellicer, J., & Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *New Phytologist*, *226*(2), 301–305. https://doi.org/10.1111/nph.16261

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*, 304. https://doi.org/10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Pieri, C. (2019). *Metodologie per la conservazione e valorizzazione delle collezioni degli Orti Botanici: Il progetto della collezione del genere Salvia dell'Orto Botan- ico di Brera*. Università degli Studi di Milano.

Pignatti, S. (2018). *Flora d'Italia. Volume 3* (Seconda edizione in 4 volumi). Edagricole.

Portale della Flora d'Italia—Portal to the Flora of Italy. (n.d.). Retrieved 13 November 2024, from http://dryades.units.it/floritaly

Pourkheirandish, M., Golicz, A. A., Bhalla, P. L., & Singh, M. B. (2020). Global Role of Crop Genomics in the Face of Climate Change. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00922

Powder, K. E. (2020). Quantitative Trait Loci (QTL) Mapping. In X. M. Shi (Ed.), *eQTL Analysis* (Vol. 2082, pp. 211–229). Springer US. https://doi.org/10.1007/978-1-0716-0026-9_15

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics*, *47*(1). https://doi.org/10.1002/0471250953.bi1112s47

Ramos, A., Barbena, E., Mateiu, L., del Mar González, M., Mairal, Q., Lima, M., Montiel, R., Aluja, M. P., & Santos, C. (2011). Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies. *Mitochondrion*, *11*(6), 946–953. https://doi.org/10.1016/j.mito.2011.08.009

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Ranjbar, M., Pakatchi, A., & Babataheri, Z. (2015). Chromosome number evolution, biogeography and phylogenetic relationships in Salvia (Lamiaceae). *Webbia*, *70*(2), 293–312. https://doi.org/10.1080/00837792.2015.1057982

Ratcliffe, O. J., Kumimoto, R. W., Wong, B. J., & Riechmann, J. L. (2003). Analysis of the Arabidopsis MADS AFFECTING FLOWERING Gene Family: MAF2 Prevents Vernalization by Short Periods of Cold. *The Plant Cell*, *15*(5), 1159. https://doi.org/10.1105/tpc.009506

Reiser, L., Bakker, E., Subramaniam, S., Chen, X., Sawant, S., Khosa, K., Prithvi, T., & Berardini, T. Z. (2024). The Arabidopsis Information Resource in 2024. *GENETICS*, *227*(1), iyae027. https://doi.org/10.1093/genetics/iyae027

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02134-9

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460. https://doi.org/10.1186/s12859-018-2485-7

Romano, S., Campo, G., & Colombo, P. (1996). Numeri cromosomici per la flora italiana. *Informatore Botanico Italiano*, *28*(2), 251–255.

Salamov, A. A., & Solovyev, V. V. (2000). Ab initio Gene Finding in *Drosophila* Genomic DNA. *Genome Research*, *10*(4), 516–522. https://doi.org/10.1101/gr.10.4.516

Salmeri, C. (2019). Karyological data of some plant species native to South Italy. *Flora Mediterranea*, *29*. https://doi.org/10.7320/FlMedit29.334

Salvia clandestina L. | Plants of the World Online | Kew Science. (n.d.). Plants of the World Online. Retrieved 24 December 2024, from http://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:455962-1

Salvia saccardiana (Pamp.) Del Carr. & Garbari, Acta Plantarum. (2010). https://www.actaplantarum.org/galleria_flora/galleria1.php?id=6848

Salvia saccardiana (Pamp.) Del Carr. & Garbari—Portale della Flora d'Italia / Portal to the Flora of Italy. (n.d.). Retrieved 9 December 2024, from https://dryades.units.it/floritaly/index.php?procedure=taxon_page&tipo=all&id=4675

Samach, A., Onouchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F., & Coupland, G. (2000). Distinct Roles of CONSTANS Target Genes in Reproductive Development of Arabidopsis. *Science*, *288*(5471), 1613–1616. https://doi.org/10.1126/science.288.5471.1613

Sangiovanni, M., Granata, I., Thind, A. S., & Guarracino, M. R. (2019). From trash to treasure: Detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics*, *20*(4), 168. https://doi.org/10.1186/s12859-019-2684-x

Schoof, G. P. D. H. (2024). *Groupschoof/AHRD* [Java]. https://github.com/groupschoof/AHRD (Original work published 2012)

Seemann, T. (n.d.). *tseemann/barrnap: Bacterial ribosomal RNA predictor*. Retrieved 21 October 2024, from https://github.com/tseemann/barrnap

Sensalari, C., Maere, S., & Lohaus, R. (2022). ksrates: Positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics*, *38*(2), 530–532. https://doi.org/10.1093/bioinformatics/btab602

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*(9), 1044–1053. https://doi.org/10.1038/s41587-020-0503-6

Shore, P., & Sharrocks, A. D. (1995). The MADS-Box Family of Transcription Factors. *European Journal of Biochemistry*, *229*(1), 1–13. https://doi.org/10.1111/j.1432-1033.1995.0001l.x

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Siljak-Yakovlev, S., Pustahija, F., Šolić, E. M., Bogunić, F., Muratović, E., Bašić, N., Catrice, O., & Brown, S. C. (2010). Towards a Genome Size and Chromosome Number Database of Balkan Flora: C-Values in 343 Taxa with Novel Values for 242. *Advanced Science Letters*, *3*(2), 190–213. https://doi.org/10.1166/asl.2010.1115

Siri, J. N. (2024). *Jorgenunezsiri/accusyn* [JavaScript]. https://github.com/jorgenunezsiri/accusyn (Original work published 2019)

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*(1), 31. https://doi.org/10.1186/1471-2105-6-31

Smit, A., Hubley, R., & Green, P. (n.d.). *RepeatMasker Open-4.0 2013-2015*. https://www.repeatmasker.org/

Spampinato, G., Laface, V. L. A., Ortiz, A. C., Canas, R. Q., & Musarella, C. M. (2019). *Salvia ceratophylloides* Ard. (Lamiaceae): A Rare Endemic Species of Calabria (Southern Italy). In *Endemic Species*. IntechOpen. https://doi.org/10.5772/intechopen.84905

Stange, M., Barrett, R. D. H., & Hendry, A. P. (2021). The importance of genomic variation for biodiversity, ecosystems and people. *Nature Reviews Genetics*, *22*(2), 89–105. https://doi.org/10.1038/s41576-020-00288-7

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*(1), 62. https://doi.org/10.1186/1471-2105-7-62

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, *19*(suppl_2), ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Štorchová, H., & Krüger, M. (2024). Methods for assembling complex mitochondrial genomes in land plants. *Journal of Experimental Botany*, *75*(17), 5169–5174. https://doi.org/10.1093/jxb/erae034

Sun, H., Ding, J., Piednoël, M., & Schneeberger, K. (2018). findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*, *34*(4), 550–557. https://doi.org/10.1093/bioinformatics/btx637

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., & Wang, Y. (2023). OrthoVenn3: An integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Research*, *51*(W1), W397–W403. https://doi.org/10.1093/nar/gkad313

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, *19*(1), 131. https://doi.org/10.1186/s13059-018-1520-3

Tajima, F. (1983). EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS. *Genetics*, *105*(2), 437–460. https://doi.org/10.1093/genetics/105.2.437

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015). ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biology*, *16*(1), 3. https://doi.org/10.1186/s13059-014-0573-1

Temsch, E. M., Temsch, W., Ehrendorfer-Schratt, L., & Greilhuber, J. (2010). Heavy Metal Pollution, Selection, and Genome Size: The Species of the Žerjav Study Revisited with Flow Cytometry. *Journal of Botany*, *2010*(1), 596542. https://doi.org/10.1155/2010/596542

The Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, *119*(4), e2115642118. https://doi.org/10.1073/pnas.2115642118

Theißen, G. (2001). Development of floral organ identity: Stories from the MADS house. *Current Opinion in Plant Biology*, *4*(1), 75–85. https://doi.org/10.1016/S1369-5266(00)00139-4

Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P. R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G. R., Godoy, J. A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Secomandi, S., … Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in Genetics*, *39*(7), 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. https://doi.org/10.1093/nar/gkx391

Treuren, R. V., Bijlsma, R., Delden, W. V., & Ouborg, N. J. (1991). The significance of genetic erosion in the process of extinction. I. genetic differentiation in salvia pratensis and scabiosa columbaria in relation to population size. *Heredity*, *66*(2). https://doi.org/10.1038/hdy.1991.23

Uliano-Silva, M., Ferreira, J. G. R. N., Krasheninnikova, K., Blaxter, M., Mieszkowska, N., Hall, N., Holland, P., Durbin, R., Richards, T., Kersey, P., Hollingsworth, P., Wilson, W., Twyford, A., Gaya, E., Lawniczak, M., Lewis, O., Broad, G., Martin, F., Hart, M., … Darwin Tree of Life Consortium. (2023). MitoHiFi: A python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics*, *24*(1), 288. https://doi.org/10.1186/s12859-023-05385-y

Vaser, R., & Šikić, M. (2021). Time- and memory-efficient genome assembly with Raven. *Nature Computational Science*, *1*(5), 332–336. https://doi.org/10.1038/s43588-021-00073-4

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, J., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2023). The

conserved domain database in 2023. *Nucleic Acids Research*, *51*(D1), D384–D388. https://doi.org/10.1093/nar/gkac1096

Wang, Y. (2024). *Wyp1125/MCScanX* [CAP CDS]. https://github.com/wyp1125/MCScanX (Original work published 2011)

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*(1), 35. https://doi.org/10.1186/s13742-015-0076-3

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*(1), 116. https://doi.org/10.1186/1471-2105-11-116

Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biology*, *17*(1), 37. https://doi.org/10.1186/s13059-016-0908-1

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Will, M., & Claßen-Bockhoff, R. (2017). Time to split Salvia s.l. (Lamiaceae) – New insights from Old World Salvia phylogeny. *Molecular Phylogenetics and Evolution*, *109*. https://doi.org/10.1016/j.ympev.2016.12.041

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P. A. J., Le Béguec, C., Fieten, H., Johnson, J., Alföldi, J., André, C., … Derrien, T. (2017). FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, *45*(8), e57. https://doi.org/10.1093/nar/gkw1306

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

Yang, L., Wang, Z., & Hua, J. (2019). Measuring Cell Ploidy Level in Arabidopsis thaliana by Flow Cytometry. In W. Gassmann (Ed.), *Plant Innate Immunity: Methods and Protocols* (pp. 101–106). Springer. https://doi.org/10.1007/978-1-4939-9458-8_11

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, *107*(20), 9264–9269. https://doi.org/10.1073/pnas.0913022107

Zare, T., Paril, J. F., Barnett, E. M., Kaur, P., Appels, R., Ebert, B., Roessner, U., & Fournier-Level, A. (2024). Comparative genomics points to tandem duplications of SAD gene clusters as drivers of increased α-linolenic (ω-3) content in S. hispanica seeds. *The Plant Genome*, *17*(1), e20430. https://doi.org/10.1002/tpg2.20430

Zhang, B., & Claßen-Bockhoff, R. (2019). Sex-differential reproduction success and selection on floral traits in gynodioecious *Salvia pratensis*. *BMC Plant Biology*, *19*(1), 375. https://doi.org/10.1186/s12870-019-1972-y

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(S6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, X., Wu, R., Wang, Y., Yu, J., & Tang, H. (2020). Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal*, *18*, 66–72. https://doi.org/10.1016/j.csbj.2019.11.011

Zhao, D., Chen, Z., Xu, L., Zhang, L., & Zou, Q. (2021). Genome-Wide Analysis of the MADS-Box Gene Family in Maize: Gene Structure, Evolution, and Relationships. *Genes*, *12*(12), 1956. https://doi.org/10.3390/genes12121956

Zhou, C. (2024). *C-zhou/oatk* [C]. https://github.com/c-zhou/oatk (Original work published 2023)

Zhou, C., McCarthy, S. A., & Durbin, R. (2023). YaHS: Yet another Hi-C scaffolding tool. *Bioinformatics*, *39*(1), btac808. https://doi.org/10.1093/bioinformatics/btac808

Ziarsolo, P., Hasing, T., Hilario, R., Garcia-Carpintero, V., Blanca, J., Bombarely, A., & Cañizares, J. (2021). K-seq, an affordable, reliable, and open Klenow NGS-based genotyping technology. *Plant Methods*, *17*(1), 30. https://doi.org/10.1186/s13007-021-00733-6

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology*, *16*(6), e1007981. https://doi.org/10.1371/journal.pcbi.1007981

Zonneveld, B. J. M. (2019). The DNA weights per nucleus (genome size) of more than 2350 species of the Flora of The Netherlands, of which 1370 are new to science, including the pattern of their DNA peaks. *Forum Geobotanicum*, *8*, 24–78. https://doi.org/10.3264/FG.2019.1022

Zuntini, A. R., Carruthers, T., Maurin, O., Bailey, P. C., Leempoel, K., Brewer, G. E., Epitawalage, N., Françoso, E., Gallego-Paramo, B., McGinnie, C., Negrão, R., Roy, S. R., Simpson, L., Toledo Romero, E., Barber, V. M. A., Botigué, L., Clarkson, J. J., Cowan, R. S., Dodsworth, S., … Baker, W. J. (2024). Phylogenomics and the rise of the angiosperms. *Nature*, *629*(8013), 843–850. https://doi.org/10.1038/s41586-024-07324-0

Appendix

- Paleni, C., Puricelli, C., Pieri, C., De Paola, L., Babolin, N., Colombo, L., Bombarely, A., Lambertini, C., Kater, M. M. Mining the *Salvia pratensis* genome to study the MADS-box gene family and their role in flower development. *In preparation*

The manuscript presents the *S. pratensis* reference genome and focuses on gene family analysis of the MADS-box transcription factors. C.P. performed the bioinformatic analyses, which are also in part described in the main body of this thesis, and wrote the article.

- Deglialberti, M., Paleni, C., Fainelli. F., Lambertini, C., Urpì, S. Genomic Analysis of Botanical Collections: opportunities and challenges. *Submitted to Teaching Tools (The Plant Cell)*

Botanical collections are repositories of plant biodiversity and analyzing them with genomic techniques can help investigate species evolution and extinction and contribute to biodiversity conservation. However, a lack of awareness of available methods or funding can limit the use of these resources. In this primer, aimed mainly at university students as a learning tool, we introduce genomic analysis of botanical collection by outlining several factors impacting success of a study, from the sample collection to the choice of sequencing method. C.P. contributed as an author particularly in the sections related to Next Generation Sequencing methods and reference genomes and developed illustrations and class activities.

- Rieu, P., Beretta, V. M., Caselli, F., Thévénon, E., Lucas, J., Rizk, M., Franchini, E., Caporali, E., Paleni, C., Nanao, M. H., Kater, M. M., Dumas, R., Zubieta, C., Parcy, F., & Gregis, V. (2024). The ALOG domain defines a family of plant-specific transcription factors acting during *Arabidopsis* flower development. In Proceedings of the National Academy of Sciences (Vol. 121, Issue 10). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2310464121

The paper investigates the role of ALOG (Arabidopsis LIGHT-DEPENDENT SHORT HYPOCOTYLS 1 (LSH1) and Oryza G1) genes in *Arabidopsis thaliana* development. C. P. performed RNA-seq data analysis and differential gene expression analysis of *lsh1 lsh3 lsh4* triple mutant compared to wild type *Arabidopsis thaliana* plants to identify a list of putative targets for regulation by the LSH transcription factors.

- Cannavò, S., Paleni, C., Costarelli, A., Valeri, M. C., Cerri, M., Saccomanno, A., Gregis, V., Chini Zittelli, G., Dobrev, P. I., Reale, L., Kater, M. M., & Paolocci, F. (2024). Co-cultivating rice plants with *Azolla filiculoides* modifies root architecture and timing of developmental stages. *Accepted for publication in Rice Science.* Preprint: https://doi.org/10.1101/2024.10.02.615985
- Consorti, E., Costarelli, A., Cannavò, S., Cerri, M., Valeri, M. C., Reale, L., Saccomano, A., Paleni, C., Gregis, V., Kater, M. M., Brilli, F., Paolocci, F., & Ghirardo, A. (2024). Co-cultivation with Azolla affects the metabolome of whole rice plant beyond canonical inorganic nitrogen fertilization. *Submitted to Rice.* Preprint: https://doi.org/10.1101/2024.10.02.615589
- Bazihizina, N., Paleni, C., Caparrotta, S., Macchiavelli, T., Guardigli, G., Colzi, I., Petrillo, M., Gonnelli, C., Saccomanno, A., Gregis, V., Mancuso, S., Comparini, D., Kater, M. M., & Pandolfi, C. (2024). *Azolla* Mediated Alterations in Grain Yield and Quality in Rice. *Accepted for publication in Physiologia Plantarum*

In Southern-East Asian countries, rice farming benefits from the use of *Azolla* spp., either as green manure or as co-cultivated plants, for the supply of nitrogen. *Azolla* spp. are ferns that, in virtue of their symbiosis with the nitrogen-fixing cyanobacterium *Trichormus azollae*, fix atmospheric nitrogen and release it to the environment upon decomposition of their biomass. These companion papers investigate the effects of

actively growing *Azolla* plants on rice and the hypothesis that *Azolla* is not just a N-supplier for rice. C. P. contributed with RNA-seq and differential gene expression analysis of root and inflorescence meristem samples from rice co-cultivated with *A. filiculoides* compared to control plants. The studies highlight potential roles in iron metabolism, root system architecture changes and several metabolomic alterations in the root, as well as upregulation of multiple transcription factor and possible defence priming in the inflorescence meristem.