University of Modena and Reggio Emilia

PhD in AGRI-FOOD SCIENCES, TECHNOLOGIES AND BIOTECHNOLOGIES - UNIMORE

Cycle XXXVII

Genetic variation underlying the adaptation of maize to the environment

Helga Cassol

Tutor: Prof. Nicola Pecchioni

Co-tutor: Dr. Chiara Lanzanova

Coordinator: Prof. Fabio Licciardello

Summary

The present thesis represents a study of the genomic diversity in maize correlated to the environmental adaptation. Genomic diversity was studied considering the single-nucleotide polymorphisms (SNPs), already well known for their role in the regulation of phenotypic traits, and the structural variations, in particular genic copy number variations (CNVs). The first part of the thesis regards the association of SNPs to adaptation traits in a Italian core collection preserved at the germplasm Genebank at CREA in Bergamo. The core collection belongs to a large panel of maize inbred lines originated from the Italian traditional landraces collected by CREA from the early 50s years of 20th century, and it was genotyped and characterized for linkage disequilibrium (LD), genetic diversity, population structure, and genetic relationships. The results of this study are described in the paper entitled “Genetic Diversity within a Collection of Italian Maize Inbred Lines: A Resource for Maize Genomics and Breeding” reported in the thesis. Successively, the core collection was characterized for different phenotypic traits in open field focusing in particular on flowering time-related traits which are highly influenced by drought and heat stress in maize. To associate the phenotypic traits data collected in 2022 and 2023, a genome-wide association mapping study (GWAS) was carried out with a special attention to male flowering time, female flowering time and anthesis-silking interval (ASI).

From the GWAS analysis 102 marker-associated traits were identified. Between them, 15 MTAs showed an overlap with flowering-related QTLs, already known from the literature.

The second part of the thesis regards the detection of structural variations, in particular CNVs. In the context of understanding the more accurate and efficient sequencing approach to detect CNVs, a bioinformatic simulation on three public available genomes of maize inbred lines were conducted. In the simulation have been compared whole-genome (WGS), whole-exome (WES) and ddRAD sequencing approach. To detect and associate the CNVs to drought and heat stress tolerance, another inbred lines panel called DROPS was identified. The panel includes more than 240 European and American inbred lines that have been crossed to a tester. The hybrids as obtained, as described in the study of Millet et al., 2016, were phenotyped in different times and locations and they were also tested in rainfed and watered regimes. The lines were sequenced with an Illumina paired-ends short reads sequencing and analyzed to identify SNP markers. The total of 131,954,634 SNPs were filtered and a final set of 89,198,737 SNP markers were retained. To set an appropriate bioinformatic pipeline, different software were considered and it was chosen the platform Hecaton, a framework that combines different specific-CNVs tools and a strict filtering conducted moreover, with a random forest model. A validation of Hecaton was performed on five NAM maize lines already studied for structural variants, to define advantages and limitations of the software. The next steps of the association of genic CNVs to adaptation traits in maize, in particular to traits connected to drought and heat stress tolerance, will regard the calling of the structural variants and their association with significant effects on the phenotypic traits already recorded.

Index

Section A - Study of the genomic diversity and its association to adaptation traits in an Italian maize core collection

Chapter 1: General Introduction

Maize global and national cultivation and significance

Maize cultivation is crucial for global cereal production, serving both animal and human consumption. Globally maize for dry grain production covers approximately 200 million hectares annually, making it the second most widely cultivated cereal globally, after wheat. Worldwide, its production is estimated around 1 billion and 240 million of tons and currently, its yield is quite around 6 tons/ha (FAOStat, 2023). Worldwide, the dominion for maize production is competed among Asia, in particular China, with 288 millions of tons, and North America, especially United States of America with 389 millions of tons; these two followed by South America and Europe. Eastern Europe, especially Ukraine, Russian Federation, Romania, Poland and Hungary, is the area with the higher value of harvested area and higher production (71 million tons), while in Western Europe prevail France and Italy (18 million tons) (Faostat, 2023). Notwithstanding the recent significant decline of maize cultivation in Italy, this crop still accounts for 8% of Italy's total arable land and a production of 5 millions of tons in 2023 (Istat Statistics, 2023). The maize uniqueness is characterized by the large area of cultivation, which implies a large capacity of adaptation to the most various environments, and by its morphological, nucleotide and structural diversity (Tenaillon & Charcosset, 2011). The importance of this crop is highlighted by the breeders' effort to increase the maize yield from the Green Revolution until now. Indeed, in the US corn belt, one of the major cultivation regions for this crop, the maize genetic grain yield raised from 74 to 123 kg ha-1 year-1 between 1930 and 2001(L. Echarte, 2013). The uses of maize grain include mainly the animal feed for 77% of the production, while for the starch sector is used the 16% of it, and for milling and human consumption only the 7%. Furthermore, maize is also used in biogas production and as silage for livestock feeding (Istat Statistics, 2023).

The introduction of maize in Europe

From the spreading of maize in Europe, different plant populations adapted to different environments constituting the European germplasm. The introduction of maize in Europe explained in the work of Rebourg et al., 2003 has two main origins. The first maize seeds introduced had Caribbean origin, but the adaptation of this plant was also guided by the presence of Northern American flint lines. From these two origins were born the European flint lines which form one of the main heterotic group, and which were

adapted to the cooler regions of Europe thanks to the selection process started from the farmers. Interestingly, the Italian flint lines might be originated from the South American Cateto maize, which ears are characterized by small hard orange flint kernels and high number of rows (Rebourg et al., 2003). Dent type, the other main heterotic group used for breeding, was cultivated only in Italy and Spain before the introduction of hybrids in the mid-20th century. The main differences between *Zea mays* var. *indentata* (dent) and *Z. mays* var. *indurata* (flint) are the morphology and the grain type. Var. *indentata* is characterized by flinty sides and soft cores of starch that cause the end of the kernels to collapse or dent during drying, while var. *indurata* is characterized by hard, glossy endosperms with smooth, hard seed coats pericarps (Dickerson, et al., 2003). It is known that European landraces and lines could represent the diversity of the European maize germplasm and the study of them can lead to new insights about the history of maize. The genotyping approach which belongs to the category of reduced-representation sequencing (RRS) methods, is considered a cost-saving and effective approach to screen the diversity inside the landraces or lines collections as demonstrated by the works of Diaw et al., 2021 and Ganal et al., 2011. The work of Diaw et al., 2021 focuses on landraces cultivated in the South of France and their connection to important genetic groups identified in different studies. For the constitution of France East South-West landraces (E-SWF) it was essential the contribution of Northern Flint landraces and partially, the contribution of Caribbean, Andean and Italian landraces. These information related to the population structure brought to the definition of two scenarios regarding the origin of the South-Western France landraces, in which are highlighted the relationships with European Northern Flint and with Spanish hybridized maize landraces (Caribbean x Northern Flint landraces, known as Pyrenees-Galicia landraces).

The Italian inbred lines collection of CREA

During the 50s years of 20th century, Regional Inspectorates of Agriculture (Ispettorati Provinciali dell'Agricoltura) decided to collect the maize varieties diffused in Italy and Experimental Station for Maize Cropping (Stazione Sperimentale per la Maiscoltura) of Bergamo sampled "Indentata" and "Indurata" maize types creating a germplasm collection, of landraces and of inbred lines in which is present the Italian inbred lines collection (Mastrangelo et al., 2024). In the work of Mastrangelo et al., 2024, a subset of the entire inbred lines collection was studied, and from the given pedigree information was defined a phylogenetic tree with four groups: Insubrian, Microsperma, Insubrian/Microsperma and Elite/White/USA. Successively, the analysis of population structure was conducted leading to the subdivision into four subpopulations. Interestingly the lines belonging to the group 3 "Scagliolino" have some correspondence with the

name of the lines and on phenotypic traits. In the context of European maize lines, Italian lines are isolated like demonstrated by the work of Gouesnard et al., 2016 and the line “Nostrano dell’Isola”, an important historical line of the collection, shows significant distance in the genetic structure based on the PAVs (presence-absence variations) and the SNPs (single nucleotide polymorphisms) found in a set of European lines belonging to Corn Belt Dent, European Flint, Northern Flint and Stiff Stalk group (Darracq et al., 2018). Considering the genetic diversity and the limited knowledge on the Italian inbred lines conserved at CREA, this panel can be considered a good resource to discover useful alleles/genes connected to biotic or abiotic stress tolerance or to reveal regulation mechanisms of important phenotypic traits, as already reported in the genome-wide association analyses of Maldonado et al., 2019 on flowering-related traits and in the genome-wide association analyses of Revilla et al., 2016 for cold tolerance.

Chapter 2: Study of Genetic Diversity in a collection of Italian maize inbred lines

plants

MDPI

Type of the Paper (Article, Review, Communication, etc.)

Genetic diversity within a collection of Italian maize inbred lines: a resource for maize genomics and breeding

Anna Maria Mastrangelo [1,*], Hans Hartings [2], Chiara Lanzanova [2], Carlotta Balconi [2], Sabrina Locatelli [2], Helga Cassol [2], Paolo Valoti [2], Giuseppe Petruzzino [1], Nicola Pecchioni [1,2]

Abstract: Genetic diversity is fundamental for studying complex architecture of traits of agronomic importance, controlled by major and minor loci. Moreover, well-characterized germplasm col-lections are an essential tool to dissect and analyse genetic and phenotypic diversity in crops. A panel of 360 entries, a subset of a larger collection maintained within the Genebank at CREA Bergamo, and which includes inbreds derived from traditional Italian maize open-pollinated (OP) varieties, and advanced breeding ones (Elite Inbreds), was analyzed to identify SNP markers using the tGBS® Genotyping by Sequencing technology. A total of 797,368 SNPs were found during the initial analysis. Imputation and filtering processes were carried out based on the percentage of missing data, redundant markers, and rarest allele frequencies, resulting in a final dataset of 15,872 SNP markers for which a physical map position was identified. Using this dataset, the inbred panel was characterized for linkage disequilibrium (LD), genetic diversity, population structure, and genetic relationships. LD decay at a genome wide level indicates the collection as a suitable re-source for association mapping. Population structure analyses, carried out with different clustering methods, showed stable grouping statistics for four groups, broadly corresponding to 'Insubria', 'Microsperma', and 'Scagliolino' genotypes, respectively, with a fourth group composed prevalently of elite accessions derived from Italian and U.S. breeding programs. Based on these results, the CREA Italian maize collection, here genetically characterized, can be considered an important tool for the mapping and characterization of useful traits and associated loci/alleles, to be used in maize breeding programs.

Keywords: maize; germplasm collections; genetic diversity

1. Introduction

Maize (*Zea mays* subsp. *mays*) is one of the most important agricultural crops worldwide. Northern Italy is one of the core areas for maize cultivation and production in Europe, with a grain yield and corn-cob mix of more than 4.7 million tons per year [1]. In historical times, maize cultivation was boosted by the diffusion of maize germplasm better adapted to European conditions, especially to longer photoperiods, after the first introduction of genetic materials from the Caribbean [2]. Since then, farmers developed numerous landraces by crossing different ecotypes, adapted to specific environmental conditions and traditional farming systems [3]. At the end of the eighteenth century, maize reached a very similar production to that of wheat in many regions of Northern Italy, where most of the maize genotypes were grown from open-pollinated (OP) seeds, till the first part of the last century. Then, after the discoveries on inbreeding and heterosis in the early 1900's, open-pollinated varieties were gradually replaced by double- and single-cross hybrids, which played a major role in increasing grain yield since the late thirties [1,4]. In order to avoid the extinction of landraces and loss of precious germplasm, a sur-vey of maize varieties, diffused in Italy, was carried out in 1949-50 by the Regional Inspectorates of Agriculture (Ispettorati Provinciali dell'Agricoltura). In 1954, the Experimental Station for Maize Cropping (Stazione Sperimentale per la Maiscoltura) of Bergamo (Director: L. Fenaroli) started a systematic acquisition of Italian maize germplasm, through a national sampling program of "Indentata" and "Indurata" types, under the aegis of the Italian Ministry of Agriculture. With the co-operation of the Ispettorati Provinciali dell'Agricoltura, samples of different populations were taken and moved to Bergamo for reproduction and classification studies [3,5].

Studying the extent of genetic diversity of germplasm collections is of paramount importance to understand their potential deployment in breeding programs. Moreover, it is the basis for genetic association studies aimed to understand the complex architecture of quantitative traits of agronomic importance. However, until now, the Italian maize Genebank has only been partially characterized, genotypically [6], leaving its genetic diversity largely unexplored.

Genotyping technologies have greatly improved in the last years due to the development of next-generation sequencing (NGS) procedures, allowing a high-throughput and relatively cheap and rapid analysis of large maize collections. Two common approaches have been mostly used so far: single nucleotide polymorphism (SNP) array platforms, and genotyping-by-sequencing (GBS), with other NGS methods growing on the market. In maize, an Illumina Infinium HD 50,000 SNP-array, named MaizeSNP50 array was

developed by Ganal et al. [7] and has been extensively used for diversity and association studies [8-10]. Nevertheless, the maize genome size (2.4 Gb), the high level of diversity and the low LD extent have favored the spread of platforms with a higher marker density. An Affymetrix Axiom 600,000 SNP-array was therefore developed and used in association genetics [9,11,12], as for detection of selective sweeps [13]. SNP arrays offer great ad-vantages in genetic analysis as they are fast and provide results on markers which can be easily compared across different germplasm collection studies. On the other hand, they present the drawback of being 'closed' tools, that do not allow the discovery of de novo SNPs. Moreover, they are characterized by some ascertainment bias in diversity analyses since the SNPs selected for developing arrays come from a fixed set of individuals which are different, for example, from those present in the present panel. Therefore, some SNPs on the array can be non-informative on the panel under study, or can show different allelic frequency profiles, compromising the ability of the SNP-arrays to provide an exact evaluation of the genetic diversity [14,15]. This limit is overcome by GBS, in which SNPs are determined in the exact genotype panel under study, and usually are available at a lower cost with respect to the SNP-arrays [16]. However, a good balance between the cost of the assay and the sequencing depth needs to be found, as the higher the coverage, and the higher the sequencing depth, and hence, the higher the quality of the markers discovered, the higher the cost of the assay. Moreover, needed sequencing depth can vary across the genome, and even between individuals, so that large portions of the genome could remain without successful SNP calls if the right read depth is not considered [17,18]. Fortunately, a part of the missing data present after GBS and filtering can be recovered through imputation with ad hoc methods [18-20].

A genotyping-based analysis of the genetic diversity of a panel of 360 lines, a subset of a larger collection preserved at CREA Bergamo Genebank which includes accessions derived from traditional Italian maize varieties, and advanced breeding lines, is presented in this study. The SNP dataset obtained following GBS has been subjected to analyses including measures of genetic diversity, linkage disequilibrium and genotypic clustering with Bayesian methods and methods based on supervised machine learning approaches. The results herein reported show the genetic structure of this panel of traditional Italian lines which can represent an important genetic tool for the future identification and study of loci and alleles associated to agronomically relevant traits through genome-wide association mapping studies (GWAS) and for their use in maize breeding.

2. Results

2.1. Genotyping of the maize collection

Genotypic characterization of the maize collection was carried out through tunable Genotyping by Sequencing (tGBS®), a technology which, compared to conventional GBS, produces better results in terms of number of on-target reads and amount of missing data thanks to additional steps of reduction of genomic DNA [21]. The experiment was con-ducted with the restriction enzyme Bsp1286I (Freedom Markers, Data2Bio, Iowa, USA). Samples were sequenced using an Illumina HiSeq X instrument, and reads were aligned to the Zea mays AGPv4 (GCA_000005005.6) reference. A total of 2 x 671,777,289 reads were obtained, with 2 x 1,749,420 average reads per sample, and a number of reads per sample ranging from 2 x 541 to 2 x 5,742,482. A total of 53.1% of reads showed a single unique alignment to the maize genome allowing to establish an initial dataset of 797,368 SNPs. Details about the number of quality trimmed sequence reads and aligned reads obtained for each sample are provided in Table S1.

After several quality filtering steps (20% missing data, 20% heterozygous calls, and minimum allele frequency > 0.05), a total of 15,872 SNP markers were retained, a mean of 1,587.2 SNP markers per chromosome (from a minimum of 1,189 SNPs on chromosome 6 to a maximum of 2,315 SNPs on chromosome 1), and an average density of 7.56 SNPs per Mbp (Table 1).

Table 1. Maize chromosomal distribution of filtered SNP markers retained for the study.

Chromosome	SNPs	SNPs/Mb	Chromosome coverage (%)
1	2,315	7.57	99.61
2	1,864	7.66	99.57
3	1,719	7.32	99.70
4	1,631	6.62	99.68
5	1,742	7.79	99.85
6	1,189	6.85	99.67
7	1,391	7.64	99.85
8	1,529	8.46	99.78
9	1,233	7.73	99.90
10	1,259	8.37	99.60
Total	**15,872**	-	-
Mean	**1,587.2**	**7.60**	**99.72**

SNPs showed a rather uniform distribution along chromosomes. In general, marker density results lower in pericentromeric regions compared to proximal and distal portions (Fig. 1).

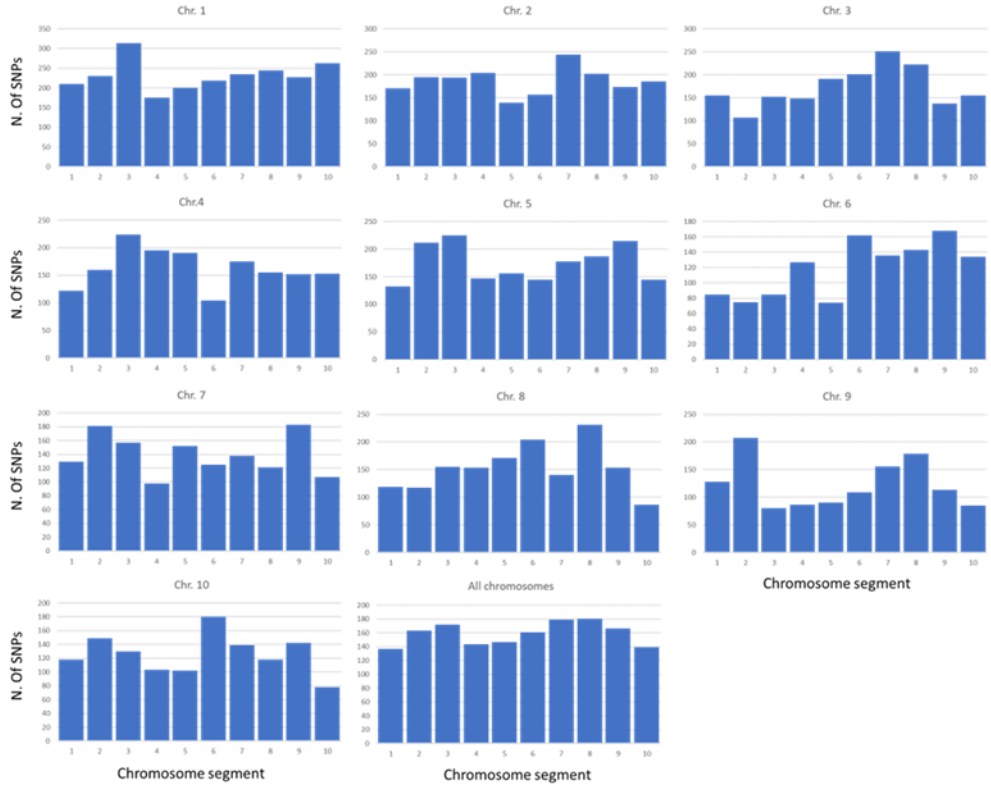


Figure 1. Number of SNPs per chromosome segment (each segment corresponding to the total length of the chromosome divided by ten), from proximal (1) to distal (10) regions, both for single chromosomes and as an average across all chromosomes.

2.2. Linkage disequilibrium analysis

LD decay was evaluated both genome wide and at the single chromosome level. LD patterns changed for the different maize chromosomes and showed, some variations for particular chromosomal regions. On average, considering the collection of 360 maize accessions, LD decayed at a distance of approximately 12.3 kb, when applying a cut-off value of r2 = 0.1.

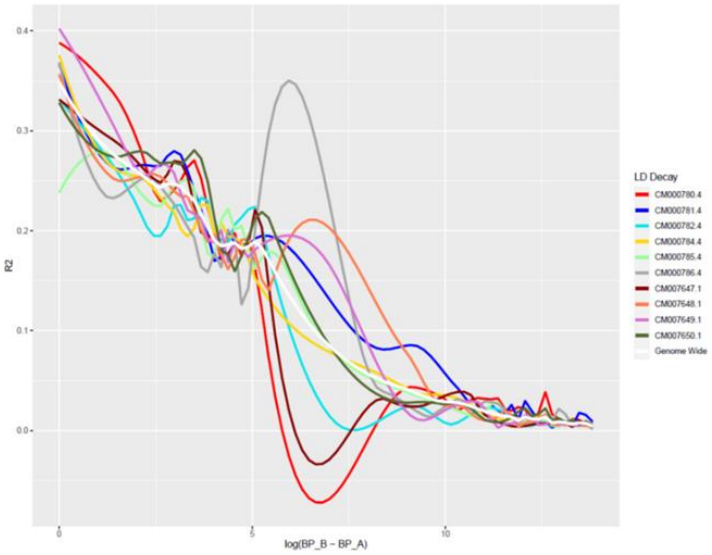


Figure 2. Linkage disequilibrium (LD) decay pattern for all chromosomes and genome wide in the entire panel of 360 genotypes. Linkage disequilibrium r2 is reported on the y-axis, while the log of the distance between two SNPs for each SNP couple is on the x-axis.

2.3. Stratification analysis of the maize collection

The inbreds under study were assigned to groups based on available pedigree data, including ‘Insubria’, ‘Microsperma’, elite lines derived from US germplasm, and white grain genotypes.

Subsequently, a phylogenetic tree was constructed. As shown in Fig. 3, the majority of lines belonging to the ‘Insubria’ type, as well as the ‘Microsperma’ lines, were subdivided into 2 separate groups. Furthermore, a third, more spread group including elite lines derived from the US germplasm and white grain genotypes, was also present.

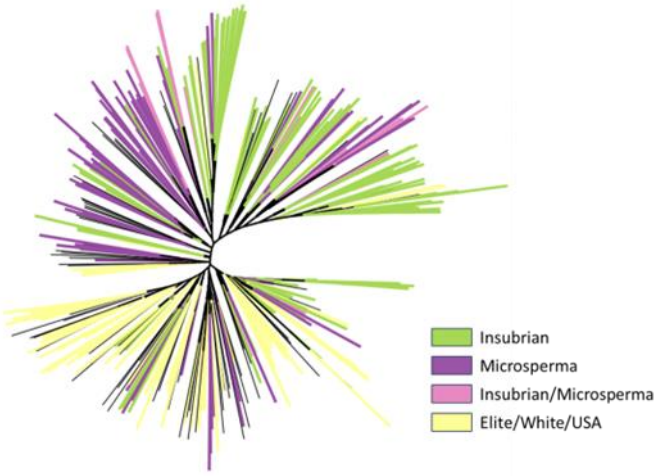


Figure 3. Phylogenetic tree of the maize panel.

Population stratification was analyzed by ADMIXTURE sub-population member-ship from k = 2 up to k = 20 based on the SNP dataset pruned at r2 = 0.8. Grouping statistics, in particular the cross-validation error rate for ADMIXTURE, stabilized at k > 4 and highlighted a certain differentiation between inbred lines derived from the same cultivar/landrace. Stratification of the collection was also analyzed at K=4 for the other grouping methods, based on the within-cluster sum of square for K-means, and the cluster dendrogram with a cut at four groups for hierarchical clustering (Fig. 4).

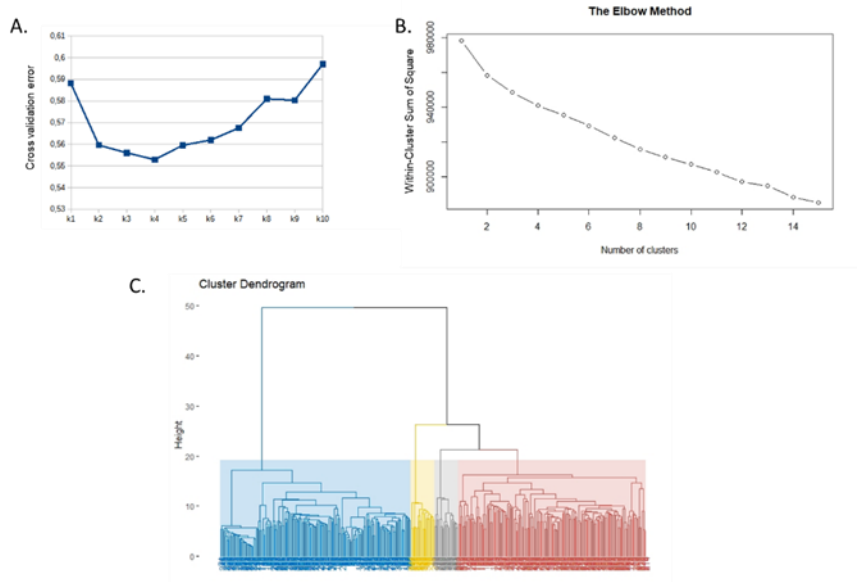


Figure 4. Clustering grouping statistics: A. Cross validation error rate for ADMIXTURE. B. With-in-Cluster Sum of Square for K-means. C. Cluster dendrogram with a cut at four groups for hierarchical clustering.

A first analysis based on PCA revealed that the two first components collectively explained 5.07% of the total variance (3.5% for the first component and 1.6% for the second). It was possible to identify a certain differentiation between the lines plotting the genotypes based on these two components, even though the maize lines did not form clearly separate groups (Fig. 5).

The results of the three clustering methods, alone or combined with PCA (first five components) or LDA, were compared using a support vector machine algorithm. Results of the comparison showed that the non-parametric clustering algorithms seem to respond better in the prediction phase than ADMIXTURE: accuracy levels of 0.90 in ADMIXTURE, 0.92 in K-means and 0.92 in hierarchical clustering for raw data, 0.90 in ADMIXTURE, 0.94 in K-means and 0.88 in hierarchical clustering for PCA, 0.77 in ADMIXTURE, 0.84 in K-means and 0.79 in hierarchical clustering for LDA were estimated. Considering the kappa coefficient, we estimated that the best performing

methods are K-means and hierarchical clustering for raw data (0.86), K-means for PCA (0.9), and K-means for LDA (0,77). According to the comparison made with the support vector machine algorithm, in this specific case K-means appeared the best clustering method, as it differentiated the groups formed by ‘Insubria’, ‘Microsperma’, and elite/USA/white types better.

Based on the above-mentioned method, group 1 (119 lines) collects mainly ‘Microsperma’ lines and some ‘Insubria’ lines of the ‘Pignoletto’ type, a few lines derived directly from ‘Nostrano dell’Isola’ or from crosses between one ‘Insubrian’ and one ‘Microsperma’ parent. Group 1 also includes 34 lines of miscellaneous type. Group 2 (63 lines) is mostly composed of ‘Insubria’ lines belonging to the ‘Nostrano dell’Isola’ type and lines derived from crosses in which one or both parents were ‘Nostrano dell’Isola’. Very few lines of the ‘Cinquantino’ type, lines derived from crosses with US germplasm, and lines belonging to the miscellaneous group are present in Group 2. Group 3 (22 lines) contains specific genotypes belonging to the ‘Scagliolino GV’ type, one ‘Nostrano dell’Isola’ and two ‘Marano’. Finally, group 4 (156 lines) contains elite lines, white grain genotypes, lines derived from US germplasm, many ‘Insubria’ lines, few ‘Microsperma’ lines, and 50 lines from the miscellaneous group (Fig. 5.A). Interestingly, a certain correspondence between k-means clustering and the phylogenetic tree was observed, as illustrated in Fig. 5.B, and 5.C.

The other two clustering methods gave similar results, particularly for the very well-separated group of the ‘Scagliolino GV’ type, but also some different aspects were observed (Fig. S1).

Based on ADMIXTURE analysis, the first group (68 lines) included a majority of ‘Insubria’ lines, in particular, ‘Nostrano dell’Isola’, ‘Isolabasso’, ‘Scagliolo’, and lines de-rived from crosses between ‘Nostrano dell’Isola’ and ‘Isolabasso’ or ‘Marano’. Some ‘Microsperma’ lines, such as ‘Sacra famiglia’ and ‘Cinquantino’, and very few elite and white lines were also included in this group). The second group (87 lines) was composed mainly of ‘Insubria’ lines, in particular, the group ‘Nostrano dell’Isola’, and some ‘Microsperma’ as ‘Cinquantino’ and ‘Sacra Famiglia’. Group 3 (68 lines) put together a very well-defined set of ‘Insubria’ lines, the ‘Scagliolino GV’, lines derived from the ‘Pignoletto’ group and crosses in which ‘Pignoletto’ is one of the parents, and ‘Microsperma’ lines belonging to the types ‘Sacra Famiglia’, ‘Marano’, ‘San Pancrazio’ and ‘Cinquantino’. The fourth group (137 lines) comprised in particular elite lines, lines derived from US breeding programs, and lines with white grain. Additionally, some ‘Insubria’ lines, one ‘Nostrano dell’Isola’, some lines derived from crosses between ‘Nostrano dell’Isola’ and other lines in which

those derived from USA and ‘Scagliolo’, and a few lines belonging to ‘Scagliolo’ group were included in the fourth group. Finally, a few ‘Microsperma’ lines, in particular, ‘Sacra Famiglia’ and ‘Cinquantino’, were also included in this large group.

According to hierarchical clustering analysis, the first group (44 lines) was com-posed of elite lines and some ‘Insubria’ lines belonging to the ‘Nostrano dell’Isola’ type. Group 2 (198 lines) grouped mainly ‘Insubria’ and ‘Microsperma’ lines (79 and 62, respectively), besides other lines of the miscellaneous type (Fig. 4.D). Group 3 (22 lines) contained specifically the ‘Scagliolino GV’ type, and just one ‘Nostrano dell’Isola’ and few ‘Marano’ lines. Group 4 (96 lines) was largely corresponding to ADMIXTURE group 4 (except for ‘Microsperma’ lines’) and contained elite lines, lines derived from US breeding programs, white grain genotypes, and lines derived from crosses between ‘Nostrano dell’Isola’ and U.S. lines or other ‘Insubria’ lines.

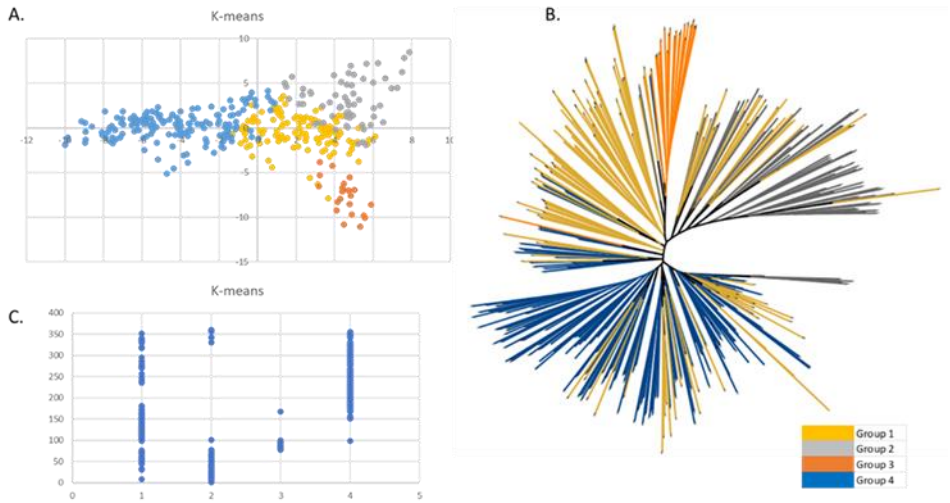


Figure 5. Clustering and phylogenetic analysis of the maize panel. A. Population structure carried out with K-means combined with PCA. B. Phylogenetic tree: colors indicate the groups identified through K-means. C. Correspondence between the grouping obtained through K-means and the phylogenetic tree.

2.4. Genetic diversity analysis

A genetic diversity analysis of the maize collection based on genotyping data was carried out. Differentiation within and among groups on the base of group classification obtained through the k-means method was also defined. AMOVA highlighted a low level (3.25%) of genetic variance distinguishing the four groups (Table 2.A), with the largest portion being observed within groups (96.75%). In group 4, containing many modern lines, concerning groups characterized by inbred lines derived from traditional landraces, a reduction of overall diversity was observed. Group 3, containing lines of the ‘Scagliolino

GV' type, was also the most differentiated group in terms of Fst values, and was shown to be the one with the highest genetic diversity. Indeed, Fst values resulted higher comparing group 3 vs. group 4 and group 2 (0.078 and 0.054, respectively - Table 2.B).

Table 2. A. AMOVA and gene diversity for five germplasm sub-sets defined according to K-means classification. B. Above diagonal elements (shades of green) of the matrix contain the average number of pairwise differences, while below diagonal elements (shades of blue) correspond to pairwise Fst values. Diagonal elements (shades of orange) report gene diversity within groups calculated as the mean number of pairwise differences. Significance was assessed upon 1,000 permutations. All values are significant at $p < 0.001$.

A.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among populations	3	18361.74	32.18	3.25
Within populations	716	685463.41	957.35	96.75
Total	719	703825.14	989.53	
F_{st}		0.032		
Maize groups	**Nº accessions**	**Nº polymorphic loci over 15872**	**Nei's gene diversity**	**Mean number of pairwise differences**
K4-1	119	15868	0.1312	2081.88
K4-2	63	15243	0.1340	2042.19
K4-3	22	12864	0.1685	2167.74
K4-4	156	15857	0.1073	1701.07
Mean value			0.1352	1998.22
Lsd (p=0,05)			0.0010	14.97
Lsd (p=0,001)			0.0016	23.58

B.

Group	1	2	3	4
1	2081.882	2090.928	2212.779	1936.086
2	0.014	2042.191	2222.115	1955.579
3	0.041	0.054	2167.744	2078.851
4	0.023	0.045	0.078	1701.074

3. Discussion

A consistent reduction in crop genetic diversity has been commonly associated with intense breeding activities aimed at developing better-performing cultivars. Even if breeding has undoubtedly improved yield and end-product quality of the most important commodities as maize, likely during the breeding process some genes/alleles useful for

agronomic traits have been lost. For this reason, maize germplasm collections composed of inbreds derived from different sources, such as traditional landraces and elite breeding lines, are of paramount importance as they represent reservoirs of genetic diversity useful to face the current agriculture challenges, mainly linked to the need of increasing production in more severe and variable environmental conditions. In this perspective, the work conducted in the early 1950s by the Regional Inspectorates of Agriculture (Ispettorati Provinciali dell'Agricoltura) and the Experimental Station for Maize Cropping (Stazione Sperimentale per la Maiscoltura) of Bergamo is of great importance. Samples of maize populations grown in different Italian regions were collected, preserved, and multiplied at Bergamo, to avoid landrace extinction and the loss of precious germplasm. More than 600 inbreds were obtained from these samples, through self-pollination or through crosses between lines derived from landraces and/or lines developed through recent breeding activities. The inbreds constitute part of a larger maize germplasm collection now conserved at CREA, Research Centre for Cereal and Industrial Crops of Bergamo. For the present study a subset of 384 inbreds belonging to different groups, was chosen for a deep genotypic characterization.

In particular, 106 inbreds were assigned to the group 'Insubria' (or 'Padani'), a racial group of inbreds deriving from the convergent adaptation to the agrosystem located in the peneplains of the Insubrian-Euganean region of Italy, where maize found a preferential habitat [3]. This class included: i) 41 inbreds derived from 'Nostrano dell'Isola', a group of landraces widely grown in many provinces of Italy and originated in the sub-Alpine region of the Bergamo province and ii) other lines of different types as 'Isola Basso', 'Scagliolo' and 'Scagliolino'. The inbred types 'Pignoletto', 'Isola Basso', 'Rostra-to', 'Scagliolo', and 'Scagliolino' resulted also belonging to the 'Insubria' group. A second racial group corresponds to the 'Microsperma' type, ('Microsperma flints') which includes landraces characterized by subcilindrical ear, small seeds with a very hard and horny texture, medium plant size and suitable for late spring or summer planting. Seventy in-bred lines belong to this group, including those derived from the landraces 'Marano', 'Nostrano dell'Isola maranizzato', Cinquantino', 'Sacra Famiglia', and 'San Pancrazio'. Additionally, 15 lines derived from a cross between an 'Insubrian' and a 'Microsperma' line. Also, some elite inbreds, developed between 2000 and 2012 years in the frame of the breeding activities carried out at CREA-Research Centre for Cereal and Industrial Crops of Bergamo (23 lines) were included in the analysis, together with 23 genotypes with white seeds comprising pearl white flints and white dents, and 45 inbreds derived from lines of the U.S. breeding programs. The remaining lines could not be assigned to a particular group (Table S2).

Choosing the genotyping method is crucial when evaluating large collections of accessions. In this context an optimal balance between costs, number, and quality of markers has to be considered. SSR markers are still used in genetic diversity studies in maize, as they are codominant and highly informative having many alleles per locus. Nevertheless, the more recent availability of reference genomes also for species with very complex genomes, and of next-generation sequencing techniques to produce sequence variation data, has made it possible to develop large sets of SNPs at a relatively low cost for this kind of studies. Although SNP arrays show several advantages, such as the presence of exonic and intronic SNPs, markers differentiated between heterotic groups, and markers associated with known genes, on the other hand, GBS has been improved for SNP discovery and mapping thanks to the two-enzyme approach, and sequence data software and pipelines have recently been developed [22]. GBS has been extensively used for genetic diversity studies, GWAS, and genomic selection approaches in many crops, including maize [23]; it offers the advantage of not showing ascertainment bias, although it can be characterized by a high rate of missing data when the sequencing depth is not optimal; hence strongly reducing the number of usable markers when a strict filtering process is applied to retain high-quality markers [22]. The strong reduction of markers was con-firmed in the present study, in which nearly 16,000 SNPs were retained starting from an initial pool of 797,368. Nonetheless, the coverage of the maize genome was good, and an average of 99.72% of the chromosomal extension was tagged by SNPs. Additionally, the physical coverage was rather uniform, apart from the gene-depleted pericentromeric regions (Fig. 1), with an average of 7.60 SNPs/Mbp. With a reported gene density in maize of 0.5-10.7 genes per 100 Kbp [24], our dataset could be used for the identification of candidate genes for traits, identified in association studies. All the SNPs used for the different analysis are biallelic, and a certain percentage of heterozygous individuals have been identified for each marker. As the accessions of our germplasm collection are inbreds, a filter based on heterozygous loci (20%) has been applied to obtain the final SNP dataset, according to the scientific literature [8]. The SNP dataset used in the present study appears suitable for a deep genetic characterization of the germplasm maize collection under study in terms of percentage of missing data, allelic frequency and heterozygous loci. Moreover, it presents the great advantage of the lack of ascertainment bias usually presented by SNP arrays. Indeed, the SNPs used in the present study have been developed directly in the accessions of the germplasm collection analyzed, allowing to obtain coherent results with different methods of analysis, as described later.

The observed LD decay in the maize collection is 12.3 Kb at r2 = 0.1. This value is comparable to other studies, as values comprised between a few kb and hundreds of Kb are usually identified at the same value of r2. As an example, the average decay distance of the LD across all chromosomes was about 5.2 kb at r2 = 0.1 for a panel of 80 maize inbred lines covering more than 80 % planting area in Jilin Province (China) [25]. Similarly, an LD-decay as 2.65 Kb at r2 = 0.1 was found in the CAAM panel (419 tropical/ sub-tropical lines) assembled by CIMMYT to map resistance to northern corn leaf blight [26]. On the other hand, higher values, such as 41.5 kb in a panel of 226 inbreds from China [10], and 310 kb in another Chinese panel of 292 inbred lines [8], were also identified.

With a total expected heterozygosity of 0.12 and a mean PIC value of 0.225, the 360 entries of the CREA maize collection show a good level of genetic diversity considering their restricted geographical origin. These values are in line with those identified in previous studies carried out with SNP markers. In fact, it is known that SSR markers are characterized by higher PIC values due to their multi-allelic nature. Aci et al. [27] found a value of 0.622 in an Algerian maize collection (47 landraces) from Saharan Oasis genotyped with 18 SSRs. Lu et al. [28] analyzed a set of 287 tropical and 160 temperate maize inbred lines, genotyped with 1943 high-quality SNPs and found a PIC value of 0.251 for the entire set with small differences when the tropical and the temperate sets were considered separately. Wu et al. [10] found a much higher PIC value (0.60), probably due to the integration of the temperate set with many lines from the Suwan region, which are characterized by a larger genetic diversity. Chittò et al. [29] found PIC values between 0.28 and 0.36 for a series of AFLP markers in a set of 71 Italian inbred lines. Similarly, Losa et al. [30] found a PIC value of 0.34 in a collection of 144 Italian inbred lines considered representative of the breeding material developed at the Bergamo Maize Breeding Station and genotyped with 811 AFLP loci.

Grouping statistics stabilized at k > 4, as reported in Fig. 4; so, at this k value it is possible to identify the main landrace types corresponding to the groups ‘Insubria’, ‘Microsperma’, a large group including more recent elite lines, in which are present those de-rived from the US germplasm and white grain genotypes, and a smaller group well separated from other lines corresponding to the ‘Scagliolino’ type.

The different methods used in this study provided comparable results in grouping the lines, with some differences, in particular for ‘Microsperma’ lines. These lines are well separated by the k-means method from the other groups. Accordingly, the comparison between the three methods done by the SVM algorithm also suggests a better

performance for k-means. K-means and hierarchical clustering are non-parametric machine learning methods [31-33] that are not based on the assumptions of the Hardy–Weinberg principle and use external dimension reduction techniques, such as principal component analysis [34], commonly used in several data-intensive biological fields [35]. Similarly to our results, other authors found that non-parametric methods are more effective than ADMIX-TURE in assigning individuals to groups [35].

In general, a clear and complete separation of genotypes based on landrace group name was not observed. An example is the ‘Nostrano dell’Isola’ type; it traces back its origin to the Caribbean cylindrical maize types, and some lines are characterized by a medium-late growing cycle and by a typical long ear with an enlarged butt and isodiametric orange flint grains [3]. It represents a very large and diverse group in Italy, and similar types are endemic in other maize countries of southern Europe: Portugal, Spain, and Romania. In our study, lines belonging to the ‘Nostrano dell’Isola’ racial type span at least three groups identified by k-means and the other clustering methods; additionally, they show a certain degree of differentiation based on their origin. In the maize collection described in the present study, lines derived from ‘Nostrano dell’Isola’ accessions, lines derived from crosses in which both parents are ‘Nostrano dell’Isola’, and lines derived from crosses in which one of the parents is ‘Nostrano dell’Isola’, and the other parent a line belonging to ‘Microsperma’, elite or lines from USA breeding programs, are included. More in detail, most of the lines derived from crosses in which one of the parents is ‘Nostrano dell’Isola’ and the other parent is a line belonging to ‘Microsperma’, elite or a line from USA breeding programs, cluster in group 4 identified by K-means, together with elite or lines from USA breeding programs. In the same group, a unique ‘Nostrano dell’Isola’ line, which is probably the parent of the other lines, is also included. Inbreds of ‘Microsperma’ type are mainly clustered in group 1 identified by k-means, but also in this case there are some exceptions, as the accessions Lo440 (‘San pancrazio 85’), Lo495 (‘Marano Vicentino’) and lo425 (‘Cinquantino San fermo’), which cluster in group 4, despite they derive from samples with the same name as other ‘Microsperma’ inbreds of group 1.

The lack of correspondence between clustering and the name of the different Italian lines was also observed in previous studies in which a small subset of this collection was characterized with AFLP markers [29,30]. Another example is given by some lines belonging to the ‘Pignoletto d’oro’ type: they are positioned with lines of the ‘Microsperma’ group despite ‘Pignolo’ lines being considered of ‘Insubria’ type. Some ‘Marano’ lines, belonging to the ‘Microsperma’ type, are very similar to ‘Nostrano dell’Isola’. This finding can be explained by the fact that ‘Marano’ lines are considered as derived from crosses

between ‘Nostrano locale’ lines derived from ‘Nostrano dell’Isola’, grown in 1916 in the Bergamo province, and ‘Pignoletto d’oro’ [36].

The AMOVA analysis revealed a low level (3.25%) of genetic variance between groups (Table 2), compared to the largest portion observed within groups (96.75%). Simi-lar results have been shown for other maize collections, even if with different proportions of variance, depending on the diversity and heterogeneity of the lines included in the different panels. As an example, in recent studies, values of genetic variance between groups varying from 3% [37] to 17% [38] were reported. Based on Nei's genetic diversity, a reduction was observed in group 4, containing many modern lines, with respect to groups characterized by inbred lines derived from traditional landraces. This result is according to the reduction in diversity observed in general following events of domestication and breeding observed not only in maize but also in other crops such as wheat [39].

The ‘Scagliolino’ type represents a well-defined group for which a clear correspondence is found between clustering and the name of the lines. Moreover, white grain lines also tend to be grouped together, as previously observed in [3]. Genetic diversity analysis is also useful to identify duplicate inbreds in a germplasm collection. The different analyses carried out in the present study allowed to identify some very similar inbreds, as in the case of Lo3 and Lo16, both derived from ‘Nostrano dell’isola’, and Lo579 and Lo580, both derived from ‘Nostrano dell’isola maranizzato’ samples. This information will be undoubtedly useful to develop a core collection from the current germplasm panel.

A set of 49 Lo Italian lines belonging to our collection and provided by CREA to INRAE, 28 of which in common with our present study, were previously included in a large panel of maize accessions derived from different countries by Gouesnard et al. [6]. The Italian lines were derived from different landraces of the ‘Insubria’ group as ‘Nostra-no dell’Isola’, ‘Isolabasso’, and ‘Scagliolo’, and they grouped together in PCoA analysis. Accordingly, the lines of that study resulting in common with our present study also grouped together, except in some cases. Of course, the Italian origin of most of the lines in our collection restricts its genetic basis compared to a wider collection composed of lines with a worldwide provenience such as that described by Gouesnard and colleagues [6]. Nonetheless, the genetic variation in the Bergamo collection remains significant. For this reason, the collection of inbred lines described in the present study represents a valuable source of genetic diversity to contribute to maize European breeding programs. Moreover, the available genotypic characterization allows us to exploit the collection in genome-wide association studies. These could be aimed to identify, other than new

QTLs, new alleles at loci of interest for specific traits, eventually left behind during the modern breeding process, but which can still be useful in breeding, as genes for resistance to dis-eases.

4. Materials and Methods

4.1. Plant material

The panel involves 384 maize inbreds (coded as 'Lo' followed by a numeric value), including 353 lines from a wider panel of lines preserved at CREA Bergamo Genebank and 31 additional advanced breeding lines (elite inbreds - EILs) derived from crosses between different Lo lines (Table S2). These inbreds derive from samples of landraces cultivated in different regions of Northern Italy before the diffusion of hybrids. All the lines be-long to 'Indentata' and 'Indurata' (dent and vitreous, flint) groups.

4.2. Genotyping and data processing

The lines were grown in 2018 in Bergamo in open field. Each genotype was sown in 4-meter-long rows (20-25 plants/row). Single plants were self-pollinated, and a sample from the flag leaf of each line was collected and air-dried in a vacuum dryer. All the samples were then shipped to Freedom Markers company (Ames, Iowa, USA), for DNA ex-traction and subsequent analysis. The lines have been subject to extensive SNP search through the tunable Genotyping by Sequencing (GBS®) technology [21] conducted with the restriction enzyme Bsp1286I (Freedom Markers, Data2Bio, Iowa, USA). Samples were sequenced using an Illumina HiSeq X instrument, and reads were aligned to the Zea mays AGPv4 (GCA_000005005.6) reference genome downloaded from NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_000005005.2/). SNP calling was conducted using only those reads that aligned to a single location in the reference genome.

Based on SNP data, 24 genotypes were removed from subsequent analyses due to a high missing rate. The dataset was then filtered based on 20% missing data, 20% heterozygous calls, and a minimum allele frequency > 0.05 [8]. Using the software PLINK [40], redundant markers were taken into account by pruning the dataset by applying a linkage disequilibrium threshold r2 = 0.99 genome-wide. Redundant markers were then merged into one unique SNP call. Moreover, two additional pruned datasets were produced by applying two different thresholds at r2 0.8 and 0.5 to run the population structure analysis. The dataset was subject to missing data imputation using Beagle 5 software applying de-fault parameters [41].

4.3. LD Decay

Pairwise marker correlations (r2 values) were determined using Plink v.1.9 considering the SNP data for each chromosome. LD decay curves were fitted through the non-linear model described in Rexroad and Vallejo [42]. The fitted regression curves were used to determine critical parameters of marker distances at r2 = 0.3 and 0.5. The r2 of un-linked markers (background noise) was estimated as the 95th quantile of r2 values of markers on different chromosomes (unlinked set). LD was calculated for each marker using the mean r2 with the 50 nearest markers, and then smoothed as one value using the step-sliding window, in order to evaluate the local LD value along chromosomes.

4.4. Population structure and cluster analysis

A bayesian method, ADMIXTURE [43], and two non-parametric methods, K-means (KM) [44] and hierarchical clustering (HC) [45] were used for clustering analysis. Dataset results pruned at r2 = 0.99 are shown. For ADMIXTURE the block relaxation algorithm and the quasi-Newton convergence acceleration method were used [43]. For both methods, be-longing to a sub-population was defined for k values increasing from 2 to 10. The optimal number of clusters was estimated based on ADMIXTURE's cross-validated error rate and minimum group size. The clustering methods were used alone or in combination with external dimension reduction techniques, such as principal component analysis (PCA) and linear discriminant analysis (LDA), calculated with Tassel 5.2 [46] and R package MASS, keeping the first 5 components; PC's 1 and 2 are visualized in a 2D scatter plot with each PC explaining 3.4 and 1.6 percent of variance, respectively.

A machine learning approach based on the Support Vector Machine model (R pack-age e1071) with a linear kernel has been used to compare the results of the clustering analysis, assessing how dimensionality reduction influences the prediction accuracy and what combination best fits the data. While the clustering approaches used are unsupervised (no information on clusters is given a priori), the SVM algorithm is a supervised model for classification and regression analysis. A label has been associated with each element of the dataset, that is a cluster membership class; the entire dataset has been then divided into training and test sets (with a ratio of 0.8); the training set then has been trained through SVM producing a classifier, which applied to the test set provided predictions on the clustering classes. Comparing these results with the expected ones, the confusion matrix with the accuracy values of the predictions and Cohen's kappa coefficients [47] for all the data combinations (raw, PCA, LDA) and clusters (ADM, KM, HC) were elaborated. To better generalize the predictive results and balance the random effect, it was decided to use the k-fold cross-validation technique

[48] with k=5 (R package caret), i.e. the dataset was divided into 5 parts, and alternately a part was chosen as a test set, with the rest as training; the results are represented as the average of the 5 fold predictions.

4.5. Phylogenetic tree

IQ-TREE v2.1.2 [49] was used for phylogenetic analyses; the PHYLIP format was used for input files. ModelFinder was used to define the model that minimizes the Bayesian information criterion score, then an ascertainment bias correction model was applied with an ultrafast bootstrap for 10,000 replicates. The command used was iqtree -s input.phy -m MFP+ASC -bb 10000 -seed 1701, where MFP is a model finder, ASC is an ascertainment bias correction, and bb is an ultrafast bootstrap approximation. The online tool Iroki was used to visualize the tree [50].

4.6. Analysis of genetic diversity within the maize collection

AMOVA was carried out to calculate genetic diversity among and within populations, together with fixation index (Fst, [51]), and the polymorphism information content (PIC, [52]). The within populations total number of polymorphic loci (N), Nei's gene diversity [53], and mean number of pairwise differences were calculated, and statistical significance was assessed based on least significance difference (LSD) at $P < 0.05$. Population differentiation was evaluated based on Nei's genetic distance [54] and population pair-wise Fst. All calculations were made using the Arlequin 3.5 software [55], and significance levels for variance components and Fst statistics were estimated based on 10,000 and 1,000 permutations, respectively.

Chapter 3: Genomic associations to adaptation traits in the Italian maize core collection

Introduction

Drought and heat stress and their impact on maize cultivation

Currently, climate change is heavily affecting the agriculture sector, in particular drought stress represents one of the biggest challenges of climate change and for most of the crops cultivation. Maize is highly affected by drought stress and the yield loss can be of 50% and 21% during the silking and grain filling stages respectively (Széles et al., 2023). In maize drought stress can vary based on the frequency and on the geographical scenarios. The water deficit can occur at early or late stage of plant development and it can last shorty or longer with different consequences. Considering the climate change previsions for 2050 in Europe, the scenarios characterized by early stress deficit will increase their frequency in most of the countries. Southern Europe is typically characterised by longer period of water deficit and this tendency are confirmed by the climate change previsions, especially for Spain, Italy and Greece (Harrison et al., 2014). In addition, drought is a complex trait where is highly present the impact of the environment to the genetics. Many physiological mechanisms and agronomic traits are involved in the response to drought stress. Osmotic adjustment, dehydration tolerance, and reduction in photosynthetic activity are one of the most common and important physiological mechanisms and in particular, the reduced photosynthetic activity has as a consequence the stomatal closure induced by ABA hormone concentration increasing, and a decreased photosynthetic enzyme production (Sheoran, 2022). The adaptability of the maize plant to the water scarcity is also evaluated considering how rapidly the plant restore its biological processes after the rewatering. This ability depends on drought duration, intensity and genotypes and it is connected to the “plant memory” of pre-stress growth (Sun et al., 2016). The flowering is the vegetative stage more altered by the deficiency of water and many studies demonstrated that anthesis-silking interval (ASI) is key trait for the maize selection under drought conditions. ASI exhibits a significant negative correlation with grain yield and relatively high heritability characterizing it as an excellent secondary trait for this purpose (Araus et al., 2012) . It is important to highlight that a trait impacting the drought tolerance in some specific conditions cannot work on a different scenario. For example, the reduction of anthesis-silking time impacts on yield in condition of late and low stress while for early stress with early relieved it is more determinant for yield the early maturity (Harrison et al., 2014). ASI, maturity, kernels per ear and ears per plant are specifically related to crop yield but other secondary traits useful to select for drought stress can be early vigor, root architecture and stay green

(Sheoran, 2022) (Harrison et al., 2014).The modification of root architecture in maize regards the increasing/reduction of lateral root branching length and density. Lateral roots start from the primary roots and the genotype with reduced later root density and high later root distribution performs better under drought stress. Following this strategy, the plant can allocate more resources on the axial elongation to reach the deeper soil and only later, reached a soil zone containing more greater moisture, the plant develops lateral roots (Zhan et al., 2015) (Hazman & Kabil, 2022). The process of stele lignification represents another physiological mechanism correlated with the drought tolerance because can improve the root mechanical strength and increase the capability to penetrate hard drying soil preserving the ability of lignified roots to work as conduits from the high-moisture deep soil to the rest of the plant (Hazman & Kabil, 2022). The effects on climate change regards also the increasing of the temperature and the consequently waves of heat. Heat stress affect strongly maize cultivation and it is usually connected to the water deficiency. The most vulnerable stages for the growth of maize plants undergone to heat stress are the early vegetative stage and the reproductive stage. Particularly during the reproductive stage heat damages the plant and its reproductive organs. The consequence of these damages have a direct and specific impact on yield and grain filling. Tassel due to its position on the plant is particularly exposed to high temperatures and the drying of the pollen causes consequently infertility. The pollen viability is the most frequent and serious effect of heat but also the stigma receptivity can be damaged. The kernel set reduction associated with the plant infertility caused by heat stress is up to 57% and 80% compared to hybrids cultivated in optimal condition (Alam et al., 2017). The decreasing in water potential and lack of heat-shock protein (HSPs) production in pollen exposed to heat stress might be some of the reasons for its susceptibility. Heat-shock proteins (HSPs) are proteins expressed during the activation of heat stress responses (HSRs), homeostatic mechanisms that mitigate the effects of heat stress. Heat stress responses are controlled by specific transcription factors called heat shock factors (HSFs) and represent a cytoplasmatic response to mainly heat stress but also to other stresses. Diverse studies of transcriptome on maize lines undergone to heat treatment demonstrated the presence of HSFs in maize (Z. Li & Howell, 2021). An example is HSFTF13, transcription factor close to HSFTF13 present in *Arabidopsis thaliana* which is involved in the response to abscisic acid and in thermotolerance (Z. Li et al., 2020). To identify heat tolerance in maize, two specific secondary traits have been found significantly associated with it: seed set percentage under open pollinated condition and pollen shedding duration. Also senescence could be considered an important trait for drought and heat stress in field conditions (Alam et al., 2017).

Methods & Materials

Plant materials and field experiments

The collection of 106 inbred lines was cultivated during 2022 and in 2023 increased to a total number of 204 lines. The panel includes inbred lines derived from traditional Italian maize varieties and represents the Italian maize diversity for inbred lines, as regards the class of maturity and cob features. The lines were thus planted in open field in three growing seasons, 2022, 2023 and 2024, to evaluate a number of traits related to drought and heat stress as the anthesis-silking interval (ASI), the height of the plant and the date of flowering. After harvest, other traits were collected as the length of the cob, the number of ranks and the number of seeds for rank. The panel also included line B73, the reference line and reference genome for maize and other European lines used as tester lines for the different maturity class.

Genotype	Line name
A632	Tester line
B37	Tester line
B73	Reference line
Lo 3	Nostrano dell'isola
Lo 4	Nostrano dell'isola
Lo 5	Nostrano dell'isola
Lo 8	Nostrano dell'isola
Lo 16	Nostrano dell'isola
Lo 17	Nostrano dell'isola
Lo 18	Nostrano dell'isola
Lo 21	Nostrano dell'isola
Lo 28	Pignoletto d'oro
Lo 30	Rostrato
Lo 34	Isola basso
Lo 35	Isola basso
Lo 37	Isola basso
Lo 38	Scagliolo
Lo 38-1	Scagliolo
Lo 39	Scagliolo
Lo 46	Scagliolo
Lo 47	Scagliolo

Lo 50w	Bianco Wimberg
Lo 52w	Bianco Oderzo
Lo 54w	Bianco Oderzo
Lo 58	Marano
Lo 59	Marano
Lo 60	Marano
Lo 64	Scagliolino G.V.
Lo 65	Scagliolino G.V.
Lo 67	Scagliolino G.V.
Lo 69A	Scagliolo
Lo 71	Nostrano dell'Isola
Lo 72	Scagliolino G.V. precoce
Lo 74	Scagliolino G.V. precoce
Lo 79	Scagliolino G.V.
Lo 81	Scagliolino G.V.
Lo 82	Marano
Lo 84	Colleoni Marne
Lo 88	Colleoni Marne
Lo 89	Scagliolino G.V. precoce
Lo 90	Scagliolino G.V. precoce
Lo 92	Scagliolino G. V. precoce
Lo 107w	Bianco Piave tipo scagliolo
Lo 113	Scagliolino G.V. precoce
Lo 115	Scagliolino G.V. precoce
Lo 116	San Pancrazio
Lo 120	Pfister Papetti
Lo 134	ExLo32xLo3
Lo 143-1	Marano
Lo 143-2	Marano
Lo 144	Scagliolino G.V. precoce
Lo 145	San Pancrazio
Lo 145-1	San Pancrazio
Lo 151	F.vulgaris
Lo 151-1	F.vulgaris
Lo 151-2	F.vulgaris

Lo 154	Campascio
Lo 158	WF9^3xM14
Lo 173	Mammarth
Lo 175	P.B. 14
Lo 175-1	P.B. 14
Lo 175-2	P.B. 14
Lo 176	INIA Alcala
Lo 177	Afganistan
Lo 185	Lo32xLo38
Lo 186	MaranoxIsolaBasso
Lo 187	Lo38xLo3
Lo 188	Lo29xLo58
Lo 190	Lo32xLo58
Lo 194	Scagliolo Marne
Lo 206	Marano 28
Lo 217	Long Yellow
Lo 220	TschirpanL N96
Lo 221	O.P. 25
Lo 228	Mais Greco
Lo 235	Torinese
Lo 276	Silverqueen
Lo 280	Mais Greco
Lo 283	11x88
Lo 285	11x165
Lo 289	11x34
Lo 294	70x110
Lo 297-1	2x118
Lo 297-2	2x118
Lo 299	5x190
Lo 309	KingKo Foggia
Lo 314	Tschirpan L.N. 96
Lo 318	Tschirpan L.N. 97
Lo 328	Lo3xLo32
Lo 330	3Ax190
Lo 332	Scagliolo DK orange

Lo 333	Scagliolo DK orange
Lo 344	A73xA334^2
Lo 347-1	Lo4xLo71
Lo 347-2	Lo4xLo71
Lo 348	Nostrano
Lo 350	Nostrano
Lo 364w	Ex 4RH3
Lo 367w	Wisconsin 7 di Aquileia
Lo 379	Sacra Famiglia
Lo 388	Cinquantino di Cremona
Lo 392	Cinquantino di Cremona
Lo 394	Cinquantino San Fermo
Lo 400	Sacra Famiglia
Lo 404	Sacra Famiglia
Lo 405	Sacra Famiglia 51
Lo 406 A	Sacra Famiglia 51
Lo 407	Sacra Famiglia
Lo 408	Sacra Famiglia
Lo 409	Sacra Famiglia
Lo 412A	Sacra Famiglia 51
Lo 429	Sacra Famiglia 73
Lo 432	Sacra Famiglia
Lo 433	Cinquantino bianchi
Lo 434	Cinquantino bianchi
Lo 435	Cinquantino bianchi tipo marano
Lo 437	Sacra Famiglia 24
Lo 438	Ill.AxWisc.CC5(AxW23)
Lo 442	Scagliolo Marne
Lo 444	TV206(Lo30xW32xW187)
Lo 446	Scagliolino
Lo 449w	Dente di Cavallo
Lo 451w	Bianco dentato 4
Lo 451w-1	Bianco dentato 5
Lo 451w-2	Bianco dentato 6
Lo 452	Lo5^2xLo19

Lo 454	San Pancrazio
Lo 455	Nostrano dell'Isola
Lo 456	TV206(Lo30xW32xW187)
Lo 457	Lo43xLo58
Lo 461	Scagliolo Marne
Lo 466	Lo12^2xLo19
Lo 468	Lo18xLo32
Lo 469-1	Nostrano dell'isola
Lo 469-2	Nostrano dell'isola
Lo 471	Gran di Merano
Lo 474-1	Lo33xLo18
Lo 474-2	Lo33xLo18
Lo 480	Giallone locale
Lo 481	Spadone
Lo 487	Nostrano dell'isola
Lo 491	Nostrano dell'Isola Finardi
Lo 492	Colleoni
Lo 494	Scagliolo Marne
Lo 496	Brianzolo
Lo 505	Scagliolo
Lo 506	Lo18xLo19
Lo 507	Marano Vicentino
Lo 509	Marano Vicentino
Lo 511	Bastardello
Lo 512	5x190
Lo 516	Trentinella 8 file
Lo 527w	Zamengo
Lo 532	Tre nodi di Cecina
Lo 536	Ideale
Lo 537	Pignoletto nostrano giallo
Lo 543	Cinquantino bianchi
Lo 549wc	Calabrese
Lo 550	Locale maranizzato
Lo 555	Locale
Lo 564rc-3	Cinquantino montagnana

Lo 564wc	Cinquantino montagnana
Lo 567	Cinquantino montagnana
Lo 568	Cinquantino montagnana
Lo 570	Cinquantino montagnana
Lo 571	Cinquantino montagnana
Lo 573	Cinquantino montagnana
Lo 574	Cinquantino montagnana
Lo 576	Nostrano dell'isola maranizzato
Lo 577	Nostrano dell'isola maranizzato
Lo 578rc	Nostrano maranizzato
Lo 578wc	Nostrano maranizzato
Lo 579-1	Nostrano dell'isola maranizzato
Lo 579-2	Nostrano dell'isola maranizzato
Lo 580	Nostrano maranizzato
Lo 587	Nostrano dell' isola
Lo 589	Nostrano dell'isola
Lo 600	Nostrano dell'isola maranizzato
Lo 602	Nostrano dell'Isola
Lo 611	Mandello
Lo 615	Lierna
Lo 621	Lierna
Lo 623	Lierna
Lo 624	Lierna
Lo 627	Lierna
Lo 1270	Lo1056xLatina
Lo 1278	Lo1096xLo1058
Lo 1320A	Lo1142xP3394
Lo 1425	Lo1301xLo1106
Lo 1430	Lo1240xLo1208
Lo 1451	Lo1279xLo1183
Lo 1453	Lo1279xLo1301
Lo 1455	Lo1299xLo1263
Lo 1457	Lo1301xLo1106
Lo 1459	Lo1301xLo1303
Lo 1497	SynBGSF

Lo 1498	DKC6530
Lo 1500	Klaxon
Lo 1504	Lo1260xLo1270
Lo 1507	Lo1301xLo1255
Lo 1512	Lo1296xLo1208
Lo 1524	Syn MP
Lo 1530	PR31G98
Mo17	Tester line

Table 1. List of traditional Italian lines included in the thesis

Phenotypic data collection and analysis

The lines have been characterized by the following phenotypic traits in open field: emergence, juvenile vigor, flowering time (male and female), anthesis-silking interval (ASI), and plants height and ears insertion height. After the harvest, the ears were characterized for the ear length, fertile ear length, seed for rank, number of ranks and 1000 seed weight. The vegetative stage and the details about the phenotypes recorded are described in the table below. The phenotypic data collected from the different years were normalized with the R package *bestNormalize* and their heritability was calculated with *inti*. The Pearson correlation was calculated on BLUP value with the R package *PerformanceAnalytics*. *T*he analysis of variance (ANOVA) were conducted with R package *stats.* Phenotypic data for GWAS analysis were analyzed in R studio with two packages, *inti* and *bestNormalize*. The outliers were removed and the data were normalized following the best fitting model.

Phenotypic trait	**Unit of measurement**	**Vegetative stage**
Emergence	Days	F3-F4, 3-4 visible leaves
Juvenile vigor	Score from 1 to 5	F6-F7, 5-6 visible leaves
Anther emission	Days	Anther emission for 50% of the plot
Silk emission	Days	Anther emission for 50% of the plot
Anthesis-silking interval (ASI)	Days	
Plant height	Centimetre (cm)	Flowering stage (after pollen emission, considered without tassel)

Ear insertion height	Centimetre (cm)	Flowering stage (after pollen emission)
Ear length	Centimetre (cm)	After harvest
Fertile ear length	Centimetre (cm)	After harvest
Seeds for ranks	Count	After harvest
Ranks number	Count	After harvest
1000 seed weight	Grams (gr)	After harvest

Table 2. Description of the phenotypic traits recorded during the open field activities

GWAS analysis

A genome-wide association mapping study (GWAS) was carried out on data collected in the two years (field trials 2022 and 2023) at the farm "La Salvagna" in Bergamo. The genotyping, population structure, and LD decay distance have been described in a previous work (Mastrangelo et al., 2024). The lines were genotyped using a genotyping-by-sequencing (GBS) for 797,368 SNPs. The SNPs were filtered out using Plink and Beagle 5 for missing data and minimum allele frequency (MAF < 0,05). GWAS has been carried out in R studio using the GAPIT3 tool (J. Wang & Zhang, 2021). For the GWAS analysis, three different models were performed: Blink, Mixed Linear Model (MLM) and Multiple Locus Mixed Linear Model (MLMM). According to the Manhattan plots and QQ plots, a compromised threshold of $P \leq 0.0001$ was regarded as threshold to for the significant MTAs in this population.

QTL research

After a literature review, the work of Leng et al., 2022 was identified as a resource for the identification of QTLs correlated with MTAs found by the genome-wide association mapping study conducted in the present work. In order to liftover the QTL genomic regions of Leng et al., 2022 calculated on version 2 of B73 (https://download.maizegdb.org/B73_RefGen_v2/B73_RefGen_v2.fa.gz) on version 4 of B73 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005005.2/), the tool *bedtools* (v2.31.1) (Quinlan & Hall, 2010) with the option "*getfasta"* was applied to obtain the fasta sequences of B73 v2 corresponding to the QTLs. Successively, blastn (Nucleotide-Nucleotide BLAST 2.16.0+) (Y. Chen et al., 2015) was used to align the sequences to B73 v4. Finally, the correspondence between the significant MTAs and the QTLs updated to B73 v4 was found applying *bedtools* with the option "*closest*".

Results

The phenotypic data recorded are summarized in Table 1. It is notable that the mean of ASI and plant height were higher for 2022 which was characterized by the high temperatures and the water deficit. The data show a high coefficient of variation, in particular for juvenile vigor, ASI, plant height and ear insertion height, and this observation implies high phenotypic variation in our panel (Table 1). The distribution of each phenotypic trait is shown in Figure 1, Figure 2, Figure 3 and Figure 4. The data are normally distributed for male and female flowering and for plant height while for ASI the distribution is positive and asymmetric.

	Year	Emergence	Juvenile vigor	Male flowering time	Female flowering time	ASI	Plant height	Ear height
Mean	2022	152,3	1,82	200,1	202,9	3,654	108,8	59,27
	2023	149,6	2,457	200,3	201,6	1,711	122,9	54,72
Range	2022	152-157	1-3	186-213	191-220	0-15	52-163	25-71
	2023	135-159	1-5	189-214	125-216	0-15	65-202,5	20-105
Sd	2022	1,09	0,42	4,09	5,23	2,99	22	15,8
	2023	1,84	0,64	4,91	6,42	1,68	25,87	15,62
CV (%)	2022	0,72	23,1	2,05	2,58	81,89	20,22	26,66
	2023	1,23	26,23	2,45	3,19	98,0	21,06	28,54

Table 1. Statistic description of phenotypic data

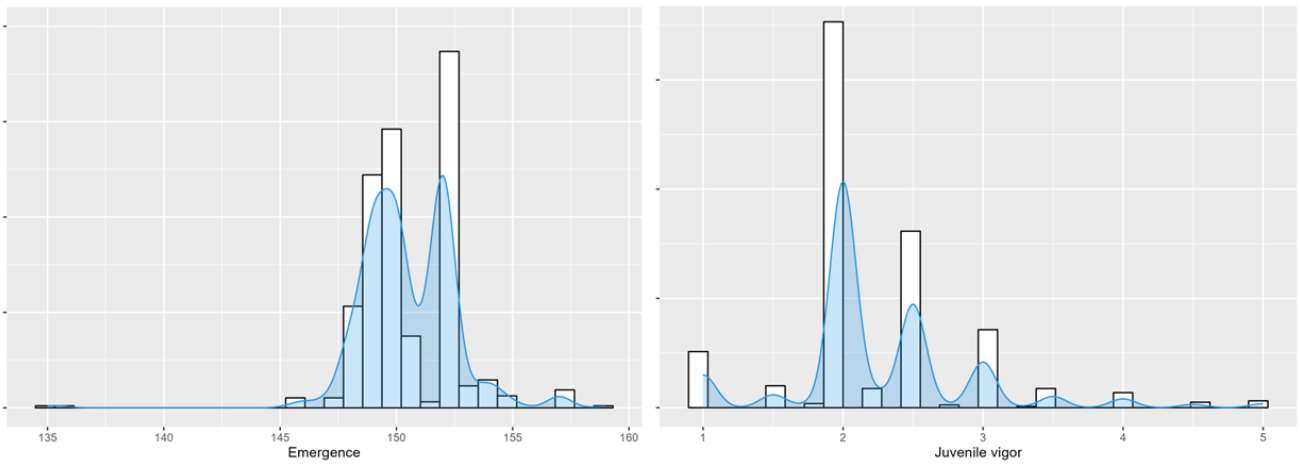


Figure 1. Emergence and juvenile vigor data distribution

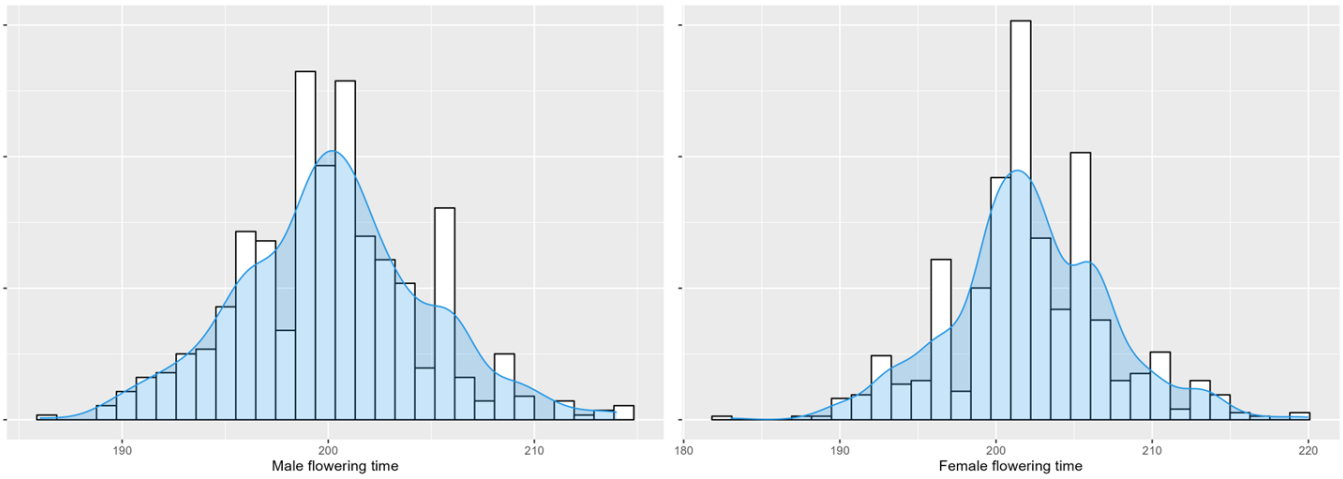


Figure 2. Male and female flowering time data distribution

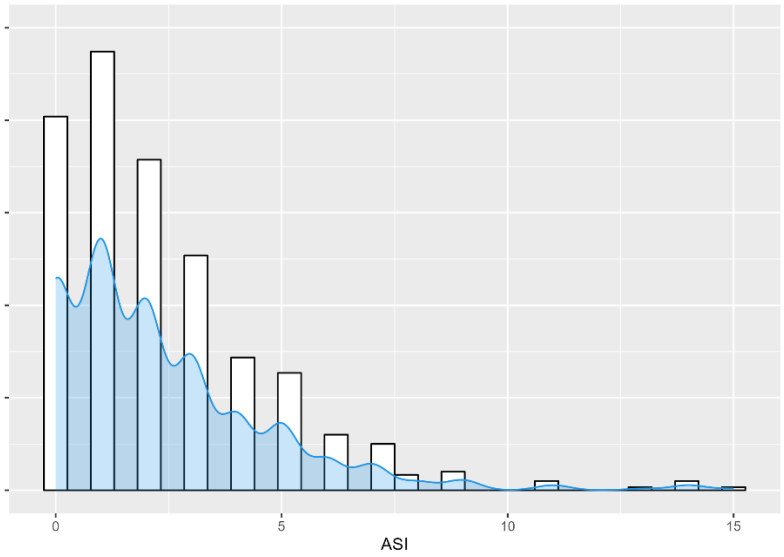


Figure 3. ASI data distribution

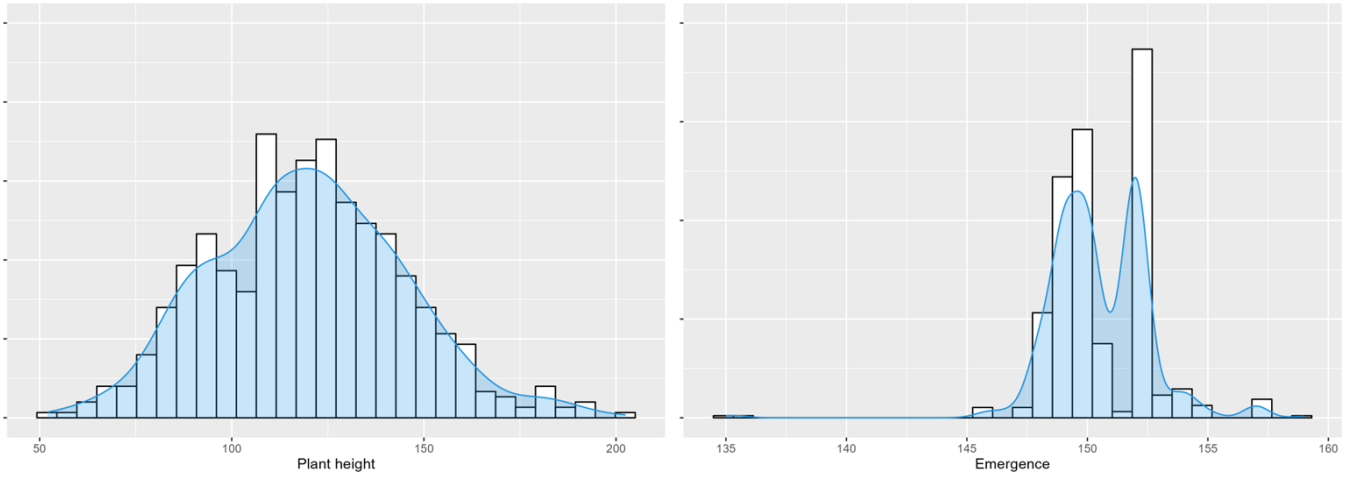


Figure 4. Plant and ear insertion height distribution

The Pearson correlation was calculated on BLUP values and reports significance for all phenotypic traits (Figure 1). In particular, male and female flowering time are highly correlated (0,88) and also female flowering time and ASI show high correlation (0,37). Plant height and ear insertion height show a high correlation of 0,71 while the correlation between plant height and juvenile vigor is 0,57. It is notable the negative correlation of -0,27 between juvenile vigor and ASI. Furthermore, emergence shows a negative correlation of -0,54 and -0,37 for juvenile vigor and plant height respectively (Figure1).

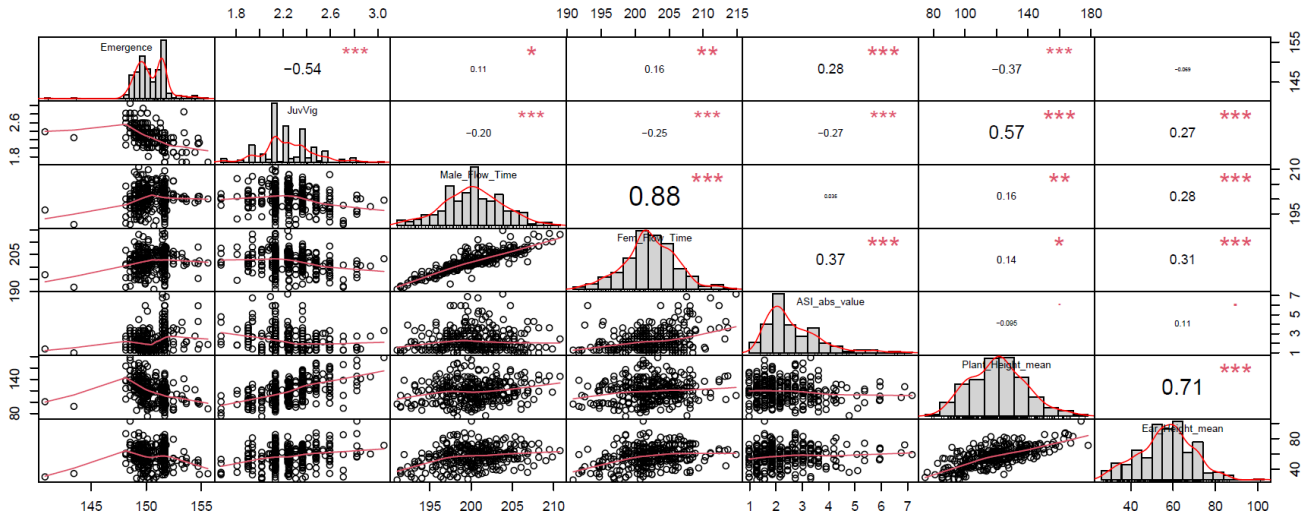


Figure 1. Correlation plot of phenotypic data

The ANOVA results highlight a statistical significance for every trait recorded among the genotypes in 2022 and 2023 field (p value < 0.001) (Table 2 and Table 3). The data did not fit a normal distribution, then a transformation was applied (Figure 1) (Table 2 and 3). The heritability reported on Table 2 and Table 3 shows high-moderate values for flowering time-related traits in both years. H^2 of male flowering time has a value of 0,74 and 0,90 for 2022 and 2023 respectively while female flowering time has a heritability value of 0,81 and 0,48 for 2022 and 2023 respectively. ASI heritability has a similar value for both years which is around 0,40. The heritability value is stable for all the other traits on both years except for juvenile vigor which decreases the H^2 value from 0,47 to 0,19 due to the different score range.

2022	Genotype significance		Rep significance	Transformation	Heritability (H^2)
Emergence	2,553E-05	***	ns	Double reversed Log_b(x + a)	0.57
Juvenile vigor	0,0007615	***	0,0119709	Asinh(x)	0.47
Male flowering time	2,18E-15	***	ns	OrderNorm	0.74
Female flowering time	2,00E-16	***	ns	/	0.81
ASI	0,0001413	***	ns	Sqrt(x + a)	0.41
Plant height	1,34E-08	***	ns	Center scale	0.66

Ear height	2,00E-16	***	0,022	Box cox	0.87

Table 2. Analysis of variance and heritability for phenotypic data of 2022

2023	**Genotype significance**		**Rep significance**	**Transformation**	**Heritability (H^2)**
Emergence	4,33E-05	***	ns	Center scale	0.56
Juvenile vigor	0,0002531	***	2,20E-16	Log_b(x + a)	0.19
Male flowering time	2,00E-16	***	ns	OrderNorm	0.90
Female flowering time	2,00E-16	***	ns	OrderNorm	0.48
ASI	0,003068	**	ns	Double reversed Log_b(x + a)	0.46
Plant height	2E-16	***	0,03099	Center scale	0.69
Ear height	2E-16	***	ns	Double reversed Log_b(x + a)	0.77

Table 1. Analysis of variance and heritability for phenotypic data of 2023

The GWAS analysis was performed with three models of GAPIT3 tool: Blink, MLM and MLMM (Figure 2 and Figure 3). The last two models show similar number of MTA while Blink shows higher number of MTA identified. The results found by every GAPIT model were compared and only MTAs present in at least two model were considered significant. Altogether 102 significant MTAs were detected with a threshold of $1.0 \times 10{-}4$. The number of significant MTAs for each phenotypic trait are reported in Table 4. From the figure 4 to the figure 7 are reported the Manhattan plots for the model MLMM for every trait for both of the years.

	Juvenile vigor	**Male flowering time**	**Female flowering time**	**ASI**	**Plant height**	**Ear height**	**Tot**
2022	18	7	5	9	10	5	54
2023	6	8	12	14	3	5	48

Table 4. Significant MTA for each trait (p value < < 9.98E-4)

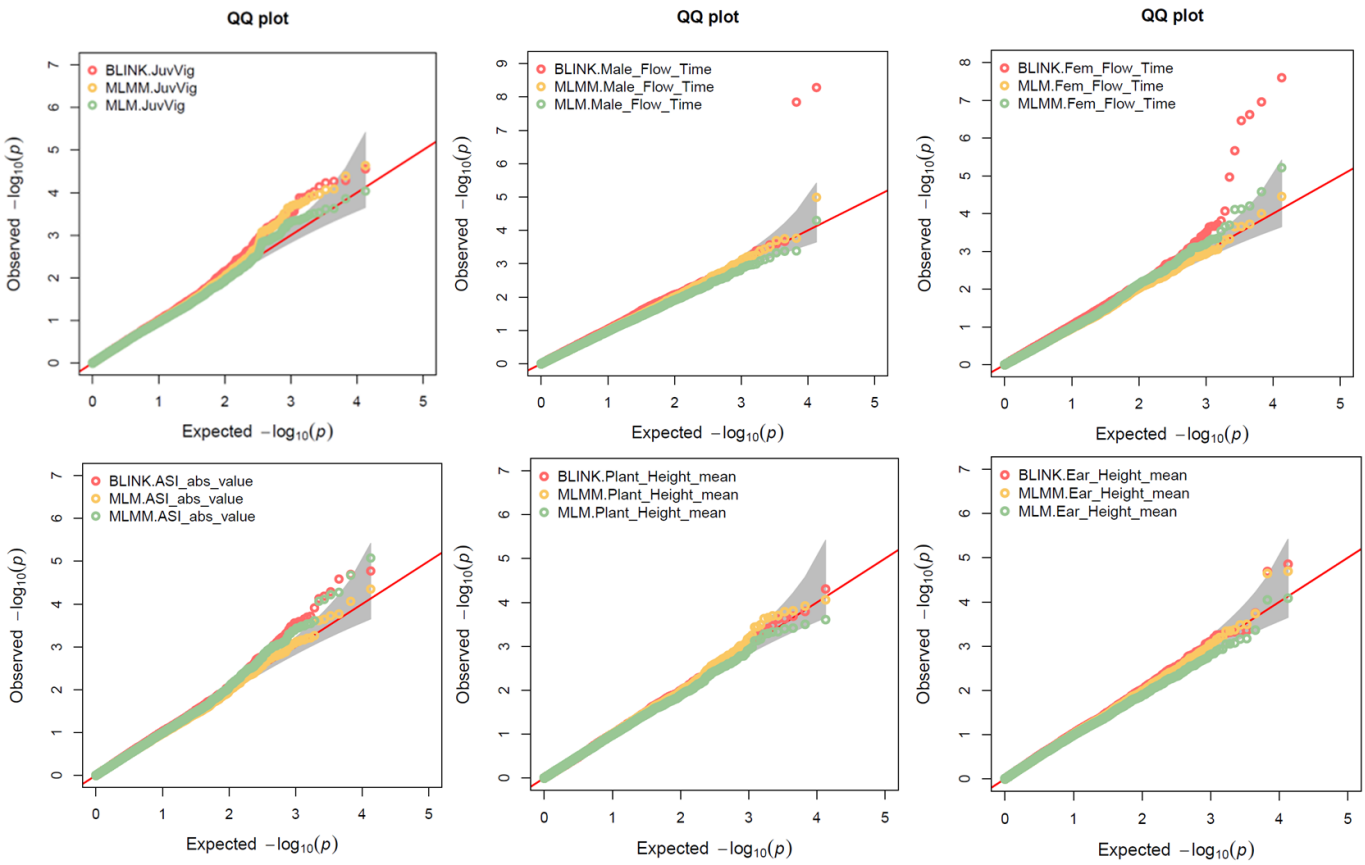


Figure 2. QQ plots for each trait in 2022

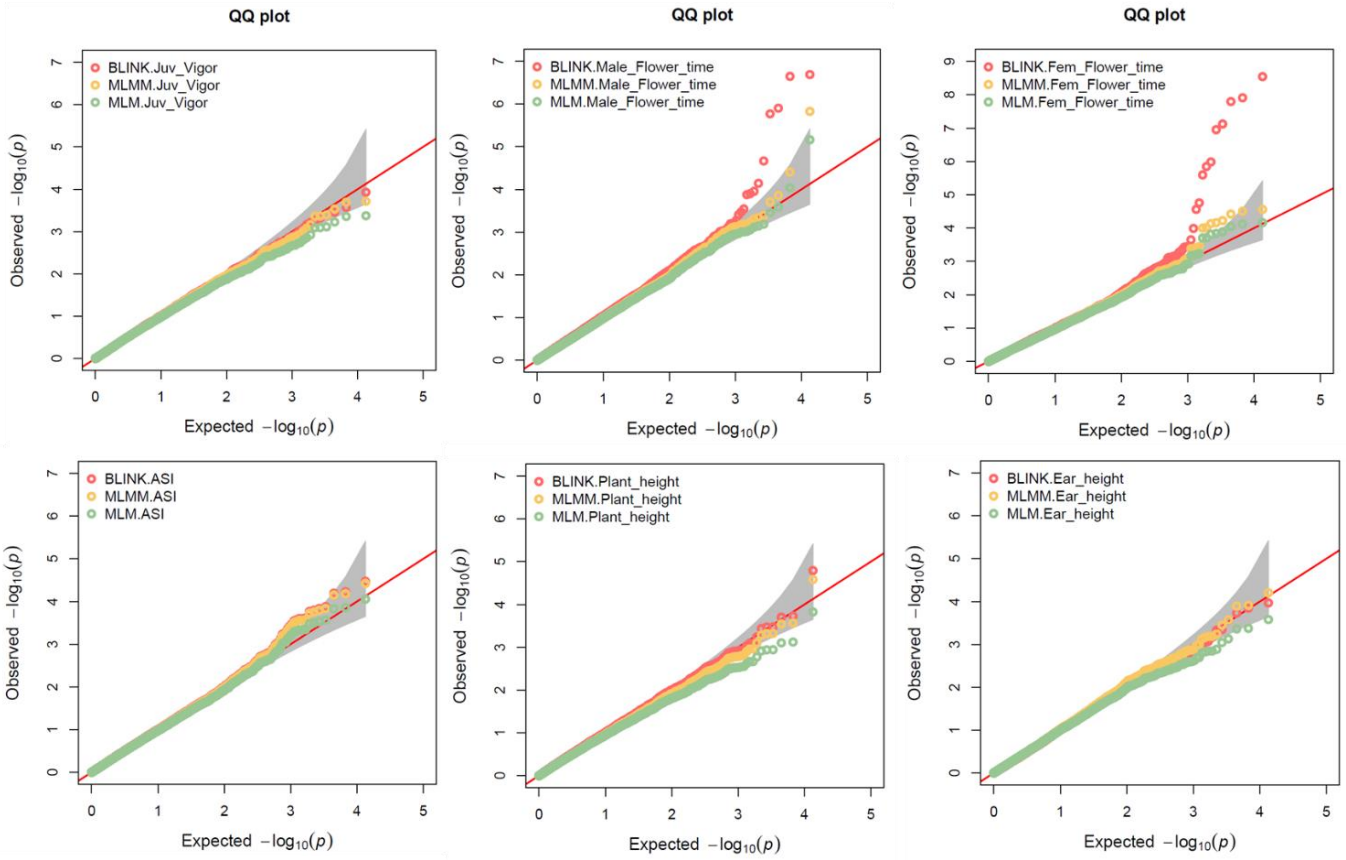


Figure 3. QQ plots for each trait in 2023

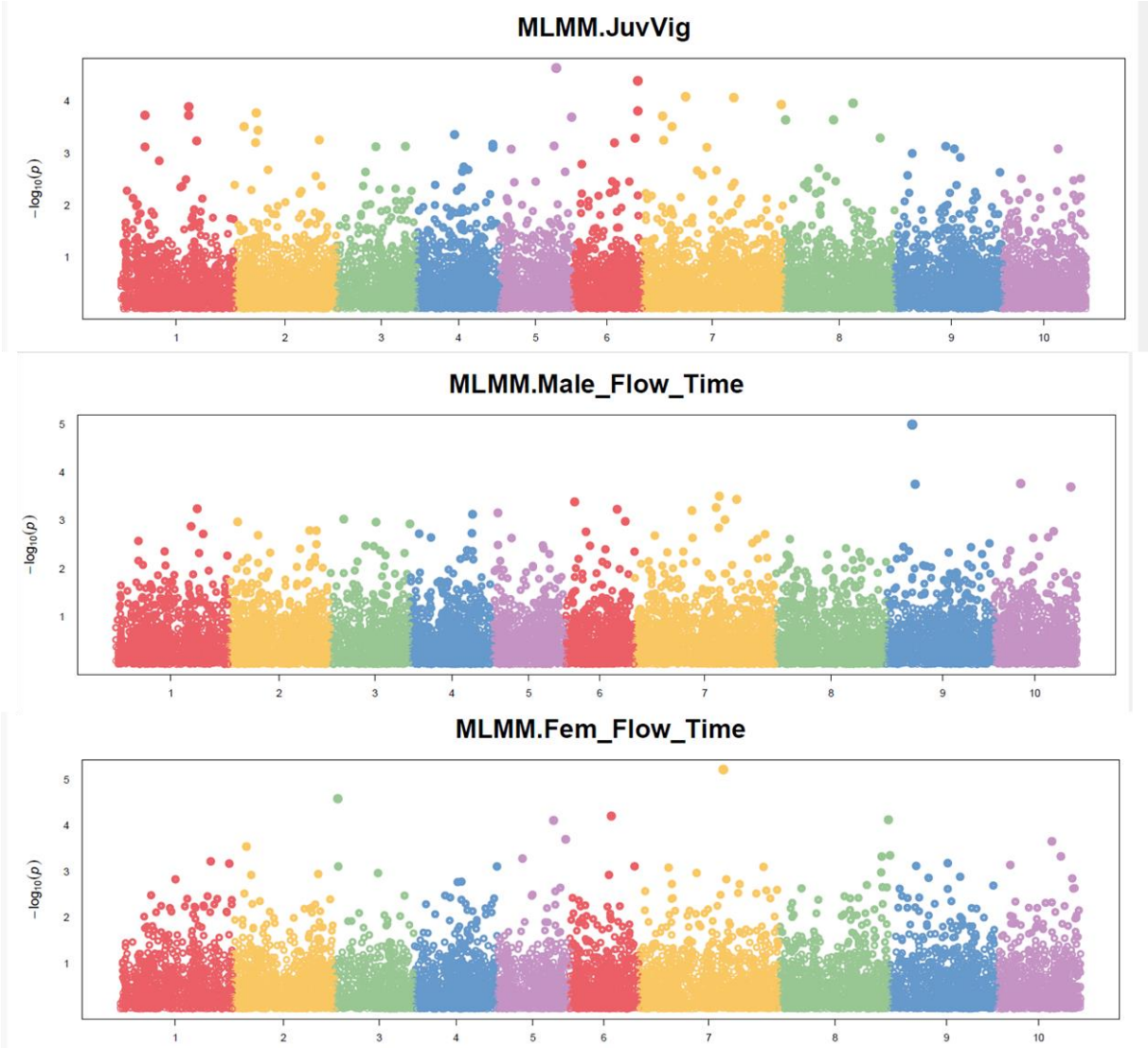


Figure 4. Manhattan plots for juvenile vigor, male flowering time and female flowering time in 2022

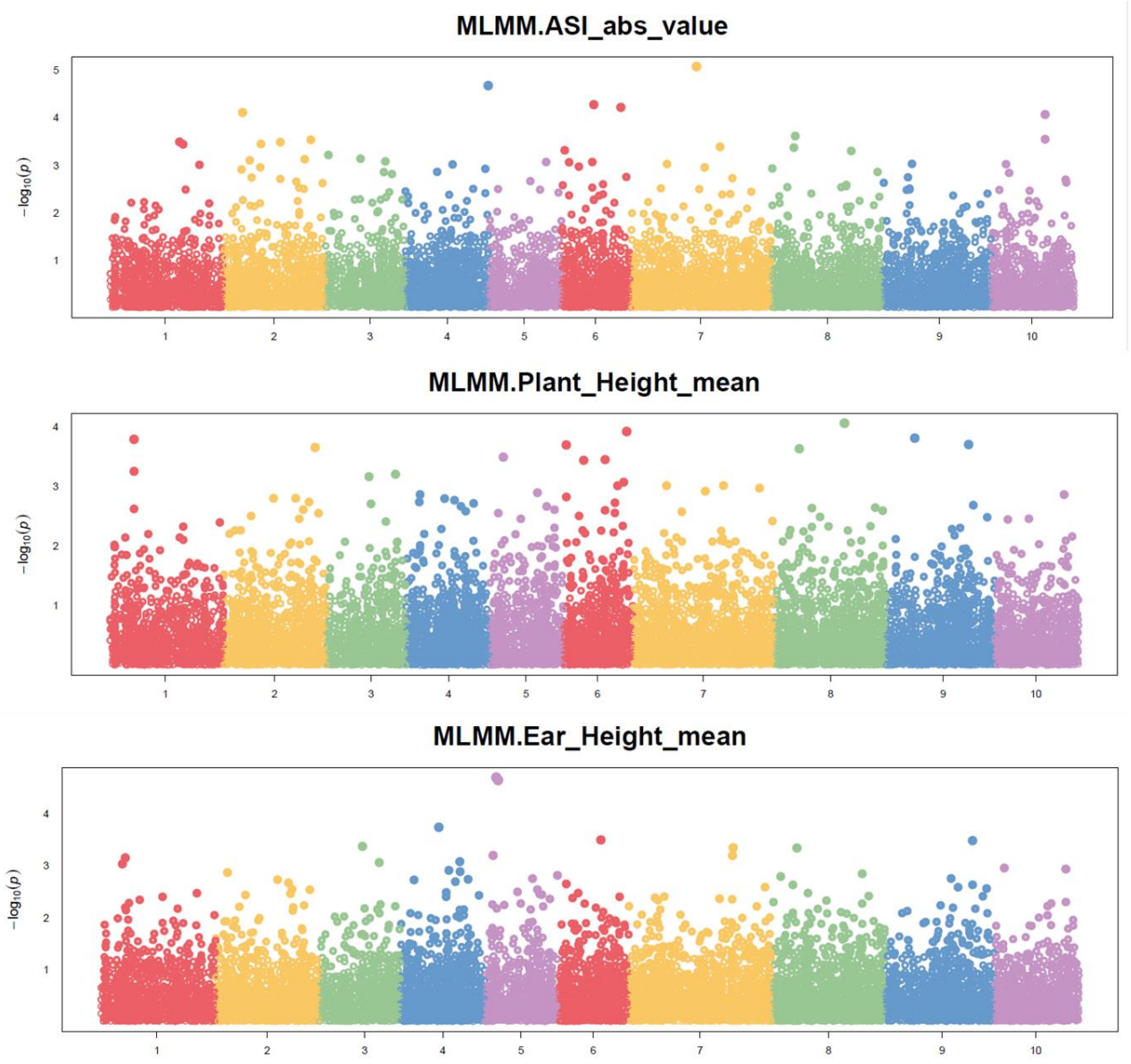


Figure 5. Manhattan plots for ASI, plant height and ear insertion height in 2022

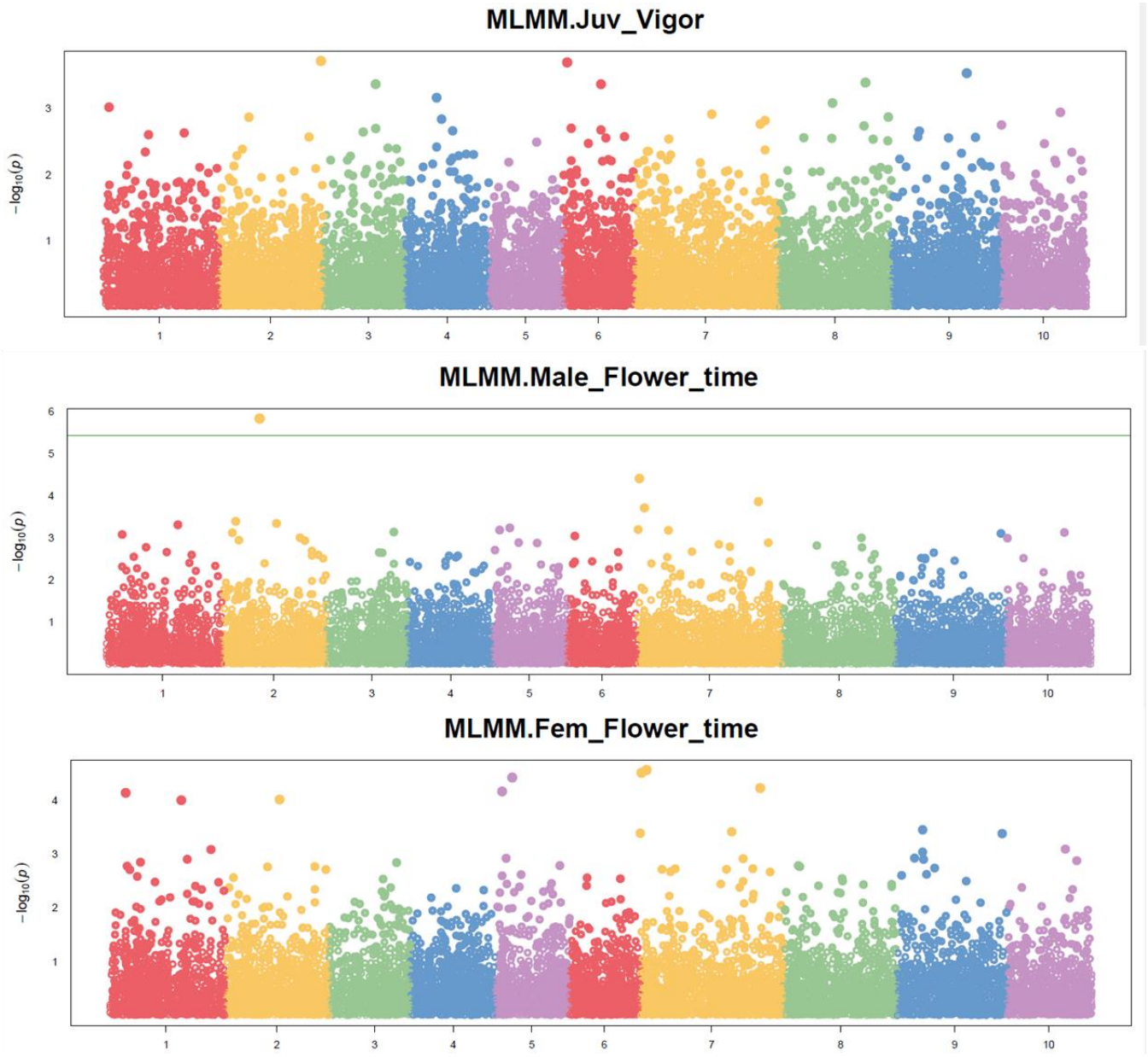


Figure 6. Manhattan plots for juvenile vigor, male flowering time and female flowering time in 2023

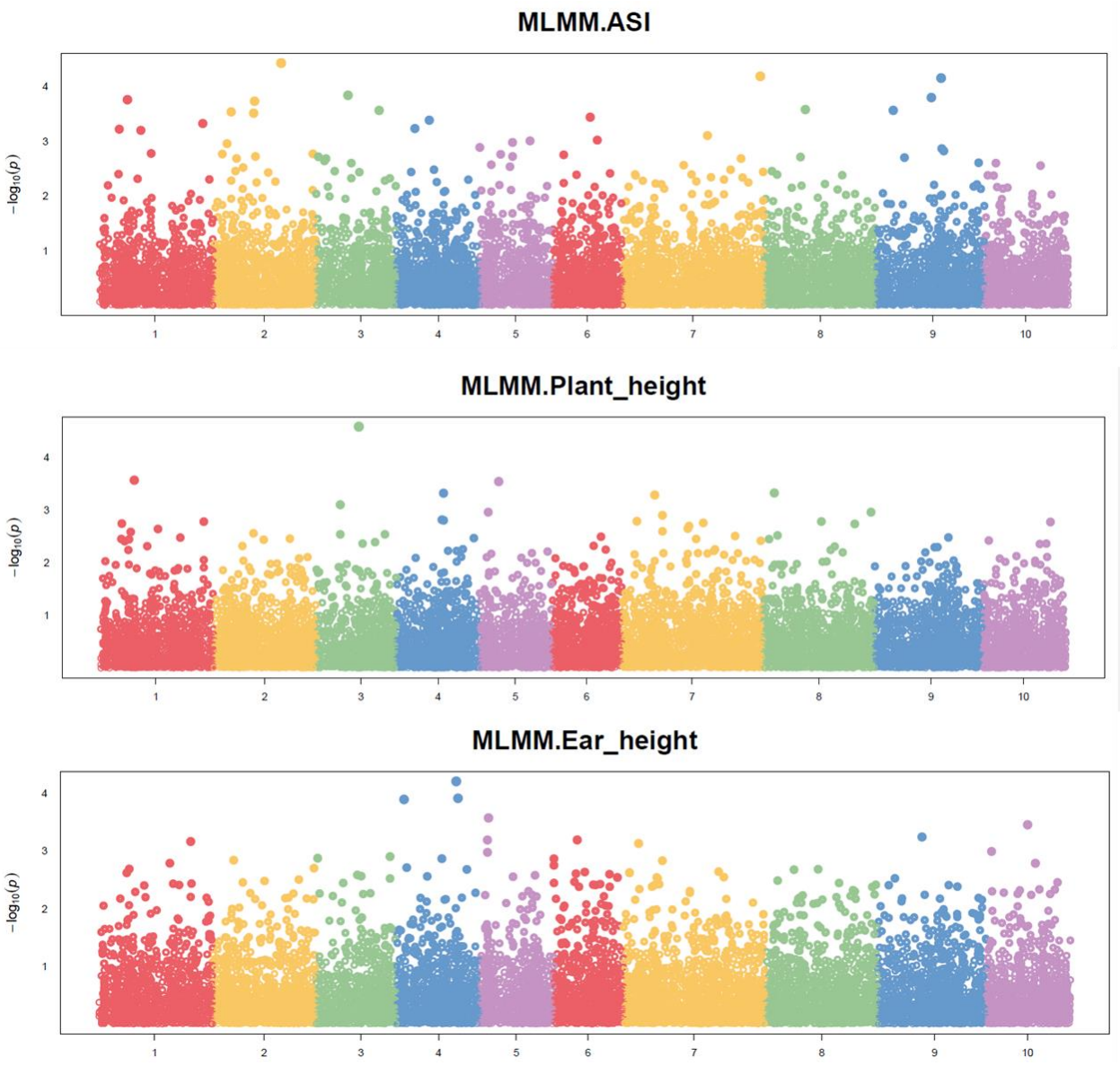


Figure 7. Manhattan plots for ASI, plant height and ear insertion height in 2023

The GWAS analysis was focused on traits related to flowering and on fifteen markers associated with QTLs already known in literature were identified (Table 5). The genomic regions of the QTLs calculated on version 2 of the reference genome B73 found in the work of Leng et al., 2022 were adjusted for the version of the refence genome used for sequencing alignment of this present study (version 4). For male flowering time (Days-to-Anthesis), five markers are associated with different QTLs, only one marker is associated with female flowering time (Days-to-Silking) while the significant MTAs for ASI are nine. Five markers were identified based on 2022 phenotypic data while the number of MTAs for the phenotypic data of 2023 is ten. Most of the MTAs are distributed on chromosome 2 while two MTAs are located on chromosome 7 and chromosome 9 and one MTA is located on chromosome 8 and chromosome 10.

Chr	Pos (Mb)	SNP	Year	Start (Mb)	End (Mb)	QTLs	Distance (Mb)
Chr2	22,74	CM000781.4-22741758	2023	25,03	28,35	qDTA2-1	2,29
Chr2	39,19	CM000781,4-39188052	2023	28,35	41,64	qASI2-2	0
Chr2	42,64	CM000781.4-42636503	2022	28,35	41,64	qASI2-2	1
Chr2	58,1	CM000781.4-58101478	2022	52,85	56,42	qASI2-4	1,68
Chr2	80,89	CM000781.4-80891600	2023	63,71	71,58	qDTA2-3	9,31
Chr2	80,89	CM000781.4-80891600	2023	63,71	71,58	qDTA2-3	9,31
Chr2	124,86	CM000781.4-124855331	2022	116,36	147,89	qASI2-1	0
Chr2	147,62	CM000781,4-147615007	2023	116,36	147,89	qASI2-1	0
Chr2	147,62	CM000781,4-147615007	2023	137,33	159,21	qASI2-1	0
Chr7	3,35	CM007647.1-3346788	2023	9,81	14,39	qDTA7-1	6,46
Chr7	14,16	CM007647.1-14159642	2023	9,81	14,39	qDTA7-1	0
Chr8	175	CM007648.1-175001475	2022	174,24	176,64	qASI8-3	0
Chr9	120,58	CM007649,1-120577810	2023	130,36	139,85	qASI9	9,79
Chr9	141,65	CM007649,1-141653826	2023	130,36	139,85	qASI9	1,81
Chr10	138,42	CM007650.1-138415119	2022	137,29	140,25	qDTS10	0

Table 5. Significant MTAs found inside the QTL region or close to QTLs

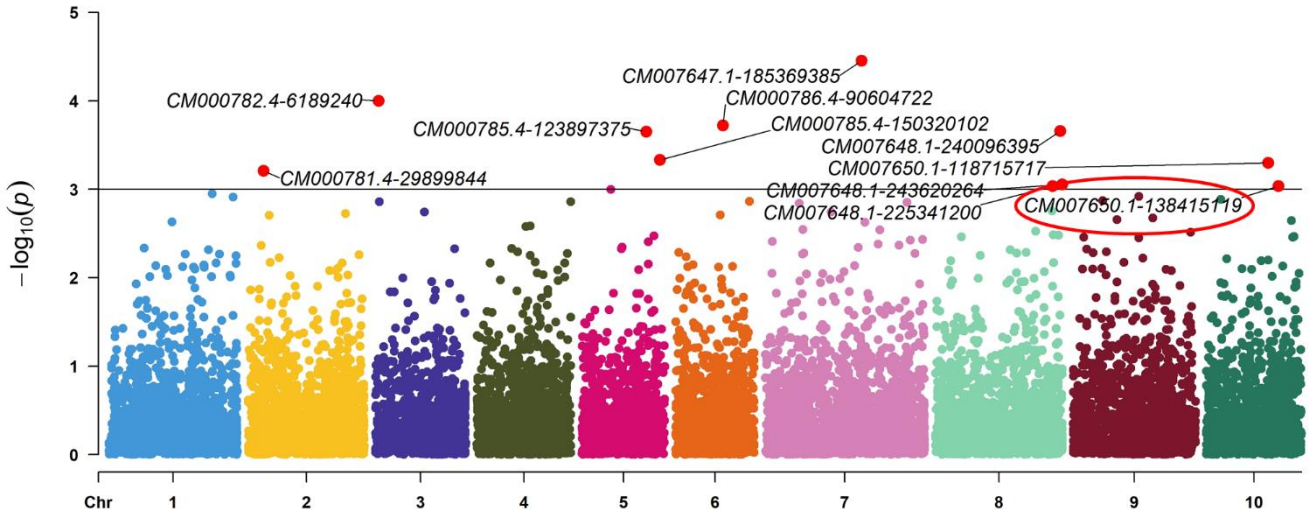


Figure 8. Manhattan plots for female flowering time in 2022

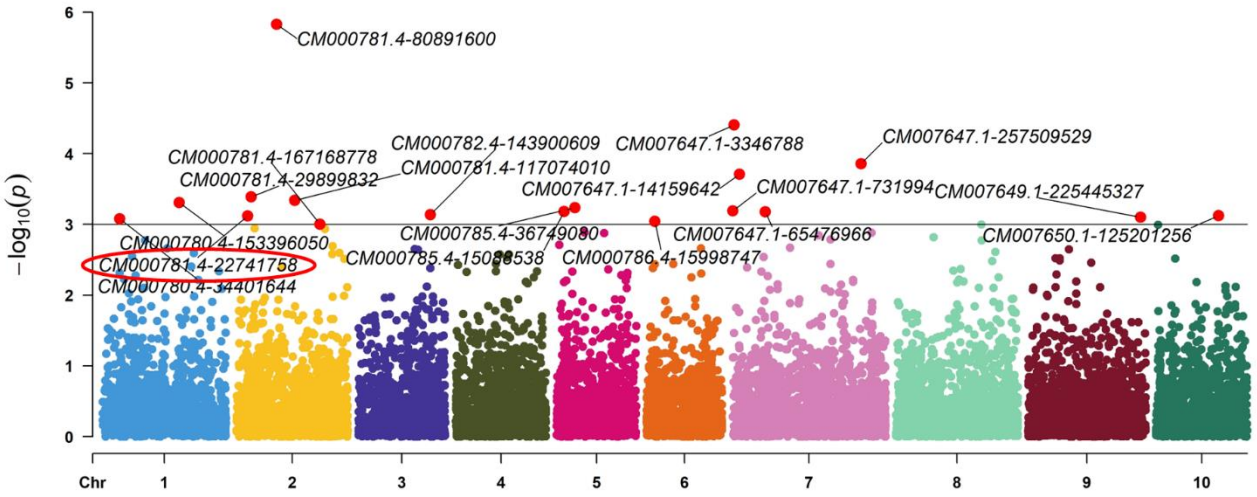


Figure 9. Manhattan plots for male flowering time in 2023

Discussion

The panel included in the present study represents a core collection of inbred lines derived from the main traditional Italian maize landraces cultivated in Italy in the post-war period, belonging to “Nostrano dell’Isola, Isola Basso, Scagliolo, Scagliolino, Marano, Sacra Famiglia”. The analysis of the population structure, carried out through different clustering methods, showed stable grouping statistics for four groups, which could be mainly referred to ‘*Insubria’*, ‘*Microsperma’*, ‘*Scagliolino’*, and a fourth one with prevalence of elite lines derived from Italian and U.S. breeding programs (Mastrangelo et al., 2024). For the understanding of the adaptation to drought stress, the phenotypic traits correlated to flowering time are key traits (Araus et al., 2012). The heritability of flowering time calculated on the Italian inbred lines belonging to this panel is quite high for both traits suggesting that are mainly controlled by genetic effects. In contrast, ASI showed a moderate heritability, around 40-45% for both year, which confirms the usefulness of this trait (Sheoran, 2022). Looking at the correlation plot, it is notable the negative correlation between ASI and juvenile vigor (-0,27) which might be linked to the greater resistance of the lines with high juvenile vigor to drought stress. Considering that juvenile vigor data was recorded using a score and the presence of missing data, it was necessary the normalization of all the phenotyping data before the GWAS analysis. The GWAS analysis was performed considering different model of the GAPIT tool, in particular the model considered were Blink, MLM and MLMM. A comparison between the three model considered highlighted that MLM and MLMM share the location of the significant MTAs found while Blink model was excluded from the successive analysis because of the limited robustness. The number of marker-trait associations (MTAs) identified for each phenotype with a *P* value thresholds < 0,0001 is 102, 54 for 2022 and 48 for 2023. Considering the importance of flowering time traits for the research of QTLs connected to drought stress, the male and female flowering time and the ASI MTAs were the focus of the study. Fifteen MTAs were found associated to QTLs already know in literature, in particular for ASI and for male flowering time. An interesting MTA connected to a QTL known in literature regards the female flowering time and it was found by Wang et al. which mapped different QTLs related to flowering time traits-related. The study identified 72 QTLs in a RIL population across four environment. The female flowering related MTA found in the present study belong to a QC21, a QTL cluster located on chromosome 10 between 137.25 and 145.65 Mb and in particular it belongs to the QTL *qDTS10* located on chromosome 10 between 137.25 and 138.90 Mb. In the region identified as QC21, it was found the gene ZFL1 (GRMZM2G098813), a homolog of FLORICAULA/LEAFY which controls the inflorescence architecture and the flower

patterning in maize (Bomblies & Doebley, 2006). Another SNP located at 22,74 Mb on chromosome 2 overlaps with a male flowering time-related QTL, named qDTA2-1.

In conclusion, the core collection composed by the traditional Italian inbred lines can be considered a good resource to identify new loci of interest for specific traits as flowering time-related traits.

Section B - Association of genic copy number variations (CNVs) with tolerance to drought and heat stress in a European panel of maize inbred lines

Chapter 1: Introduction

The increasing capacity of sequencing technologies to detect genomic variations and the decreasing costs for these technologies are conducing into new information levels for genetic studies. Between all the genomic variants, structural variants are still less studied than the most frequent point mutations (SNPs), or the short INDELs.

The following paragraphs have the aim to explore the novel knowledges about the structural variants and their impact on the family of *Poaceae* where is located the genus *Zea* and the species *Zea mays*. *Zea mays* is a species highly characterized by the presence of large and small structural variants which shaped its domestication across the *Poaceae* family and its adaptation to the environment (Springer et al., 2009; Huang et al., 2021; Nagarajan et al., 2024).

The definition of SVs and their importance

Structural variations (SVs) are genome rearrangements between individuals of the same or distinct populations of a species. Generally, SVs can be classified as a deletion, an insertion, a copy number variation (CNV), an inversion or a translocation (Yuan et al.). The dimensions of structural variations can vary notably from large deletions and insertions called Presence-Absence Variants (PAV) and Copy Number Variations (CNV) which indicated gain or loss of DNA portions ranged from 50 bp to several Mb (Springer et al., 2009; Pös et al., 2021). Diverse studies of human diseases highlight the impact of structural variants, microscopically visible or submicroscopic: the main effects are highly penetrant syndromes which can be distinguished by Mendelian sporadic disorders and complex diseases. Mendelian sporadic disorders such as Smith–Magenis syndrome (SMS), Williams–Beuren syndrome (WBS) and Potocki–Lupski syndrome (PLS) are whole or partial chromosomal abnormalities e detectable by traditional karyotyping. Besides these well-known syndromes, SVs are also involved in complex multifactorial diseases such as Crohn's disease, rheumatoid arthritis, attention deficit hyperactivity disorder, and type 1 and type 2 diabetes. Studying loci affected by structural variants involved in complex diseases leaded also into new insights on Mendelian disease (Weischenfeldt et al., 2013). The study of these kind of variants is also conducted on livestock animals and confirmed the effect on phenotypes like the CNVs affecting the coat colour in pig or influencing the innate immunity gene families in cattle, pig and chicken (Bickhart & Liu, 2014). In plants the importance of structural variants is well-

known and regards multiple aspects of the plants phenotypes such as the regulation of complex agronomic traits (Massman et al., 2024; Francia et al., 2016), the tolerance or resistance to biotic and abiotic stresses (Schmidt et al., 2024; Animasaun et al., 2024). Furthermore, these genome arrangements can be helpful in increasing the knowledge about the different polyploidization events occurred during the domestication (Huang et al., 2021; Mamidi et al., 2020).

Methods and strategies to detect SVs

The analysis of structural variants can be performed starting from short reads or long reads sequencing data and multiple specific software. The main approach used for SVs detection with short reads consists in mapping discordance between a sample read and the reference genome. This approach includes the read-depth (RD) method, the split-read (SR) method and the paired-end (PE) method.

In the read-depth method, a random distribution in mapping depth is assumed and the divergence from this distribution in the sequenced sample is interpreted as a duplication if it shows higher read depth. If it shows a reduced read depth, it is interpreted as a deletion compared to the reference genome.

The split-read method leverages alignments that map over the breakpoint of a structural variant on the basis of a “split” sequence-read.

The last method used for paired-end short reads defines the structural variants basing on the orientation and distance of paired ends compared to the reference genome. Paired-end method can map deletions, insertions, translocations and inversions while read-depth can map effectively deletions and insertions. Between all the methods previously described a different approach is the detection of SVs using the assembly of contigs. The sequence assembly can be conducted considering the presence of a reference genome or using a de novo strategy. All of these methods cannot be considered completely accurate: the resolution for breakpoints in the read-depth method is low, for paired-end method resolving ambiguous mapping assignments in repetitive regions is challenging and the reliability of split-read method is good mainly in unique regions of the genome. Assembly approach is considered robust and versatile but the SVs detection is biased by the collapsing of duplications and repeats regions. The ability of alignment methods to discover SVs is also influenced by the coverage of the sequencing: high coverage can improve the sensibility and the specificity of the analysis (Alkan et al., 2011; Ho et al., 2020; Gabur et al., 2019).

The evolution of sequencing leaded to the introduction of long read technology in the detection of structural variants which can improve the mappability in ambiguous regions of the genome and the capacity to identify the SVs not already found by the short reads (Chawla et al., 2021; Jain et al., 2018). As for the short reads technology the main methods for the screening of SVs with long reads is based on the extraction of alignment information and the contigs generation from *de novo* or reference-guided assembly (Ahsan et al., 2023). Besides the multiple methods and approaches regarding the sequencing of genome, the SVs calling process is also considered challenging due to the continuous development of new algorithms and software.

The SVs callers can vary considering the type of reads available (long or short reads) and if the sequencing is conducted with a whole-genome or whole-exome approach. To better understand the SVs callers potentialities benchmarking studies are a good strategy and allow the researchers to the individuation of the best caller considering the sequencing features. Between the software included in benchmarking studies focused on whole-genome short reads sequencing, tools like Delly, Manta, Lumpy, GRIDSS and Wham were identified for their precision and accuracy. For long reads sequencing cuteSV, pbsv, SVIM, Sniff les and SVision can be cited as callers for read-based strategy, whereas Phased Assembly Variant (PAV) and SVIM-asm are cited as callers for assembly-based strategy (Kosugi et al., 2019; Sarwal et al., 2022; Lin et al., 2022).

Location of structural variants across chromosomes and genome

To better understand the significance of structural variants in Poaceae, it's important to consider their location across chromosomes and genome. Darracq et al., 2018 sequenced the genome of the French maize inbred line F2, which involved a *de novo* assembly. This assembly revealed the presence of 10,044 F2-specific Presence-Absence Variations (PAVs) not found in the maize reference line B73. A subset of 1028 F2-specific PAVs and 691 B73-specific PAVs were analyzed to examine their distribution along the chromosome. The results show that PAVs are more abundant at the chromosome tips of the F2 line, when considering all chromosomes. The distribution of B73-specific PAVs is similar, with high densities at the chromosome tips and low densities in the pericentromeric regions. This finding is supported by the work of Hirsch et al., 2016, who highlighted genomic variation in the maize line PH207. The variation in density can be attributed to the differing recombination rates between pericentromeric regions and distal regions. Chromosome tips exhibit a higher recombination rate compared to pericentromeric regions. A similar trend is observed for deletions found in 270 maize

inbred lines analysed by Zhang et al., 2013. De Oliveira et al., 2020 discovered a similar trend in wheat while examining various accessions of different wheat species. The researchers concentrated their screening of structural variations (SVs) on chromosome 3B, which revealed a high density of SVs (38%) in the distal regions of chromosome 3B, in contrast to the proximal regions, where the density was significantly lower (8%). The recombination phenomenon offers a plausible explanation for the elevated frequency of structural variants observed in the Poaceae family. In contrast, the absence of recombination events led to a uniform distribution of variations across the chromosomes as exemplified by the CNVs screening of 4H chromosome of barley (Muñoz-Amatriaín et al., 2013). Furthermore, the other chromosomes of barley analyzed in the Muñoz-Amatriaín et al., 2013 study, corroborate the increased frequency of variants in distal regions.

Structural variants in plant genomes are mainly located in intergenic regions, making it challenging to define their role in gene expression. This finding is supported by the analysis of Copy Number Variations (CNVs) and Presence-Absence Variations (PAVs) conducted by Springer et al. The study conducted by Springer et al., which involved gene annotation of comparative genomic hybridization (CGH) probes on maize lines B73 and Mo17, identified a higher number of probes annotated in intergenic regions. However, structural variants can also affect genic regions, consequently influencing gene expression. The effect of structural variants on gene expression is further confirmed by the percentages of structural variants (SVs) found in millet that overlap genic and flanking regions. According to Yan et al., 2023, the presence of SVs in genic or flanking regions in millet ranges between 37 and 44%. Broomcorn millet, which is characterized by two subgenomes (subgenome A and subgenome B), exhibits a higher abundance of PAVs in downstream or upstream regions (2 kb), while genic PAVs demonstrate an unbalanced proportion between subgenome A and subgenome B, as highlighted by Chen et al., 2023.

The significance of SVs related to the domestication of Poaceae

Poaceae family has been duplicated multiple times during its domestication process. The specific duplication that occurred was the whole-genome rho duplication. This event leaded to subsequent duplication among the different subfamilies *Pooideae*, *Panicoideae* and *Oryzoideae*. Whole-genome duplication, chromosome inversions, and gene losses resulting from fractionation provide a solid foundation for the generation of structural variants, particularly copy-number variations (T. Zhang et al., 2024). The genes generated by the rho duplication include those related to environment adaptation

and stress response. For instance, the gene COLD1 in rice is an example of a rho-duplication-derived gene (Ma et al., 2015;L. Zhang et al., 2022). Structural variants are useful in elucidating the domestication process of Poaceae. An example is the structure of Rubisco activase (RCA) gene which is highly conserved, and it derives from a common ancestor in the grass species. It is formed by six exons and six introns and it is present in double copies Rca1 e Rca2. (Nagarajan et al., 2024) results suggest that the tandem duplication of the gene has happened after the creation of the Poales subfamily, and it is specific of Poaceae. For example, the locus HvTB1 is responsible of the high tillering in barley. Two allelic variants can be found, one form connected to six-rowed barley and the other form connected to two-rowed barley. Analyzing the six-rowed barleys belonging to the barley pangenome, diverse copies of HvTB1 were found in a segment of 21 kb (Jayakodi et al., 2024). This locus belongs to transcription factors called TCP (*Teosinte-branched 1 Cycloidea Proliferating*) that induce modifies in the plant development, in stress response and in hormonal pathways. It has been demonstrated that the rapid expansion of TPC transcription factors in land plant is primarily caused by the whole-genome duplications, events happened during the evolution of monocots like the plants in the *Poaceae* family (Ren et al., 2023). Evidence of this phenomenon is represented by rye (Secale cereale) in which has been found an high level of TPC proteins, higher than *Oryza sativa* and *Hordeum vulgare*, and the presence of a pair of fragment duplications in TPC genes. This evidence suggests that the duplication has been the evolutionary mechanism to increase the number of TPC genes and increase the adaptability to harsh environment in rye (Ren et al., 2023). Structural variants played a role also in the control of vernalization in barley modifying the gene responsible of this process and leading to the creation of spring cultivars which are defined for the lack of vernalization necessity during the crop growth. The sensitivity vernalization model considered by the scientific community in barley is formed by the gene HvBM5A located in VHR-H1 locus which is repressed by VHR-H2 leading to vernalization sensitivity. Massman et al. discovered that an intronic deletion in the gene HvBM5A in the spring cultivar Woody-1 is the basis for the loss of vernalization sensitivity, event present also in the reference cultivar Morex which is also a spring barley. Besides the whole-genome rho duplication, the BOP clade (*Bambusoideae*, *Oryzoideae*, and *Pooideae*) occurs in other more recent multiple polyploidy or whole-genome duplication (WGD) events. The BOP clade is composed by important genera like *Oryza*, *Avena*, *Hordeum* and *Triticum*. The work of Liu et al., 2023 focused on rice, *Brachipodium* and *Avena* spp. to characterize the genome expansion in BOP. *Avena* genome has occurred in more expansion than *Brachipodium* and rice and it can be mainly explained by the presence of the annotated repetitive elements. An analysis of different diploid Avena species (*A.*

longiglumis, *A. atlantica*, *A. strigose* and *A. eriantha*) reveals the presence of translocation on the chromosomic terminal regions. The hexaploid genome of *Avena sativa* supports the hypothesis that the genome underwent distal translocation events (from 10% to 38% of chromosome arms) that caused variation in the chromosome structure probably leading to a reproductive isolation.

Variations in copy number play a role in the creation of new phenotypes, which can subsequently influence the direction of domestication as demonstrated by the findings of Xue et al., 2019. A cross between Teosinte and the maize inbred line W22 produced plants that exhibited sick phenotypes. Xue et al., 2019 investigated why backcrossing these plants failed to restore a normal phenotype. Upon analyzing the read coverage, it becomes evident that certain regions on chromosome 9 display increased read depth. These regions overlap with genes annotated in W22, and the expression data reveal higher copy numbers in the sick plants compared to the normal ones. Seed shattering (SHT) is an important trait that leaded the domestication of *Poaceae*. It is a well-known trait subject to strong selection pressure in *Poaceae* as it happened in broomcorn millet and in foxtail millet. Chen et al., 2023 examined over 1000 presence-absence variations (PAVs) under 225 regions under selection in broomcorn millet. The study focused on key traits connected to the domestication of broomcorn millet like the seed shattering trait which is controlled by multiple genes. Chen et al., 2023 identified different deletions and one PAVs that affect the seed losses in millet. Specifically, *longmi058828* is truncated by a PAV that advantages seed shattering. The mutation of *longmi058828* appears more frequently in wild population. Regarding foxtail millet (*Setaria italica*) the study of H. Liu et al., 2022 considers the sh1 gene responsible for the loss of seed shattering in species like sorghum, maize, Asian and African rice. Inside the second exon of sh1, it is present a transposon that was inserted during the domestication process of foxtail millet. The gain of this transposon was also used as an indicator to classify the domesticated millets versus the wild millets. For example, the green millet cultivar E28 previously considered wild is re-classified as a domesticated cultivar for the presence of the transposon and the consequently loss of seed shattering (H. Liu et al., 2022). Also, during the domestication of *Sorghum*, the seed shattering trait was lost. This loss is attributed to a 2.2 kilobase deletion in the Shattering1 gene. Tao et al. analyzed a panel consisting of wild sorghum and improved inbred lines for the presence of PAVs in *Sorghum*. A pairwise comparison between these two accession types reveals that approximately 92 out of 120 genes underwent gain or loss during the domestication process in wild accessions, whereas only 15 out of 20 genes underwent gain or loss during the improvement process. Instead, Li et al., 2019 focused their work on understanding the genetics

mechanism behind the seed shattering in rice. The rice cultivar “Oonari” derives from a mutant of the easy shattering cultivar “Takanari” and is characterized by a moderate seed shattering. The analysis of the progenies derived from “Oonari” cultivar revealed a high presence of structural variations and indels. The study focused on a duplication occurred in the same region where is located TO20 and TO21, two SNPs connected to the seed shattering trait. The duplication regards the mini-RNA defined osa-miR172d which then can be considered a gene involved in the seed shattering domestication process in rice.

The ratio between structural variants in wild species and in domesticated species can vary significantly. Huang et al., 2021 identified a large presence-absence variation (PAV) of 3.2 megabases that affects *Zea* species and originated from *Tripsacum*. *Tripsacum* is an ancient grass species notable for its high adaptability to stressful environmental conditions. During the domestication of the *Zea* genus, a common ancestor allowed the migration of the large PAV from *Tripsacum* into the *Zea* genome. The PAV contains 70 actively expressed genes, which are associated with traits related to environmental adaptation, such as heading date and ear height. PAV are most prevalent structural variant in wild and cultivated oats. Specifically, cultivated oats show inversions on 4D chromosome while these inversions are absent in the ancestral oat *Avena sterilis*. The genes affected by PAVs are correlated with agronomical traits connected to the domestication process as SHORT VEGETATIVE PHASE (SVP) gene affected by a 132 bp insertion in the D subgenome of three cultivated oats. A higher expression of SVP gene is responsible of a negative regulation of spikelet number in wild oats compared to cultivated ones (He et al., 2024).

The origin of SVs and their impact on Poaceae phenotypic traits

The origin of structural variants resides in the mechanisms of reparation of DNA. DNA faces different damages during its replication and the Double Strand Breaks (DSBs) have endogenous causes like reactive oxygen species, improperly repaired single nucleotide lesions, unrepaired single strand breaks and exogenous causes like chemical mutagens (Currall et al). The most prevalent events that generate SVs are non-allelic homologous recombination (NAHR), DNA break repair errors (NHEJ) or replication errors like, for example, fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR). NHEJ is the most present mechanism of repair in plants and the effect of NHEJ on plants genome is related to the formation of insertions and/or deletion (Sall et al., 2024; Zheng et al., 2021). NAHR is well-known for being the main mechanism of generation of copy number variations as result of misalignment in

genomic regions that contain highly identical sequences such as repetitive DNA (Kuo et al., 2024; Zmienko et al., 2016).

Copy number variations (CNVs) are among the most prevalent structural variants influencing phenotypic traits in the *Poaceae* family. Numerous examples of CNVs pertain to pigment expression in *Poaceae* grains. For instance, carotenoid content can be affected by copy variations, as observed in the maize line A188. This line is characterized by white seed color, in contrast to the yellow seeds of the maize reference line B73. The investigation conducted by G. Lin et al., 2021 identified CNVs in A188, revealing 13 copies of the gene ccd1. High expression levels of ccd1 are presumably associated with carotenoid degradation. Genic CNVs can also result in ectopic expression patterns, as seen in the rice cultivar N22. This cultivar has a duplication of the Kala4 gene, and interestingly, one border of the duplicated DNA fragment (containing the entire Kala4 gene) is close to the start codon of Kala4. This sequence modification results in a distinct promoter sequence for the derived copy from the parent Kala4 copy. The N22 cultivar is characterized by black spikelets at the top, with Kala4 being activated in anthocyanin development, leading to the "black pericarp" phenotype. The hypothesis posited is that the ectopic expression derived from the duplication of Kala4 induces black spikelets only at the top of N22 (Qin et al., 2021). Similarly, presence/absence variations (PAVs) also modify grain development, as demonstrated by Tao et al. in sorghum. A PAV of 3,216 bp was identified in the sequence of Yellow seed1, a MYB gene controlling phlobaphene accumulation in the pericarp. Copy number variations can be highly complex mechanisms, exemplified by the process of carotenoid esterification in wheat. The XAT-7D gene is responsible for esterifying carotenoids in common wheat. However, Rodríguez-Suárez et al., 2023 discovered an orthologous gene, XAT-7A1, located on chromosome 7A in durum wheat. The XAT-7A1 gene exhibits varying copy numbers in the analysed landraces, ranging from 2.07 to 8.68. Among these Spanish landraces, Rodríguez-Suárez et al., 2023 identified four genotypes associated with carotenoid production. The first and second genotypes produce diesters and monoesters, the third produces only monoesters, and the fourth does not produce esters. Consequently, the genotypes were sequenced and classified as Type 1, 2, or 3 according to their association with a particular genotype. Notably, XAT-7A1 Type 1 is of specific interest due to its relevance to the production of carotenoid diesters and monoesters and its high expression levels during grain development. These findings support the hypothesis that XAT-7A1 is a multicopy locus with linked copies that affect the genotype containing XAT-7A1 Type 1, leading to the production of carotenoid esters. The determination of red glume colour in wheat is an interesting example of how the presence of a specific gene

copy can modify expression. The study of the R2R3 MYB gene demonstrated the presence of five duplicated genes with distinct structures, but only one (G57) was present in every red glume accession (Mao et al., 2024).

Certain copy number variations (CNVs) are implicated in modifying plant development. In *Zea mays*, a mutant for inflorescence architecture contains the ZmWUS1 gene in a large tandem duplication. The ZmWUS1 gene plays a role in the development of shoot apical meristems. An analysis of the expression of this mutant, containing the tandem duplication, revealed alterations in vegetative development, as demonstrated by Z. Chen et al., 2021. Modifications of plant development induced by CNVs affect not only maize but also other Poaceae species. The polyploidy of wheat facilitates the presence of copy number variations. A pertinent example is the control of awn length governed by copies of the B1 gene. A comparison between the awnless accession SY20 and the awned accession 7D12 revealed that SY20 possessed five copies of B1, whereas 7D12 had only one copy. The inhibition of awn formation in SY20 correlates with a high level of B1 promoter expression. This result is further confirmed by the presence of B1 promoters in the same regions for 77 other awnless accessions (J. Li et al., 2023). The B1 gene is also affected by another interesting CNV in durum wheat. This CNV, located on chromosome B2, influences the heading time in durum wheat lines. Hexaploid lines harbouring a higher copy number have an earlier heading time and are more prevalent in regions with frequent heat and drought events (Würschum et al., 2019) . Kirby T. Nilsen, 2020 provides a good example of modification of another trait connected to plant development, stem solidity. This trait is crucial for wheat growth as solid stems provide resistance to a pathogen called wheat stem sawfly. The researchers identified the gene SSt1 located on chromosome arm 3BL and the transcription factor involved in the development of the solid stem genotype. The transcription factor named TdDof shows copy number variation in the hollow-stemmed mutant line “pithless1”. Specifically, “pithless1” presents a deletion of 673.9 kb in the SSt1 region, while the solid-stem lines CDC Fortitude and W9262-260D3 carry additional copies of TdDof. The line CDC Fortitude has three copies of TdDof, and TdDof2 and TdDof3 show a unique insertion compared to hollow stem lines. The transcription factor TdDof encodes a Dof zinc-finger protein with a highly conserved domain in vascular plants. A higher copy number of TdDof positively correlates with its expression during early stem elongation, inducing a solid stem phenotype. Another example of a complex mechanism connected to CNVs is the OsMTD1-located CNV. The study by Q. Liu et al., 2021 focused on a gene already reported to control tillering in rice. This gene, cloned from *Oryza sativa* spp. *Japonica* cultivar Nipponbare, is located on chromosome 8. An analysis revealed a tandem

duplication in the Nipponbare cultivar that is absent in subspecies indica. In the OsMTD1-located CNV, a mi-RNA named osa-miR156f was also identified, which differs between *indica* cultivars and is associated with rice tillering. A screening of CNVs among 190 rice cultivars highlighted that some cultivars have only one copy, while others have two. The difference in copies is associated with a lower level of tillering for the duplication of OsMTD1 and a higher level of tillering for genotypes carrying only one copy. The osa-miR156f and the OsMTD1 gene play opposite roles in controlling tillering: overexpression of osa-miR156f induces more tillers, whereas OsMTD1 overexpression also induces more tillers. Thus, the control of this phenotypic trait is primarily modulated by the presence of the OsMTD1 gene and its duplication, and secondarily by the expression levels of the two genic elements within the OsMTD1-located CNV. Q. Liu et al., 2021 hypothesized that OsMTD1 acts as a transcription regulator suppressing osa-miR156f and consequently lowering the number of tillers.

In *Poaceae*, the connection of structural variants (SVs) with transposable elements (TEs) activity in upstream genomic regions, where promoters and regulatory motifs are located, indicates their effect on abiotic and biotic stress-related mechanisms (C. D. Hirsch & Springer, 2017). These variants can modify the expression of genes involved in the mechanisms of tolerance or resistance to stresses. Yan et al. studied a family of transcription factors called RWP-RK, which is abundant in pearl millet and enhances the plant's tolerance to heat stress. Pearl millet is well-known for its high resistance to heat, and the work focused on genes enriched in stress-related pathways. The data showed an expansion of RWP-RK genes associated with the presence of long terminal repeat (LTR) TEs. The role of these transcription factors is confirmed by RNA-seq data, which also highlighted another type of gene involved in heat stress tolerance: endoplasmic reticulum (ER)-related genes. ER-related genes are affected by numerous structural variations, and particularly, one SV is associated with HSP70, an ER-related gene which, when upregulated, plays a role in the degradation of misfolded proteins as a response to heat stress. The co-regulation guided by ER-related genes and RWP-RK genes modulates expression to improve the response to heat stress. SVs in pearl millet regulate ER-related genes, supported by the presence of SVs correlating with different expression levels of ER-related genes between heat-resistant and susceptible pearl millet accessions. Another example of pearl millet's stress tolerance is the work of Animasaun et al., 2024, which highlights the role of DREB1A, a member of the DREB subfamily A-1 of the ERF/AP2 transcription factor family. These transcription factors are known to interact with the Dehydration-Responsive Element (DRE/CRT) and are involved in stress regulation in Arabidopsis. The study focused on drought stress and compared drought-

tolerant (NGB00886) and susceptible (NGB00885) millet accessions, revealing a higher copy number of the DREB1A gene in the most tolerant accession. The upregulated DREB1A may enhance the osmotic adaptation ability of the tolerant millet accession NGB00886, modulating drought and water-stress resistance. This mechanism might also be co-regulated by the presence of the P5CS gene, involved in proline accumulation, for which a lower copy number correlates with higher drought tolerance, as supported by high expression of this gene in the stressed accession NGB00885. Heat stress also impacts photosynthesis and its enzymes. Rubisco activase tends to lose its activity above 35°C, affecting photosynthesis efficiency. The connection with heat stress is strong, and heat stress induces different expressions in rice, maize, and wheat. Nagarajan et al., 2024 focused on Transposon Elements (TE) called Heat Shock Elements (HSE). These elements can modify expression, increasing tolerance to high temperatures of RCA. The integration of heat-responsive expression may have occurred in the Rca1 promoter of the grass ancestor. Later, species-specific HSE have been integrated into the promoter region, altering expression. Heat shock elements have been identified in *Oryza brachyantha*, rice, wheat, maize, and sorghum. In each promoter sequence of the cited species, insertion events of TEs suggest that TEs lead to different expression patterns for heat stress in the Rca1 gene.

Pathogens are not excluded from the effects of structural variations. Among the *Barley yellow dwarf* (BYDV) *virus* species, BYD-PAV is the most widespread worldwide. The need to study BYD-PAV-resistant maize inbred lines led to the creation of a bi-parental mapping population to understand the putative genes and their structural variations. The results found by Schmidt et al., 2024 showed the presence of two putative genes: Zm00001eb428010 and Zm00001eb428020, respectively a candidate gene for BYDV-PAV resistance already identified, and a P-loop containing nucleoside triphosphate hydrolases superfamily protein. The analysis of structural variants in the genic regions showed 9 SVs, including a 54 bp deletion in the 5`-UTR, a 91 bp insertion in intron 6, and a 362 bp deletion in intron 7 in gene Zm00001eb428010. These SVs are shared by three different inbred lines that are resistant to BYDV-PAV. The presence of these three SVs may influence post-transcriptional regulatory mechanisms and modify the abundance and/or properties of the proteins. In barley, the class of R-proteins is involved in pathogen recognition and signaling initiation. Muñoz-Amatriaín et al., 2013 discovered that R-genes are affected by CNVs, especially near the ends of the 1H and 7H chromosome short arms, supporting the influence of SVs on proteins involved in disease resistance mechanisms. Other examples of disease resistance are reported by Song et al., 2021, highlighting the presence of two large SVs: an expansion region of 30.75–

31.57 Mb and an insertion region of 31.90–32.76 Mb in the rice variety Minghui 63. These SVs may contribute to its marked resistance to blast and bacterial blight.

Transposon elements and their connection with SVs

The variation in Poaceae genomes is also driven by the presence of transposon elements. Indeed, transposable elements can induce Double Strand Breaks (DSBs) with their excision or their reintegration playing a role in Non-Homologous End Joining (NHEJ) repair and serving as template for Non-Allelic Homologous Recombination (NAHR) (Currall et al., 2013). One of the ways they can generate structural variations regard the Reversed Ends Transposition (RET), one of the two types of the Alternative Transposition. This mechanism involves Class II of TEs and regards the reversely-oriented 5' and 3' termini of two elements located nearby each other on the same chromatid. In rice the most present structural variations are dominated by the DNA TEs and LTRs insertions (55 % and 38.3 % respectively). Interestingly the mechanism of formation of the SVs are NHEJ and NAHR which are known as non-TEI mechanisms, but a high percentage of the SVs formed with these mechanisms have one breakpoint overlapped a TE. This information found by the work of Quin et al. highlights the connection between structural variations and transposable elements; the latest can induce DNA breaks for NHEJ and homologous sequences for NAHR. The importance of TEs related to SVs in rice is also confirmed by Fuentes et al., 2019: the sequences that are affected by the TEs activity have also a high abundance of SVs. Other species belonging to *Poaceae* are highly characterized by transposons. Some examples are maize and wheat which have more than 70% of transposable elements. The effect of TEs on maize finds confirmation in the work of J. Zhang et al., 2013, with the identification of TEs-originated tandem duplication ranging in size from 8157 bp to ~ 5.3 Mbp. Another study on maize TEs conducted by Munasinghe et al., 2023 has investigated the role of TEs in relation with the structural variants. The analysis of the 26 NAM inbred founder lines has led to the concept of polymorphic TEs which are located in large structural variations. The work focused on structural variations bigger than 50 kb which are more than 2.4 million. A classification of polymorphic SVs was created based on the overlap between TEs and SVs. Interestingly 23% of the SVs represent more than 95% of the TEs ("SV = TE" events) while only 13% of SVs found in B73 do not overlap TEs. This trend is also confirmed by the 87,43 % of TEs-overlapped PAVs in oat (He et al., 2024). The different observations about the polymorphic TEs suggest that the maize genome is facing a contraction and is trying to mitigate the spread of TEs because the polymorphic TEs results as deletions rather than insertions. Rye during the domestication process underwent into an expansion process which resulted in the formation of transposed

duplicated genes (TrDGs) which interestingly involved also the starch biosynthesis-related genes (SBRGs) (G. Li et al., 2021). Also in wild oat, PAVs derived from LTR-RTs largely contributed to plant growth and environmental adaptation, such as terpene synthase activity, signal transduction and regulation of gene expression (He et al., 2024).

Chapter 2: A bioinformatic simulation for the detection of CNVs

The detection of copy number variations (CNVs) is challenging due to technical limitations such as coverage needed and sequencing approach. Sequencing approaches for the CNVs detection include genome skimming as ddRAD sequencing or RAD sequencing, exome sequencing and whole-genome sequencing. Considering the moderate dimensions of maize genome, which is around 2,3 Gb, the cost of sequencing might be onerous, in particular if the study conducted includes a high number of genotypes. To better understand the limitations and advantages of the different sequencing approaches, and to identify the best approach, a bioinformatic simulation was performed on three genomes of inbred lines already sequenced and publicly available.

Materials and Methods

The samples included in the simulation derive from the work of Wang et al., 2020 in which Chinese and American maize lines were sequenced with Illumina X-ten sequencer to obtain paired-end short reads. The sequencing approach was whole-genome sequencing (WGS). The three samples used for the simulation are reported in the table below (Table 1) which reports also the Sequence Read Archive (SRA) number for the download of the sequences from NCBI.

Line	**NCBI SRA**
Mt42	SRR11301724
PHV53	SRR11302177
PHP85	SRR11302179

Table 1. Samples used for the present simulation correlated with their SRA number from NCBI

The software ddRadSeqTools (https://github.com/GGFHF/ddRADseqTools) was considered to design *in silico* and to test double digest RADseq (ddRADseq) experiments. After a literature survey, the restriction enzymes was chosen for the ddRADseq digestion from the work of Yang et al., 2016. The enzymes chosen are AvaII and MspI.

To obtain exome sequences from the whole-genome samples, the WGS sequences were intersected to the annotated file of B73, the reference genome for maize, using the tool *bedtools* v2.31.1 (Quinlan & Hall, 2010) with the option "*intersect*".

The calling of CNVs for WGS and ddRAD sequences were performed using three different software specifically built for the screening of structural variants, called Manta (https://github.com/Illumina/manta) (X. Chen et al., 2016), Delly (https://github.com/dellytools/delly) (Rausch et al., 2012) and Lumpy (https://github.com/arq5x/lumpy-sv) (Layer et al., 2014). Instead, the exome sequences was analysed with a specific tool for exome sequencing approach called DeCoN (https://github.com/RahmanTeam/DECoN).

To calculate the overlap between the WGS samples and the ddRAD and WES samples, the tools *bedtools* v2.31.1 (Quinlan & Hall, 2010) with the option "*intersect*" was considered.

The variants found by the different tools included were annotated in three categories defined as genic elements, gene proximity regions (2-3 kb) and intergenic regions, using the R package *ChIPseeker* (https://www.bioconductor.org/packages/devel/bioc/vignettes/ChIPseeker/inst/doc/ChIPseeker.html) (Yu et al., 2015) appropriately settled for maize sequences.

Results

In Table 1 is reported the coverage value obtained for every sequences for each approach. As expected, the coverage value is uniformly distributed and varies between a maximum of 20X to a minimum of 15X.

Sequencing approach	SRR11301724	SRR11302177	SRR11302179
ddRad	19.11	16.32	14.81
WES	20.08	17.00	15.92
WGS	20.68	17.93	16.73

Table 1. Coverage value for each sample and for each sequencing approach

The number of variants detected in the different samples for each sequencing approach are reported in Table 2. The results highlight that WES variants represent the 4% of the WGS variants while ddRAD variants represent the 0,04% of WGS variants. The overlap between the WGS variants, WES variants and ddRAD variants was calculated and the

match is quite complete considering ddRAD approach, while for WES the overlap varies from 57 to 64 %, as shown in Table 3.

Sample	WGS	WES	ddRADSeq
SRR11301724	45338	1867	23
SRR11302177	41162	1775	18
SRR11302179	25768	2043	21

Table 2. Number of variants for each sample for each sequencing approach

Sample	WES	ddRADSeq
SRR11301724	1079 (58%)	22 (95,6%)
SRR11302177	1145 (64%)	18 (100%)
SRR11302179	1166 (57%)	21 (100%)

Table 3. Number of WES and ddRAD variants overlapping WGS variants

The variants of each sample for each sequencing approach were annotated considering three categories. The first category is called genic elements and includes exons, promoters less distant than 2 kb, UTRs and downstream regions (<300 bp). The second category is called genic proximity regions and includes promoter located 2-3 kb distant from genic elements. The last category is called intergenic regions and includes introns and distal intergenic regions. The percentages of each category are reported in the pie charts reported below (Figure 1, Figure 2 and Figure 3). Whole-Genome Samples (WGS) report all the annotation categories, mainly intergenic regions, followed by genic elements and a low percentages of gene proximity regions. Instead, Whole-Exome Sequencing (WES) samples report only genic elements and gene proximity regions as expected. Proximity regions are absent in ddRAD samples.

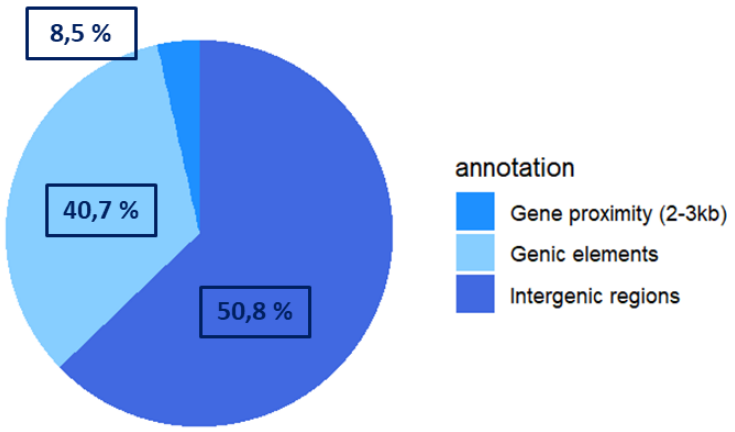


Figure 1. Annotation of CNVs for WGS samples

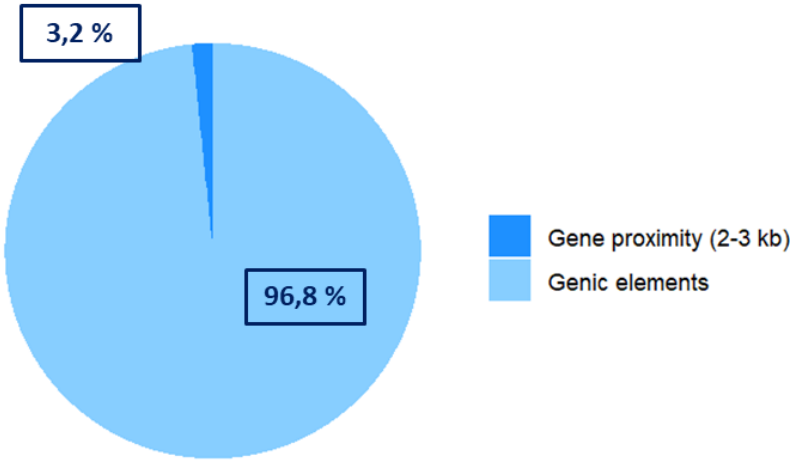


Figure 2. Annotation of CNVs for WES samples

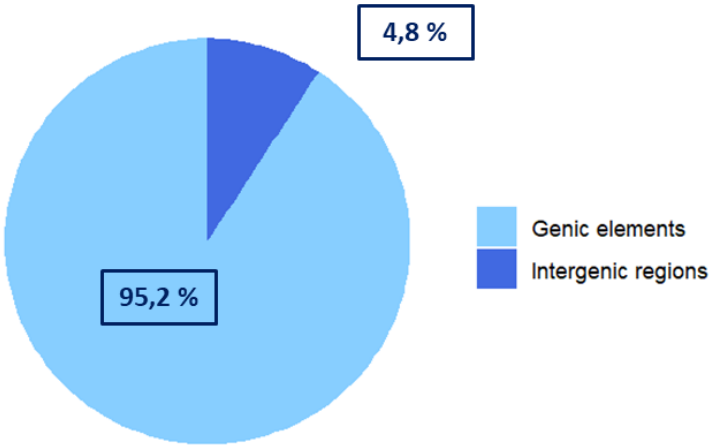


Figure 3. Annotation of CNVs for ddRAD samples

Conclusion

The bioinformatic simulation conducted on three different maize genomes of inbred lines to compare the efficacy of the whole-genome, whole-exome and ddRAD sequencing approaches highlights high variability between such approaches for number of variants found and for their hits / annotation categories. In particular, variant numbers of the samples sequenced at the same coverage show a low detection for WES and in particular for ddRADseq. Moreover, ddRAD sequencing excludes an important category for the study of CNVs like the gene proximity regions. Whole exome approach appears more accurate for the detection of CNVs, although only a whole-genome approach can be considered accurate and it can extensively detect variants in each annotation category. In conclusion, the present results suggest the whole-genome approach as the best method for sequencing maize samples in order to detect CNVs.

Chapter 3: The detection of genic CNVs on inbred lines DROPS panel and the association to traits connected to drought and heat stress

In order to detect the genic copy number variations and to associate the genic CNVs to traits connected to drought and heat stress, a maize inbred lines panel named DROPS was identified. The panel is composed by more than 240 European lines which was crossed with a flint line and phenotyped in multiple years and locations for the main traits, including important traits correlated to drought and heat stress as male and female flowering time and anthesis-silking interval (ASI). The panel was also tested in open field in 2023 at CREA (Centro di Ricerca Cerealicoltura e Colture Industriali) in Bergamo and flowering time related traits were recorded. To achieve the detection of genic CNVs, the panel was sequenced with Illumina short reads and the single nucleotide polymorphism (SNPs) were identified. Moreover, a validation on five public lines belonging to NAM maize panel was carried out to test the CNVs caller Hecaton.

Methods & Materials

The inbred lines DROPS panel

The panel, referred to as DROPS panel, contains different maize inbred lines. It includes some lines from a wider panel of them from Europe and America, and some additional lines derived from public breeding programs in Hungary, Italy and Spain, plus recent lines free of patent from the USA (Negro et al., 2019). Considering the 242 lines belonging to the DROPS panel considered in the present study, 92 have WGS data publicly available on NCBI and ENA databases, while the remaining 160 lines had to be newly sequenced. The sequencing project included 12 lines with publicly available genome, resequenced for their low coverage, and the tester line UH007 used for the DROPS phenotyping. Furthermore, in order to compare the sequencing performance of the different study which sequenced the 92 lines, 2 lines for each study, for a total of 12, were resequenced. In the table below are summarized the lines involved in the present study, their origin and the eventually public database where the sequences are stored (Table 1). The statistical description on DROPS lines was performed in R with *ggplot2*.

Variety	Origin	Public Database	Paper
11430	USA	None	None
A3	France	None	None
A310	France	None	None
A340	France	NCBI	Qiu et al. 2021

A347	France	None	None
A374	France	NCBI	Qiu et al. 2021
A375	France	NCBI	Wang et al. 2010
A554	France	NCBI	Bukowski et al. 2018
A654	France	NCBI	Bukowski et al. 2018
AS5707	USA	NCBI	Qiu et al. 2021
B100	Germany	NCBI	Wang et al. 2010
B103	France	NCBI	Bukowski et al. 2018
B104	France	NCBI	Qiu et al. 2021
B105	France	NCBI	Qiu et al. 2021
B106	France	ENA	Grzybowski et al. 2023
B107	France	NCBI	Qiu et al. 2021
B108	Germany	NCBI	Qiu et al. 2021
B109	Germany	NCBI	Qiu et al. 2021
B110	Germany	NCBI	Qiu et al. 2021
B113	Germany	None	None
B14a	France	NCBI	Bukowski et al. 2018
B37	France	NCBI	Qiu et al. 2021
B73	France	NCBI	Qiu et al. 2021
B84	France	NCBI	Qiu et al. 2021
B89	France	None	None
B97	USA	NCBI	Qiu et al. 2021
B98	Germany	NCBI	Wang et al. 2010
C103	France	NCBI	Qiu et al. 2021
CO109	France	ENA	Brandenburg et al. 2017
CR1Ht	USA	None	None
D09	Germany	NCBI	Unterseer et al. 2014
DK4676A	USA	NCBI	Qiu et al. 2021
DK78010	USA	NCBI	Qiu et al. 2021
DK78371A	USA	NCBI	Qiu et al. 2021
DKFAPW	USA	NCBI	Qiu et al. 2021
DKFBHJ	USA	NCBI	Qiu et al. 2021
DKIB02	USA	None	None
EA1027	France	None	None

EA1163	France	None	None
EA3076	Spain	None	None
EC136	Spain	None	None
EC140	Spain	None	None
EC169	Spain	NCBI	Unterseer et al. 2014
EC175	Spain	None	None
EC232	Spain	None	None
EC242C	Spain	None	None
EC334	Spain	None	None
EP10	Spain	None	None
EP2	Spain	None	None
EP2008-18	Spain	None	None
EP29	Spain	None	None
EP51	Spain	None	None
EP55	Spain	None	None
EP56	Spain	None	None
EP67	Spain	None	None
EP72	Spain	None	None
EP77	Spain	None	None
EZ11A	Spain	None	None
EZ18	Spain	None	None
EZ31	Spain	None	None
EZ35	Spain	None	None
EZ36	Spain	None	None
EZ37	Spain	None	None
EZ38	Spain	None	None
EZ40	Spain	None	None
EZ42	Spain	None	None
EZ47	Spain	None	None
EZ48	Spain	None	None
EZ5	Spain	NCBI	Unterseer et al. 2014
F04401	France	None	None
F04402	France	None	None
F04701	France	None	None

F04702	France	None	None
F05101	France	None	None
F05404	France	None	None
F1808	France	None	None
F1890	France	None	None
F218	France	None	None
F252	France	NCBI	Unterseer et al. 2014
F353	France	NCBI	Unterseer et al. 2014
F354	France	None	None
F608	France	None	None
F618	France	NCBI	Unterseer et al. 2014
F7019	France	None	None
F7028	France	None	None
F7057	France	None	None
F7081	France	None	None
F7082	France	None	None
F748	France	None	None
F752	France	None	None
F838	France	None	None
F874	France	None	None
F888	France	None	None
F894	France	None	None
F908	France	None	None
F912	France	None	None
F918	France	None	None
F922	France	None	None
F924	France	None	None
F98902	France	NCBI	Unterseer et al. 2014
FP1	France	None	None
FR19	USA	NCBI	Qiu et al. 2021
H99	France	NCBI	Qiu et al. 2021
I198	France	None	None
I238	France	None	None
I242	France	None	None

I261	France	None	None
I267	France	None	None
IDT	France	None	None
LAN496	France	None	None
LH123Ht	USA	NCBI	Qiu et al. 2021
LH145	USA	NCBI	Qiu et al. 2021
LH38	USA	NCBI	Qiu et al. 2021
LH59	USA	NCBI	Qiu et al. 2021
LH60	USA	NCBI	Qiu et al. 2021
LH65	USA	NCBI	Qiu et al. 2021
LH74	France	NCBI	Qiu et al. 2021
LH82	France	NCBI	Qiu et al. 2021
LH85	USA	NCBI	Qiu et al. 2021
LH93	USA	None	None
Lo1016	Italy	None	None
Lo1026	Italy	None	None
Lo1035	Italy	None	None
Lo1038	Italy	None	None
Lo1056	Italy	None	None
Lo1063	Italy	None	None
Lo1087	Italy	None	None
Lo1094	Italy	None	None
Lo1095	Italy	None	None
Lo1101	Italy	None	None
Lo1106	Italy	None	None
Lo1123	Italy	None	None
Lo1124	Italy	None	None
Lo1172	Italy	None	None
Lo1180	Italy	None	None
Lo1187	Italy	None	None
Lo1199	Italy	None	None
Lo1203	Italy	None	None
Lo1223	Italy	None	None
Lo1242	Italy	None	None

Lo1251	Italy	None	None
Lo1253	Italy	None	None
Lo1261	Italy	None	None
Lo1266	Italy	None	None
Lo1270	Italy	None	None
Lo1273	Italy	None	None
Lo1274	Italy	None	None
Lo1280	Italy	None	None
Lo1282	Italy	None	None
Lo1284	Italy	None	None
Lo1288	Italy	None	None
Lo1290	Italy	None	None
Lo1301	Italy	None	None
Lo904	Italy	None	None
Lp5	USA	NCBI	Qiu et al. 2021
ML606	USA	NCBI	Qiu et al. 2021
Mo15W	France	ENA	Grzybowski et al. 2023
MO17	France	NCBI	Qiu et al. 2021
MS153	France	NCBI	Qiu et al. 2021
Ms71	Germany	NCBI	Qiu et al. 2021
Mt42	France	NCBI	Bukowski et al. 2018
N16	France	None	None
N192	USA	NCBI	Qiu et al. 2021
N22	France	ENA	Brandenburg et al. 2017
N25	France	None	None
N6	France	NCBI	Qiu et al. 2021
NC290	Germany	None	None
NC358	Germany	NCBI	Bukowski et al. 2018
NDB8	France	None	None
NK764	USA	None	None
NK807	USA	NCBI	Qiu et al. 2021
NQ508	USA	NCBI	Qiu et al. 2021
NS701	USA	NCBI	Qiu et al. 2021
Oh02	France	None	None

Oh33	France	NCBI	Qiu et al. 2021
Oh40B	France	NCBI	Qiu et al. 2021
OH43	France	NCBI	Qiu et al. 2021
Os426	France	ENA	Grzybowski et al. 2023
P465P	France	None	None
Pa36	France	None	None
Pa405	Germany	ENA	Grzybowski et al. 2023
Pa91	France	NCBI	Qiu et al. 2021
PB116	France	None	None
PB98TR	France	None	None
PH207	USA	NCBI	Qiu et al. 2021
PHB09	USA	NCBI	Qiu et al. 2021
PHG35	USA	NCBI	Qiu et al. 2021
PHG39	USA	NCBI	Qiu et al. 2021
PHG47	USA	NCBI	Qiu et al. 2021
PHG50	USA	NCBI	Qiu et al. 2021
PHG71	USA	NCBI	Qiu et al. 2021
PHG80	USA	NCBI	Qiu et al. 2021
PHG83	USA	None	None
PHG84	USA	NCBI	Bukowski et al. 2018
PHG86	USA	NCBI	Bukowski et al. 2018
PHH93	USA	NCBI	Qiu et al. 2021
PHJ40	USA	NCBI	Qiu et al. 2021
PHK29	USA	NCBI	Qiu et al. 2021
PHK76	USA	NCBI	Qiu et al. 2021
PHR36	USA	NCBI	Qiu et al. 2021
PHT77	USA	NCBI	Bukowski et al. 2018
PHV63	USA	NCBI	Qiu et al. 2021
PHW65	USA	NCBI	Qiu et al. 2021
PHZ51	USA	NCBI	Qiu et al. 2021
PP147	France	ENA	Brandenburg et al. 2017
SC-Malawi	Mexico	None	None
UH007	Germany	None	None
UH250	Germany	NCBI	Unterseer et al. 2014

UH304	Germany	NCBI	Unterseer et al. 2014
UH6102	Germany	None	None
UH6179	Germany	None	None
UHP024	Germany	None	None
UHP060	Germany	None	None
UHP064	Germany	None	None
UHP074	Germany	None	None
UHP087	Germany	None	None
UHP089	Germany	None	None
UHP104	Germany	None	None
UHP115	Germany	None	None
UHP148	Germany	None	None
UHS018	Germany	None	None
UHS020	Germany	None	None
UHS025	Germany	None	None
Va26	France	NCBI	Qiu et al. 2021
W117	France	NCBI	Unterseer et al. 2014
W153Rht	France	None	None
W153Rht	Germany	None	None
W182B	France	NCBI	Bukowski et al. 2018
W182E	France	NCBI	Qiu et al. 2021
W23	Germany	NCBI	Qiu et al. 2021
W602S	Germany	NCBI	Qiu et al. 2021
W604S	Germany	NCBI	Qiu et al. 2021
W64A	USA	NCBI	Qiu et al. 2021
W9	France	NCBI	Qiu et al. 2021
W95115	France	None	None
W95115	Germany	None	None
WF9	France	NCBI	Qiu et al. 2021

Table 1. List of the maize inbred lines object of the present study for which are reported the public database in which are stored

Paper	N° lines	Coverage	Database	Type of sequencing	Paired/single end	Technology
Qiu et al. 2021	69	~20X	NCBI	Short reads	Paired end	Illumina HiSeq X Ten
Unterseer et al. 2014	6	11,75 X	NCBI	Short reads	Paired end	Illumina HiSeq 2000
Bukowski et al. 2018	8	4-10X	NCBI	Short reads	Paired end	Illumina
Brandenburg et al. 2017	1	~15X	ENA	Short reads	Paired end	Illumina
Wang et al. 2020	3	12,55X	NCBI	Short reads	Paired end	Illumina X Ten
Grzybowski et al. 2022	4	~22X	ENA	Short reads	Paired end	Illumina HiSeq X Ten

Table 2. List of DROPS lines publicly available already sequenced

Sequencing and filtering of variants

For sequencing, the lines were cultivated in a growth chamber at 25°C until the emission of the second leaf. Then, the first leaf (excluded cotyledon) of each line was cut and shipped to Igatech (Udine, Italy) preserved on dry ice. The DNA was extracted from the leaves using the Sbeadex kit (LGC, Teddington, UK). Celero DNA-Seq Library Prep' kit (Tecan, Männedorf, Switzerland) has been used for library preparation following the manufacturer's instructions. Both input and final libraries were quantified by Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) and quality tested by Agilent 2100 Bioanalyzer High Sensitivity DNA assay (Agilent technologies, Santa Clara, CA, USA). Libraries were then prepared for sequencing and sequenced on Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA) in paired-end 150 mode. The sequencing performed on the lines is based on a whole-genome short reads approach and the goal coverage was set to 5X. The base calling and demultiplexing was performed with Illumina BCL Convert v3.9.3. The adapter sequences were masked during the demultiplex with BCL convert. The reads were mapped to the Zea mays Zm-B73-REFERENCE-NAM-5.0 assembly downloaded from MaizeGDB (https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/). The sequences were aligned with BWA-MEM1 0.7.17-r1188. Duplicated sequences were marked by picard2 and removed from downstream analysis and only

sequences to unique positions were kept for the alignment. The variant calling was performed using GATK3 v4.3.0.0 for the identification of SNPs and small indels. The output file obtained after the alignment was filtered using BCFtools v1.21 (H. Li, 2011) with the following command:

bcftools filter -i "(FORMAT/DP >= 1)" --set-GTs "." input_file.vcf.gz | bcftools view -i "QUAL >= 30" | bcftools view -i "(N_PASS(FORMAT/DP >= 2) >= 80)" | bcftools view -i "(N_PASS(FORMAT/GT == 'RA') + N_PASS(FORMAT/GT == 'AA') >= 1)" --output-type z > output_file_filtered.vcf.gz

Structural variant calling and validation

The structural variant calling was performed with the software Hecaton (https://github.com/raul-w/hecaton), an open source software specifically designed for plant genomes that detects copy number variants (CNVs) using short paired-end Illumina reads. CNVs calling is performed integrating existing structural variant callers through a machine-learning model and several custom post-processing scripts. The NAM lines considered for the validation are available through ENA BioProject IDs PRJEB31061. The overlap between the output obtained by Hufford et al. and the output obtained by Hecaton in the present study was calculated using *bedtools* v2.31.1 with the option "*intersect*".

Results

1. Phenotypic analysis of DROPS panel

The DROPS panel was phenotyped in ten locations, in two years (2012 and 2013) and in two water regimes rainfed and irrigated, as reported in Millet et al., 2016. To carry out the open field evaluation, the lines were crossed with a flint line used as tester called UH007. The figures reported below represent a statistic description of the hybrids (Figure 1, Figure 2, Figure 3 and Figure 4). The data shown in the histograms highlights the normal distribution for every trait. The mean value of anthesis is around 70 days while for silking the mean is 73 days. The anthesis silking interval mean is 3 days but the value range from a delay of 27 days for the male flower to a delay of 23 days for the female flower.

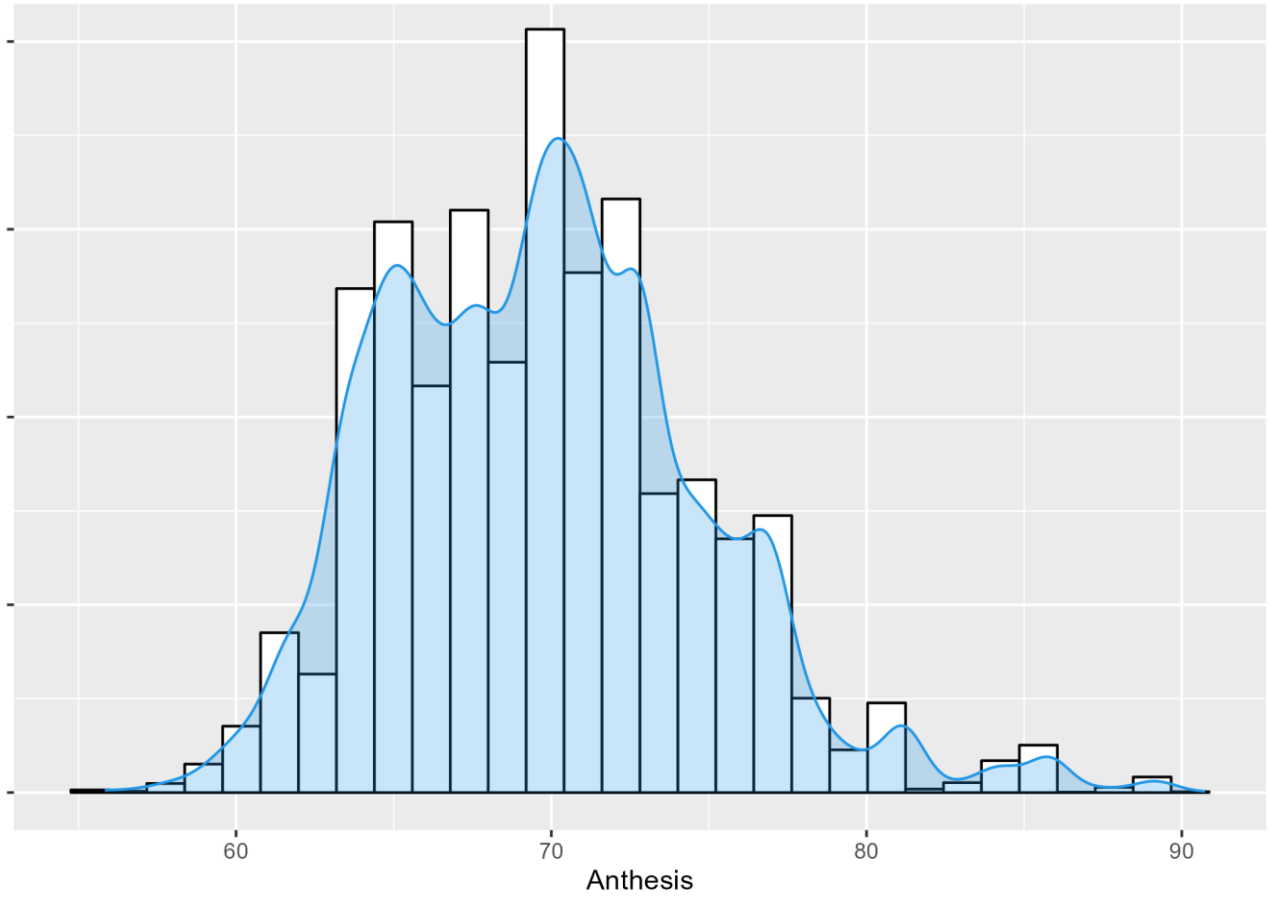


Figure 1. Anthesis data distribution for DROPS hybrids (Days)

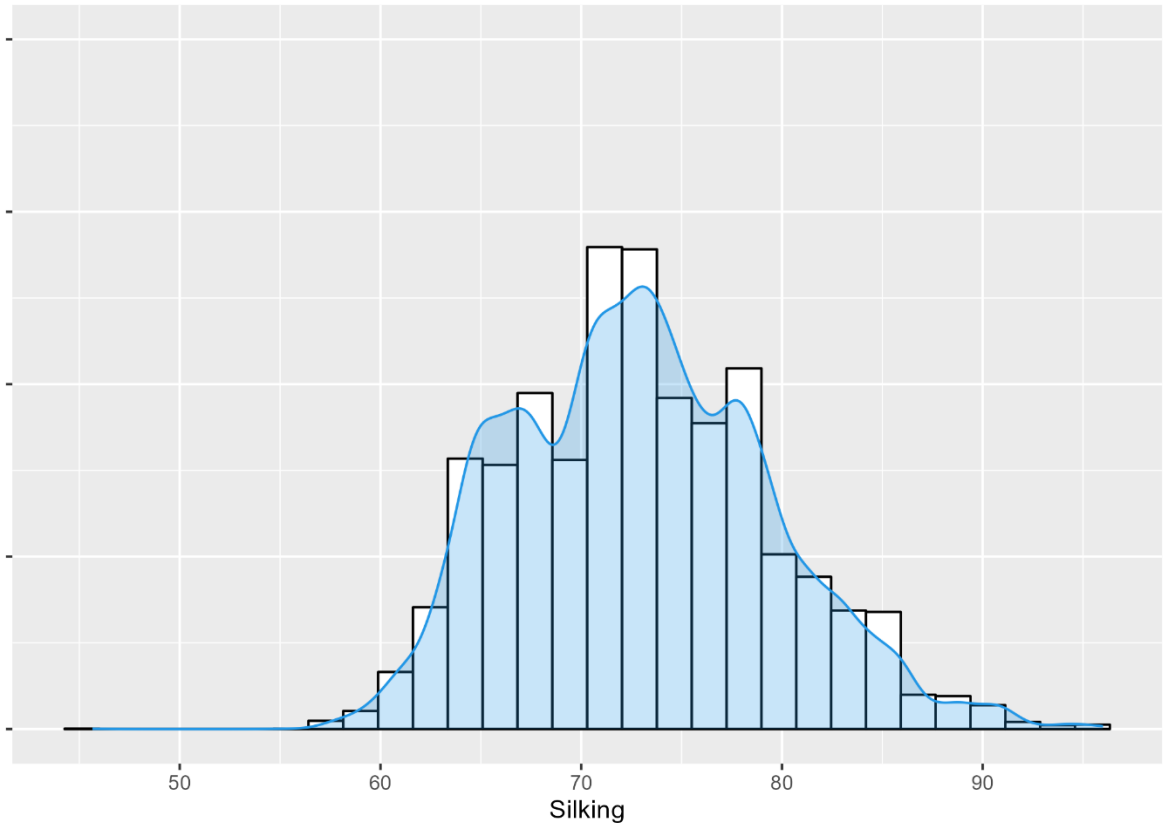


Figure 2. Silking data distribution for DROPS hybrids (Days)

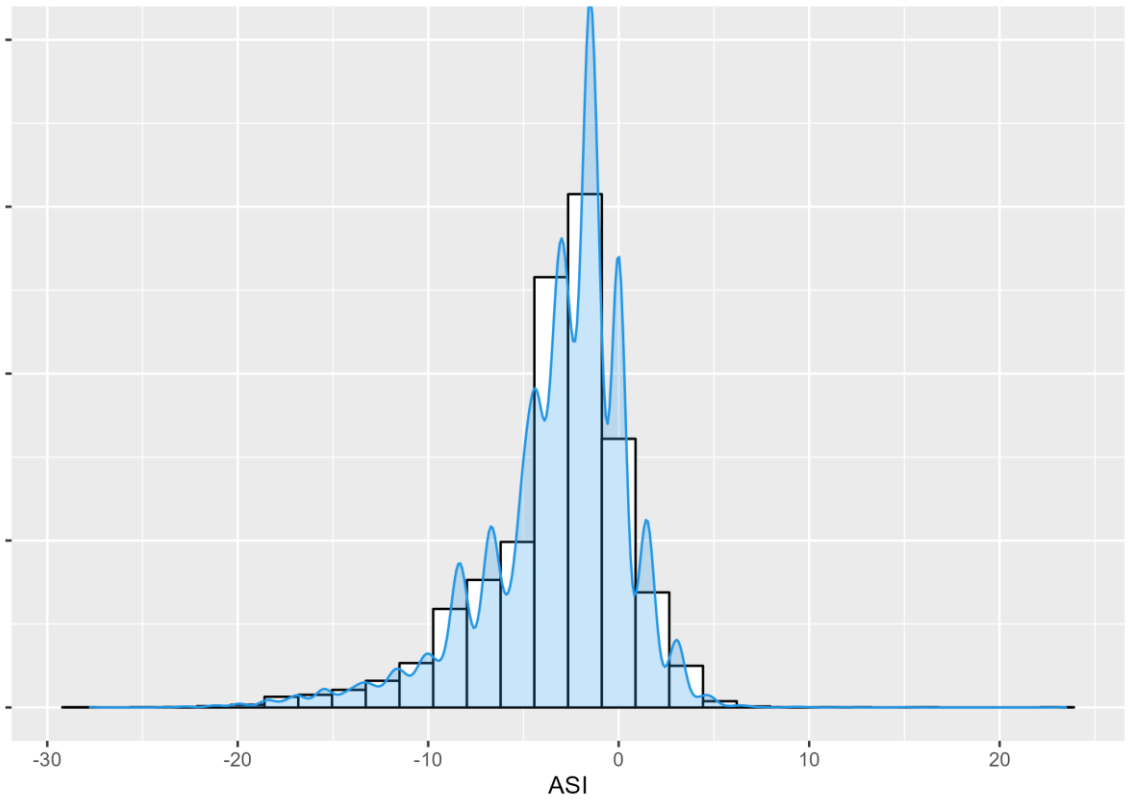


Figure 3. ASI data distribution for DROPS hybrids (Days)

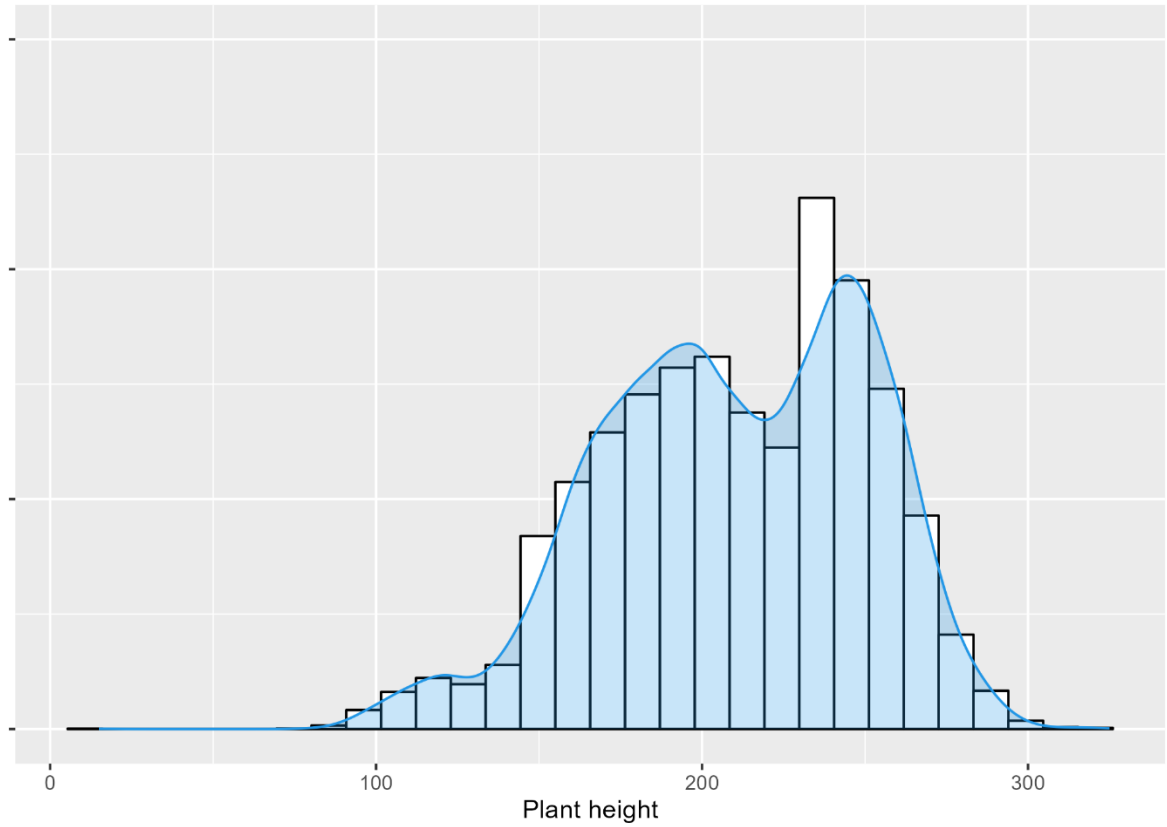


Figure 4. Plant height data distribution for DROPS hybrids (cm)

A Pearson correlation analysis was performed for the main traits and the results are shown in Figure 5. The correlation was calculated on BLUEs values of phenotypic traits. As expected, the higher positive correlation is present for anthesis and silking (0,81) while both flowering traits are negative correlated with plant height (-0,54 and -0,28).

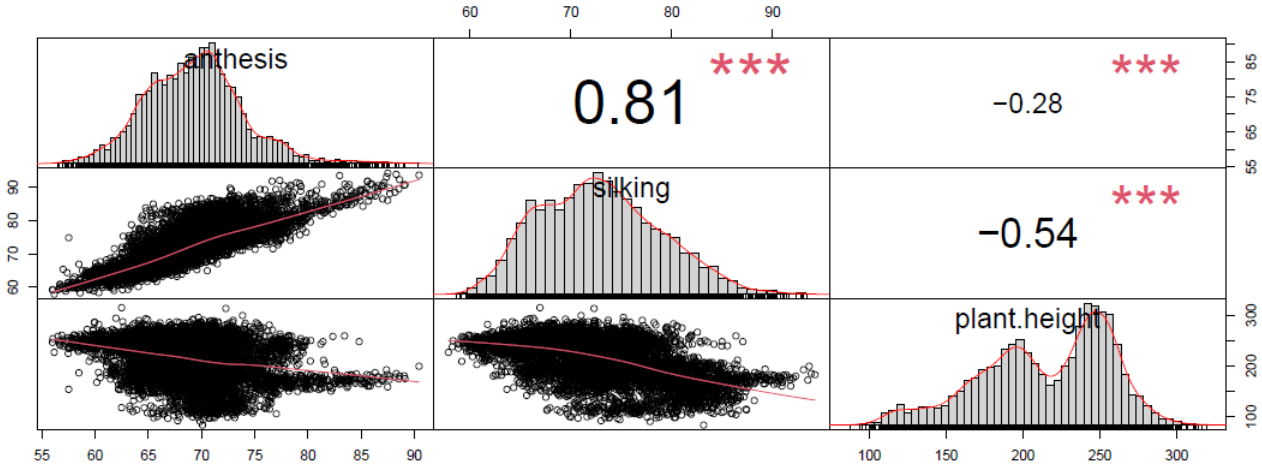
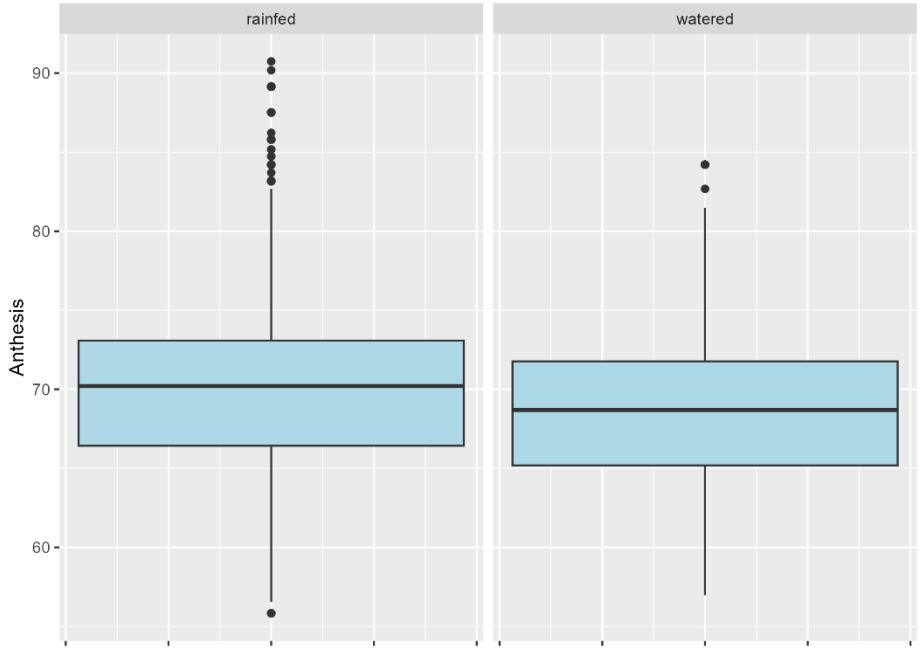


Figure 5. Pearson correlation analysis calculated on BLUEs for anthesis, silking and plant height

The boxplots below (Figure 6, Figure 7, Figure 8 and Figure 9) represent the effects of the two water regimes, rainfed and irrigated, on flowering traits and on plant height. The boxplots confirm the negative effect of water scarcity on maize during flowering and the plant growth reduction during rainfed regime as shown in Figure 9,The statistical significance between the two water regimes included in the study are confirmed by the t-test reported in Table 1.



F*igure 6. Boxplot for anthesis divided into rainfed and watered regime*

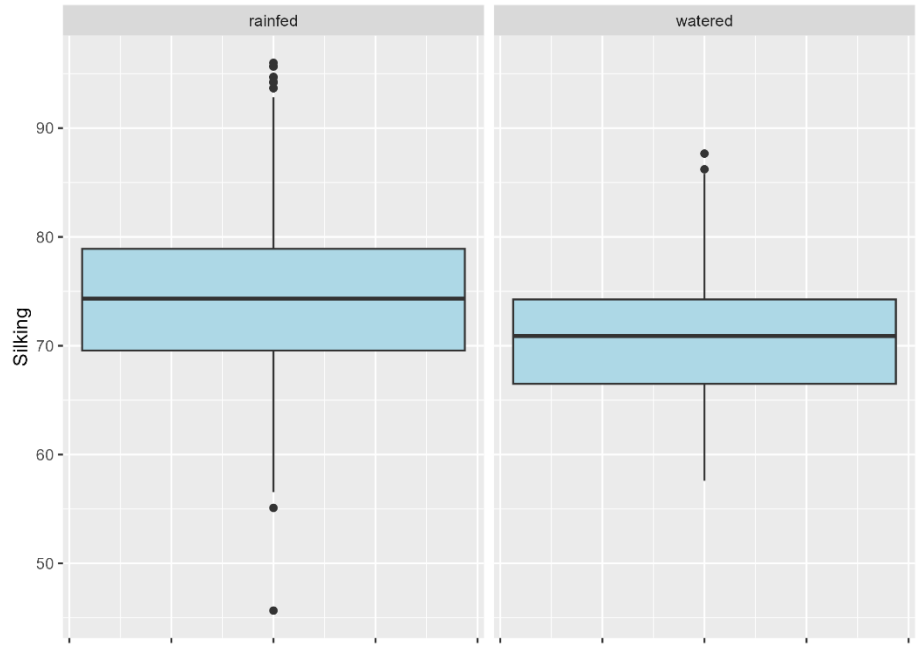


Figure 6. Boxplot for silking divided into rainfed and watered regime

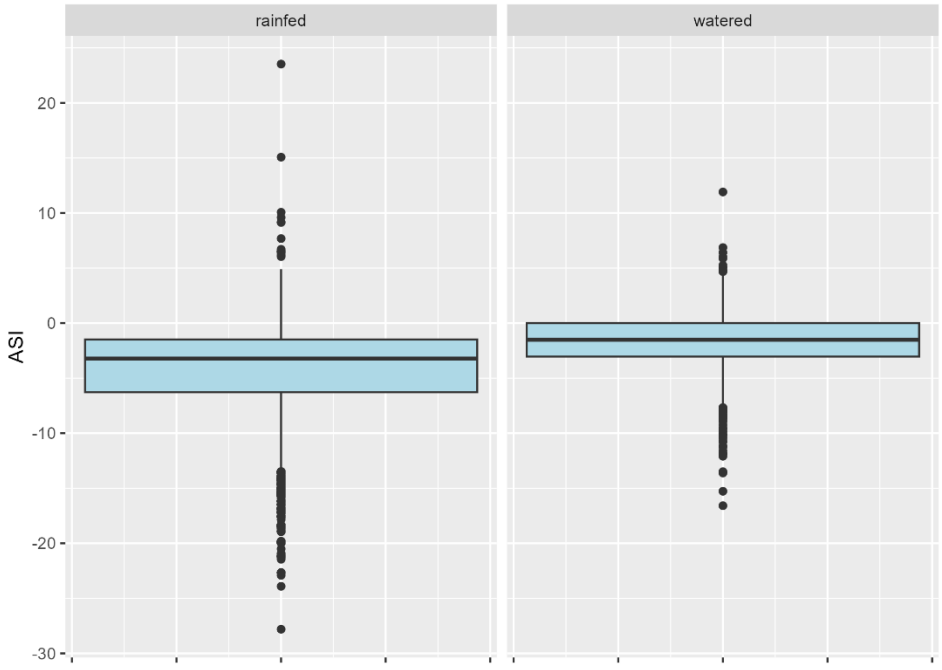


Figure 8. Boxplot for ASI divided into rainfed and watered regime

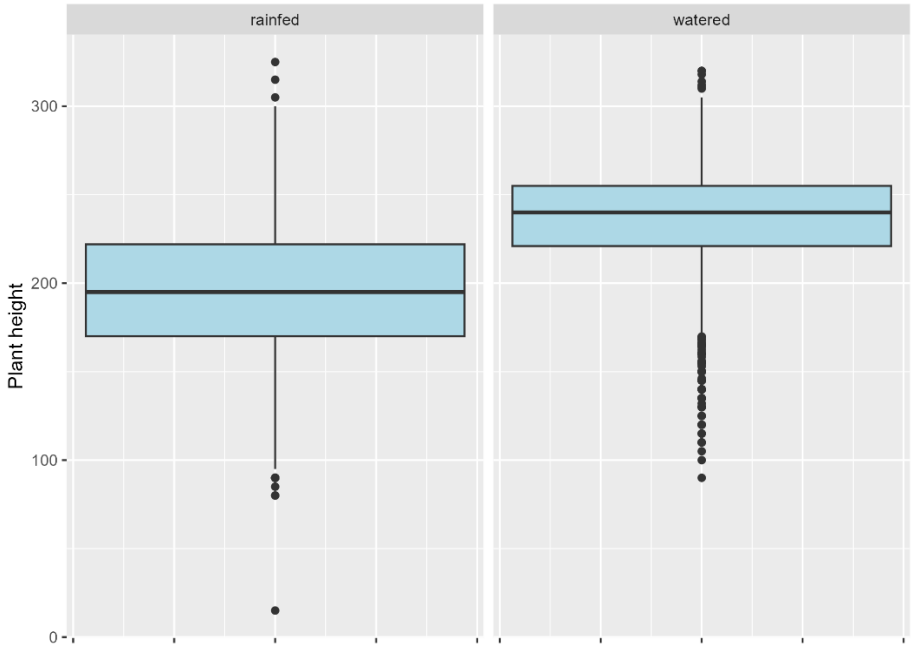


Figure 9. Boxplot for plant height divided into rainfed and watered regime

Trait	Rainfed mean	Watered mean	t-test
Anthesis (days)	70,27	68,89	< 2,2e-16
Silking (days)	74,34	70,59	< 2,2e-16
ASI (days)	-4,07	-1,70	< 2,2e-16
Plant height (cm)	197,07	233,42	< 2,2e-16

Table 1. Means and t-test results calculated for anthesis, silking, ASI and plant height for rainfed and watered regime

The data evaluation on open field was also conducted at CREA in Bergamo during the 2023 season on the inbred lines. Histograms below represent the distribution of open field data which show a normal distribution and distribution intervals comparable to the hybrids. It is notable the reduction of the DROPS lines plant height in comparison to the data recorded by Millet et al., 2016.

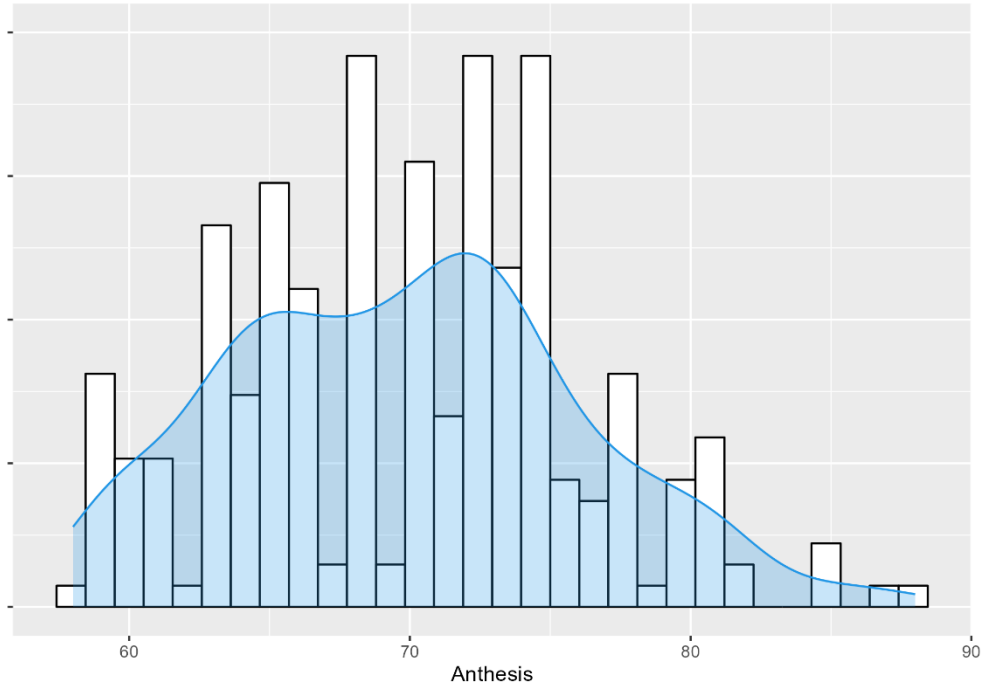


Figure 10. Anthesis data distribution for DROPS lines (Days)

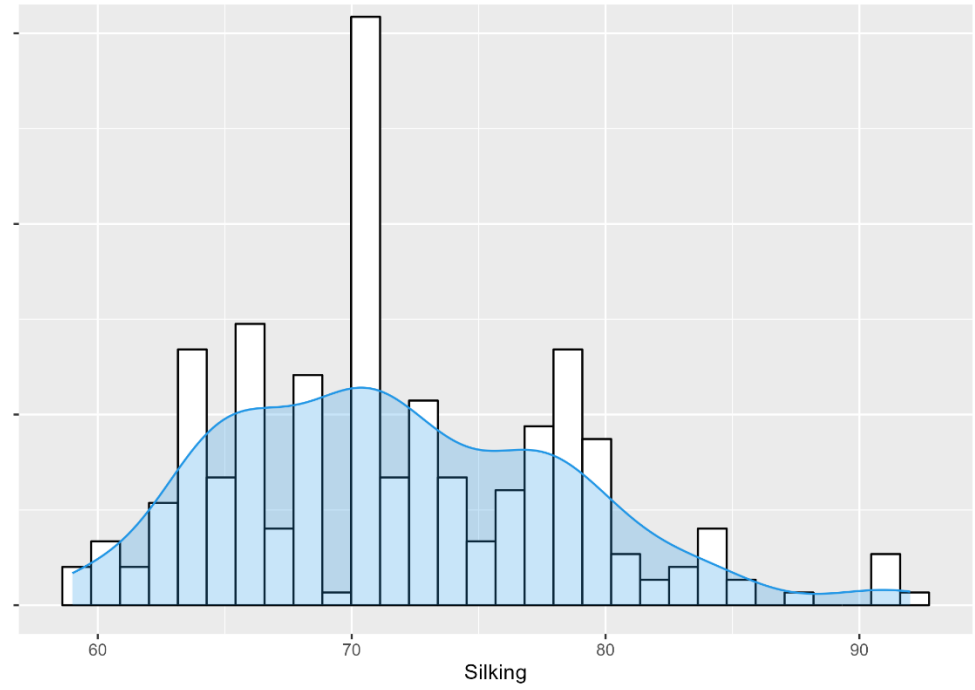


Figure 11. Silking data distribution for DROPS lines (Days)

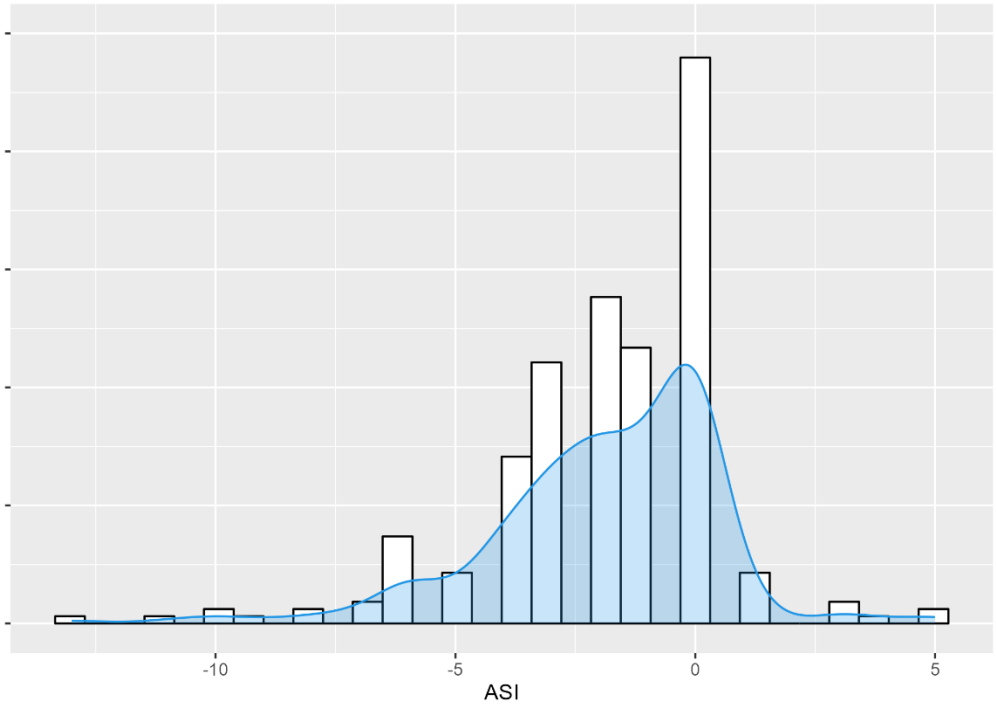


Figure 12. ASI data distribution for DROPS lines (Days)

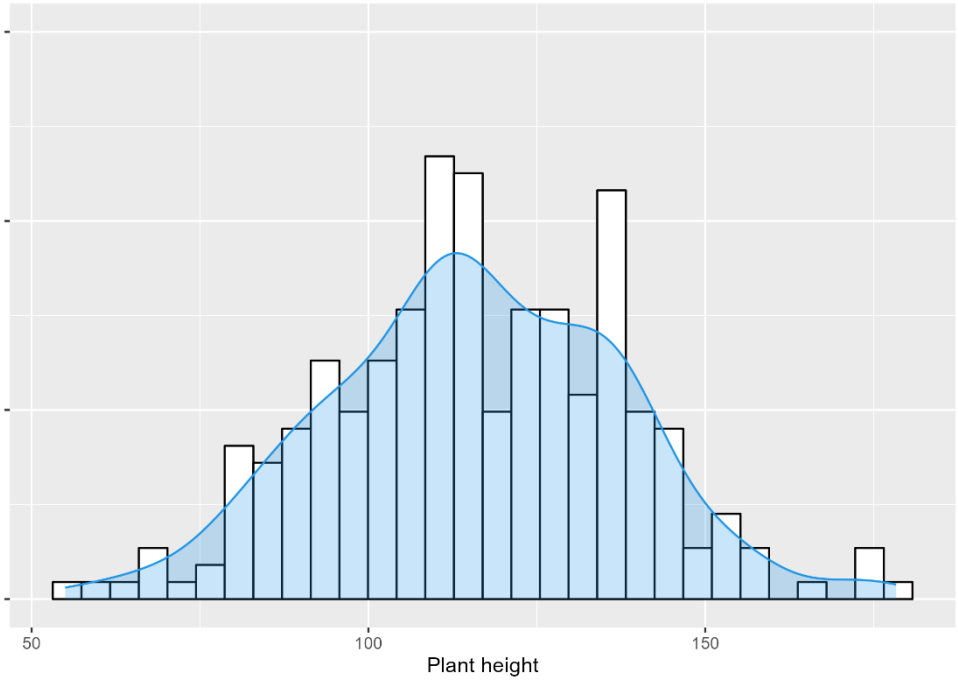


Figure 13. Plant height data distribution for DROPS lines (Days)

In conclusion, maize hybrids derived from DROPS inbred lines show an interesting variability for phenotypic data and the comparison between rainfed and watered regime confirms the response of the lines to drought stress, in particular for flowering time traits. The data recorded in open field in 2023 at CREA Bergamo on DROPS inbred lines confirms the phenotypic variability and show comparable distribution intervals to DROPS hybrids.

The DROPS panel sequencing: preliminary results on SNPs calling

After a quality-filtering step, a total of 89,198,737 SNP markers were retained, with a mean of 9,792,093 SNP markers per chromosome (from a minimum of 7,169,600 SNPs on chromosome 10 to a maximum of 14,151,061 SNPs on chromosome 1). Regarding indels, a total number of 10,014,799 indels were identify. The distribution of the SNPs for every chromosome is shown in Figure 1.

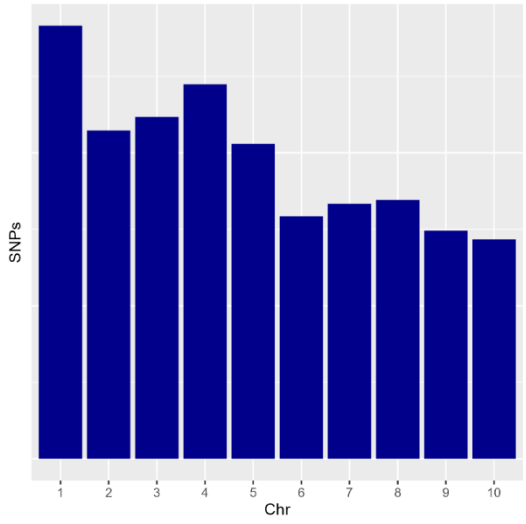


Figure 1. Distribution of SNPs per chromosome

The validation of structural variant caller Hecaton on five maize NAM lines

The software chosen for the structural variant calling is called Hecaton which performs the analysis with four different tools and creates a merged file containing the outputs of each tool. To better understand the potential and the limitations of Hecaton it was performed a validation of the results. The validation regards a comparison between the variants found in a true set of lines already screened for the structural variants and the variants found by Hecaton. The true set of lines used are the NAM parents and the work of Hufford et al., 2021 was used as golden standard for the structural variants information of these lines. Hufford et al., 2021 used an approach based on long reads and chromosomal genome assemblies. For this validation, all the tools were considered and the Hecaton analysis was performed for five samples. In the tables below are summarized the number of variants found by Hufford et al., 2021 and the number of

variants found with Hecaton (Table 1) (Table 2). The results in Table 1 show a huge difference between the deletions found by Hufford et al., 2021 and those found by Hecaton. For the samples B97 and Ki3, Hecaton found around 10% of the total variants found by Hufford et al. while for the samples MS71 and Tx303 it founds around 20% of the variants. Interestingly, the percentage for the sample CM103 is higher, around 40%. The overlap between the variants of the two approaches has been calculated and it is around 50%.

Sample	**Number of variants Hufford et al.**	**Number of variants found by Hecaton**	**Overlap (%)**
B97	24592	2682	49
CML103	14233	5223	27
MS71	25159	5119	53
Ki3	25203	2542	51
Tx303	26776	4751	53

Table1. Comparison between the deletions found in Hufford et al., 2021 *and the deletions found by Hecaton*

The results in Table 2 show a huge difference between the duplications found by Hufford et al., 2021 and those found by Hecaton. The samples show a higher duplication value for Hecaton than for the Hufford et al. approach which increases between 50 and 80% for all samples except for CML103 which has a 10-fold increase and for B97 where Hecaton finds only 46% of the variants. The overlap between the variants of the two approaches has been calculated and is significantly low or in some samples completely absent.

Sample	**Number of variants Hufford et al.**	**Number of variants found by Hecaton**	**Overlap (%)**
B97	39	18	5
CML103	11	107	0
MS71	33	54	2
Ki3	28	50	0
Tx303	43	65	0

Table 2. Comparison between the duplications found in Hufford et al., 2021 *and the deletions found by Hecaton*

Discussion

The validation conducted in the present study highlights the advantages and the limitations of the use of Hecaton to detect structural variants. The validation focused on deletions and duplication considering that the main interest in the present study are copy-number variations. Considering the number of deletions, the study of Hufford et al., 2021 reports a significant higher value of deletions in comparison to Hecaton. This difference might be explained by the use of PacBio long reads and higher coverage for Illumina PE reads (from a range of 26x to 73x) (Hufford et al., 2021) In particular, it is known that the use of long reads improves the mapping and also increase the potentiality to capture larger SVs better compared to short reads alone (Mahmoud et al., 2019). For the same value of coverage, the long reads sequencing methods are considerable expensive compared to short reads sequencing methods and for this reason an approach that includes long reads and high coverage short reads is still valuable (Zhao et al., 2021). Besides this consideration, the overlap between the variants found by Hufford et al., 2021 and the variants found by Hecaton is around 50% for most of the samples which confirms the good sensitivity and precision of Hecaton. It is notable that Hecaton output files are strictly filtered by the use of a random forest model implemented in the software and trained by *Arabidopsis thaliana* and *Oryza sativa Japonica* samples (Wijfjes et al., 2019). In contrast, Hecaton found higher value of duplication compared to Hufford et al., 2021 with a low overlapping of the genomic regions, suggesting a potential overcalling. The higher number of deletions compare to the number of duplication found highlights Hecaton major sensitivity to deletions which is confirmed in the work of Boatwright et al., 2022 conducted on sorghum accessions. The analysis of Boatwright et al., 2022 also confirmed the better efficiency of Hecaton in detecting CNVs compared to the performance of other tools reported in a previous study. Considering the results obtained, it is necessary to extend the validation on others lines belonging to the NAM panel to better achieve the limitations of Hecaton, especially regarding duplications.

Future perspectives

In order to identify the structural variants, especially the deletions and the duplications present in the DROPS panel, the entire panel including the sequenced lines and the publicly available sequenced lines will be analyzed by Hecaton. An appropriate filtering will be applied to identify the true variants excluding true and false positives. To study the DROPS lines genetic association with stress tolerance traits, on a genomic scale, it will be necessary to associate phenotypic variation with variation in type and number of

(genic) CNVs of the inbred maize lines. This will be performed with specific bioinformatic tools, some of them to be appropriately set up.

References

Ahsan, M. U., Liu, Q., Perdomo, J. E., Fang, L., & Wang, K. (2023). A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. In *Nature Methods* (Vol. 20, Issue 8, pp. 1143–1158). Nature Research. https://doi.org/10.1038/s41592-023-01932-w

Alam, M. A., Seetharam, K., Zaidi, P. H., Dinesh, A., Vinayan, M. T., & Nath, U. K. (2017). Dissecting heat stress tolerance in tropical maize (Zea mays L.). *Field Crops Research*, *204*, 110–119. https://doi.org/10.1016/j.fcr.2017.01.006

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. In *Nature Reviews Genetics* (Vol. 12, Issue 5, pp. 363–376). https://doi.org/10.1038/nrg2958

Animasaun, D. A., Mustapha, K. A., Akinbobola, A. M., Bakare, A. T., Ogunjobi, J. T., Adedoyin, K. A., & Awujoola, K. F. (2024). Morphological screening and expression of droughtrelated genes P5SC1 and DREB1A in water-stressed pearl millet (Pennisetum glaucum) at the pre-fruiting stage. *Bragantia*, *83*. https://doi.org/10.1590/1678-4499.20230270

Araus, J. L., Serret, M. D., & Edmeades, G. O. (2012). Phenotyping maize for adaptation to drought. In *Frontiers in Physiology: Vol. 3 AUG*. https://doi.org/10.3389/fphys.2012.00305

Bickhart, D. M., & Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. In *Frontiers in Genetics* (Vol. 5, Issue FEB). Frontiers Research Foundation. https://doi.org/10.3389/fgene.2014.00037

Boatwright, J. L., Sapkota, S., Jin, H., Schnable, J. C., Brenton, Z., Boyles, R., & Kresovich, S. (2022). Sorghum Association Panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *Plant Journal*, *111*(3), 888–904. https://doi.org/10.1111/tpj.15853

Bomblies, K., & Doebley, J. F. (2006). Pleiotropic effects of the duplicate maize FLORICAULA/LEAFY genes zfl1 and zfl2 on traits under selection during maize domestication. *Genetics*, *172*(1), 519–531. https://doi.org/10.1534/genetics.105.048595

Chawla, H. S., Lee, H. T., Gabur, I., Vollrath, P., Tamilselvan-Nattar-Amutha, S., Obermeier, C., Schiessl, S. V., Song, J. M., Liu, K., Guo, L., Parkin, I. A. P., & Snowdon, R. J. (2021). Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnology Journal*, *19*(2), 240–250. https://doi.org/10.1111/pbi.13456

Chen, J., Liu, Y., Liu, M., Guo, W., Wang, Y., He, Q., Chen, W., Liao, Y., Zhang, W., Gao, Y., Dong, K., Ren, R., Yang, T., Zhang, L., Qi, M., Li, Z., Zhao, M., Wang, H., Wang, J., … Diao, X. (2023). Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nature Genetics*, *55*(12), 2243–2254. https://doi.org/10.1038/s41588-023-01571-z

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Research*, *43*(16), 7762–7768. https://doi.org/10.1093/nar/gkv784

Chen, Z., Li, W., Gaines, C., Buck, A., Galli, M., & Gallavotti, A. (2021). Structural variation at the maize WUSCHEL1 locus alters stem cell organization in inflorescences. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22699-8

Currall, B. B., Chiangmai, C., Talkowski, M. E., & Morton, C. C. (2013). Mechanisms for Structural Variation in the Human Genome. *Current Genetic Medicine Reports*, *1*(2), 81–90. https://doi.org/10.1007/s40142-013-0012-8

Darracq, A., Vitte, C., Nicolas, S., Duarte, J., Pichon, J. P., Mary-Huard, T., Chevalier, C., Bérard, A., Le Paslier, M. C., Rogowsky, P., Charcosset, A., & Joets, J. (2018). Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-4490-7

De Oliveira, R., Rimbert, H., Balfourier, F., Kitt, J., Dynomant, E., Vrána, J., Doležel, J., Cattonaro, F., Paux, E., & Choulet, F. (2020). Structural Variations Affecting Genes and Transposable Elements of Chromosome 3B in Wheats. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00891

Diaw, Y., Tollon-Cordet, C., Charcosset, A., Nicolas, S. D., Madur, D., Ronfort, J., David, J., & Gouesnard, B. (2021). Genetic diversity of maize landraces from the South-West of France. *PLoS ONE*, *16*(2 February). https://doi.org/10.1371/journal.pone.0238334

Dickerson, G. W. (n.d.). *Cooperative Extension Service Specialty Corns Guide H-232*.

Francia, E., Morcia, C., Pasquariello, M., Mazzamurro, V., Milc, J. A., Rizza, F., Terzi, V., & Pecchioni, N. (2016). Copy number variation at the HvCBF4–HvCBF2 genomic segment is a major component of frost resistance in barley. *Plant Molecular Biology*, *92*(1–2), 161–175. https://doi.org/10.1007/s11103-016-0505-4

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., Grigoriev, A., Mauleon, R., & Alexandrov, N. (2019). Structural variants in 3000 rice genomes. *Genome Research*, *29*(5), 870–880. https://doi.org/10.1101/gr.241240.118

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. In *Theoretical and Applied Genetics* (Vol. 132, Issue 3, pp. 733–750). Springer Verlag. https://doi.org/10.1007/s00122-018-3233-0

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., Clarke, J. D., Graner, E. M., Hansen, M., Joets, J., Le Paslier, M. C., McMullen, M. D., Montalent, P., Rose, M., Schön, C. C., Sun, Q., Walter, H., Martin, O. C., & Falque, M. (2011). A large maize (zea mays L.) SNP genotyping array: Development and germplasm

genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0028334

Gouesnard, B., Zanetto, A., & Welcker, C. (2016). Identification of adaptation traits to drought in collections of maize landraces from southern Europe and temperate regions. *Euphytica*, *209*(3), 565–584. https://doi.org/10.1007/s10681-015-1624-8

Harrison, M. T., Tardieu, F., Dong, Z., Messina, C. D., & Hammer, G. L. (2014). Characterizing drought stress and trait influence on maize yield under current and future conditions. *Global Change Biology*, *20*(3), 867–878. https://doi.org/10.1111/gcb.12381

Hazman, M. Y., & Kabil, F. F. (2022). Maize root responses to drought stress depend on root class and axial position. *Journal of Plant Research*, *135*(1), 105–120. https://doi.org/10.1007/s10265-021-01348-7

He, Q., Li, W., Miao, Y., Wang, Y., Liu, N., Liu, J., Li, T., Xiao, Y., Zhang, H., Wang, Y., Liang, H., Yun, Y., Wang, S., Sun, Q., Wang, H., Gong, Z., & Du, H. (2024). The near-complete genome assembly of hexaploid wild oat reveals its genome evolution and divergence with cultivated oats. *Nature Plants*. https://doi.org/10.1038/s41477-024-01866-x

Hirsch, C. D., & Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1860*(1), 157–165. https://doi.org/10.1016/j.bbagrm.2016.05.010

Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A. G., Fields, C. J., Wright, C. L., Koehler, K., Springer, N. M., Buckler, E., Buell, C. R., de Leon, N., Kaeppler, S. M., Childs, K. L., & Mikel, M. A. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell*, *28*(11), 2700–2714. https://doi.org/10.1105/tpc.16.00353

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. In *Nature Reviews Genetics* (Vol. 21, Issue 3, pp. 171–189). Nature Research. https://doi.org/10.1038/s41576-019-0180-9

Huang, Y., Huang, W., Meng, Z., Braz, G. T., Li, Y., Wang, K., Wang, H., Lai, J., Jiang, J., Dong, Z., & Jin, W. (2021). Megabase-scale presence-absence variation with Tripsacum origin was under selection during maize domestication and adaptation. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02448-2

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Coletta, R. Della, Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., … Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373*(6555), 655–662. https://doi.org/10.1126/science.abg5289

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jayakodi, M., Lu, Q., Pidon, H., Timothy Rabanus-, M., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A., Brar, G. S., Braune, K., Bunk, B., Chalmers, K. J., Egevang Jørgensen, M., Feng, J.-W., Feser, M., Fiebig, A., Gundlach, H., Guo, W., … Mascher, M. (2024). *Adaptive diversification through structural variation in barley*. https://doi.org/10.1101/2024.02.14.580266

Kirby T. Nilsen, S. W. D. X. (2020). *Copy number variation of TdDof controls solid-stemmed architecture in wheat*. *117*, 28708–28718. https://doi.org/10.1073/pnas.2009418117/-/DCSupplemental

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1720-5

Kuo, W. H., Wright, S. J., Small, L. L., & Olsen, K. M. (2024). De novo genome assembly of white clover (Trifolium repens L.) reveals the role of copy number variation in rapid environmental adaptation. *BMC Biology*, *22*(1). https://doi.org/10.1186/s12915-024-01962-6

L. Echarte, L. N. J. D. M. M. C. M. R. A. D. M. (2013). *Grain yield determination and resource use efficiency in maize hybrids released in different decades*. Agricultural chemistry.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6). https://doi.org/10.1186/gb-2014-15-6-r84

Leng, P., Khan, S. U., Zhang, D., Zhou, G., Zhang, X., Zheng, Y., Wang, T., & Zhao, J. (2022). Linkage Mapping Reveals QTL for Flowering Time-Related Traits under Multiple Abiotic Stress Conditions in Maize. *International Journal of Molecular Sciences*, *23*(15). https://doi.org/10.3390/ijms23158410

Li, F., Numa, H., Hara, N., Sentoku, N., Ishii, T., Fukuta, Y., Nishimura, N., & Kato, H. (2019). Identification of a locus for seed shattering in rice (Oryza sativa L.) by combining bulked segregant analysis with whole-genome sequencing. *Molecular Breeding*, *39*(3). https://doi.org/10.1007/s11032-019-0941-3

Li, G., Wang, L., Yang, J., He, H., Jin, H., Li, X., Ren, T., Ren, Z., Li, F., Han, X., Zhao, X., Dong, L., Li, Y., Song, Z., Yan, Z., Zheng, N., Shi, C., Wang, Z., Yang, S., … Wang, D. (2021). A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. *Nature Genetics*, *53*(4), 574–584. https://doi.org/10.1038/s41588-021-00808-z

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, J., Xin, X., Sun, F., Zhu, Z., Xu, X., Yang, J., Xie, X., Yu, J., Wang, X., Li, S., Tian, S., Li, B., Xie, C., & Ma, J. (2023). Copy number variation of B1 controls awn length in wheat. *Crop Journal*, *11*(3), 817–824. https://doi.org/10.1016/j.cj.2022.10.007

Li, Z., & Howell, S. H. (2021). Heat stress responses and thermotolerance in Maize. In *International Journal of Molecular Sciences* (Vol. 22, Issue 2, pp. 1–19). MDPI AG. https://doi.org/10.3390/ijms22020948

Li, Z., Tang, J., Srivastava, R., Bassham, D. C., & Howell, S. H. (2020). The transcription factor bZIP60 links the unfolded protein response to the heat stress response in maize. *Plant Cell*, *32*(11), 3559–3575. https://doi.org/10.1105/TPC.20.00260

Lin, G., He, C., Zheng, J., Koo, D. H., Le, H., Zheng, H., Tamang, T. M., Lin, J., Liu, Y., Zhao, M., Hao, Y., McFraland, F., Wang, B., Qin, Y., Tang, H., McCarty, D. R., Wei, H., Cho, M. J., Park, S., … Liu, S. (2021). Chromosome-level genome assembly of a regenerable maize inbred line A188. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02396-x

Lin, J., Jia, P., Wang, S., & Ye, K. (2022). *Comparison and benchmark of long-read based structural variant detection strategies*. https://doi.org/10.1101/2022.08.09.503274

Liu, H., Fang, X., Zhou, L., Li, Y., Zhu, C., Liu, J., Song, Y., Jian, X., Xu, M., Dong, L., & Lin, Z. (2022). Transposon Insertion Drove the Loss of Natural Seed Shattering during Foxtail Millet Domestication. *Molecular Biology and Evolution*, *39*(6). https://doi.org/10.1093/molbev/msac078

Liu, Q., Xu, J., Zhu, Y., Mo, Y., Yao, X. F., Wang, R., Ku, W., Huang, Z., Xia, S., Tong, J., Huang, C., Su, Y., Lin, W., Peng, K., Liu, C. M., & Xiao, L. (2021). The Copy Number Variation of OsMTD1 Regulates Rice Plant Architecture. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.620282

Liu, Q., Ye, L., Li, M., Wang, Z., Xiong, G., Ye, Y., Tu, T., Schwarzacher, T., & Heslop-Harrison, J. S. (Pat). (2023). Genome-wide expansion and reorganization during grass evolution: from 30 Mb chromosomes in rice and Brachypodium to 550 Mb in Avena. *BMC Plant Biology*, *23*(1). https://doi.org/10.1186/s12870-023-04644-7

Ma, Y., Dai, X., Xu, Y., Luo, W., Zheng, X., Zeng, D., Pan, Y., Lin, X., Liu, H., Zhang, D., Xiao, J., Guo, X., Xu, S., Niu, Y., Jin, J., Zhang, H., Xu, X., Li, L., Wang, W., … Chong, K. (2015). COLD1 confers chilling tolerance in rice. *Cell*, *160*(6), 1209–1221. https://doi.org/10.1016/j.cell.2015.01.046

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. In *Genome Biology* (Vol. 20, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1828-7

Maldonado, C., Mora, F., Bengosi Bertagna, F. A., Kuki, M. C., & Scapim, C. A. (2019). SNP-And haplotype-based GWAS of flowering-related traits in maize with network-assisted gene prioritization. *Agronomy*, *9*(11). https://doi.org/10.3390/agronomy9110725

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., Sreedasyam, A., Shu, S., Lovell, J. T., Feldman, M., Wu, J., Yu, Y., Chen, C., Johnson, J., Sakakibara, H., Kiba, T., Sakurai, T., Tavares, R., Nusinow, D. A., … Kellogg, E. A. (2020). A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. *Nature Biotechnology*, *38*(10), 1203–1210. https://doi.org/10.1038/s41587-020-0681-2

Mao, W., Wang, X., Chen, Y., Wang, Y., Ma, L., Xie, X., Wu, X., Xu, J., Zhang, Y., Zhao, Y., Tian, X., Guo, W., Hu, Z., Xin, M., Yao, Y., Ni, Z., Sun, Q., & Peng, H. (2024). Map-based cloning and characterization reveal that an R2R3 MYB gene confers red glume in wheat. *The Crop Journal*. https://doi.org/10.1016/j.cj.2024.03.002

Massman, C., Maughan, P. J., Nandety, R. S., Clare, S. J., Fiedler, J. D., & Hayes, P. M. (2024). Exploratory genomic sequence analysis reveals structural differences at key loci for growth habit, seed dormancy, and rust resistance in barley. *Genetic Resources and Crop Evolution*. https://doi.org/10.1007/s10722-024-01875-x

Mastrangelo, A. M., Hartings, H., Lanzanova, C., Balconi, C., Locatelli, S., Cassol, H., Valoti, P., Petruzzino, G., & Pecchioni, N. (2024). Genetic Diversity within a Collection of Italian Maize Inbred Lines: A Resource for Maize Genomics and Breeding. *Plants*, *13*(3). https://doi.org/10.3390/plants13030336

Millet, E. J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S. D., Laborde, J., Bauland, C., Praud, S., Ranc, N., Presterl, T., Tuberosa, R., Bedo, Z., Draye, X., Usadel, B., Charcosset, A., Van Eeuwijk, F., & Tardieu, F. (2016). Genome-wide analysis of yield in Europe: Allelic effects vary with drought and heat scenarios. *Plant Physiology*, *172*(2), 749–764. https://doi.org/10.1104/pp.16.00621

Munasinghe, M., Read, A., Stitzer, M. C., Song, B., Menard, C. C., Ma, K. Y., Brandvain, Y., Hirsch, C. N., & Springer, N. (2023). Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. *PLoS Genetics*, *19*(12). https://doi.org/10.1371/journal.pgen.1011086

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., Mayer, K. F. X., Taudien, S., Platzer, M., Jeddeloh, J. A., Springer, N. M., Muehlbauer, G. J., & Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, *14*(6). https://doi.org/10.1186/gb-2013-14-6-r58

Nagarajan, R., Kahlon, K. S., & Mohan, A. (2024). *Tandemly Duplicated Rubisco Activase Genes of Cereals Show Differential Evolution and Response to Heat Stress*. https://doi.org/10.21203/rs.3.rs-4676428/v1

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., Tardieu, F., Charcosset, A., & Nicolas, S. D. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1926-4

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., & Szemes, T. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. In *Biomedical Journal* (Vol. 44, Issue 5, pp. 548–559). Elsevier B.V. https://doi.org/10.1016/j.bj.2021.02.003

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., Li, X., Li, Y., Liao, Y., Gao, Q., Tu, B., Yuan, H., Ma, B., Wang, Y., Qian, Y., … Li, S. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, *184*(13), 3542-3558.e16. https://doi.org/10.1016/j.cell.2021.04.046

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18). https://doi.org/10.1093/bioinformatics/bts378

Rebourg, C., Chastanet, M., Gouesnard, B., Welcker, C., Dubreuil, P., & Charcosset, A. (2003). Maize introduction into Europe: The history reviewed in the light of molecular data. *Theoretical and Applied Genetics*, *106*(5), 895–903. https://doi.org/10.1007/s00122-002-1140-9

Ren, Y., Ma, R., Lian, Q., & Jiang, T. (2023). *Evolutionary analysis of TCP gene family and its response to hormonal and abiotic stress in rye (Secale cereale L.)*. https://doi.org/10.21203/rs.3.rs-2771554/v1

Revilla, P., Rodríguez, V. M., Ordás, A., Rincent, R., Charcosset, A., Giauffret, C., Melchinger, A. E., Schön, C. C., Bauer, E., Altmann, T., Brunel, D., Moreno-González, J., Campo, L., Ouzunova, M., Álvarez, Á., de Galarreta, J. I. R., Laborde, J., & Malvar, R. A. (2016). Association mapping for cold tolerance in two large maize inbred panels. *BMC Plant Biology*, *16*(1). https://doi.org/10.1186/s12870-016-0816-2

Rodríguez-Suárez, C., Requena-Ramírez, M., Hornero-Méndez, D., & Atienza, S. (2023). Towards carotenoid biofortification in wheat: identification of XAT-7A1, a multicopy tandem gene responsible for carotenoid esterification in durum wheat. *BMC Plant Biology*, *23*(1). https://doi.org/10.1186/s12870-023-04431-4

Sall, S. O., Alioua, A., Staerck, S., Graindorge, S., Pellicioli, M., Schuler, J., Raffy, Q., Rousseau, M., & Molinier, J. (2024). *Characterization of radiations-induced genomic structural variations in* Arabidopsis thaliana. https://doi.org/10.1101/2024.07.25.605065

Sarwal, V., Niehus, S., Ayyala, R., Kim, M., Sarkar, A., Chang, S., Lu, A., Rajkumar, N., Darfci-Maher, N., Littman, R., Chhugani, K., Soylev, A., Comarova, Z., Wesel, E., Castellanos, J., Chikka, R., Distler, M. G., Eskin, E., Flint, J., & Mangul, S. (2022). A comprehensive benchmarking of WGS-based deletion structural variant callers. *Briefings in Bioinformatics*, *23*(4). https://doi.org/10.1093/bib/bbac221

Schmidt, M., Guerreiro, R., Baig, N., Habekuß, A., Will, T., Ruckwied, B., & Stich, B. (2024). Fine mapping a QTL for BYDV-PAV resistance in maize. *Theoretical and Applied Genetics*, *137*(7). https://doi.org/10.1007/s00122-024-04668-z

Sheoran, S. (2022). Recent Advances for Drought Stress Tolerance in Maize (Zea mays L.): Present Status and Future Prospects. In *Frontiers in Plant Science* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fpls.2022.872566

Song, J. M., Xie, W. Z., Wang, S., Guo, Y. X., Koo, D. H., Kudrna, D., Gong, C., Huang, Y., Feng, J. W., Zhang, W., Zhou, Y., Zuccolo, A., Long, E., Lee, S., Talag, J., Zhou, R., Zhu, X. T., Yuan, D., Udall, J., … Chen, L. L. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant*, *14*(10), 1757–1767. https://doi.org/10.1016/j.molp.2021.06.018

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C. T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, *5*(11). https://doi.org/10.1371/journal.pgen.1000734

Sun, C., Gao, X., Chen, X., Fu, J., & Zhang, Y. (2016). Metabolic and growth responses of maize to successive drought and re-watering cycles. *Agricultural Water Management*, *172*, 62–73. https://doi.org/10.1016/j.agwat.2016.04.016

Széles, A., Horváth, É., Simon, K., Zagyi, P., & Huzsvai, L. (2023). Maize Production under Drought Stress: Nutrient Supply, Yield Prediction. *Plants*, *12*(18). https://doi.org/10.3390/plants12183301

Tenaillon, M. I., & Charcosset, A. (2011). A European perspective on maize history. In *Comptes Rendus - Biologies* (Vol. 334, Issue 3, pp. 221–228). Elsevier Masson SAS. https://doi.org/10.1016/j.crvi.2010.12.015

Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G., Ma, X., Wang, H., Xie, Y., Li, Q., Song, G., Kong, D., Zheng, Z., Wei, H., Shen, R., Wu, H., Chen, C., Meng, Z., Wang, T., … Wang, H. (2020). Genome-wide selection and genetic improvement during modern maize breeding. *Nature Genetics*, *52*(6), 565–571. https://doi.org/10.1038/s41588-020-0616-3

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics and Bioinformatics*, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. In *Nature Reviews Genetics* (Vol. 14, Issue 2, pp. 125–138). https://doi.org/10.1038/nrg3373

Wijfjes, R. Y., Smit, S., & De Ridder, D. (2019). Hecaton: Reliably detecting copy number variation in plant genomes using short read sequencing data. *BMC Genomics*, *20*(1). https://doi.org/10.1186/s12864-019-6153-8

Würschum, T., Rapp, M., Miedaner, T., Longin, C. F. H., & Leiser, W. L. (2019). Copy number variation of Ppd-B1 is the major determinant of heading time in durum wheat. *BMC Genetics*, *20*(1). https://doi.org/10.1186/s12863-019-0768-2

Xue, W., Anderson, S. N., Wang, X., Yang, L., Crisp, P. A., Li, Q., Noshay, J., Albert, P. S., Birchler, J. A., Bilinski, P., Stitzer, M. C., Ross-Ibarra, J., Flint-Garcia, S., Chen, X., Springer, N. M., & Doebley, J. F. (2019). Hybrid decay: A transgenerational epigenetic decline in vigor and viability triggered in backcross populations of teosinte with maize. *Genetics*, *213*(1), 143–160. https://doi.org/10.1534/genetics.119.302378

Yan, H., Sun, M., Zhang, Z., Jin, Y., Zhang, A., Lin, C., Wu, B., He, M., Xu, B., Wang, J., Qin, P., Mendieta, J. P., Nie, G., Wang, J., Jones, C. S., Feng, G., Srivastava, R. K., Zhang, X., Bombarely, A., … Huang, L. (2023). Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nature Genetics*, *55*(3), 507–518. https://doi.org/10.1038/s41588-023-01302-4

Yang, G. Q., Chen, Y. M., Wang, J. P., Guo, C., Zhao, L., Wang, X. Y., Guo, Y., Li, L., Li, D. Z., & Guo, Z. H. (2016). Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. *Plant Methods*, *12*(1). https://doi.org/10.1186/s13007-016-0139-1

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, *31*(14), 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

Zhan, A., Schneider, H., & Lynch, J. P. (2015). Reduced lateral root branching density improves drought tolerance in maize. *Plant Physiology*, *168*(4), 1603–1615. https://doi.org/10.1104/pp.15.00187

Zhang, J., Zuo, T., & Peterson, T. (2013). Generation of Tandem Direct Duplications by Reversed-Ends Transposition of Maize Ac Elements. *PLoS Genetics*, *9*(8). https://doi.org/10.1371/journal.pgen.1003691

Zhang, L., Zhu, X., Zhao, Y., Guo, J., Zhang, T., Huang, W., Huang, J., Hu, Y., Huang, C. H., & Ma, H. (2022). Phylotranscriptomics Resolves the Phylogeny of Pooideae and Uncovers Factors for Their Adaptive Evolution. *Molecular Biology and Evolution*, *39*(2). https://doi.org/10.1093/molbev/msac026

Zhang, T., Huang, W., Zhang, L., Li, D.-Z., Qi, J., & Ma, H. (2024). Phylogenomic profiles of whole-genome duplications in Poaceae and landscape of differential duplicate retention and losses among major Poaceae lineages. *Nature Communications*, *15*(1), 3305. https://doi.org/10.1038/s41467-024-47428-9

Zhao, X., Collins, R. L., Lee, W. P., Weber, A. M., Jun, Y., Zhu, Q., Weisburd, B., Huang, Y., Audano, P. A., Wang, H., Walker, M., Lowther, C., Fu, J., Gerstein, M. B., Devine, S. E., Marschall, T., Korbel, J. O., Eichler, E. E., Chaisson, M. J. P., … Talkowski, M. E. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *American Journal of Human Genetics*, *108*(5), 919–928. https://doi.org/10.1016/j.ajhg.2021.03.014

Zheng, Y., Li, S., Huang, J., Fu, H., Zhou, L., Furusawa, Y., & Shu, Q. (2021). Identification and characterization of inheritable structural variations induced by ion beam radiations in rice. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, *823*. https://doi.org/10.1016/j.mrfmmm.2021.111757

Zmienko, A., Samelak-Czajka, A., Kozlowski, P., Szymanska, M., & Figlerowicz, M. (2016). Arabidopsis thaliana population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-3221-1