Genetic Diversity in Food Legumes: Novel Approaches of Investigation

Ph.D Supervisor

Prof. Roberto Papa

Ph.D School director

Prof. Bruno Mezzetti

Ph.D Co-supervisor

Dr. Elisa Bellucci

Ph.D candidate

Simone Papalini

UNIVERSITÀ POLITECNICA DELLE MARCHE

Department of Agricultural, Food and Environmental Sciences

Ph. D. in Agricultural, Food and Environmental Sciences

N° Cycle XXXVII

Table of Contents

Chapter I 1

A General Overview 1

Introduction and the project INCREASE 1

Genetic diversity in modern agriculture 2

Domestication and Its Impact on Plant Genetic Diversity 4

PGR conservation 10

Local Adaptation: A Driving Force Behind Plant Diversity 11

Adaptive role of structural variation 13

Molecular characterization of Plant genetic resources 14

Genomic Complexity Reduction Approaches 15

The genus Lens 17

Origin and domestication 17

Cultivated Lentil 19

The Lentil genome 20

Italian lentil landrces form Castelluccio di Norcia 20

***Herbaria* and Ancient Samples in Genomics Studies** 22

Research questions and objectives 24

Transition to Chapter II 26

Chapter II 27

Genetic Diversity in Lentil Italian Landraces Genotyped By a CRISPR-Cas9-Based Depletion Approach 27

Materials and Methos 28

Plant material 28

DNA extraction and Genotyping and VCF filtering 29

VCF filtering 30

Heterozygote Calling Errors Resulting from Misalignment 33

Characterization of population structure 37

Detection of loci under selection 37

Detection of Inversions through Local PCA 40

Identification of Genes Potentially Under Selection 40

GO enrichment analysis 42

Genotype imputation 42

Results 44

Characterization of genetic groups 44
Genetic Diversity within the Castelluccio Subpopulation 47
Characterization of genetic groups by Imputed data 51
Detection of loci under selection 53
Genes under selection 56
Pcadapt: comparison of results between pruned and non-pruned data 58
Discussion 59
Characterization and Genetic Diversity of the Castelluccio Population 59
Adaptation to high altitudes 60
Conclusion 62
Transition to Chapter III 63
Chapter III 64
General discussions 79
General Conclusion 81
Bibliography 82

Abstract (English version)

Genetic diversity studies allow us to catch the existing genomic variation among and within populations and are crucial in guiding our understanding of crops' adaptation, facilitating their use, exploitation, and improvement. For this reason, investigating genetic diversity is particularly crucial in addressing contemporary and future challenges related to global warming and food security. This aspect is even more critical for legume crops due to their significant nutritional and agronomic importance. However, genetic diversity investigations often face several challenges. These include the complexity of analyzing large and highly repetitive genomes in studies involving extensive germplasm collections and the effective utilization of museum materials, such as herbarium specimens. In this thesis, our objective was to address these challenges specifically in legume crops by implementing specific approaches. To achieve this, two different studies are proposed, in the first a genetic diversity study was conducted on an Italian lentil landrace population genotyped using a CRISPR-Cas9-based genome depletion system. In this case, an innovative method for genotypic characterization was combined with the investigation of the genetic diversity of the population, which also allowed us to explore signs of adaptation to high altitudes. Additionally, a comprehensive literature review was performed to assess the challenges and opportunities associated with employing herbarium specimens in genetic studies. Our findings contribute to advancing solutions to these challenges, for a more effective exploration and utilization of genetic diversity in legume crops.

Abstract (versione italiana)

Studiare la diversità all'interno delle popolazioni rappresenta un aspetto importante al fine di comprendere come le specie di interesse agricolo si siano adattate all'ambiente, facilitandone così lo sfruttamento e miglioramento. Di conseguenza, indagare la diversità genetica risulta particolarmente importante nell'affrontare le sfide odierne e future legate al riscaldamento globale, il rapido cambiamento degli agro-ambienti e più in generale quelle relative alla sicurezza alimentare. Questo aspetto acquisisce maggiore interesse se si vanno a considerare le specie di legumi ad uso alimentare, a grazie al loro significativo valore nutrizionale e agronomico. Tuttavia, indagare la diversità genetica presenta molteplici sfide. Tra queste risultano rilevanti l'inclusione e l'analisi di grandi genomi altamente ripetuti in studi che coinvolgono ampie collezioni di germoplasma, così come l'inclusione di materiali preservati nel tempo come i campioni d' erbario. In questo lavoro di tesi, il nostro obiettivo è stato quello di, attraverso l'applicazione di approcci innovati, provare ad affrontare e rispondere a queste sfide. Per raggiungere gli obbiettivi preposti, sono stati implementati due diversi studi. Nel primo, è stato condotto uno studio sulla diversità genetica di una popolazione di lenticchie italiane proveniente da Castelluccio di Norcia (PG), genotipizzata utilizzando un sistema di deplezione del genoma basato su CRISPR-Cas9. In questo caso, l'approccio per la caratterizzazione genotipica è stato combinato con lo studio della diversità genetica, consentendo l'investigazione e la detezione di segnali genetici correlabili all'adattamento ad alte altitudini. Per quanto riguarda lo sfruttamento dei campioni d'erbario invece, è stata effettuata un'esaustiva revisione della letteratura per valutare le sfide e le opportunità associate al loro utilizzo. In fine i risultati ottenuti attraverso queste analisi hanno contribuito a sviluppare soluzioni concrete applicabili alle sfide preposte, permettendo l'esplorazione e un utilizzo più efficace della diversità genetica e delle risorse genetiche vegetali di alcune specie di leguminose ad uso alimentare.

List of Figures

Figure 1. *The main evolutionary forces and their general impact on genetic variability extracted from C. K. Khoury et al., 2022.. _____3*
Figure 2*. Centres of plant domestication extracted from Purugganan & Fuller, 2009. _____5*
Figure 3. *The relationship between competitive trait allocation and productivity in wild versus crop plants, from Gao et al., 2023. _____7*
Figure 4. *Processes of domestication and improvement from Moyers et al., 2018._____8*
Figure 5. *The footprint of natural selection. From Gao et al., 2023_____12*
Figure 6*. Early diffusion of domesticated lentil (Lens culinaris) reconstructed through archaeological evidence from Sonnante et al 2009._____18*
Figure 7*. Plant material distribution _____29*
Figure 8. *A) Coverage distribution of all called variants, B) Coverage distribution of variants located within annotated genes._____30*
Figure 9. Proportion of Heterozygous Sites per Individual. _____31
Figure 10*. The histograms display the missingness values per individual before and after filtering (a, c) and the average depth values per individual before and after filtering (b, d). _____33*
Figure 11.*Misalignment to the incorrect calling of a heterozygous genotype. _____35*
Figure 12*. Distribution of homozygous genotype counts in relation to the increase in heterozygous genotypes. _____36*
Figure 13. *Scree plot. _____39*
Figure 14. The score plot. _____39
Figure 15. *LD decay. _____41*
Figure 16*. Mechanism for Selecting Candidate Genes Under Selection._____41*
Figure 17. *Principal Component Analysis PCA. _____45*
Figure 18. *Genetic structure of the lentil panel by performing sNMF._____46*
Figure 19. *Cross-entropy criterion values as a function of the number of populations in sNMF_____47*
Figure 20*. Pie charts of Q values obtained at K = 5 for the CS population. _____49*
Figure 21. *Correlation of Q1 for Imputed and Non-Imputed Data. _____51*
Figure 22. *The genetic structure of the lentil panel by performing sNMF performed on imputed data. _52*
Figure 23*. Detection of markers under selection._____53*
Figure 24*. Manhattan Plot Highlighting Selection Signals, and Structural Variant Regions. _____55*
Figure 25*. GO enrichment analysis._____56*
Figure 26*. Selected Marker Associated with the Gene Lcu.2RBY.6g066820 on Chromosome 6._____57*
Figure 27. *Comparison of markers under selection between pruned and non-pruned data. _____58*

List of Tables

Table 1. *A: Statistics on the coverage distribution of all variants, B: Statistics on the coverage distribution of variants within gene sequences.*_____*30*
Table 2. *Statistics on heterozygosity levels per individual.* _____*31*
Table 3. *The tables e, f, g, and h respectively present the summaries of the values described in the histograms.* _____*32*
Table 4. *Purity levels contained within the material provided by local farmers from Castelluccio di Norcia.*_____*50*
Table 5. *positions of the 22 regions identified as inversions through detection by local PCA.* _____*54*

Glossary of Key Acronyms

Acronym	Full Name
INCREASE	Intelligent Collections of Legumes for European Agriculture, Food and Environment (H2020 project)
CRISPR-Cas9	Clustered Regularly Interspaced Short Palindromic Repeats - CRISPR-associated protein 9
PGR	Plant Genetic Resources
GEA	Genotype-Environment Association
GWAS	Genome-Wide Association Study
CNV	Copy Number Variation
SV	Structural Variation
RFLP	Restriction Fragment Length Polymorphism
LD	Linkage Disequilibrium
PCR	Polymerase Chain Reaction
NGS	Next-Generation Sequencing
WGS	Whole-Genome Sequencing
SNP	Single Nucleotide Polymorphism
GBS	Genotyping-by-Sequencing
RAD-Seq	Restriction site-Associated DNA Sequencing
WES	Whole-Exome Sequencing
PGI	Plant Genetic Improvement
SSD	Single Seed Descent
VCF	Variant Call Format
MAF	Minor Allele Frequency
PCA	Principal Component Analysis
PCs	Principal Components
sNMF	Sparse Nonnegative Matrix Factorization
FDR	False Discovery Rate
FST	Fixation Index
GO enrichment	Gene Ontology Enrichment
ANOVA	Analysis of Variance

Chapter I

A General Overview

Introduction and the project INCREASE

In the face of escalating global challenges, a key issue is the lack of food availability, food accessibility, and food affordability due to the climate crisis and the associated loss of biodiversity (Kang et al., 2009; Van De Wouw et al., 2010). Despite the progress made in improving human welfare worldwide over the past century, much remains to be done to achieve the vision of creating a world free of hunger and malnutrition. Food and agriculture must contribute to improving the living standards of all, especially the poorest, in an economically, socially, and environmentally sustainable manner.

In this context understanding the origin, evolution, dissemination, and adaptation of crop species is a crucial aspect for their effective exploitation and conscious conservation (Gepts, 2006; Cortinovis et al., 2020). These aspects are explored within the H2020 INCREASE "Intelligent Collections of Food-Legume Genetic Resources for European Agrofood Systems" project (Bellucci et al., 2021), under which my PhD research, has been carried out. In particular, With 28 partners from 14 different countries, coordinated by Marche Polytechnic University UNIVPM Prof. Roberto Papa, INCREASE H2020 is an international project that has the general objective of improving the sustainable use of Plant Genetic Resources (PGR) in food and agriculture. To achieve this goal, the

INCREASE partners work to enhance the genotypic and phenotypic characterization of food legumes, in chickpeas (*Cicer arietinum* L.), common beans (*Phaseolus vulgaris* L.), lentils (*Lens culinaris* M.), and lupin (*Lupinus albus* L. *and L. mutabilis*) by developing a new approach to conserve, manage, and characterize PGR. This involves generating, analyzing, and integrating extensive genotypic and phenotypic information by developing advanced databases, innovative visualization tools, and molecular approaches. The project also aims to improve gene bank standards for data management, establish a central data management infrastructure and an expert portal, and connect PGR with precise, high-quality information and homogenized passport data. Ultimately, these efforts will facilitate a comprehensive exploration of plant genetic resources and contribute to the sustainable use and management of theme

Genetic diversity in modern agriculture

The concept of PGR it is closely related to the concept of genetic diversity, which refers to the amount of genetic variability present among individuals of a variety, population, or species (Brown, 1983; Rao & Hodgkin, 2002; Salgotra & Chauhan, 2023). Genetic diversity is driven by two fundamental processes: mutation and recombination. Mutations are essential to introduce new alleles into a population's gene pool (Lynch et al., 2016). Recombination, occurring during sexual reproduction, facilitates the shuffling of these alleles, resulting in novel genetic combinations that enhance the variability (Felsenstein, 1974). In addition to these, different evolutionary forces act upon the alleles present in different populations, further influencing genetic diversity. Selection favors alleles that confer advantageous traits, enhancing an organism's reproductive success in a given environment (Wade & Kalisz, 1990). Genetic drift, particularly in small populations, can lead to random fluctuations in allele frequencies, potentially reducing genetic variation over time. Gene flow differently, can introduce new genetic material and lead to adaptive introgression (Bellucci et al., 2023). Collectively, these evolutionary forces interact to

shape the genetic landscape of populations, resulting from the accumulation of numerous genetic differences in DNA sequences among individuals within the populations over time and space.

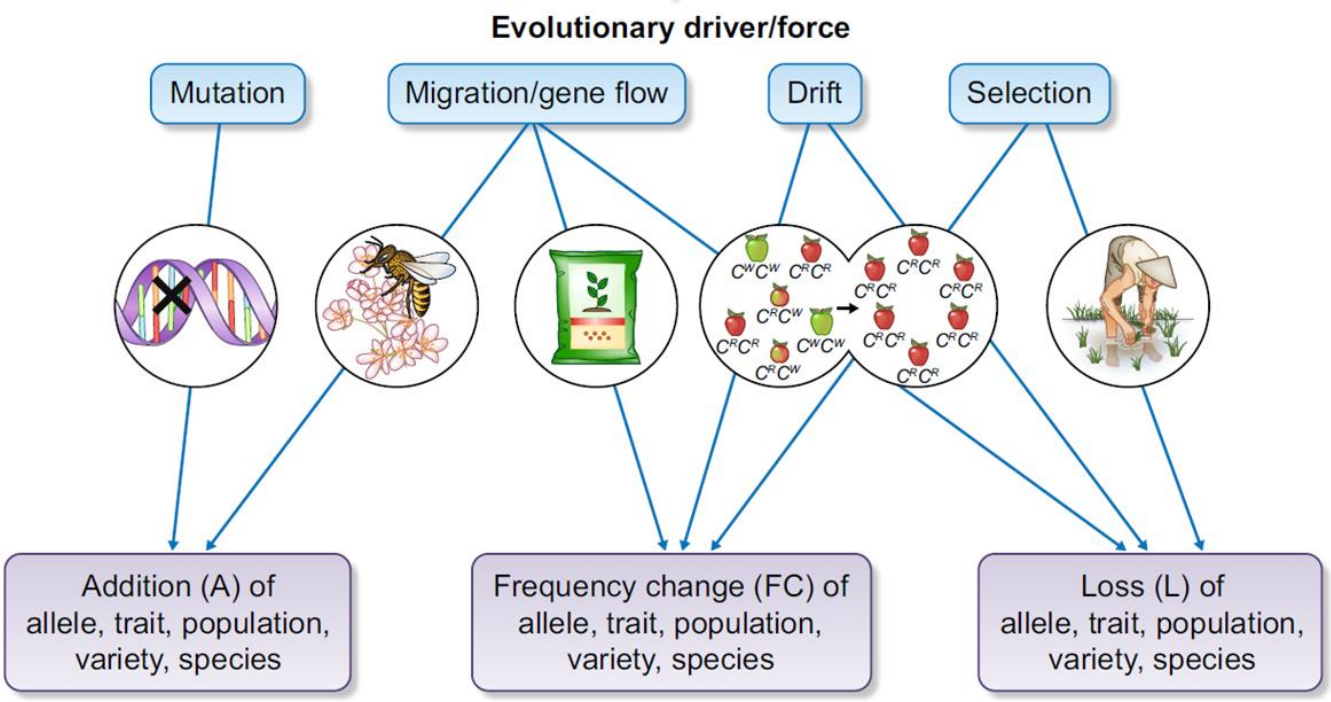


Figure 1. *The main evolutionary forces and their general impact on genetic variability extracted from C. K. Khoury et al., 2022..*

The forces considered include mutation, natural selection, genetic drift, and migration, each influencing allele frequencies within populations over time.

Today, agriculture faces a significant decline in the genetic diversity of cultivated crops (Hoban et al., 2020). In addition to the reduction in species diversity, plant breeding has contributed to decreased genetic variation within species (Gepts, 2006; Condón et al., 2008). This reduction also occurs through developing highly adapted breeding populations, selecting the most favorable genotypes, and promoting a few widely grown homogeneous cultivars (Viana et al., 2022). The development of homogeneous cultivars was followed by a gradual replacement of landraces and significantly contributed to the

erosion of genetic diversity (Van De Wouw et al., 2010). In traditional agroecosystems, genetically heterogeneous landraces were often cultivated in mosaics of different varieties and species, creating spatial diversification that served as a safeguard against genetic losses (Zeven, 1998). This diversity was maintained through farmers' cultivation and selection practices, with local exchange and gene flow between landraces promoting genetic variation, and continuous cultivation driving local adaptation (Thomas et al., 2011). However, the introduction of modern cultivars, selected for uniform traits, has reduced this variability, further exacerbating “genetic erosion” alongside other anthropogenic factors (Rogers L., 2004). More generally, the reduction in both inter- and intraspecific genetic diversity can result in several risks, including Increased genetic vulnerability to pests and diseases (Zhu et al., 2000), lack of adaptation to environmental stressors, and limited genetic variability for important traits (Hajjar et al., 2008). Another risk associated with the decline in genetic diversity is reaching performance plateaus, where crops no longer improve despite continued selection efforts. The 1960s shift in global agricultural practices through the implementation of the phenomenon that has become known as the 'Green Revolution' (Goron & Raizada, 2015). This systematic transformation was characterized by the widespread adoption of high-yielding cultivars and modernized agricultural methodologies. The initiative was predominantly advocated as a dual solution: first, as a catalyst for economic development in agrarian societies, and second, as a mechanism for maintaining sociopolitical stability through enhanced food security. However, subsequent empirical observations began to reveal the unintended consequences of this agricultural intensification. Among themes has been observed a significant reduction in genetic diversity and genetic vulnerability (Smale, 1997).

Domestication and Its Impact on Plant Genetic Diversity

The genetic diversity of cultivated plants has been significantly impacted by the process of domestication (Gupta, 2004; Reif et al., 2005; Bitocchi et al., 2017). Domestication represents a co-evolutionary pathway driven by human selection, characterized by

progressive morphological and physiological changes that distinguish cultivated species from their wild ancestors (Fuller, 2007). This complex biological transformation stands as one of humanity's most profound technological achievements, serving as the foundational catalyst for the Neolithic Revolution approximately 13,000-10,000 years ago (Kislev et al., 1992; Weisdorf, 2003). This transition facilitated the shift of hunter-gatherer societies into sedentary agricultural communities, marking a watershed moment in human cultural evolution (Harlan, 1975). The emergence of reliable food surpluses through domesticated species precipitated a cascade of sociocultural innovations. The process of plant domestication occurred independently in at least 24 regions (***Figure 2***) around the world and is marked by significant changes in plant traits driven by human intervention (Purugganan & Fuller, 2009).

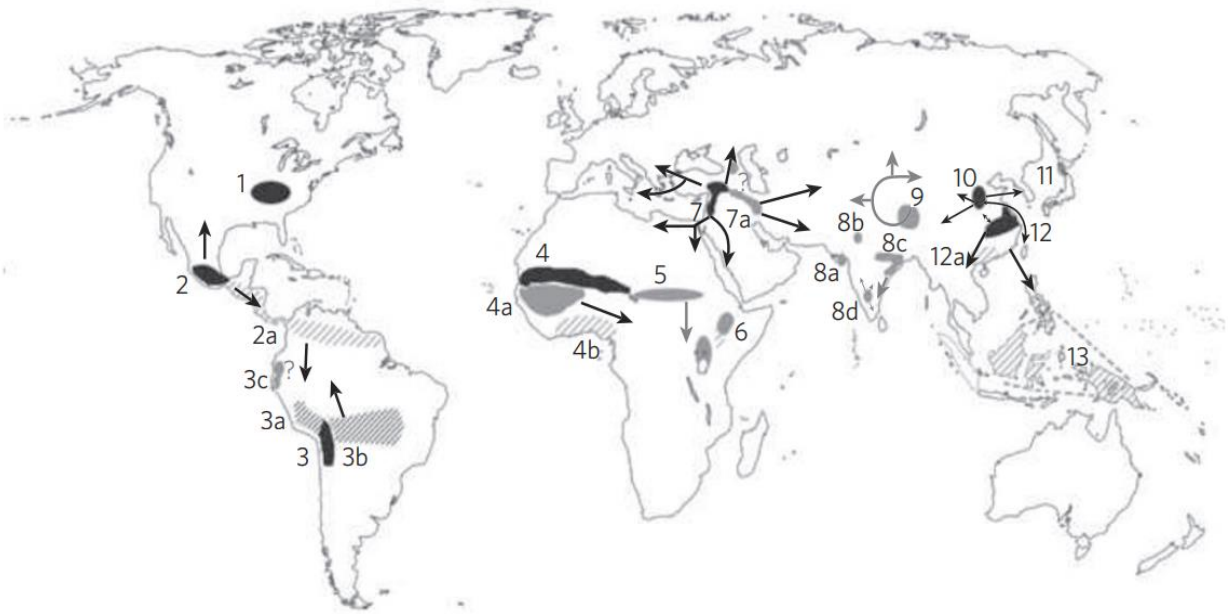


Figure 2. *Centres of plant domestication extracted from Purugganan & Fuller, 2009.*

"Solid-shaded areas and hatched areas indicate regions of important seed-crop domestication and vegecultural crops, respectively. Accepted primary domestication centres are shown in black, and potentially important secondary domestication centres are shown in grey. Arrows indicate major trajectories of spread of agriculture and crops out of some centres. Areas are numbered, and examples of crop species and the year by which they were domesticated in each area are as follows: 1, eastern North America (Chenopodium berlandieri, Iva annua and Helianthus annuus, 4,500–4,000 years before present (yr bp)); 2, Mesoamerica (Cucurbita pepo, 10,000 yr bp; Zea mays, 9,000–7,000

yr bp); 2a, northern lowland neotropics (Cucurbita moschata, Ipomoea batatas, Phaseolus vulgaris, tree crops, 9,000–8,000 yr bp); 3, central mid-altitude Andes (Chenopodium quinoa, Amaranthus caudatus, 5,000 yr bp); 3a, north and central Andes, mid-altitude and high altitude (Solanum tuberosum, Oxalis tuberosa, Chenopodium pallidicaule, 8,000 yr bp); 3b, lowland southern Amazonia (Manihot esculenta, Arachis hypogaea, 8,000 yr bp); 3c, Ecuador and northwest Peru (Phaseolus lunatus, Canavalia plagiosperma, Cucurbita ecuadorensis, 10,000 yr bp; the question mark indicates that there is some question of the independence of crop origins of this centre from 3, 3a and 3b); 4, West African sub-Sahara (Pennisetum glaucum, 4,500 yr bp); 4a, West African savanna and woodlands (Vigna unguiculata, 3,700 yr bp; Digitaria exilis, Oryza glaberrima, 4,000 yr bp?); 6, East African uplands (Eragrostis tef, Eleusine coracana, 4,000 yr bp?) and lowland vegeculture (Dioscorea cayenensis, Ensete ventricosum, poorly documented); 7, Near East (Hordeum vulgare, Triticum spp., Lens culinaris, Pisum sativum, Cicer arietinum, Vicia faba, 13,000–10,000 yr bp); 7a, eastern fertile crescent (additional Hordeum vulgare and, 9,000 yr bp, also goats); 8a, Gujarat, India (Panicum sumatrense, Vigna mungo, 5,000 yr bp?); 8b, Upper Indus (Panicum sumatrense, Vigna radiata, Vigna aconitifolia, 5,000 yr bp); 8c, Ganges (Oryza sativa subsp. indica, 8,500–4,500 yr bp); 8d, southern India (Brachiaria ramosa, Vigna radiata, Macrotyloma uniflorum, 5,000–4,000 yr bp); 9, eastern Himalayas and Yunnan uplands (Fagopyrum esculentum, 5,000 yr bp?); 10, northern China (Setaria italica, Panicum miliaceum, 8,000 yr bp; Glycine max, 4,500 yr bp?); 11, southern Hokkaido, Japan (Echinochloa crusgalli, 4,500 yr bp); 12, Yangtze, China (Oryza sativa subsp. japonica, 9,000–6,000 yr bp); 12a, southern China (Colocasia, Coix lachryma-jobi, poorly documented, 4,500 yr bp?); 13, New Guinea and Wallacea (Colocasia esculenta, Dioscorea esculenta, Musa acuminata, 7,000 yr bp).”

Archaeological evidence reveals that early hunter-gatherer groups gradually transitioned to cultivating food plants, leading to profound modifications in various species depending on their use (Wright, 1994). Domestication traits vary considerably across plants, largely influenced by whether they are grown for fruits, vegetative organs, or seeds. Together, these changes underscore how human cultivation has shaped the evolutionary pathways of domesticated plants, ultimately driving their divergence from their wild ancestors (Özkan et al., 2011). These changes involve significant morphological and physiological alterations including also the mode of reproduction, yield of edible portion, ease of harvesting, defensive adaptations, timing of production, and plant architecture (Denham et al., 2020). For example, The domestication of wild grass teosinte (*Zea mays* ssp. *parviglumis*) into modern maize (*Z. mays* ssp. *mays*) involved the selection of key traits

such as larger ears and seeds and the suppression of axillary branching. An important evolutionary step in this process was the release of the kernel from the hardened casing that originally encased it in teosinte (H. Wang et al., 2005). This modification exposed the kernel on the ear's surface, making it more accessible for human consumption and marking a critical moment in maize domestication. As a result, domesticated organisms undergo structural and functional modifications that distinguish them from their wild ancestors, this is referred to as 'domestication syndrome', where traits that increase competition and minimize individual production are selected against in favor of traits that reduce competition and enhance community-level productivity (Di Vittori et al., 2017).

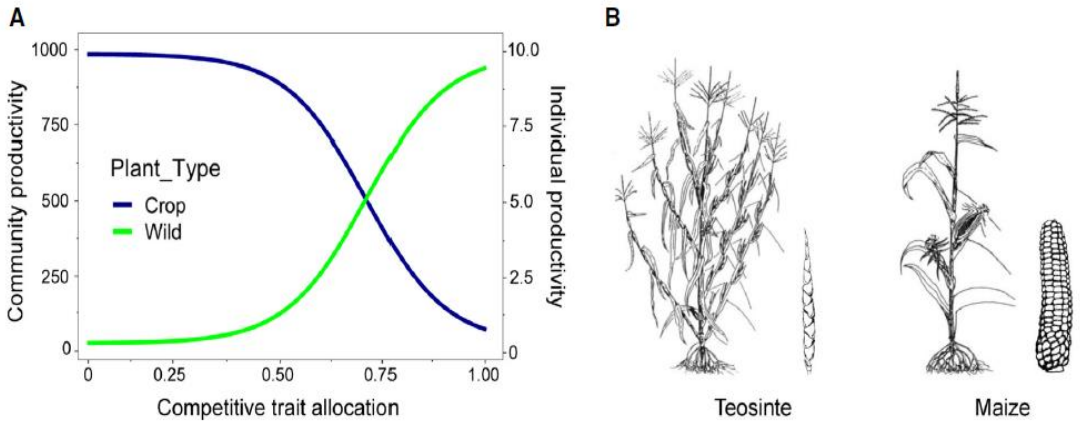


Figure 3. *The relationship between competitive trait allocation and productivity in wild versus crop plants, from Gao et al., 2023.*

(A) *The productivity of individual wild plants increases with greater competitive trait allocation, but community productivity declines due to resource limitations (green line). Conversely, reduced competitiveness in crops permits more efficient resource utilization and increased group productivity (dark-blue line). The x axis represents competitive trait allocation (ranging from 0 to 1), where higher values indicate increased competitiveness. The y axis represents yield in kg/ha (for details of logistic model, see supplemental information).* (B) *The many branches of teosinte make it a better competitor in the wild than its cultivated derivative, maize, which has fewer branches and thus can be grown at much higher densities in cultivated fields.*

This predictable pattern of traits shared across different species highlights how domestication exerts a common evolutionary pressure, shaping cultivated forms in similar ways (Di Vittori et al., 2019). Another notable consequence is the "domestication bottleneck" of genetic diversity, which refers to the drastic reduction in genetic variation

during the domestication process (Kumar et al., 2021). Z. M. Li et al., 2011 analyzed nucleotide variation in 14 nuclear genes in species-wide samples of African rice. The authors observed that African rice (*O. glaberrima*) retained only 30% of the genetic diversity detected in its wild ancestor *O. barthii*. These genetic bottlenecks are not limited to the initial domestication phase but continue in modern crops ***Figure 4A***, with the diffusion of crops to other locations outside of their centers of origin.

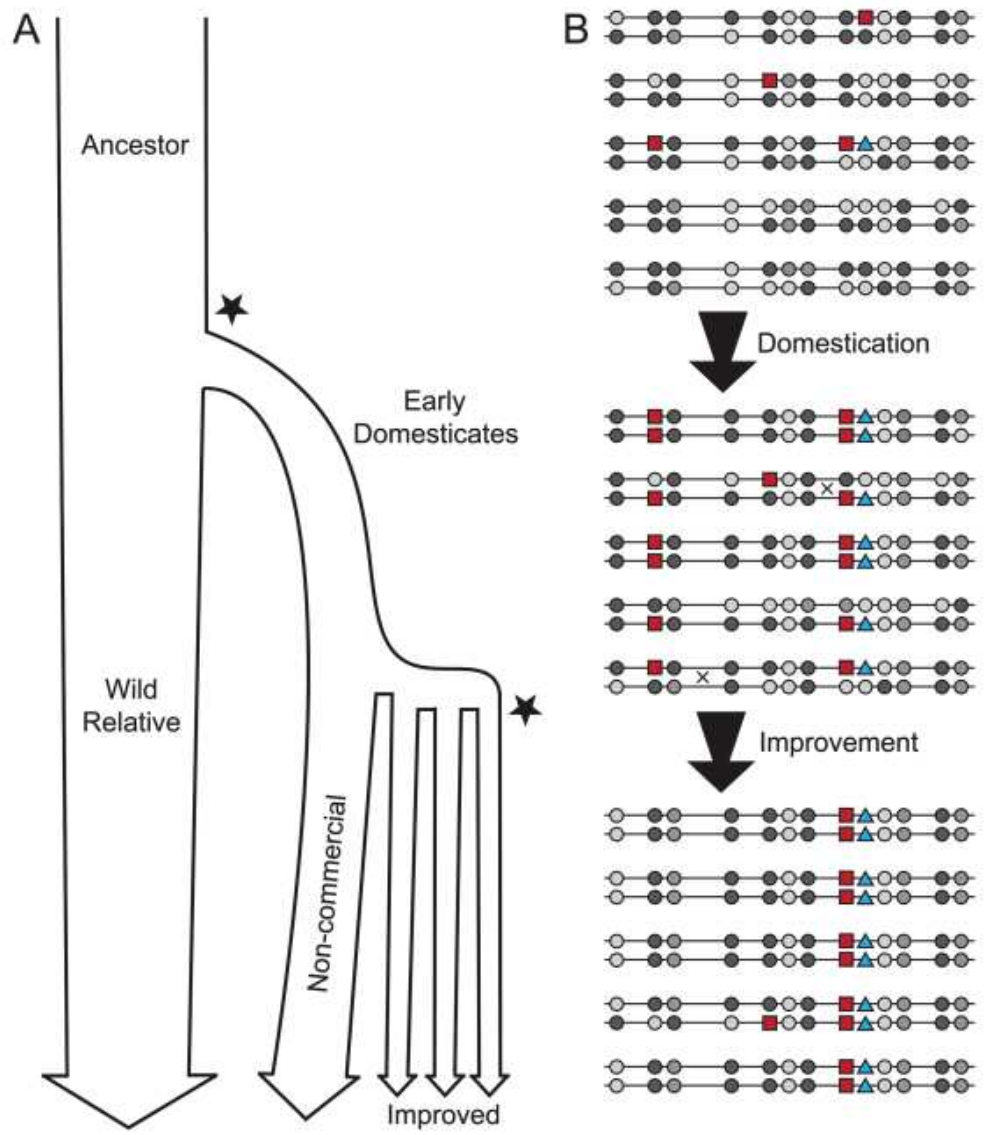


Figure 4. *Processes of domestication and improvement from Moyers et al., 2018.*

(A) Typical changes in effective population size through domestication and improvement. Stars indicate genetic bottlenecks. These dynamics can be reconstructed by examining

patterns of genetic diversity in contemporary wild relative, domesticated noncommercial, and improved populations. (B) Effects of artificial selection (targeting the blue triangle variant) and linkage disequilibrium on deleterious (red squares) and neutral variants (grey circles, shades represent different alleles). In the ancestral wild population, 4 deleterious alleles are at relatively low frequency (mean = 0.10) and heterozygosity is high (H0 = 0.51). After domestication, the selected blue triangle and linked variants increase in frequency (3 remaining deleterious alleles, mean frequency = 0.46), heterozygosity decreases (H0 = 0.35), and allelic diversity is lost at 2 sites. Recombination may change haplotypes, especially at sites less closely linked to the selected allele (Xs). After improvement, further selection for the blue triangle allele has: lowered heterozygosity (H0 = 0.08), increased deleterious variant frequency (2 remaining deleterious alleles, mean = 0.55), and lost allelic diversity at 6 additional sites. For color, please see the figure online.

Indeed, elite cultivars are often selected from previous domestication events to enhance crop improvement, where the diversity loss can be mainly attributed to the limited diversity introduced and then the subsequent selective breeding process (Hyten et al., 2006; Van De Wouw et al., 2010).

The reduction in genetic diversity during domestication and the subsequent selective breeding process also leads to an accumulation of deleterious mutations (Liu et al., 2017). This accumulation occurs in different mechanisms; thanks to the effect of genetic drift, which becomes stronger relative to selection due to reduced effective population size, and genetic hitchhiking ***Figure 4B***, where deleterious variants linked to artificially selected beneficial alleles increase in frequency despite their negative fitness effects (Moyers et al., 2018). However, the transition from outcrossing to selfing mating systems increases the recombination, leading to the purging of highly deleterious recessive alleles due to their increased exposure in homozygous states (Arunkumar et al., 2015). These mechanisms are collectively known as the “cost of domestication” and can constrain selection efficiency, consequently limiting breeding gains in domesticated species and reducing the effectiveness of breeding programs (Lu et al., 2006).

PGR conservation

Plant genetic resources (PGR) represent plants and their alleles contained with actual or potential economic, scientific, or social value for humanity (Rao & Hodgkin, 2002). Their conservation and sustainable use are crucial for future food security, providing the genetic diversity necessary for developing new cultivars with improved yields, quality, and resistance to both biotic and abiotic stresses (C. Khoury et al., 2010; C. K. Khoury et al., 2022). Two fundamental concerns have emerged regarding PGR, their finite nature requiring conservation, and the increased crop vulnerability due to genetic uniformity. To mitigate these issues, both *in-situ* and *ex-situ* conservation strategies have been implemented. In situ conservation maintains genetic resources in their natural habitat, allowing evolutionary forces to impact genetic diversity. This process can lead to the development of landraces, varieties that tend to be more resilient, with enhanced capacity to tolerate both biotic and abiotic stresses, resulting in high yield stability and intermediate yield levels under low-input agricultural systems (Sanchez et al., 2023; Zeven, 1998). Differently, *Ex-situ* conservation involves preserving genetic resources outside their natural habitats through various methods including seed gene banks, tissue culture, cryopreservation, and greenhouses. This strategy is particularly crucial for protecting endangered and overexploited genetic resources that face potential extinction due to habitat destruction, climate change, and human interference. Among the various *ex-situ* techniques, seed storage represents the most efficient method for long-term conservation. The effective utilization of these conserved resources requires proper characterization, which can be conducted through morphological assessment or molecular approaches (Laurentin, 2009). While morphological characterization has limitations due

to environmental influences and the limited number of observable traits, molecular characterization, particularly DNA-based techniques, offers more reliable and comprehensive genome analysis, independent of environmental factors.

Local Adaptation: A Driving Force Behind Plant Diversity

Previously (see “Domestication and Its Impact on Plant Genetic Diversity”) we described how crops originating from different centers of domestication and spread globally. The adaptation of the agro-environment that occurred in plants to become cultivation was fundamental to the transition from hunter-gatherer societies to agricultural civilizations, ultimately enabling urban development and modern civilization. Nowadays, these crops face unprecedented challenges due to severe climate changes that are taking place (Calleja-Cabrera et al., 2020). Indeed in recent decades, anthropogenic climate change has manifested through increased greenhouse gas emissions and more frequent extreme weather events such as droughts, heatwaves, and storms. According to the Intergovernmental Panel on Climate Change (IPCC), 2023: Summary for Policymakers (Calvin et al., 2023), measurements from 2011-2020 show that global temperatures have increased by 1.1°C above 1850-1900 levels. Scientific projections indicate we are approaching a critical threshold, with temperatures expected to reach 1.5°C in the coming years, a level that will further intensify multiple climate hazards and risks worldwide. These environmental pressures are significantly impacting crops' morphology, physiology, phenology, production capacity, distribution ranges, and species richness (Chapman et al., 2012; Raza et al., 2019).

Local adaptation is a typical characteristic of natural plant populations, a trait also found in crop varieties, especially in landraces that have evolved in situ while maintaining genetic exchange with local wild populations (Broccanello et al., 2023; Mercer & Perales, 2010). However, modern breeding programs have shifted focus toward developing cultivars with broad environmental adaptability, potentially reducing local

adaptation strength. The genomic basis of adaptation can be investigated through different approaches (Ross-Ibarra et al., 2007). The Genome-environment Association (GEA) analyses, allowed to map loci contributing to environmental adaptation by utilizing environmental parameters as response variables (Cortés et al., 2022). This methodology parallels traditional genome-wide association studies (GWAS) but substitutes conventional trait measurements with environmental parameters, such as climate and soil characteristics from the population's origin site. This approach proves particularly valuable for identifying alleles contributing to stress resistance, especially for cryptic physiological features or traits that are challenging to phenotype conventionally. An alternative approach is based on the identification of genes showing genetic adaptation signatures (Ross-Ibarra et al., 2007). Selection signatures in the genome can be detected through several distinct patterns ***Figure 5***. These include local reduction in genetic variation referred to as selective sweep, altered polymorphism frequency distributions, enhanced linkage disequilibrium, and increased genetic differentiation (Gao et al., 2023).

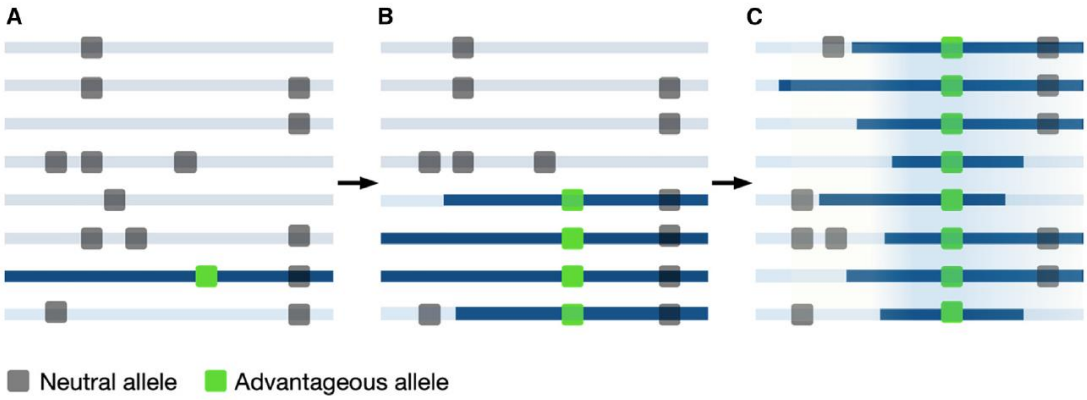


Figure 5. *The footprint of natural selection. From Gao et al., 2023*

A selective sweep is a process whereby a beneficial mutation (green) rapidly increases in frequency in a population experiencing positive selection. (A) Eight chromosomal regions (light-blue rectangles) with segregating neutral alleles (gray boxes). The dark-blue rectangle is the chromosome carrying the adaptive allele (green box). This adaptive allele provides a selective advantage to the individuals carrying it. (B) During a selective sweep, the beneficial mutation (green) increases in frequency. Owing to recombination with other chromosomes in the population, only some nearby linked alleles hitchhike with

the advantageous allele. C) The selective sweep has been completed, and the advantageous mutation (green) is now highly frequent in the population. Reduced recombination between the adaptive and linked alleles creates a selective sweep signature in the chromosomal region where diversity is low, linkage disequilibrium is high, and Tajima's D is negative (shared area).

While these genomic signatures provide crucial insights into adaptive processes, their interpretation requires careful consideration. Indeed, similar genomic patterns can also result from demographic factors, such as population bottlenecks, genetic admixture, and population structure (Booker et al., 2021; Ross-Ibarra et al., 2007). The practical implications of these genomic insights are substantial for crop improvement strategies, as selected alleles can be incorporated into cultivated germplasm through marker-assisted or genomic selection approaches, potentially enhancing crop resilience to climate change. This integration of evolutionary understanding with modern genomic tools offers promising pathways for developing climate-resilient varieties capable of sustaining agricultural productivity in the face of growing environmental challenges.

Adaptive role of structural variation

Structural variants (SVs) have emerged as crucial contributors to functional genomic variation. These variants, defined as genomic alterations covering more than 1 kb in length, encompass a spectrum of modifications including insertions/deletions, inversions, translocations, and copy number variations (CNVs). SVs can arise through various cellular mechanisms during both meiotic and mitotic cell division and can be led by recombination, replication, or DNA break repair errors (Gorkovskiy & Verstrepen, 2021). The classification of SVs can be approached through their impact on DNA content balance. Balanced SVs, such as inversions and translocations, maintain the overall genetic information while rearranging its position within or between chromosomes. In contrast, unbalanced SVs alter the genomic content through deletions, insertions, or duplications,

leading to copy number variations. These structural alterations can occur throughout the genome, affecting both coding and non-coding regions, including highly repetitive elements like transposons. Notably, certain forms of SVs, particularly large inversions, can significantly influence genomic dynamics by reducing recombination rates (Galludo et al., 2018; Hämälä et al., 2021). This reduction in recombination can facilitate the maintenance of adaptive polymorphisms in Linkage Disequilibrium LD (Rane et al., 2015). This has important implications for evolution and adaptation, as it allows multiple beneficial alleles to be inherited together as a single genetic unit (Kapun et al., 2016, 2023).

Molecular characterization of Plant genetic resources

The field of genotyping has undergone remarkable advancement over the past decades. This progression represents a fundamental shift in our ability to analyze genetic variation, with each new technological phase bringing significant improvements in accuracy, throughput, and applications (Kockum et al., 2023). Molecular techniques such as Restriction Fragment Length Polymorphism (RFLP) analysis and PCR-based methods, introduced in the mid-1980s, represented a significant innovation in molecular genetics (Williams, 1989). Microarray technology marked the next substantial leap in genotyping capabilities, introducing unprecedented levels of throughput in variant analysis. Traditional solid-phase arrays with fixed probe locations were initially developed, followed by more advanced bead-based systems offering increased probe density. These platforms revolutionized genetic analysis by enabling the simultaneous examination of numerous variants while significantly reducing the cost per data point (Gunderson et al., 2005). Contemporary genotyping has been further revolutionized by sequencing-based approaches. The advent of Next-Generation Sequencing (NGS) has dramatically improved these capabilities through massively parallel analysis, significantly reduced per-base costs, and enhanced throughput and accuracy (Gunderson et al., 2005). These advanced methodologies facilitate large-scale population genomics studies, enable

comprehensive genome-wide association analyses, and provide unprecedented insights into genetic diversity. The progression toward sequencing-based methodologies has markedly reduced ascertainment bias compared to traditional genotyping approaches, enabling more accurate and comprehensive genetic analyses. The continuous advancement in genotyping technologies has enabled the implementation of large-scale genetic studies across extensive cohorts (Bonin et al., 2004). An example of that is the EU H2020 INCREASE project. Such large-scale initiatives demonstrate how modern genotyping approaches facilitate the comparative analysis of extensive germplasm collections with sufficient statistical power. These technological developments have proven particularly valuable for projects requiring cost-effective, high-throughput genotyping methods to process large sample numbers in an economically feasible manner, thereby advancing both fundamental research and practical applications in crop improvement programs.

Genomic Complexity Reduction Approaches

Whole Genome Sequencing (WGS) represents the most comprehensive form of genetic analysis available today, enabling the determination of complete DNA sequences. However, while WGS provides exhaustive genomic information, its application faces significant challenges when studying species with large, highly repetitive genomes across numerous samples. This is especially true for applications that require a large number of molecular markers primarily due to economic constraints and computational demands (Ganal et al., 2011). The challenge becomes particularly acute when dealing with organisms possessing extensive genomic content. This is evident in some legume crops, such as lentil (*Lens culinaris* M.), pea (*Pisum sativum*), and faba bean (*Vicia faba*), which are characterized by large genome sizes (Cooper et al., 2017; Kreplak et al., 2019; Ramsay et al., 2021). In these species, the genomes often contain substantial portions of repetitive sequences, which are typically under less selective pressure than non-repetitive, coding regions. These repetitive sequences, while part of the genome's architecture, are often not the primary focus of most genotyping projects, suggesting the need for more targeted

approaches. Various alternative methods have been developed to achieve cost-effective high-throughput genotyping, enabling the processing of increased sample numbers in an economically feasible manner while maintaining high accuracy rates (Bonin et al., 2004). Single-Nucleotide Polymorphism (SNP) microarray-based platforms represent one such alternative to WGS. However, these arrays are limited to detecting known SNPs with high population frequency, potentially missing rare or novel variants that could be of biological significance. To address this limitation, several reduced-representation sequencing methods have emerged. These include Genotyping by Sequencing (GBS) (Elshire et al., 2011), Restriction Site-associated DNA Sequencing (RAD-Seq) (Gamble et al., 2015), and double-digest RAD-Seq (ddRADseq) (Burns et al., 2017). These approaches leverage restriction endonucleases' specificity to concentrate sequencing efforts on specific genomic regions, thereby reducing costs while maintaining analytical power (Peterson et al., 2012). Differently, Whole Exome Sequencing (WES) has also emerged as a cost-effective alternative to WGS, focusing specifically on coding regions that comprise a limited region of the genome. While this approach maintains high efficiency in detecting functional variants, it shares similar limitations with microarray-based methods regarding the detection of novel variants (Belkadi et al., 2015). Recent innovations have introduced even more sophisticated approaches, notably the application of CRISPR-Cas9 technology for selective depletion of unwanted genome fractions from sequencing libraries (Rossato et al., 2023a). This system employs the Cas9 enzyme, programmed through specific guide-RNA sequences, to eliminate unwanted sequences from the library. This technology has been further refined and commercialized as "CRISPRclean" by Jumpcode Genomics (2021), offering new possibilities for targeted sequencing approaches.

The genus Lens

Origin and domestication

The genus *Lens* belongs to the family Fabaceae (Schaefer et al., 2012). The classification of the Lens genus has long been subject to debate. Earlier classifications recognized up to seven distinct taxa, including the cultivated lentil *L. culinaris* Medik. subsp. *culinaris*, its wild progenitor, and several other species (Van Oss et al., 1997). The current, widely accepted classification recognizes four species: *L. culinaris* (encompassing four subspecies: subsp. *culinaris,* subsp. *orientalis*, subsp. *odemensis*, and subsp. *tomentosus*), *Lens lamottei*, *Lens ervoides*, and *Lens nigricansn* (Ferguson et al., 2000; Koul et al., 2017; Wong et al., 2015). The wild *Lens* species are primarily distributed across the Mediterranean region, extending from Southern Europe and North Africa to parts of the Middle East, with some species, like *L. culinaris* ssp. *orientalis* reaching Central Asia and *Lens ervoides* extending to East Africa (Sonnante et al., 2009a). Numerous early indications strongly suggests that *L. culinaris* subsp. *orientalis* is the wild progenitor of the cultivated *L. culinaris* subsp. *Culinaris* (Ladizinsky, 1993; Zohary, 1972). This hypothesis is supported by archaeological findings, with the oldest carbonized lentil remains discovered in the Franchthi cave in Greece, dating back to approximately 11,000 BC, and in Tell Mureybit, Syria, from around 8,500-7,500 BC ***Figure 6*** (Hansen & Renfrew, 1978; Sonnante et al., 2009a; Zohary, 1972) . Carbonized lentil remains from 7th millennium BC settlements in the Near East do not conclusively indicate their domestication status, as wild types and small-seeded cultivated forms are virtually indistinguishable (Sonnante et al., 2009a). However, a significant discovery at Yiftah-el in Israel provides compelling evidence of early agricultural practices (Caracuta et al.,

2017; Kislev ME., 1985). A substantial deposit of lentils, dated to approximately 6,800 BC, was unearthed at this site providing evidence of an established agricultural system.

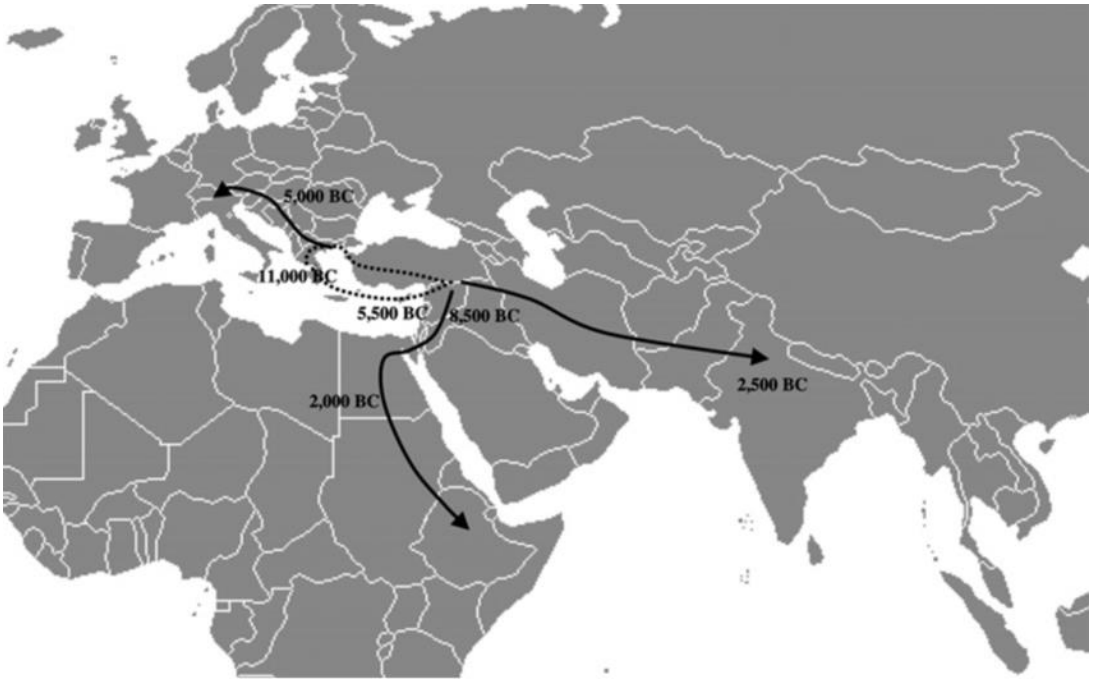


***Figure 6**. Early diffusion of domesticated lentil (Lens culinaris) reconstructed through archaeological evidence from Sonnante et al 2009.*

Molecular-marker-based studies confirmed that *Orientalis* is the sole wild progenitor of cultivated lentil (Dissanayake et al., 2020; Liber et al., 2021; Pavan et al., 2019). Indeed, a genetic bottleneck has been revealed in Southwest Asia, between southern Turkey in the north and Jordan in the south, that led to the emergence of the cultivated gene pool from a single sub-population of wild *orientalis* approximately 10,000 years ago (Liber et al., 2021). Following its domestication lentils underwent a gradual and incremental diffusion to South Asia, Iran, Central Asia, Eastern Africa, the Mediterranean, and the Americas, potentially experiencing introgression and hybridization with wild relatives

which contributed to shaping its genetic diversity and population structure (Guerra-Garcia et al., 2022).

Cultivated Lentil

Cultivated Lentil (*Lens culinaris* M. subsp. *culinaris*) is one of the earliest domesticated crop legumes of significant nutritional and agricultural importance. Global production of lentils has seen substantial growth over the last fifteen years, reaching approximately 6.5 million tons in 2022 (*FAOSTAT*, 2024), highlighting its increasingly important role in the global economy and trade networks. Lentils have garnered increased interest from both environmental sustainability and food security perspectives (Warne et al., 2019). Indeed, Lentils are renowned as an important source of plant-based protein, offering high fiber and low-fat content (Jarpa-Parra, 2018; Tosh & Yada, 2010). Additionally, lentil cultivation, like that of other legumes, offers the added benefit of enhancing soil fertility through nitrogen fixation, which can reduce the need for nitrogen fertilizer use in agricultural systems (Abi-Ghanem et al., 2011). The crop is broadly categorized into two varietal types based on morphological traits: the large-seeded macrosperma and the small-seeded microsperma (Chahota & Sharma, 1993), each exhibiting a wide range of seed colors and nutrient compositions (Palacios et al., 2023; Rawal & Navarro, 2019). Lentil cultivation has undergone significant geographical shifts over the past three decades. Currently, Canada leads global production, contributing approximately 40% of the world's lentils, followed by India at 22%. Other significant producers include Turkey, Australia, Nepal, and the United States (Joshi et al., 2017; Rawal & Navarro, 2019). This shift reflects not only changes in agricultural practices but also in global trade patterns and the increasing importance of lentils in international markets.

The Lentil genome

The lentil genome is characterized by large size and a remarkably high percentage of repetitive sequences making both its assembly and characterization significant challenges (Rossato et al., 2023a). Recently, a high-quality genome assembly of CDC Redberry (Ramsay et al., 2021), a small red lentil variety bred at the University of Saskatchewan, Canada, has been released as a reference. The assembly, generated using 54x long reads and polished with both long and short reads, covers 3.76 Gbp of the estimated 3.92 Gbp genome size (based on K-mer distribution analysis), closely aligning with the 4.06 Gbp estimate obtained from flow cytometry for this species. A striking feature of the CDC Redberry genome is its high repeat content, with repetitive elements constituting 82.6% (3.2 Gb) of the estimated 3.9 Gb assembly. This proportion is significantly higher than other diploid legume assemblies such as Medicago, chickpea, and common bean (Parween et al., 2015; Schmutz et al., 2014; Tang et al., 2014), but it's comparable to the pea genome (Kreplak et al., 2019). The repetitive regions observed in the genome assembly primarily consist of Class I transposable elements (83.6%), mostly Long Terminal Repeat (LTR) retrotransposons of the Ty3/gypsy (63.8%) and Ty1/copia (15.4%) types (Ramsay et al., 2021).

Italian lentil landrces form Castelluccio di Norcia

The cultivation of Castelluccio di Norcia lentils represents a significant case study in agrobiodiversity conservation and geographical indication protection in the Apennine region of central Italy. This landrace, which received Protected Geographical Indication (PGI) status from the European Union in 1997, is cultivated at approximately 1,300 masl in a distinct geographical area of 20 km² within the Sibillini Mountains National Park (European Commission, 1997). The cultivation zone, characterized by an open-field agricultural landscape devoid of physical boundaries between plots, exhibits specific

climatic conditions that have contributed to the development of the phenotypic traits in the local lentil variety. These traits include reduced seed diameter (approximately 2 mm), heterogeneous seed coloration ranging from streaked green to light brown (Ceccobelli et al., 2019), enhanced resistance to cold stress, drought and salt tolerance, and pest resistance, along with elevated iron content and high cooking properties (Acito et al., 2023; M. Sidari et al., 2008; Muscolo et al., 2014). Molecular and biochemical characterization studies have provided evidence for genetic differentiation among Italian landraces (Scippa et al., 2010; Sonnante et al., 2009b; Sonnante & Pignone, 2001; Torricelli et al., 2012). Notably, (Zaccardelli et al., 2012) employed 16 microsatellite loci to assess genetic diversity, demonstrating that Castelluccio di Norcia genotypes cluster according to their geographical origin, supporting the genetic identity of this germplasm. The agricultural system has evolved significantly over time, transitioning from diverse crop rotation practices to a specialized lentil cultivation system, with cultivated areas expanding from 250 to 500 hectares between 1998 and the present day (Fatichenti F, 2019). This transformation reflects both market dynamics and the recognition of the crop's economic potential. More generally, the landrace's distinctive characteristics result from dual selection pressures: anthropogenic selection through traditional farming practices and environmental adaptation to specific agroecological conditions. Current on-farm conservation practices are maintained primarily due to the premium market value of the local variety compared to imported alternatives from Canada, the United States, and Turkey (Verza et al., 2024). This case exemplifies the complex interaction between environmental factors, human selection, and market forces in maintaining agricultural biodiversity in marginal mountain ecosystems, with genetic evidence supporting the uniqueness and the importance of the local germplasm.

Herbaria and Ancient Samples in Genomics Studies

In the previous paragraphs, we introduced the importance of studying and understanding genetic diversity and how understanding the evolutionary history of agronomic species is crucial for their conservation and proper exploitation. The study of how genetic diversity has evolved over time lies at the heart of paleogenomics. Researchers in paleogenomics employ two main approaches to investigate evolutionary changes. The first approach utilizes reference genomes from existing species to reconstruct ancestral sequences and structures across extended evolutionary timescales (Frichot et al., 2014). The second methodology focuses on direct observation of genomic modifications over shorter periods through the analysis of ancient DNA preserved in subfossil specimens (Trucchi et al., 2021). The integration of these approaches has provided insights into the evolutionary mechanisms that have shaped contemporary genome organization and regulation (Pont et al., 2019). In this context *Herbaria*, traditionally developed as repositories of preserved plant specimens, represent invaluable resources for understanding genetic and phenotypic diversity across time, having undergone significant historical evolution since their establishment (Bakker, 2018; Taylor & Swan, 1994a; W. Wang, 2018; Willis et al., 2017). Initially appearing as illustrated books documenting medicinal plants, they transformed into comprehensive repositories housing preserved specimens of plants, seeds, fungi, and algae at various phenological stages (Taylor & Swan, 1994a, 1994b). These specimens are typically preserved through drying on paper sheets or storage in folded packets or boxes, often treated with preservative agents such as solvents or pesticides. Some specimens may also be preserved in liquid solutions containing acids, aldehydes, alcohols, or other organic solvents. The establishment of modern herbaria in the 16th century was pioneered by Luca Ghini (Von Engelhardt, 2011). Today, herbaria

encompasses over 400 million specimens globally, facilitated by digital platforms that provide high-resolution images and associated data, enhancing accessibility and collaboration (Thiers et al., 2016) .Recent advances in paleogenomics and molecular biology, particularly in DNA extraction, amplification procedures, and next-generation sequencing strategies, have revolutionized the study of herbarium specimens. These technological developments have enabled researchers to obtain molecular markers and whole-genome sequences from centuries-old preserved specimens, significantly expanding their utility beyond traditional taxonomic and systematic studies .

The molecular analysis of herbarium specimens has proven instrumental in resolving ambiguous phylogenetic relationships and classifications (W. Wang, 2018). Through the study of ancient DNA from extinct species and historical samples, researchers can gain crucial insights into evolutionary histories and bridge significant gaps in our understanding of past biodiversity patterns (Särkinen et al., 2012). The molecular investigation of herbarium specimens has become particularly valuable in understanding species adaptation to novel geographic areas, as these processes typically involve both molecular adaptive mechanisms and significant phenotypic and phenological changes (Rosche et al., 2022). However, the use of *herbaria* for genomics is not without challenges. Different aspects are able to complicate the successful implementation of genomic studies using herbarium specimens. The biochemical postmortem damage in preserved samples represents a significant concern, as aDNA molecules undergo various forms of degradation over time. Environmental factors such as light exposure, uncontrolled moisture, temperature fluctuations, and exposure to reactive chemicals can lead to DNA strand breaks, base modifications, and DNA molecule crosslinking, all of which compromise sequence quality and integrity (Orlando et al., 2021). Furthermore, the limited quantity of recoverable aDNA from ancient specimens increases the risk of cross-contamination between samples and potential contamination from modern DNA during handling and processing procedures (Drabkova et al., 2002; Drábková, 2014). Advanced protocols for aDNA purification, alongside bioinformatics tools tailored to low-quality DNA, have been developed to address these issues, enabling the study of genetic diversity, evolutionary processes, and taxonomic relationships. Despite limitations, such as sample size and methodological biases in historical collections,

herbaria remains important resources for linking past and present genetic landscapes, complementing modern genotyping and phenotyping programs.

Research questions and objectives

This study addresses the critical challenge of characterizing genetic diversity in crops with large and highly repetitive genomes, specifically focusing on lentil (Lens culinaris M.), a crop of increasing importance in the context of global transition towards plant-based protein sources and food security. Within the framework of the INCREASE project, we aimed to overcome the limitations posed by complex plant genomes through the implementation of a novel molecular approach for genotypic characterization. Specifically, our objectives were to: (1) validate the first application of this innovative genotyping methodology in a crop species; (2) assess its effectiveness in characterizing locally-maintained heterogeneous material from farmers' collections; (3) investigate adaptation patterns in a high-altitude adapted lentil population; and (4) explore the potential of extending genetic diversity studies through the analysis of historical herbarium specimens.

The second chapter investigates a panel comprising 80 lentil accessions from Castelluccio di Norcia and 25 reference accessions from the INCREASE project. The research addresses three main questions:

Q1. Is the CRISPR-Cas9-based genomic complexity reduction approach effective for genotyping large lentil populations?

Q2. What is the extent and structure of genetic diversity preserved within the Castelluccio di Norcia lentil population?

Q3. What are the genetic signatures of local adaptation within the Castelluccio di Norcia population?

These research questions aim to validate our novel molecular approach while providing insights into the genetic architecture and adaptive patterns of a historically significant

Italian landrace maintained at high altitudes. The inclusion of INCREASE project reference accessions enables contextualization of the Castelluccio di Norcia population within the broader spectrum of lentil genetic diversity.

The third chapter explores the potential of expanding our understanding of temporal changes in genetic diversity through the integration of herbarium specimens in molecular studies. This approach offers a window into the historical genetic composition of crop populations. However, several challenges need to be addressed to successfully implement such studies.

The chapter evaluates multiple aspects, beginning with material preservation and characteristics. This includes assessment of *herbarium* specimen preservation status, understanding the specific physical and chemical properties of historical plant materials, and evaluation of degradation patterns and their impact on molecular analyses. Subsequently, the chapter addresses molecular manipulation challenges, focusing on DNA extraction optimization for degraded historical samples, and adaptation of genotyping approaches for fragmented historical DNA. Finally, the chapter examines analytical considerations, encompassing data quality assessment and validation procedures, integration of historical data with contemporary genetic information, and addressing biases and limitations in temporal genetic analyses.

Transition to Chapter II

(*And methodological notes*)

As previously described, this thesis focuses on the characterization of genetic diversity in lentil populations through an innovative molecular approach designed to overcome the challenges posed by complex genomes. The first chapter specifically addresses the validation of this methodology and its application to characterize the high-altitude adapted lentil landrace from Castelluccio di Norcia. Before proceeding with the detailed description of the research, it is important to note that the molecular approach employed here represents a novel application of CRISPR-Cas9 technology for genotyping purposes. The implementation of this approach required substantial effort in developing a specific variant calling filtering strategy, leading to the optimization of the genetic information obtained from sequencing. Furthermore, it enabled the development of an approach for identifying incorrect heterozygous genotype calls resulting from alignment errors. Although this approach was not applied to this panel due to coverage levels that led to the exclusion of heterozygous genotypes from the analyses, it has introduced a possibility that will be further explored and applied in future datasets.

Moreover, beyond investigating signatures of adaptation, this work has allowed for in-depth characterization of the genetic identity of the Castelluccio di Norcia population, which represents a significant socio-economic heritage that is locally preserved year after year by local farmers.

Chapter II

Genetic Diversity in Lentil Italian Landraces Genotyped By a CRISPR-Cas9-Based Depletion Approach

Abstract

Genetic diversity studies allow us to catch the existing genomic variation among and within populations and are crucial in guiding our understanding of crops' adaptation, facilitating their use, exploitation, and improvement. Modern genomics approaches rely on the availability of high-throughput and high-density genotyping information. The decrease in next-generation sequencing (NGS) costs enlarged the use of single Nucleotide Polymorphism (SNP) to describe populations' genetic diversity. However, the sequencing of large genome crops, such as lentil (approximately 4 Gbp), is still an issue in terms of sample preparation efforts, genomic redundant information management, and sequencing costs. In this study, we applied a novel sequencing approach that exploits high-throughput (Twist 96-Plex Library Prep Kit iGenomX) and CRISPR-Cas9-based depletion (Jumpcode Genomics) protocol. Getting a depletion of repetitive elements from genomic libraries avoids the sequencing of the whole genome and, contrary to Genotype-By-Sequencing (GBS) approaches, concentrates sequencing data on informative coding and regulatory regions. The described approach has been used to evaluate the genetic diversity of a Protected Geographic Indication (PGI) lentil landrace from "Castelluccio di Norcia"-Perugia, Italy. Eighty genotypes provided by sixteen local farmers from "Castelluccio di Norcia", and a panel of twenty-five Single Seed Descent (SSD) lentils accessions from the EU H2020 INCREASE project have been genotyped. The exploration of genetic

diversity and the analyses conducted have shown signs of local adaptation at both genes and structural variation levels for high-altitude agroecological conditions. This comprehensive approach has proven to be particularly suitable for genotyping genomes with these characteristics and exploring their diversity, effectively capturing adaptive patterns in a large and complex genome.

Materials and Methos

Plant material

80 genotypes of *L. culina*ris provided by 16 different local farmers from the plateau of Castelluccio di Norcia (Perugia, Italy-1452m a.s.l.), which we refer to as populations of Castelluccio (CS), plus 25 single seed descent (SSD) lentil accessions from the T-core INCREASE project (Bellucci et al., 2021) were analyzed. Lentils T-core accessions have been selected according to their geographical provenience: 10 accessions from Italy (IT), 10 accessions from the Mediterranean basin (MB), and 5 accessions spread across different continents (W) **Figure7**.

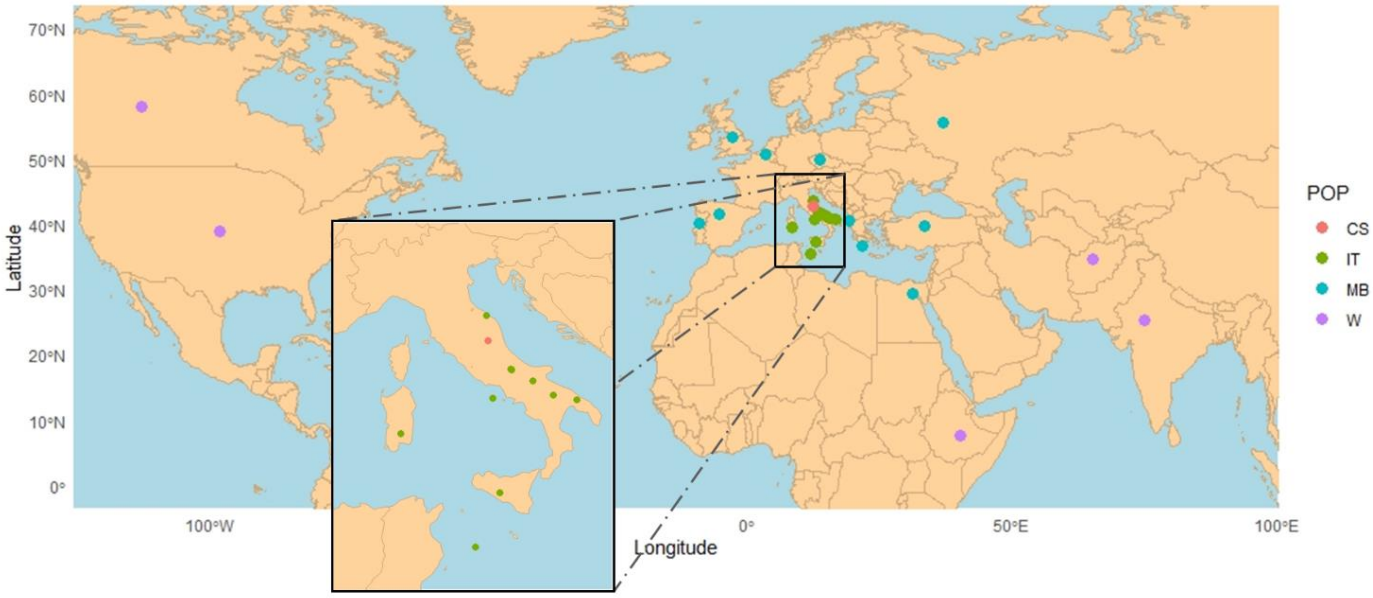


***Figure** 7. Plant material distribution*

DNA extraction and Genotyping and VCF filtering

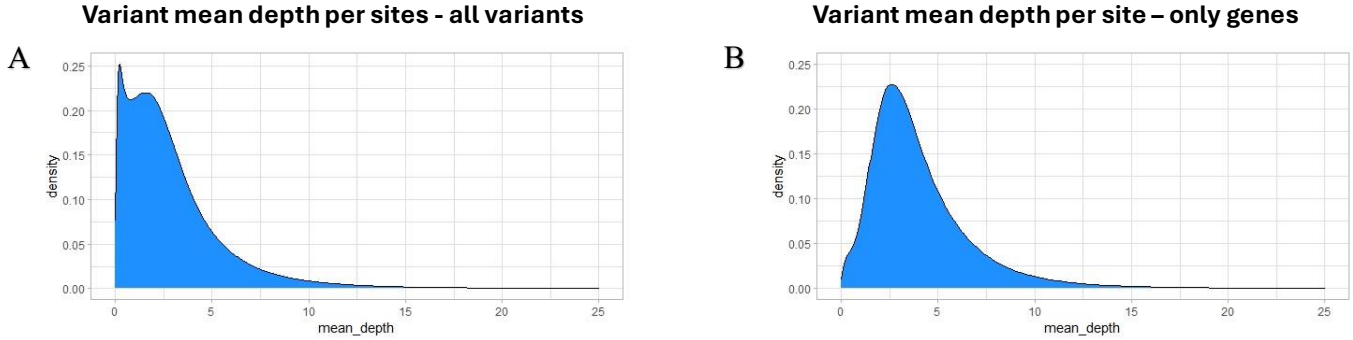
The DNA has been extracted using DNeasy 96 Plant kit (Qiagen) according to the manufacturer's instructions, except for the grinding step conducted using a TissueLyser II (Qiagen) set at 30 Hz for a total of 6 minutes, increasing the protocol recommended time. DNA was extracted from young plant leaves, previously collected, and placed into silica gel for drying. Isolated DNA has been sequenced according to the sequencing approach that exploits **high-throughput** (Twist 96-Plex Library Prep Kit – iGenomX) and **CRISPR-Cas9-based depletion** (Jumpcode Genomics) protocol by the Department of Biotechnology, University of Verona (Rossato et al., 2023b). Depleted libraries have been sequenced by NovaSeq 6000 instrument to generate 150-bp paired-end reads. For the pre-processing data analysis pipeline from demultiplexing to variant calling and variant filtration refer to (Rossato et al., 2023b). The biallelic genetic variants were further filtered using VCFtools (Danecek et al., 2011a). The total of 17,218,816 SNPs, were filtered for a maximum mean depth of 15, for individual missingness less than 0.6, resulting in the removal of two accessions from T-Core that did not meet this criterion, and for site missingness less than 0.5, as well as for a Minor Allele Frequency (MAF) less than 0.01. Heterozygous genotypes were removed. At the end of the process, 8,432,106 SNPs were obtained. The filtered data were imputed using the *impute()* function from the R package LEA (Frichot & François, 2015). The imputed data were used to conduct parallel population structure analyses, aiming to assess the impact of missingness levels on the non-imputed data. To evaluate the impact of LD on PCAdapt's ability to detect markers under selection, SNP datasets were pruned to reduce the levels of linkage disequilibrium between them. We pruned the SNP dataset based on LD in Plink v.1.9 (Purcell et al., 2007), using a threshold of $r^2 > 0.5$ ("-indep 50 5 5"). This implies that the SNPs were pruned in windows of 50 SNPs with a step size of 5. During each step, SNPs with an r^2 value greater than 0.5 were removed, ensuring that only one SNP from each

pair was retained. The pruned dataset contains a total of 1,188,835 SNPs, and the results from PCAdapt on pruned and non-pruned data were compared.

VCF filtering

A variant filtering strategy was developed and applied to the 17,218,816 SNPs obtained from the variant calling. All subsequent steps listed and described were performed using VCFtools (Danecek et al., 2011) and were applied to biallelic markers. The ***Figure 8A*** shows the distribution of the coverage per site of the called variants, which presents an average for raw data of approximately 3x for all variants, and greater than 4x when considering only variants within annotated genes (***Figure 8B***). This distribution, with mean values around 3x and median values slightly above 2x (***Table1 A,B***), suggested that no minimum depth filter should be applied. Conversely, a maximum cutoff was applied, removing sites with a median coverage greater than 15, because regions with very high read depths are likely repetitive and map to multiple parts of the genome.

Table 1. *A: Statistics on the coverage distribution of all variants, B: Statistics on the coverage distribution of variants within gene sequences.*



Min.	1stQu.	Median	Mean	3rdQu.	Max.
0,01	1,14	2,30	3,06	3,93	1533,49

Min.	1stQu.	Median	Mean	3rdQu.	Max.
0,01	2,30	3,44	4,15	5,12	1533,49

Figure 8. *A) Coverage distribution of all called variants, B) Coverage distribution of variants located within annotated genes.*

We observe an increase of more than 1 in the average coverage of variants within genes compared to the average coverage when considering all variants.

Heterozygous genotypes were removed in accordance with the average heterozygosity values per individual slightly above 2% on average (***Figure 9***), and due to the minimum depth levels accepted by the previous filtering, which were insufficient to call them accurately.

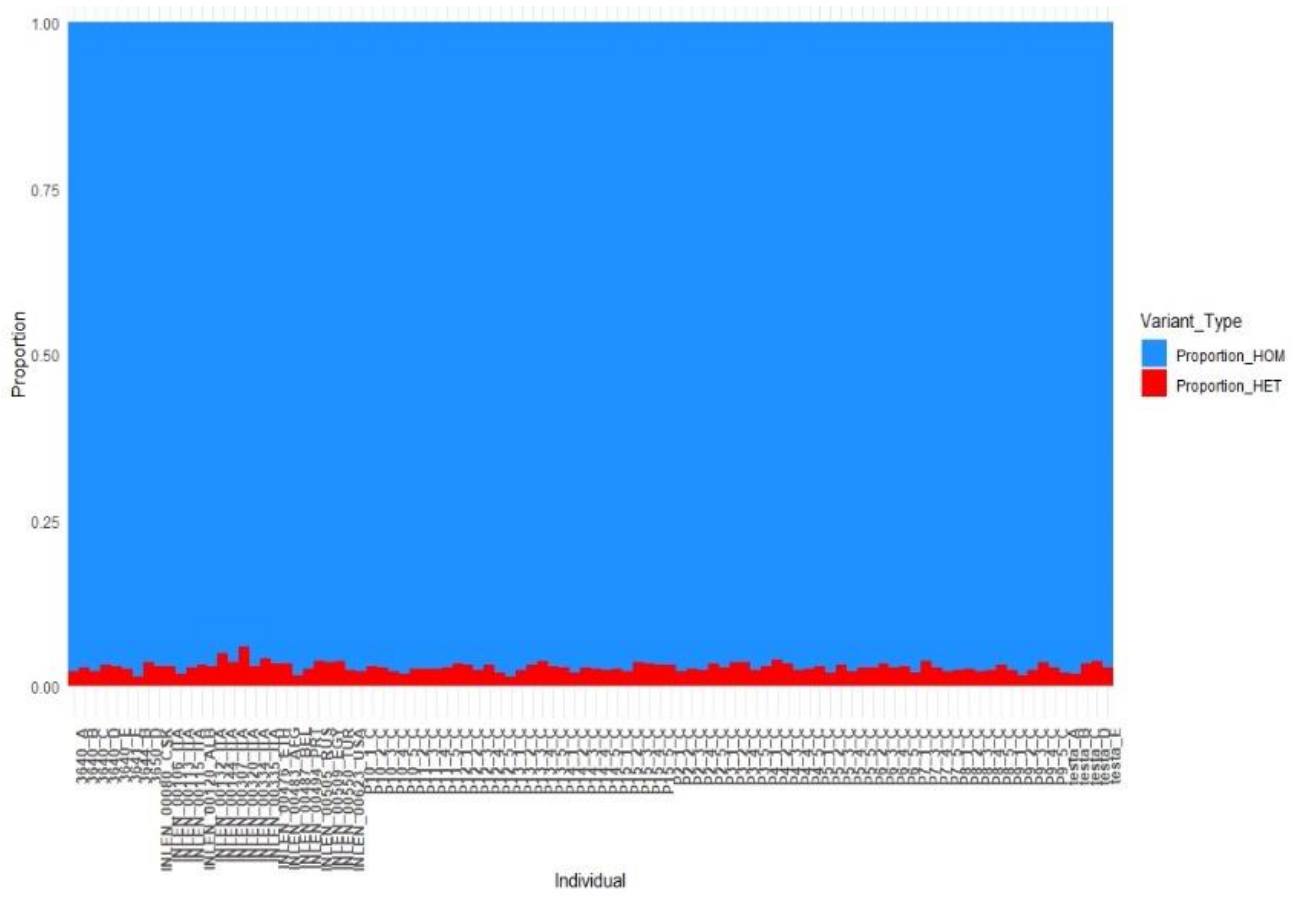


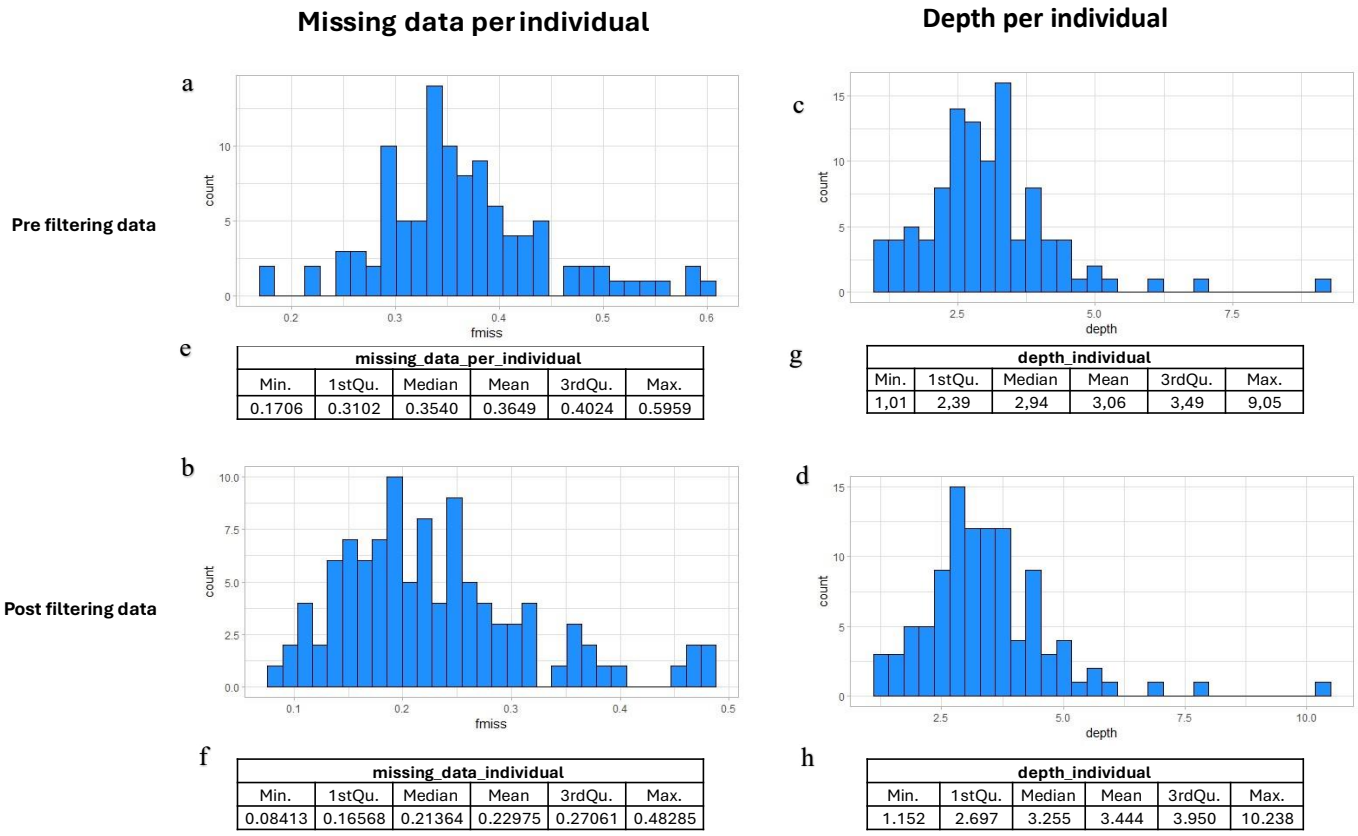
Figure 9. *Proportion of Heterozygous Sites per Individual.*

Table 2*. Statistics on heterozygosity levels per individual.*

Heterozygous_Variants_for_Each_Individual					
Min.	1stQu.	Median	Mean	3rdQu.	Max.
0,01418	0,02334	0,02727	0,02798	0,03202	0,05899

To mitigate the impact of missing data in subsequent analyses, an individual missingness filter of 0.6 was applied (individuals with an average missingness greater than 0.6 were removed). This process resulted in the exclusion from analyses of two accessions from T-core that did not meet this criterion. It should be noted that when using VCFtool to remove samples from the dataset, the sites themselves are not removed, but the genotypes of the samples, are replaced with missing data. This results in the formation of monomorphic sites. The presence of these sites can introduce artifacts during subsequent analyses and have been excluded. Additionally, a site missingness threshold of 0.5 was applied. A MAF cutoff was also applied, removing sites with a frequency lower than 0.01. Filtering the VCF resulted in a reduction of individual-level missingness, consequently improving data quality. Despite not filtering based on minimum depth, individual coverage expected remained nearly unchanged ***Figure 10 c,d***. At the end of the filtering process, we obtained 8,432,106 biallelic SNPs.

Table3. *The tables e, f, g, and h respectively present the summaries of the values described in the histograms.*



e

missing_data_per_individual					
Min.	1stQu.	Median	Mean	3rdQu.	Max.
0.1706	0.3102	0.3540	0.3649	0.4024	0.5959

g

depth_individual					
Min.	1stQu.	Median	Mean	3rdQu.	Max.
1,01	2,39	2,94	3,06	3,49	9,05

f

missing_data_individual					
Min.	1stQu.	Median	Mean	3rdQu.	Max.
0.08413	0.16568	0.21364	0.22975	0.27061	0.48285

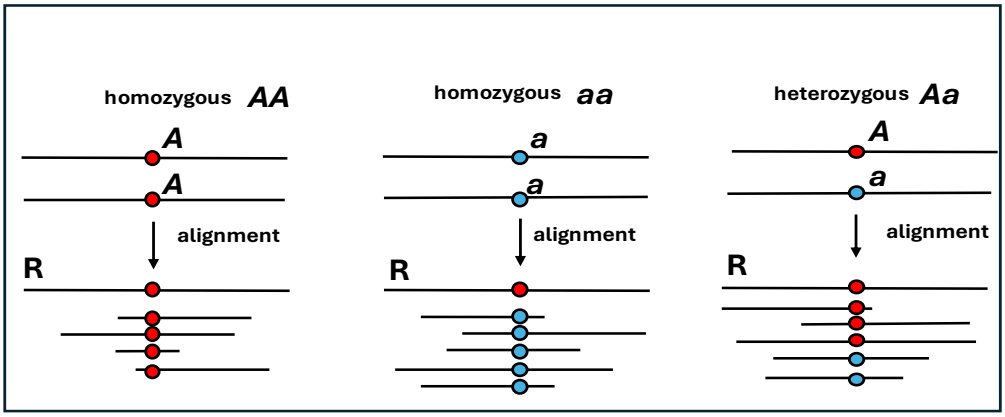
h

depth_individual					
Min.	1stQu.	Median	Mean	3rdQu.	Max.
1.152	2.697	3.255	3.444	3.950	10.238

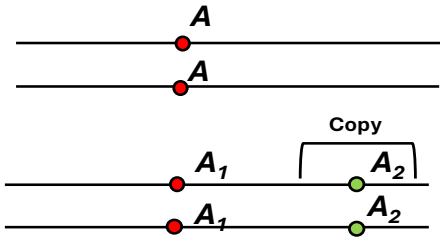
***Figure 16**. The histograms display the missingness values per individual before and after filtering (a, c) and the average depth values per individual before and after filtering (b, d).*

Heterozygote Calling Errors Resulting from Misalignment

Sequencing artifacts have been identified when using NGS technologies, potentially introducing systematic biases and technical variations that can significantly impact genomic data interpretation and analysis. A critical step in NGS analysis is the mapping of sequencing reads to a reference genome or assembly. Accurate read alignment ensures that downstream analyses, such as variant calling or population genetic inferences, are reliable (Crawford & Lazzaro, 2012; H. Li et al., 2008). Misalignments or mapping errors, however, can introduce biases that compromise the detection of true biological signals, including selection signals or demographic histories (Dallaire et al., 2023). Gene duplication events, potentially followed by nucleotide mutations in one or more of the copies, are a common source of read misalignments and mapping errors (Lee et al., 2021). When the duplicated gene copies are highly similar, sequencing reads originating from these regions may align interchangeably to either copy in the reference genome. Since these reads can carry different variants, it may appear as two different alleles exist at a single genomic location, even though the variants are actually located on separate gene copies. As a result, this misalignment can lead to incorrect genotyping of heterozygous sites, suggesting heterozygosity where none exists, as shown in ***Figure 11***. Consequently, and especially in large, highly repetitive genomes, such as the lentil genome, alignment errors are more frequent and an unexpected excess of heterozygous sites may be observed. It is important to specify that these problems are amplified when the sequencing coverage is low, since there is an inadequate number of reads to confirm the true allelic composition of a site.



Duplication events lead to an increase in the number of copies of a gene, where subsequent mutation events can result in the fixation of different allelic forms



During the alignment phase, the detection of heterozygous genotypes may be observed, which are, in fact, not present at the site but result from overlapping reads originating from another site, previously duplicated, that carries an allelic variation

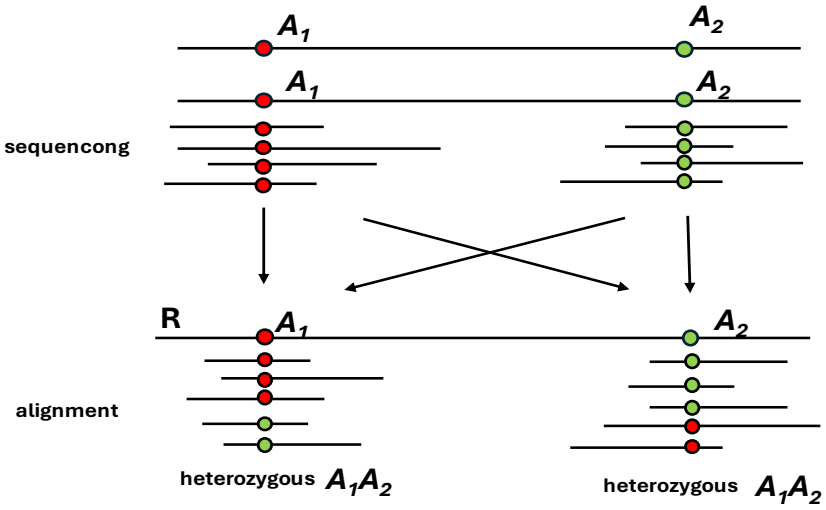


Figure 11. *Misalignment to the incorrect calling of a heterozygous genotype.*

The image illustrates an example of how an incorrect alignment in a duplicated region can occur, resulting in the miscalling of heterozygous genotypes instead of homozygous ones.

Our Castelluccio (CS) panel, composed of 80 individuals, represents a relatively large and isolated population that is predominantly self-pollinating. In such populations, heterozygous sites are not commonly expected, however, we observed that a mean higher than 2.5 % of the genotyped sites were heterozygous, which was more than expected. We hypothesize that this unexpected heterozygosity may be due to mapping errors, leading to the genotyping of "false heterozygous" sites. To test this hypothesis, we considered the following: at a segregating site (a site where at least one individual in the population is heterozygous), we assume that both homozygous genotypes are present in the population. If a site is a falsely called heterozygous site (i.e., actually homozygous for one of the two alleles), we expect that one of the homozygous genotypes (the correct genotype) would be significantly more represented than the other. This is because individuals mistakenly called as heterozygous are, in reality, homozygous for the true allele.

To analyse the data, we processed the VCF file containing Castelluccio SNPs. Biallelic sites were filtered to exclude those with more than 50% missing data or a (MAF below 0.01. Additionally, we retained only sites with a minimum coverage of 5 reads, to ensure the reliability of the called genotypes. For each site, we classified the homozygous genotypes into two categories: more represented (indicated in red) and less represented (indicated in blue) ***Figure 12***. The distributions of these two classes of homozygous genotypes were then analysed relative to the increase in heterozygous sites. As a result, for all sites with an equal number of heterozygous genotypes, we generated a distribution of the more represented homozygous genotype and a corresponding distribution for the less represented homozygous genotype.

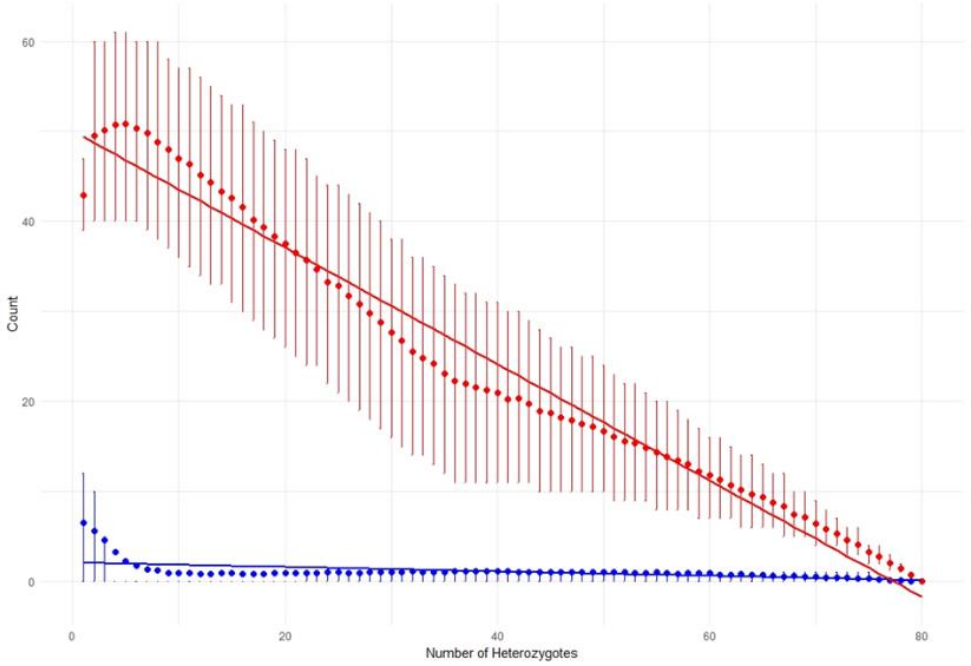


***Figure 12.**. Distribution of homozygous genotype counts in relation to the increase in heterozygous genotypes..*

The graph shows in red the distribution of the more represented homozygous genotypes and in blue the distribution of the less represented homozygous genotypes, at sites where at least one heterozygous genotype is present, according to the number of heterozygous genotypes observed. The lines represent the regression between the means of the distributions and the increase in the number of heterozygous genotypes per site.

Our observations align with what we would expect in the case of errors in the reads alignment. Specifically, we found that as the number of heterozygous genotypes increases at a site, the less represented homozygous genotype remains nearly absent instead of increasing proportionally. This trend is highlighted by the regression lines calculated between the means of the distributions of the two homozygous genotypes and the number of heterozygous genotypes. For the more abundant homozygous genotype (red line), the regression slope was -0.633, while for the less represented homozygous genotype (blue line), the slope was -0.095. These values indicate that as the number of heterozygous genotypes increases, the less represented homozygous genotype remains near zero.

This pattern can be explained by alignment errors caused by mutations that occurred after genomic duplication. As previously mentioned, when a region of the genome is duplicated, mutations may accumulate in one or more copies over time, introducing slight differences between them. During alignment, reads from these slightly divergent duplicated regions may fail to map to their true locations in the genome. Instead, overlapping reads from these regions can lead to misinterpretation of the mutations as two different alleles. Consequently, the site may be incorrectly called as heterozygous, even though it is actually homozygous for the more abundant homozygous genotype.

Characterization of population structure

The 8,432,106 biallelic SNPs obtained from VCF filtering were used to characterize the genetic groups within the lentil panel consisting of 103 accessions. The genetic groups present in the lentil panel were characterized via genetic structure analysis with sNMF (Frichot et al., 2014), using the R package LEA (Frichot & François, 2015). The optimal number of ancestral groups (K) was determined using the cross-entropy criterion over ten iterations, with K ranging from 1 to 10. We also performed a PCA in the R package pcadapt (Duforet-Frebourg et al., 2016). The variance among the 16 local farmers who contributed material to form the CS population was assessed using a one-way Analysis of Variance (ANOVA), based on each genetic group defined by membership coefficients at K = 5.

Detection of loci under selection

We used two complementary approaches to identify candidate SNPs involved in high-altitude local adaptation. We identified SNPs most strongly associated with the observed population structure using pcadapt (Duforet-Frebourg et al., 2016). Pcadapt employs Principal Component Analysis (PCA) to determine population structure and calculates

the correlation between genetic variants and significant principal components (Luu et al., 2017). The number of PCs to be retained (K) was chosen by running a PCA with a large number of PCs (20 in our case). The optimal K was selected based on the scree plot ***Figure 13***, which shows that most of the variance is described by PC1. Additionally, when plotting all the principal components, PC1 and PC2 were able to separately cluster the CS population together with the two Italian accessions living at high altitudes, distinct from the rest of the T-core (***Figure 14)***, confirming K=2 as the best choice. To assess the effect of linkage disequilibrium on the detection of selection signals, the analysis was conducted on both the entire set of variants and those filtered through pruning, and the results were compared. To identify putative markers under selection we applied a False Discovery Rate (FDR) correction with a threshold of 0.05. To evaluate the effect of LD, pcadapt was run on both unpruned data and pruned data. To mitigate the influence of missing data on this analysis, it was also repeated on a dataset where the missing data were imputed using the *impute()* function from the R package LEA. In addition to pcadapt tests, the approach of population differentiation tests using snmf() has been employed. Methods for detecting outlier loci from a genomic background focus on extreme values of the fixation index (FST) across loci. The snmf() function can compute fixation indices when population genetic structure is continuous, especially when defining subpopulations is challenging, or in the presence of admixed individuals (Martins et al., 2016). In the snmf() approach, population differentiation statistics are calculated from ancestry coefficients derived from an snmf object, and p-values are provided for all loci. Similarly to pcadapt analysis, a false discovery rate (FDR) correction with a threshold of 0.05 has been applied. The results obtained from those tests were compared.

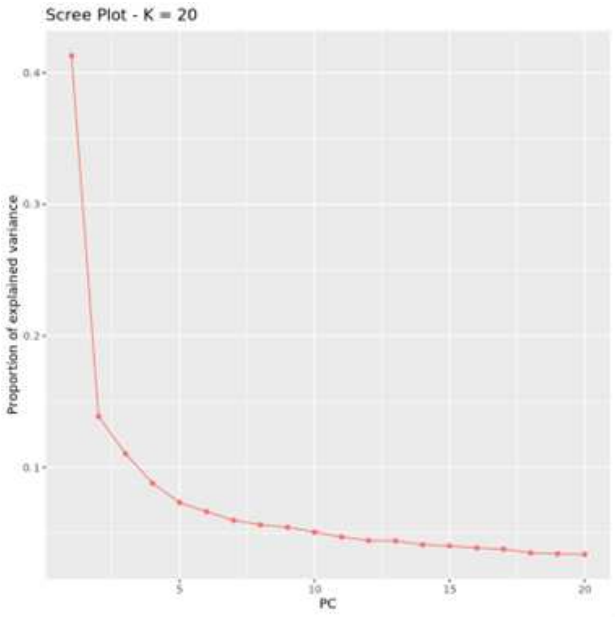


Figure 13.. *Scree plot.*

Displays the percentage of variance explained by each PC in decreasing order.

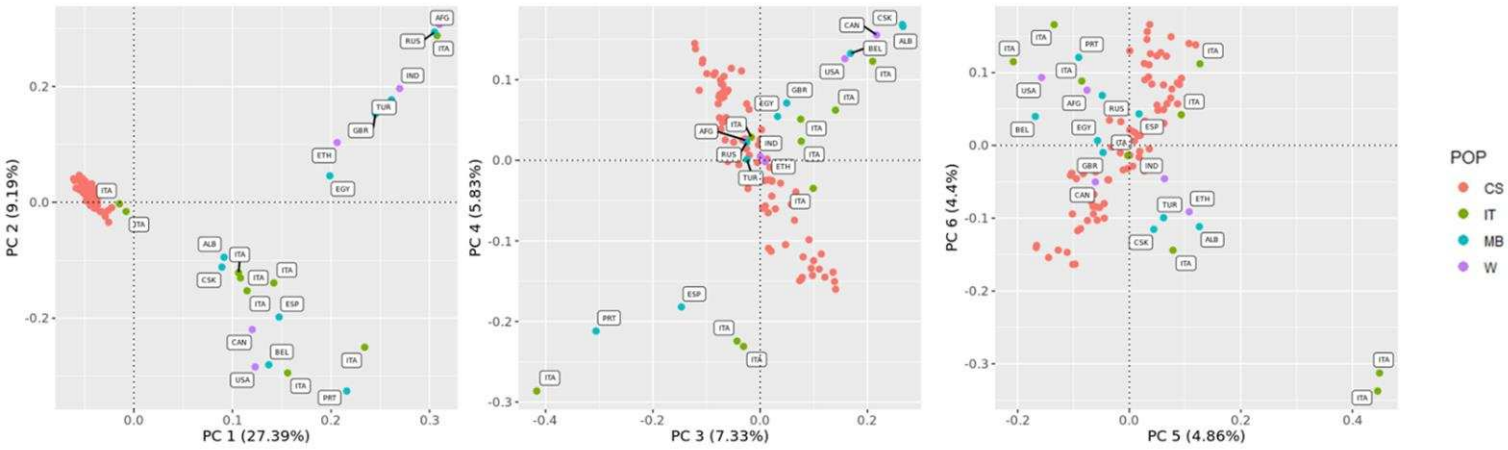


Figure 14.. *The score plot.*

Displays the projections of the individuals onto the specified principal components. In this figure, PC1 and PC2, PC3 and PC4, and PC5 and PC6 are plotted. The first quadrant, which shows PC1 and PC2, is the only one capable of separating the high-altitude population from the rest of the T-core.

Detection of Inversions through Local PCA

We employed local PCA to detect structural variations, such as chromosomal inversions, by analyzing patterns of relatedness across the genome (H. Li & Ralph, 2019). By setting windows of 300 variants, we were able to capture variation in the patterns of relatedness formed by population structure across the genomes. The results of Local PCA detection to identify inversions were compared with putative variants under selection, to attempt to detect potential associations between structural variations and adaptive genetic changes.

Identification of Genes Potentially Under Selection

The 58,433 SNPs identified by combining the results of, PCAdapt applied to imputed data, and the genome scan for selection based on FST was used to identify putative genes under selection. For each SNP, windows consisting of 100 kb upstream and 100 kb downstream (total 200kb window size) were identified to account for the effect of linkage disequilibrium, which presents an r^2 value of 0.29 at an average distance of 100 kb ***Figure 15***. When these windows overlapped, they were merged to form larger windows. Due to their non-independent segregation, regions identified as potential inversions were treated as single larger windows. For each identified window, the variant with the highest p-value from PCAdapt was selected. Finally, the genes closest to these variants were considered potential genes under selection ***Figure 16***.

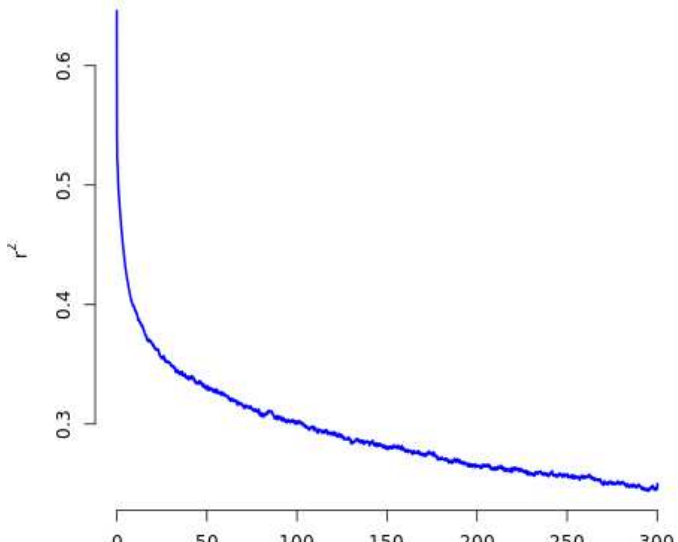


Figure 15. *LD decay.*

The LD decay using PopLDdecay (Zhang et al., 2019) has been calculated, and the average distance LD decay provided indications for selecting a window size that minimizes the effect.

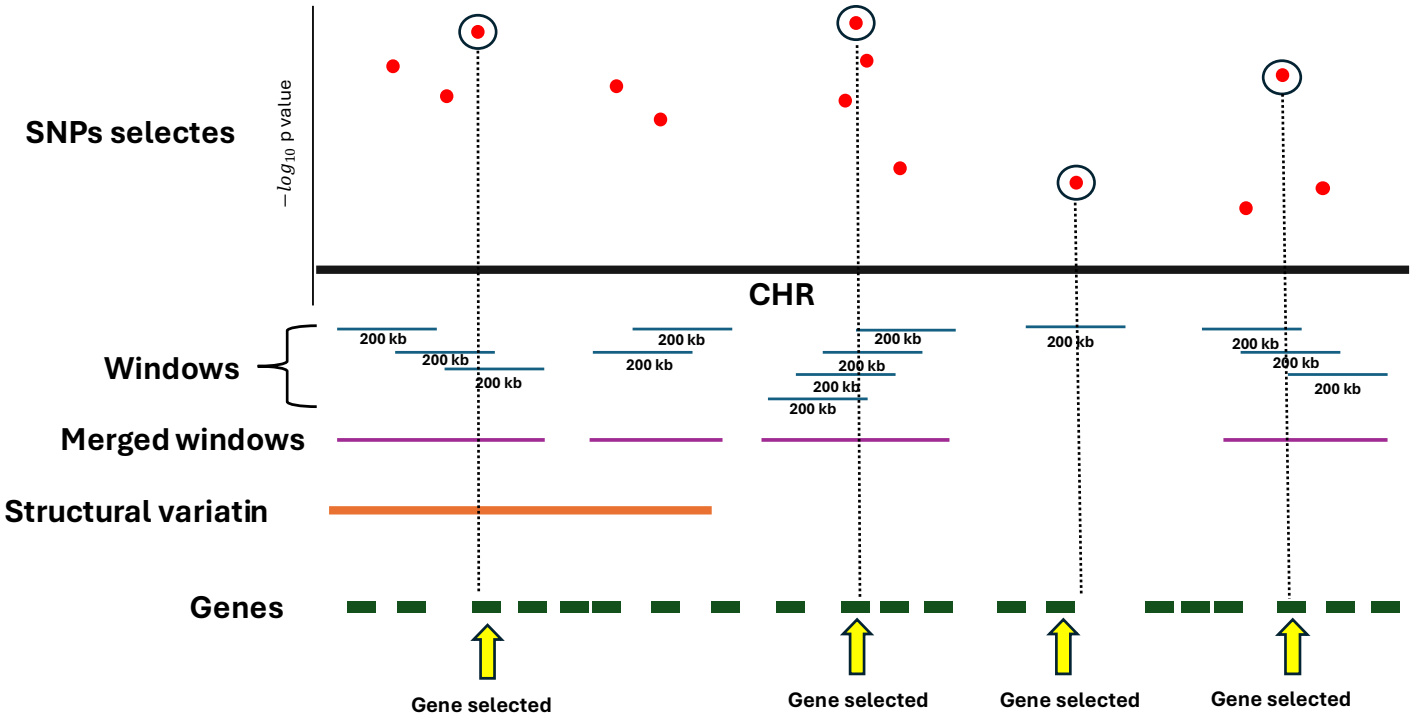


Figure 16. *Mechanism for Selecting Candidate Genes Under Selection.*

The image illustrates the strategy used to select genes considered under selection. In red, the markers shared by the applied selection tests are shown, ordered on the chromosome

and based on their PCAdapt p-value. For each marker, a 200kb window (blue, 100kb upstream and 100kb downstream) was set. If these windows overlapped, they were merged into larger windows (purple). In cases of structural variation, the corresponding genomic region was treated as a single large window. For each window, merged window, or structural variation, the marker with the highest -log10(p-value) was selected. For each of these, a single gene was identified, specifically the closest gene or the one in which the marker was located.

GO enrichment analysis

GO enrichment analysis was performed on orthologs of Arabidopsis thaliana identified through OrthoFinder (Emms & Kelly, 2015, 2019), starting from the primary transcripts of 2,433 genes detected as potentially under selection. Enriched terms among the putative genes under selection were determined using agriGO v2.0 (Tian et al., 2017). The analysis applied a hypergeometric test, adjusted for multiple hypothesis testing using the Hochberg FDR method, with a significance threshold of <0.05 and a minimum mapping entry count of 3.

Genotype imputation

Missing genotypes were imputed using ancestry and genotype frequency estimates and were conducted by an sNMF run. The accuracy of imputation was evaluated through two methods. First, the correlation of Q1 values (and similarly for Q2), obtained from the structure analysis conducted using sNMF, between imputed and non-imputed data was calculated. Second, a test was conducted where 10,000 random genotypes were hidden

by transforming them into missing data. After imputation, the imputed genotypes were compared with the original data previously stored.

Results

Characterization of genetic groups

A total of 8,432,106 high-quality biallelic SNPs were obtained from all 103 lentil samples in the panel, which consisted of 80 genotypes from Castelluccio and 23 SSD accessions from the T-core. This SNP dataset was used to assess the population structure of the CS population by performing genetic structure analysis using sNMF, and by performing a PCA.

The cross-validation technique to compute an entropy criterion showed a clear minimum is observed at K = 5 (***Figure 19 a***). The CS individuals cluster separately from the T-core accessions, forming a distinct group (K = 2), except for two Italian accessions, both originating from high-altitude environments (above 900 masl) (***Figure 17, 18***). From the structure analysis (***Figure 18***), it is possible to observe the presence of a substructure within the CS, which exhibits three distinct groups with a high level of admixture, that however, is not attributable to the material donors. The T-core population also exhibits a structure, which is more distinctly differentiated along the PC2 axis. At K = 4 (***Figure 18***), it is possible to distinguish the separation of a group, represented more externally compared to the Castelluccio population in the PCA (***Figure 17***). This group is predominantly composed of germplasm originating from the eastern regions of the population. The group described separates from the remaining part of the T-core, which is positioned intermediate between CS and the eastern material. This group consists mainly of Italian accessions, particularly those from the northwestern MB population, as well as the two American accessions, one from the United States and one from Canada. This intermediated group also exhibits higher levels of admixture compared to the other T-core group, which appears more diversified. From the PCA, we observe that the Egyptian accession is positioned between the two main groups constituting the T-core, showing greater similarity to the Ethiopian accession than to the Indian one, which in turn is more similar to the Turkish accession. The most diversified material in the T-core

consists of the Afghan and Russian accessions, along with one Italian accession which was shown to be markedly different from the other Italians.

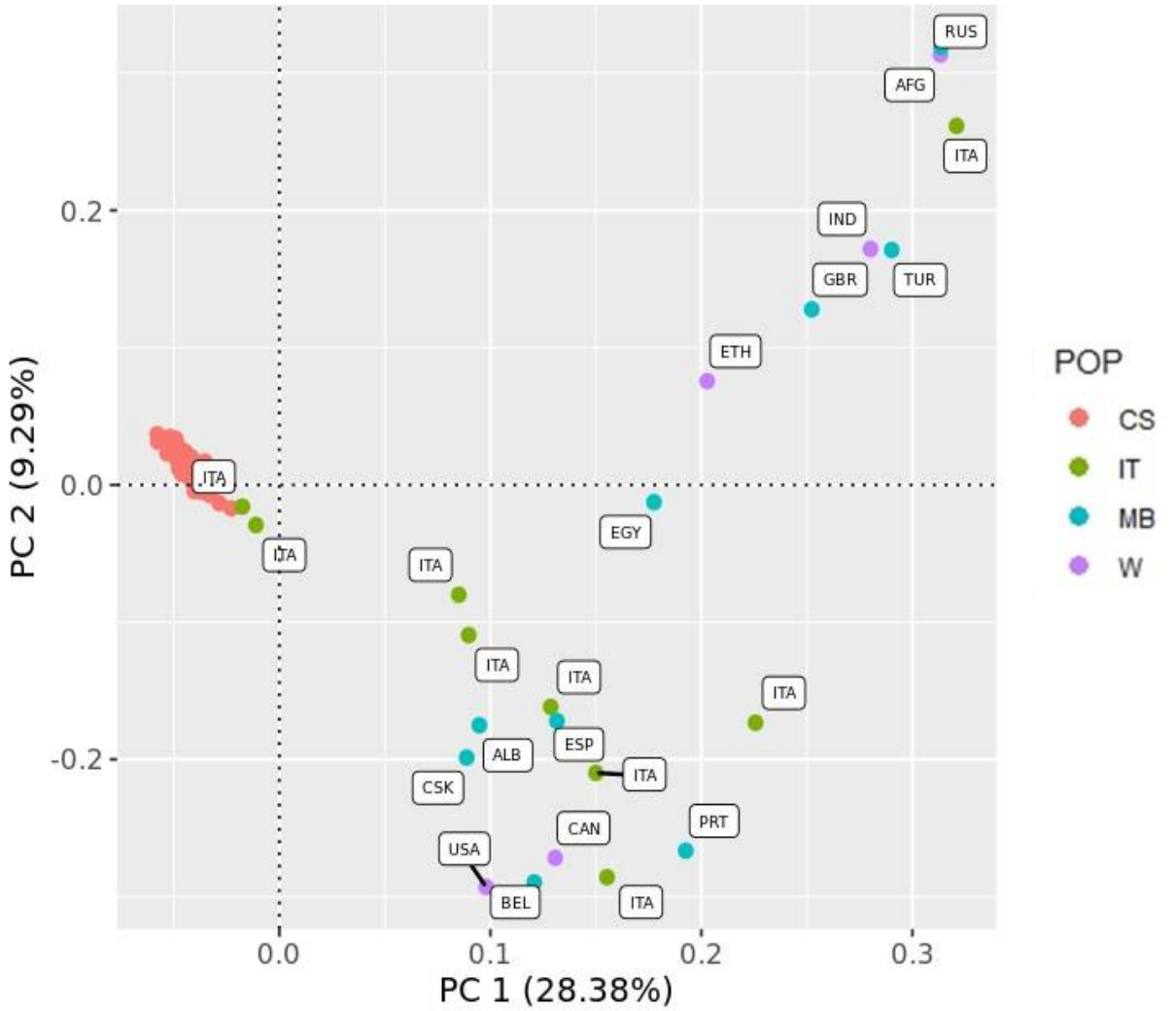


Figure 17. *Principal Component Analysis PCA.*

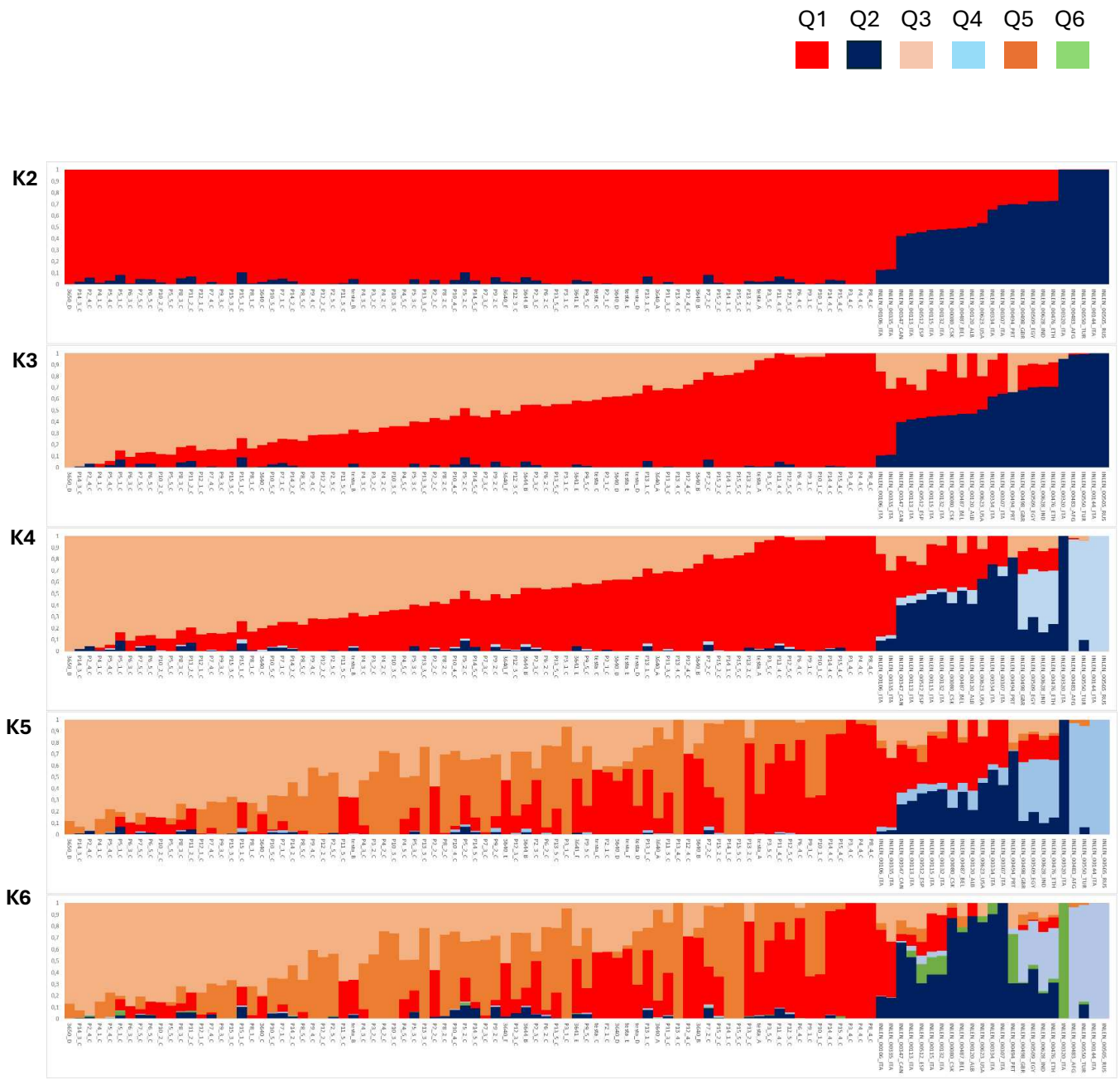


Figure 18. *Genetic structure of the lentil panel by performing sNMF.*

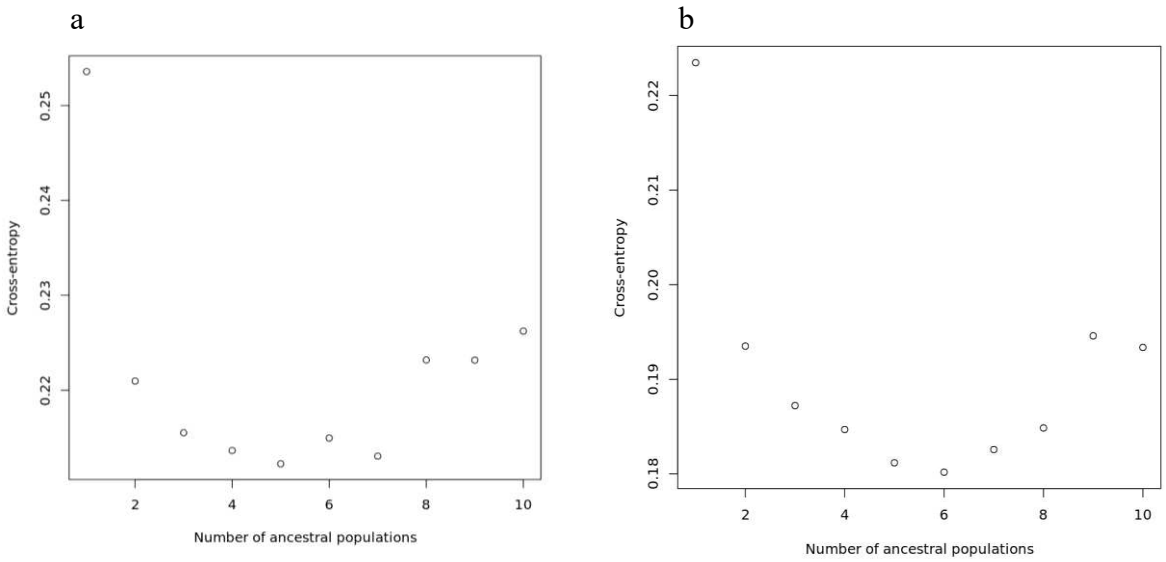
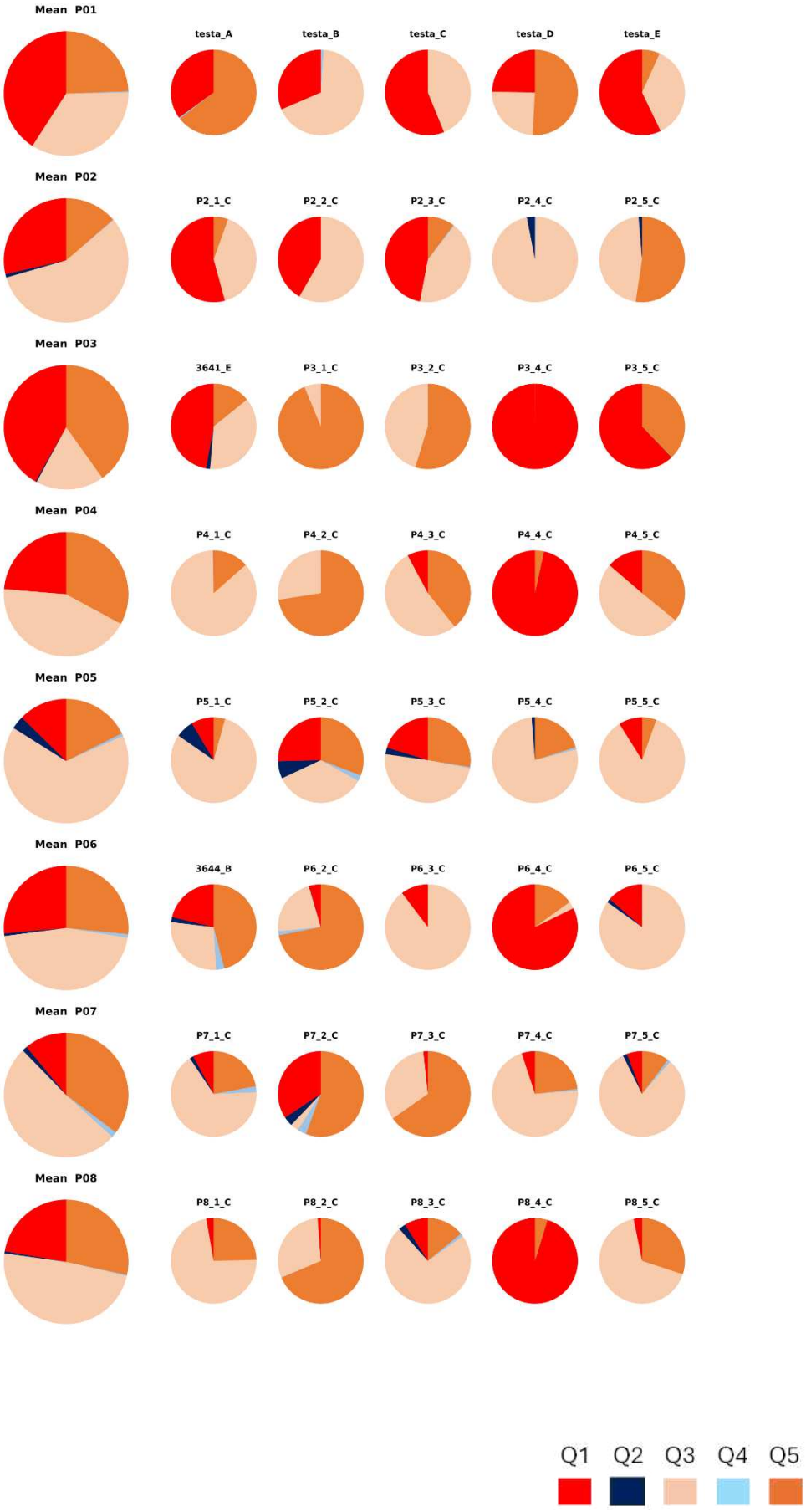


Figure 19. *Cross-entropy criterion values as a function of the number of populations in sNMF*

a) Shows the cross-entropy values as a function of the number of ancestral populations obtained from the analysis of non-imputed data, b) obtained from the imputed data.

Genetic Diversity within the Castelluccio Subpopulation

By focusing on the genetic diversity within the Cs population, we identified three distinct genetic groups emerging at K = 2, 3, and 5. These genetic groups were evenly distributed among materials from local farmers (***Figure 20***), as undelight by ANOVA results, which indicated no significant differences of them between local farmers (Q1 P-value = 0.87; Q3 P-value = 0.561; Q5 P-value = 0.571).


Mean P01
testa_A
testa_B
testa_C
testa_D
testa_E
Mean P02
P2_1_C
P2_2_C
P2_3_C
P2_4_C
P2_5_C
Mean P03
3641_E
P3_1_C
P3_2_C
P3_4_C
P3_5_C
Mean P04
P4_1_C
P4_2_C
P4_3_C
P4_4_C
P4_5_C
Mean P05
P5_1_C
P5_2_C
P5_3_C
P5_4_C
P5_5_C
Mean P06
3644_B
P6_2_C
P6_3_C
P6_4_C
P6_5_C
Mean P07
P7_1_C
P7_2_C
P7_3_C
P7_4_C
P7_5_C
Mean P08
P8_1_C
P8_2_C
P8_3_C
P8_4_C
P8_5_C
Q1
Q2
Q3
Q4
Q5

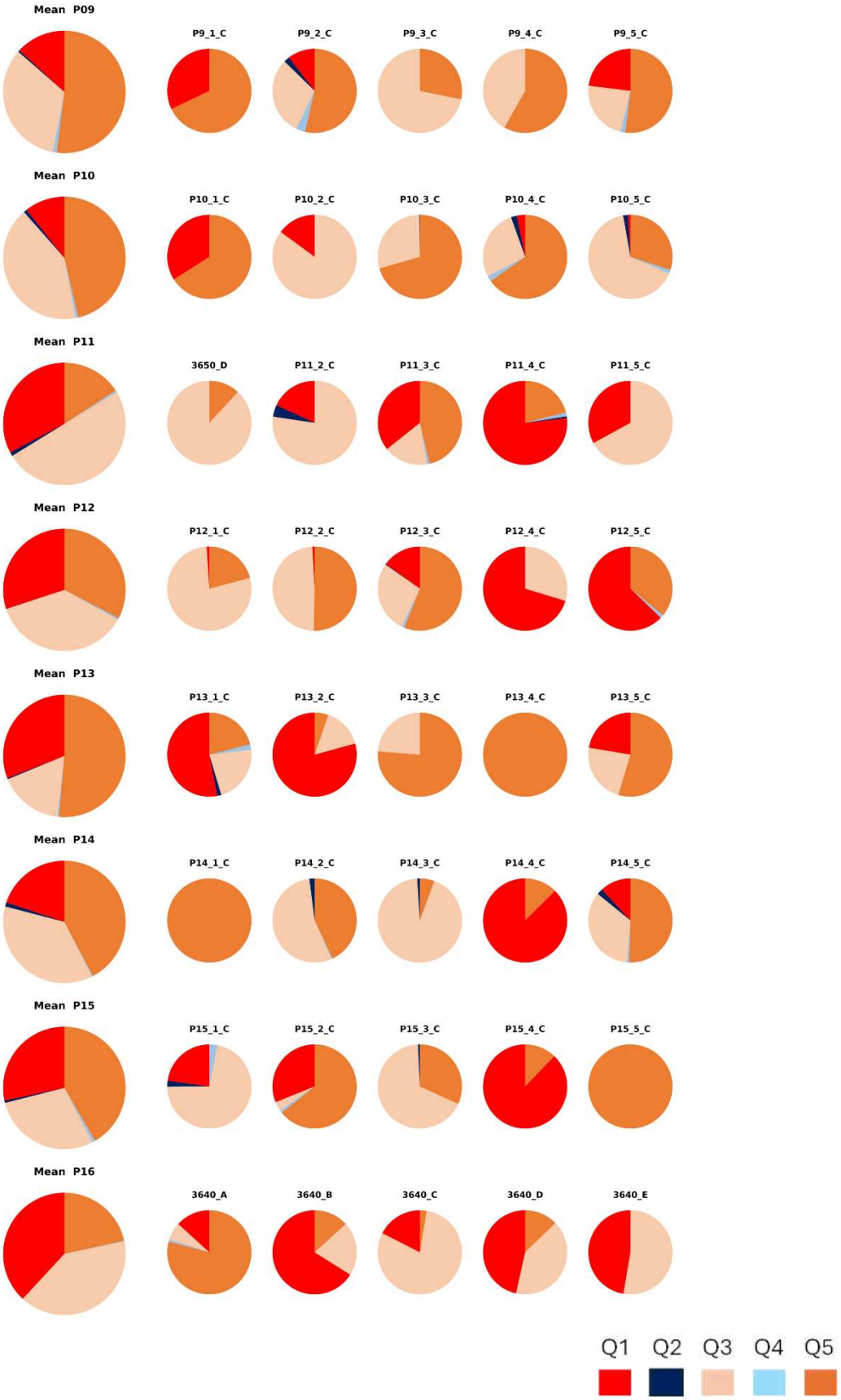


***Figure 20**. Pie charts of Q values obtained at K = 5 for the CS population.*

The pie charts are arranged in rows, each representing a local farmer. The larger pie chart at the beginning of each row shows the average Q values across all genotypes contributed by each farmer, while the smaller pie charts represent individual genotypes.

Table 4. *Purity levels contained within the material provided by local farmers from Castelluccio di Norcia.*

A, shows the donors for whom the sum of the average Q values across the 5 genotypes provided exceeds the purity threshold of 99%, while B, refers to 9 farmers. Both tables report the sums of Q1, Q3, and Q5 values.

A

Above 99%	
pop	Sum_Q1_Q3_Q5
P01	0,9969
P03	0,9967
P04	0,9997
P08	0,9922
P12	0,9946
P13	0,9927
P16	0,9980

B

Below 99%	
pop	Sum_Q1_Q3_Q5
P02	0,9896
P05	0,9588
P06	0,9840
P07	0,9720
P09	0,9849
P10	0,9844
P11	0,9849
P14	0,9870
P15	0,9857

Characterization of genetic groups by Imputed data

The correlation of Q1 values (and similarly for Q2) yields a correlation coefficient of 0.997 (***Figure 21***). The test conducted by randomly converting 10,000 genotypes into missing data resulted in a success rate of 92.9%.

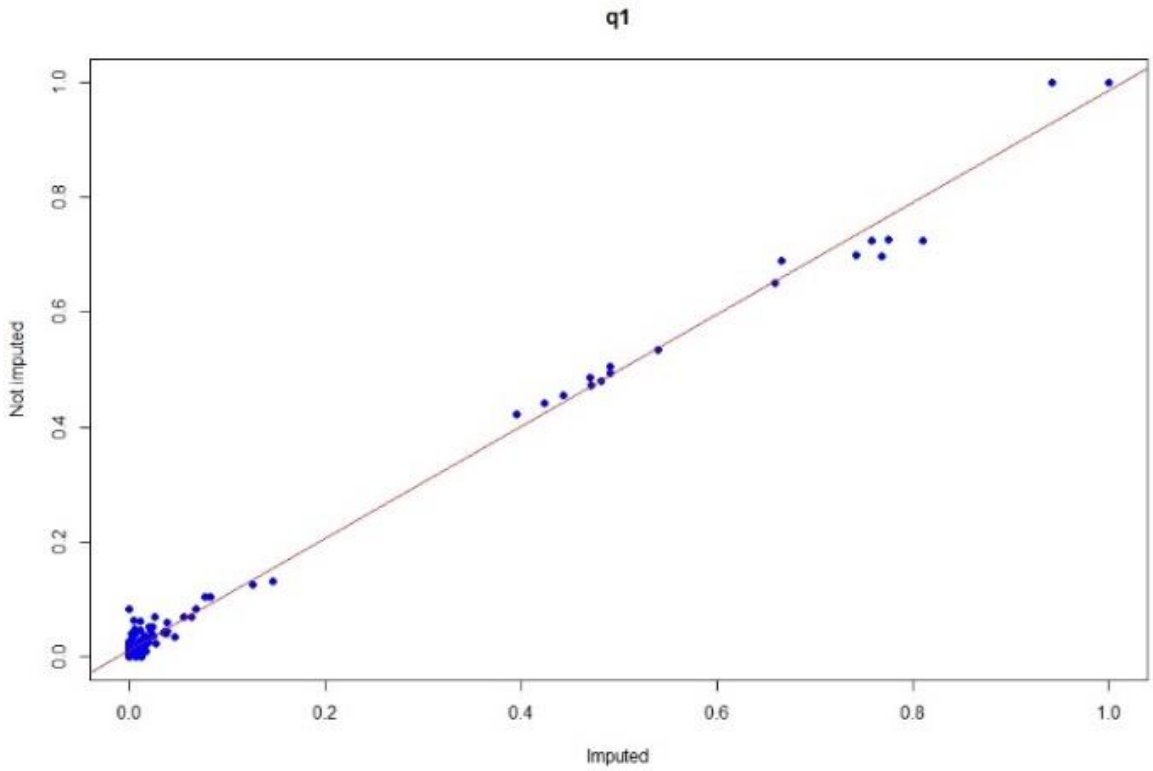


Figure 21. *Correlation of Q1 for Imputed and Non-Imputed Data.*

The imputed data yielded a cross-entropy result that indicated the optimal value of K as 6 (***Figure 19 b***). Comparing the population structure and PCA results between imputed and non-imputed data, we observe that up to K=5, the same groups and relationships are maintained among the genotypes that make up the population. At K=6 for the imputed data, the structure analysis reveals the emergence of a new group within the CS population (***Figure 22***), which is not distinguishable in the non-imputed data.

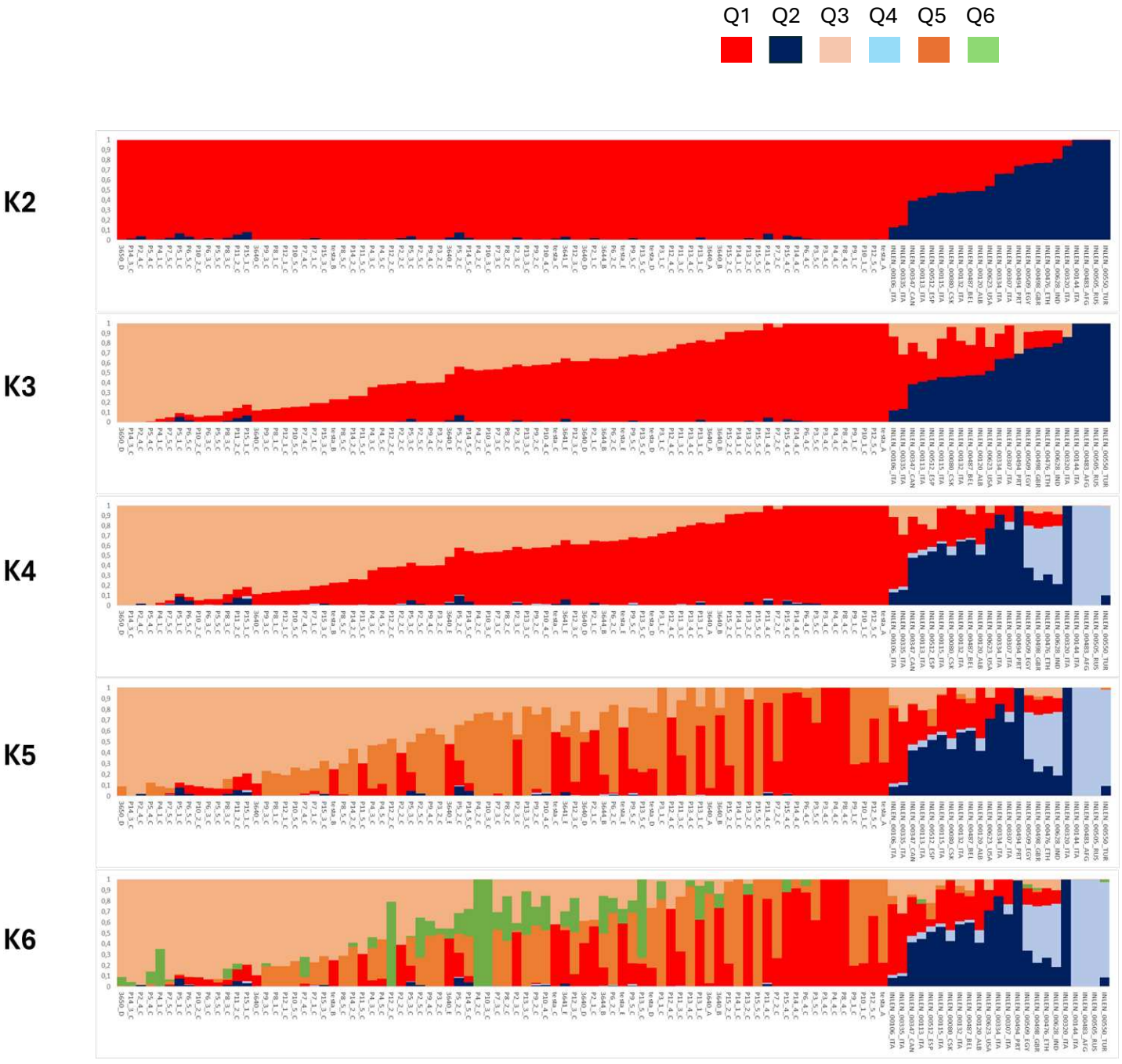


Figure 22. *The genetic structure of the lentil panel by performing sNMF performed on imputed data.*

Detection of loci under selection

Using the pcadapt R package on non-imputed data, we identified 490,034 SNPs potentially under selection. Among these, 65,292 SNPs overlapped with the 76,763 SNPs identified as potentially under selection through a Genome scan for selection using the snmf test. Furthermore, of these 65,292 SNPs, 58,433 were also detected in the results from pcadapt conducted on a dataset with imputed missing data. Therefore, these latter SNPs are considered the most reliable candidates for potential selection.

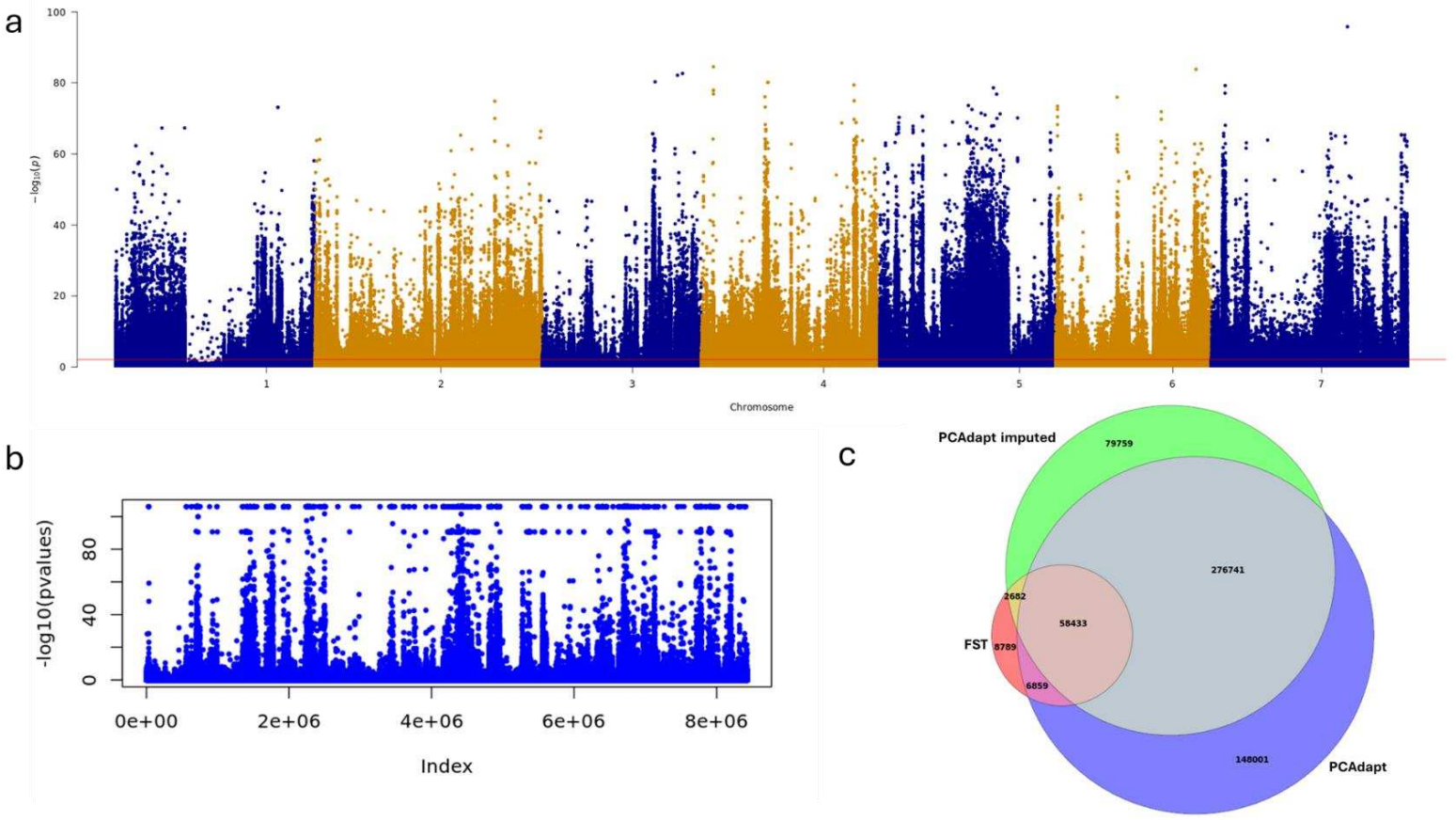


Figure 23. *Detection of markers under selection..*

a) Manhattan plot displaying the variants identified with pcadapt. This plot highlights the statistical significance of SNPs across the genome, providing insight into potential

selection signals. b) Manhattan plot showing the values obtained through a method focused on detecting outlier loci by assessing extreme values of the fixation index (FST). c) Venn diagram illustrating the overlap of variants identified across the three tests: PCAdapt applied to both imputed and non-imputed data, and the Genome scan based on FST. This diagram summarizes the common variants found by all three approaches, highlighting the consistency and reliability of the identified markers.

From the analysis of inversion detection through local PCA, 22 regions were identified across the entire genome (***Table 5***).

Table 5. *positions of the 22 regions identified as inversions through detection by local PCA.*

chr	start	end
Chr1	186,101,558	290,818,824
Chr1	292,714,029	313,223,240
Chr1	452,948,598	468,845,284
Chr1	494,929,538	505,370,178
Chr1	516,445,616	528,876,747
Chr2	204,925,386	279,501,136
Chr2	298,678,815	321,364,950
Chr3	95,580,257	120,136,728
Chr3	125,051,626	139,420,956
Chr3	148,235,606	161,124,474
Chr3	229,030,646	254,488,099
Chr3	237,587,314	256,316,283
Chr3	268,006,976	278,239,247
Chr4	24,650,195	40,221,657
Chr4	216,869,362	230,399,771
Chr4	393,053,145	422,580,658
Chr5	417,080,878	429,815,670
Chr6	359,532,623	373,770,622
Chr7	277,982,883	303,498,881
Chr7	337,547,402	359,230,892
Chr7	396,021,027	420,113,166
Chr7	499,376,870	509,395,743

To identify which of these structural variations may have played an adaptive role, the results of the adaptation tests were mapped onto genomic windows. From this analysis, we observed that certain structural variations show an enrichment in markers identified as under selection, while others appear to lack such enrichment, as illustrated in the ***Figure 24***.

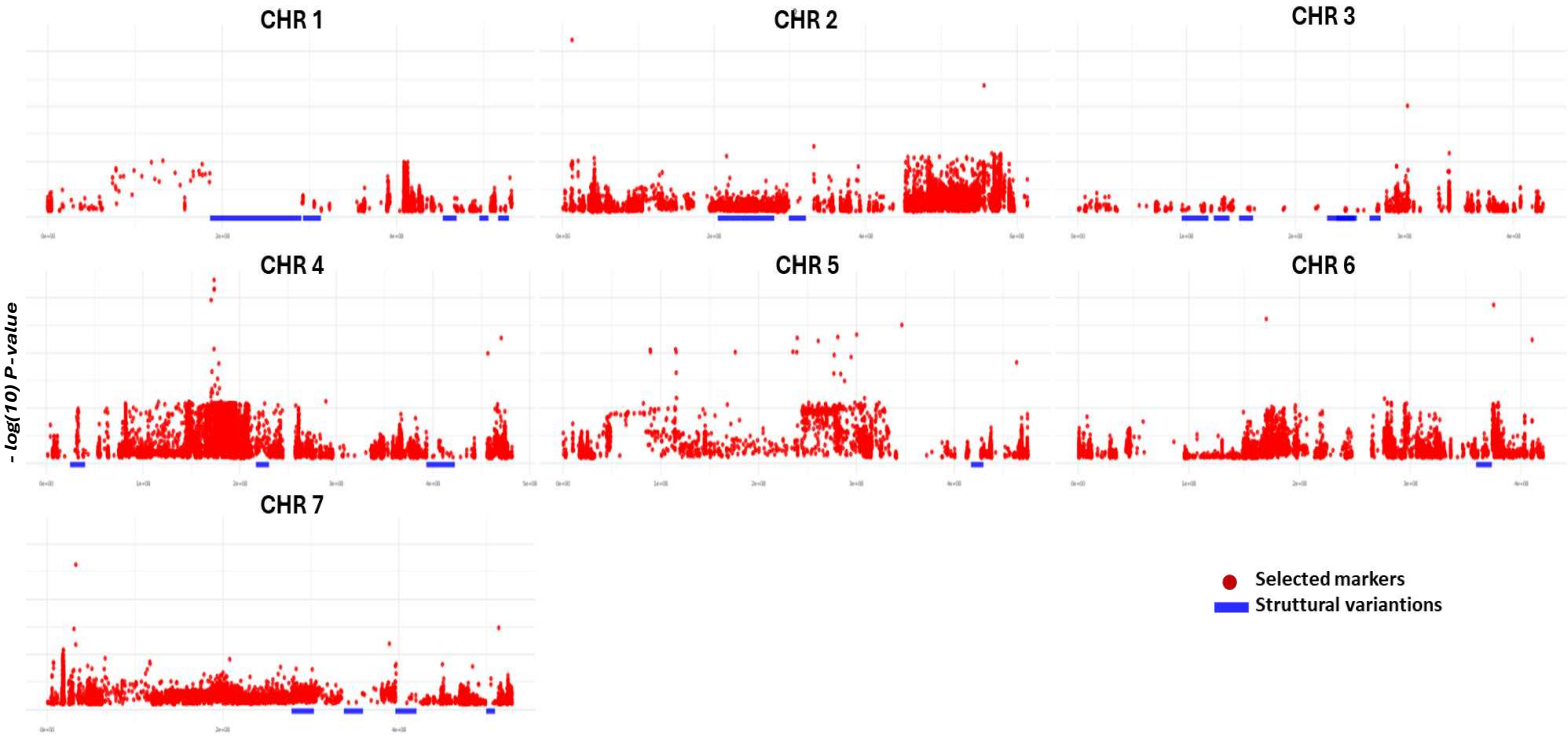


Figure 24. *Manhattan Plot Highlighting Selection Signals, and Structural Variant Regions.*

The Manhattan plot illustrates the p-values obtained from PCAdapt for the 58,433 SNPs shared between the imputed and non-imputed data, as well as the genome scan for selection using the SNNMF test. The 22 regions identified as structural variants through Local PCA are represented by the blue bar.

Genes under selection

For each window, including structural variants, a single gene was selected, resulting in a total of 2,433 genes considered potentially under selection. The primary transcripts of these genes were used to identify 1,048 *Arabidopsis thaliana* orthologs, which showed significant enrichment for Gene Ontology (GO) terms associated with high-altitude agro-environmental conditions. These GO categories include response to UV (GO:0009411), regulation of photosynthesis (GO:0010109), response to temperature stimulus (GO:0009266), regulation of photosynthesis, light reaction (GO:0042548), response to stress (GO:0006950), response to light stimulus (GO:0009416), response to oxygen-containing compound (GO:1901700), cellular response to stress (GO:0033554), response to radiation (GO:0009314), and response to abiotic stimulus (GO:0009628) ***Figure 25***.

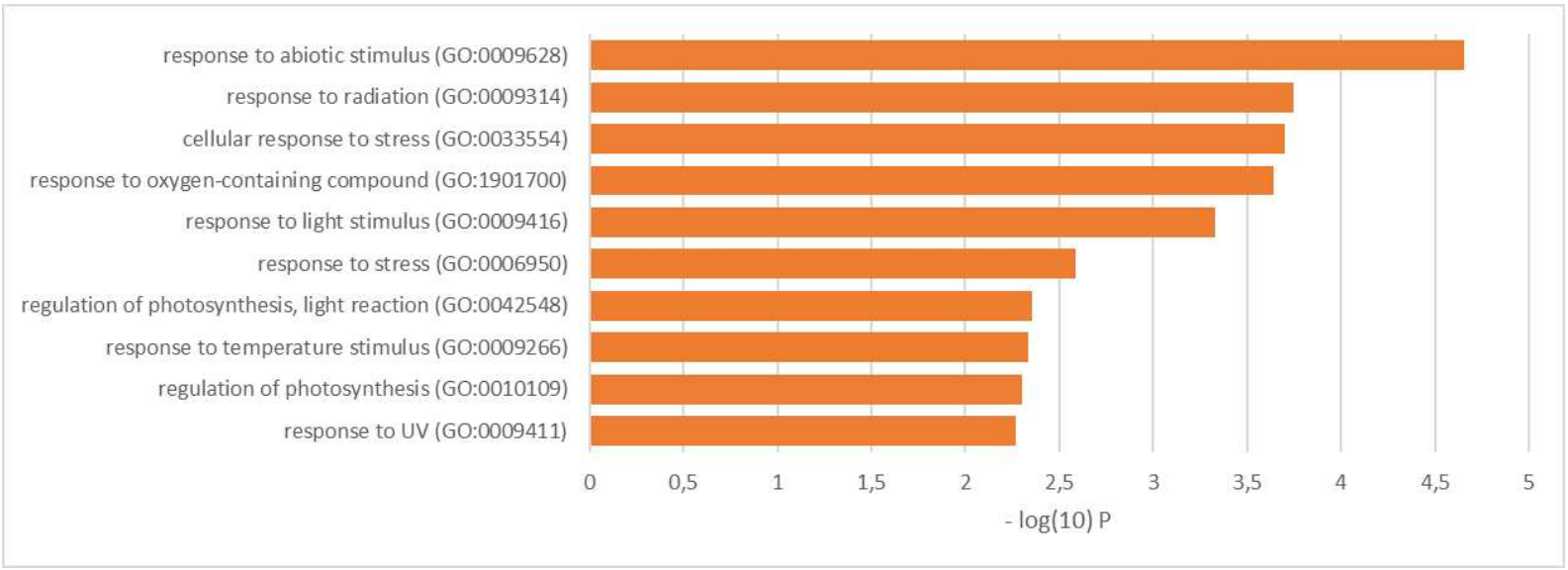


Figure 25. *GO enrichment analysis.*

Bar chart showing the GO enrichment analysis highlighting categories of interest enriched in the context of high-altitude agro-environmental adaptation, obtained from 1,048 Arabidopsis thaliana orthologs.

We identified the gene Lcu.2RBY.6g066820 in chromosome 6, as one of the strongest signals, indicating its potential role in the selection ***Figure 26***. This gene encodes Alanine--Glyoxylate Aminotransferase (AGT), a critical enzyme in photorespiration, which converts alanine and glyoxylate into glycine and pyruvate in peroxisomes. AGT has been linked to stress responses, such as low-temperature adaptation, suggesting its possible involvement in environmental adaptability (Chung et al., 2009).

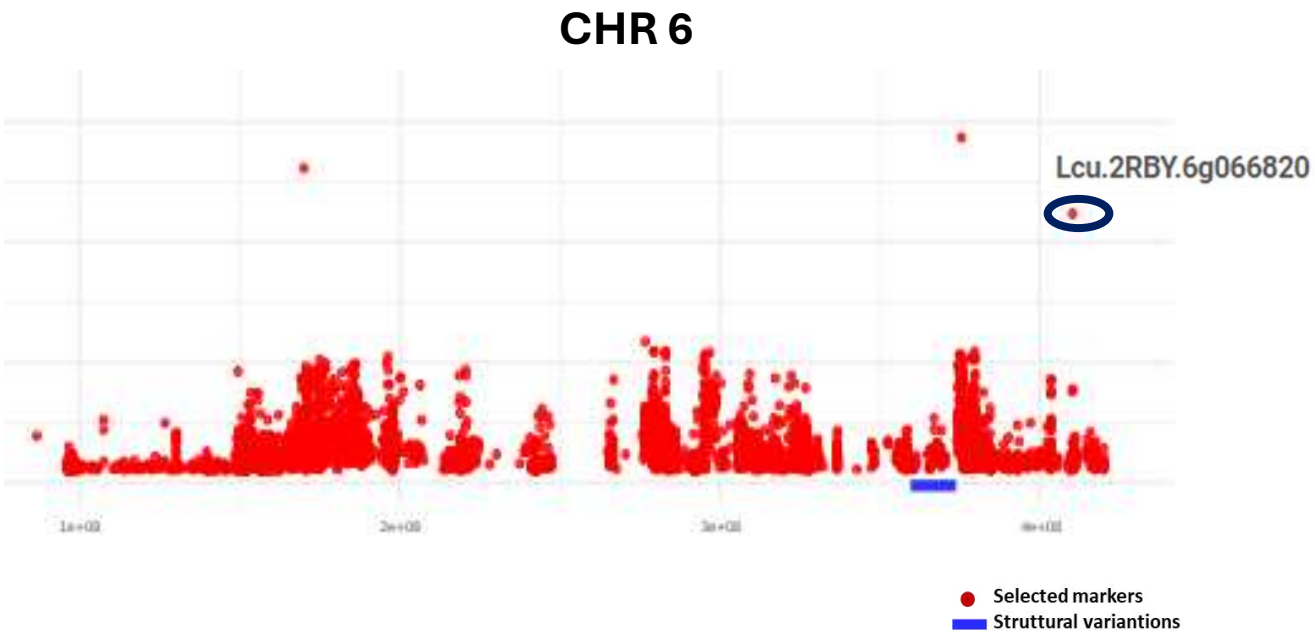


Figure 26. *Selected Marker Associated with the Gene Lcu.2RBY.6g066820 on Chromosome 6.*

The graph shows the -log10 p-values of selected markers on chromosome 6, highlighting the marker that led to the identification of the gene Lcu.2RBY.6g066820

Pcadapt: comparison of results between pruned and non-pruned data

The analysis was performed on pruned data to assess the impact of linkage disequilibrium on PCAdapt's ability to identify markers under selection. These results were then compared with those obtained from the non-pruned data to evaluate any differences. Through pruning, the LD among SNPs in the dataset is reduced by eliminating one of the two variants for each pair with an r^2 value less than 0.5 (methods). As shown in the figure, this leads to an inevitable decrease in the p-values of the identified variants, consequently reducing the sensitivity of the analysis in identifying markers as putatively under selection. These results led us to consider the non-pruned data for our analyses and to develop a different strategy to account for the effect of LD and, consequently, the non-independence of the markers.

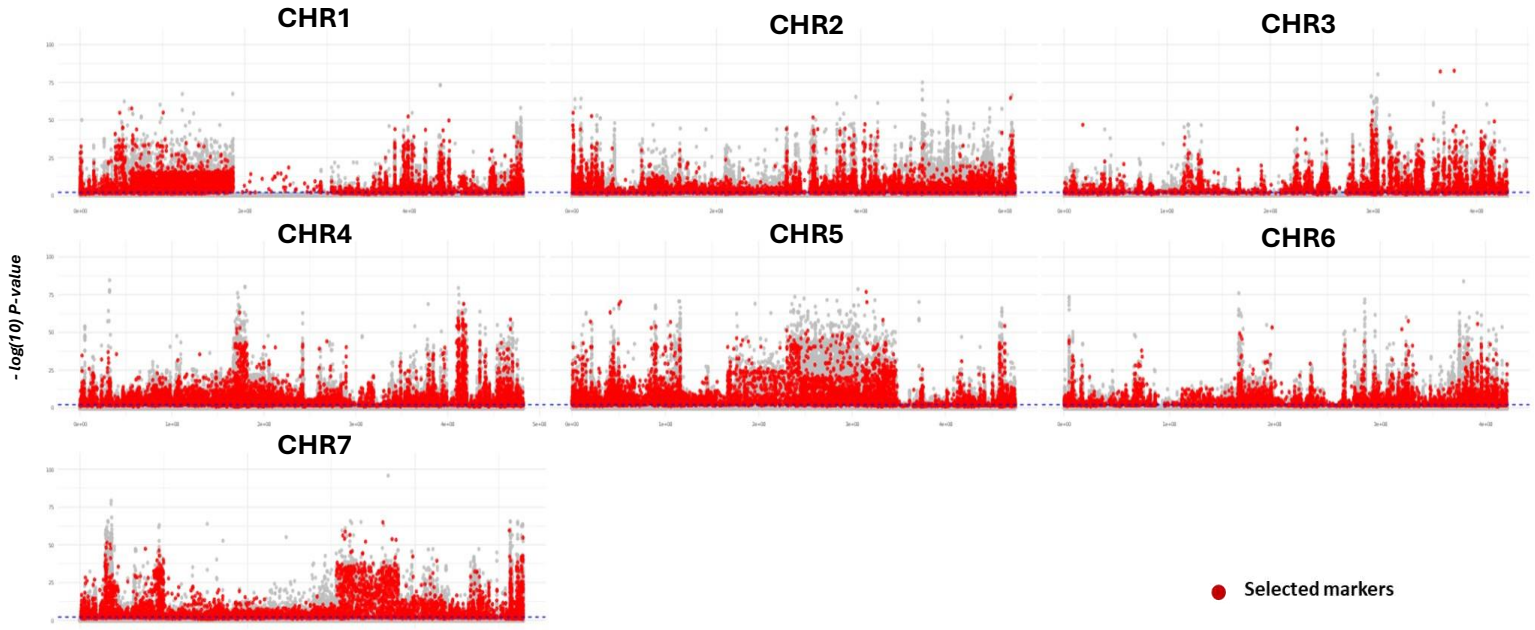


Figure 27. *Comparison of markers under selection between pruned and non-pruned data.*

The figure shows the Manhattan plot divided by chromosome. The gray points represent the -log p-values obtained from PCAdapt on the non-pruned data, while the red points represent those obtained from the pruned dataset. A general decrease in p-values was observed in the pruned dataset.

Discussion

Characterization and Genetic Diversity of the Castelluccio Population

The evidence presented in this study allows us to conclude that implementing CRISPRclean technology combined with Twist multiplexing libraries proved to be an effective approach for genotyping large germplasm collections with complex genomic features such as lentils. Indeed, it allowed us to selectively deplete repetitive elements, and successfully reduce genomic complexity, maintaining focus on informative coding regions in our lentil population. Furthermore, the application of this approach has highlighted the known problem of the incorrect calling of heterozygous genotypes following duplication “events false heterozygotes”. The dataset we used allowed us to develop a model capable of distinguishing artifactual segregating sites from those truly present in heterozygous form. This approach opens up the possibility of developing differential filtering capable of distinguishing between the two types of heterozygotes, those that are genuinely present and the 'false' ones. Since this approach and issue are sensitive to sequencing coverage, further testing and considerations are needed to finalize it, but it opens up the possibility of investigating heterozygosity levels with greater accuracy.

As expected from (Rossato et al., 2023), we observed that this sequencing approach resulted in a higher average coverage of coding regions compared to non-coding ones, leading us to conclude that improved alignment quality occurred in coding regions during the sequencing. The effectiveness of this methodology was validated through the exploration of the genetic diversity of the CS population, which revealed a distinct identity compared to the compared populations. To weigh the impact of missing data, we replicated the analyses on the imputed dataset. At the structure level up to k=5, we obtained comparable results between the two datasets. The imputed data, however, indicate K=6 as the best K; this divergence is consistent with the imputation error rate that we estimated to be approximately 7 percent. Consequently, we deemed it appropriate to proceed with the subsequent analyses using the non-imputed dataset.

Notably, as previously observed by (Zaccardelli et al., 2012), Castelluccio material shows closer genetic similarity with two accessions from the high-altitude Apennine regions compared to other Italian landraces, suggesting the occurrence of a possible adaptive convergence.

Our analysis revealed that the CS population, while genetically diverse, maintains internal homogeneity structured into three genetic groups evenly distributed among sixteen local donors. This structure was further supported by variance analysis, which showed no significant differences in genetic group distribution among donors. Our results show that the CS population originates from three distinct ancestral genetic pools. This result is consistent with the observations of (Zaccardelli et al., 2012) who reported higher diversity in the CS material compared to other Italian landraces.

From the assignment of a 99% purity threshold for the Castelluccio population, we identified both pure and non-pure material within the accessions. The identification of P01 as a pure donor through our molecular analysis strongly validates its pre-established status as reference material for the Castelluccio landrace. These seven local pure donors allow us to establish the core identity of the Castelluccio material. The remaining nine donors exhibit slight levels of introgression, the potential adaptive nature of which merits further investigation.

Adaptation to high altitudes

To evaluate high-altitude adaptation, we compared results from multiple statistical approaches, applying pcadapt to both imputed and non-imputed data to assess the impact of missing data in our dataset, alongside FST-based genome scans. The analysis of *Arabidopsis thaliana* orthologs of our candidate genes revealed enrichment in multiple GO ontology categories related to high-altitude responses, for example, response to UV (GO:0009411), regulation of photosynthesis (GO:0010109), response to temperature

stimulus (GO:0009266), and regulation of photosynthesis, light reaction (GO:0042548), suggesting these genes are indeed under selection and play adaptive roles in the specific agro-environment of the CS population. Among the top candidates in terms of signal strength, we identified also the gene *Lcu.2RBY.6g066820*, which may have a direct role in response to low-temperature conditions. Our observations collectively show that we were able to successfully detect loci truly under selection, even in the presence of linkage disequilibrium (LD), by implementing our strategic approach to genomic region analysis. Another approach possible to reduce the impact of LD on the identification of loci under selection is pruning. However, since this approach acts directly on LD through variant filtering, we observed that it drastically reduces the ability to identify the most interesting peaks in a random manner. For this reason, it proved more suitable to proceed with selection tests without considering their non-independence, and then consider the weight of LD during the screening phase of the results through the setup of windows that had a size evaluated based on LD decay. Consequently, the structural variants identified were treated as non-recombining blocks and considered as single windows with one identified gene each, as our ability to discriminate signals within these regions is limited by high LD levels inside. It is known that structural variants possess high intrinsic adaptive value, as they enable the inheritance of allelic patterns with lower recombination rates, as observed in inversions. The application of a localized PCA approach led to the identification of 22 structural variants. By cross-referencing these with adaptation signals, we observed that some variants showed enrichment of our signals, suggesting their fixation in the population due to their adaptive role. Others, particularly evident in chr1, were completely devoid of adaptation signals. These findings prompt us to consider that the adaptive role of these structural variants warrants deeper investigation, as they represent important patterns in identifying adaptive capabilities and resistance in the CS population's agro-environment.

Conclusion

In conclusion, we applied an innovative complexity reduction approach based on CRISPR-Cas9 to characterize the genetic diversity contained in a landrace population from Castelluccio di Norcia. We effectively investigated the signatures of adaptation to high altitudes that were detectable. Our results allowed us to validate this approach and additionally refine analytical strategies at various levels. This work opens up the possibility of investigating genetic diversity in large and highly repetitive genomes with greater sensitivity and depth, reducing the difficulty of incorporating such genomes into studies involving large collections.

Transition to Chapter III

The following chapter is structured differently from the previous one. Rather than presenting a specific and analytical study of genetic diversity, this chapter focuses on a comprehensive literature review, exploring the challenges and opportunities associated with the use of herbarium specimens in studies investigating genetic diversity. Published in 2023, this chapter highlights both the benefits and the challenges of working with these unique samples. In addition to discussing their potential applications, the chapter examines methodological protocols related to molecular practices, with a particular emphasis on DNA extraction, amplification, and sequencing techniques. It concludes with a discussion on data analysis, providing perspectives on the integration of herbarium specimens in modern genetic research.

The following chapter is presented in its published form. Consequently, the numbering of figures and tables does not align with the sequence used up to this point. Similarly, the bibliographic references are presented independently from the rest of the text.

plants

Chapter III

MDPI

Review

Challenges and Opportunities behind the Use of *Herbaria* in Paleogenomics Studies

Simone Papalini [1], Valerio Di Vittori [1,*], Alice Pieri [1], Marina Allegrezza [1], Giulia Frascarelli [1], Laura Nanni [1], Elena Bitocchi [1], Elisa Bellucci [1], Tania Gioia [2], Luis Guasch Pereira [3], Karolina Susek [4], Maud Tenaillon [5], Kerstin Neumann [6] and Roberto Papa [1,*]

[1] Department of Agricultural, Food and Environmental Sciences, Marche Polytechnic University, 60131 Ancona, Italy; s.papalini@pm.univpm.it (S.P.); a.pieri@univpm.it (A.P.); m.allegrezza@univpm.it (M.A.); g.frascarelli@univpm.it (G.F.); l.nanni@univpm.it (L.N.); e.bitocchi@univpm.it (E.B.); e.bellucci@univpm.it (E.B.)
[2] School of Agricultural, Forestry, Food and Environmental Sciences, University of Basilicata, 85100 Potenza, Italy; tania.gioia@unibas.it
[3] Spanish Plant Genetic Resources National Center, National Institute for Agricultural and Food Research and Technology (CRF-INIA-CSIC), 28805 Alcalá de Henares, Madrid, Spain; luis.guasch@inia.csic.es
[4] Legume Genomics Team, Institute of Plant Genetics, Polish Academy of Sciences, Strzeszynska 34, 60-479 Poznan, Poland; ksus@igr.poznan.pl
[5] Génétique Quantitative et Evolution–Le Moulon, Université Paris-Saclay, INRAE, CNRS, AgroParisTech, 91190 Gif-sur-Yvette, France; maud.tenaillon@inrae.fr
[6] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Seeland, Germany; neumannk@ipk-gatersleben.de
* Correspondence: v.divittori@univpm.it (V.D.V.); r.papa@univpm.it (R.P.)

Abstract: Paleogenomics focuses on the recovery, manipulation, and analysis of ancient DNA (aDNA) from historical or long-dead organisms to reconstruct and analyze their genomes. The aDNA is commonly obtained from remains found in paleontological and archaeological sites, conserved in museums, and in other archival collections. *Herbarium* collections represent a great source of phenotypic and genotypic information, and their exploitation has allowed for inference and clarification of previously unsolved taxonomic and systematic relationships. Moreover, *herbarium* specimens offered a new source for studying phenological traits in plants and for disentangling biogeography and evolutionary scenarios of species. More recently, advances in molecular technologies went in parallel with the decreasing costs of next-generation sequencing (NGS) approaches, which paved the way to the utilization of aDNA for whole-genome studies. Although many studies have been carried out combining modern analytic techniques and ancient samples, such as *herbarium* specimens, this research field is still relatively unexplored due to the need for improving strategies for aDNA manipulation and exploitation from ancient samples. The higher susceptibility of aDNA to degradation and contamination during *herbarium* conservation and manipulation and the occurrence of biochemical postmortem damage can result in a more challenging reconstruction of the original DNA sequence. Here, we review the methodological approaches that have been developed for the exploitation of historical *herbarium* plant materials, such as best practices for aDNA extraction, amplification, and genotyping. We also focus on some strategies to overcome the main problems related to the utilization of *herbarium* specimens for their exploitation in plant evolutionary studies.

Keywords: plant genetic resources; population genomics; crop evolution; food legumes

check for updates

Citation: Papalini, S.; Di Vittori, V.; Pieri, A.; Allegrezza, M.; Frascarelli, G.; Nanni, L.; Bitocchi, E.; Bellucci, E.; Gioia, T.; Pereira, L.G.; et al. Challenges and Opportunities behind the Use of *Herbaria* in Paleogenomics Studies. *Plants* **2023**, *12*, 3452. https://doi.org/10.3390/plants12193452

Academic Editor: Andreas W. Ebert

Received: 30 August 2023
Revised: 26 September 2023
Accepted: 26 September 2023
Published: 30 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. History and State of the Art of Herbaria

Herbarium collections represent large and not entirely explored deposits of genetic and phenotypic information [1] that are able to provide snapshots of the diversity that was present in the past. Initially, *herbaria* consisted of figurative books (Figure 1a) describing

the characteristics of medicinal plants [2]. Over time, *herbaria* evolved to contain preserved plants, seeds, fungi, and algae collected at different phenological stages [3]. *Herbarium* specimens are commonly preserved or dried on paper sheets or stored in folded packets or small boxes, which can be treated with solvents or pesticides to protect the specimens. However, it is also possible to find plant material conserved in liquid solutions that may contain acids, aldehydes, alcohols, or other organic solvents for specimen preservation [4]. More recently, *herbarium* samples are usually provided with supplementary label notes, often including details about the collection site and observations made by the collector, as shown in Figure 1b [5]. The history of modern *herbaria* began in the 16th century, when Luca Ghini (1490–1556), a professor of medical botany at the University of Bologna and Pisa, developed a method to make plant specimens transportable and preservable over time [6,7]. However, his legacy mainly consists of books, manuscripts, and letters, since the dried plants collected by Ghini have been largely lost after his death. Ghini's contribution to the field is highly recognized, also thanks to the *herbaria* established by his disciples. Some notable examples include the oldest surviving *herbarium*, the "*Erbario Cibo*", preserved at the Pontifical University Library in Rome, and the *herbarium* of Ulisse Aldrovandi (1522–1605) that is conserved in Bologna (Figure 1c) [8–11]. In the 18th century, Carl Linnaeus developed his *herbarium* "cabinet" collection, including approximately 14,000 sheets of plant specimens and several zoological specimens [12]. In addition to his *herbaria*, Linnaeus provided innovative instruction on how to establish an *herbarium* collection, including proper techniques for collecting, drying, pressing, and mounting plants onto paper. Linnaeus also emphasized the relevance of collecting closely related specimens and using standardized collecting procedures to facilitate the exchange of materials and information among botanists. His efforts laid the foundations for the establishment of modern *herbarium* collections, which are still essential resources for botanical research and education [13]. In 1935, the International Association for Plant Taxonomy (IAPT) established the "Index *Herbarium*" (http://sweetgum.nybg.org/science/ih/), a comprehensive resource that serves as a global directory of *herbaria* from different independent contributors. Nowadays, it includes almost 400 million specimens provided by more than 3500 active *herbaria* worldwide [14]. Among them, the largest number of specimens are preserved at the "Muséum National d'Histoire Naturelle" (France), at the "New York Botanical Garden" (USA), at the "Komarov Botanical Institute" (Russia), and at the "Royal Botanic Gardens" (UK) [14]. Until the mid-20th century, researchers had two options for examining *herbarium* collections: either by physically visiting the *herbaria* or by requesting specimens on loan for study. Towards the end of the 20th century, significant advances took place with the massive process of digitalization of *herbarium* collections [15], which resulted in the development of virtual *herbaria*. Thus, *herbarium* specimens were available as high-resolution images linked to electronically associated notes [16,17] that are accessible through online platforms. Such databases improved the exploitation and the exchange of plant specimens, facilitating collaborative research and conservation efforts and providing a powerful source for investigating plant diversity and evolution dynamics. However, as discussed below, aDNA can be fragmented, damaged, and present in minute amounts, posing several challenges associated with its exploitation for genomics studies.

1.2. Herbarium Genomics

In recent decades, paleogenomics has greatly benefited from significant advances in the field of molecular biology [18], including those in DNA extraction and amplification procedures and improvements in next-generation sequencing (NGS) strategies. This has enabled researchers to obtain molecular markers as well as whole-genome sequences (WGS) from specimens dating back several centuries, including those preserved in *herbaria*. Paleogenomics studies were also applied to the analysis of pathogens preserved within these specimens, which allowed detailed explorations of temporal signals of divergence among modern and ancient strains [19,20]. As stated above, the characterization and digitization processes of collections have greatly promoted the accessibility and utilization of ancient

samples in a wide range of studies [15]. Traditionally, *herbarium* specimens have been employed for taxonomy and systematics studies. These collections still represent a significant resource for both the discovery and formal description of plant taxa. Moreover, molecular approaches have revolutionized the study of taxonomy, bringing significant advancements in resolving ambiguous phylogenetic relationships and classifications [21]. By exploiting aDNA from now-extinct species or ancient samples, researchers can gain valuable insights into the evolutionary history of different organisms and species and bridge gaps in our understanding of past biodiversity and present genetic diversity, also allowing for the reconstruction of phylogenetic trees [21]. Moreover, taking into consideration that the dissemination of species in novel geographic areas is associated with molecular adaptative mechanisms as well as significant phenotypic and phenological changes [22,23], *herbaria* can be exploited to dissect the plant response to changing environmental conditions across time [24,25]. Indeed, recruiting phenotypic and genotypic information from historical materials collected in different geographical areas is crucial to reconstructing the evolutionary dynamics and adaptation processes of a species. For example, Myers et al. (2022) [26] analyzed phenotypic data acquired from illustrations of the common bean from the XVI century to provide historical support to the dynamics of the introduction of the two main gene pools of this species (i.e., Mesoamerican and Andean) from the Americas to Europe following the travels of Columbus. The application of molecular tools to ancient biological samples might provide the opportunity to investigate population dynamics across different time points as a result of migration processes, gene flow, occurrence of novel mutations, and changes at the genome level also driven by genetic drift and by natural selection, although limitations can arise depending on the quality of the aDNA. This makes it possible to investigate the complex interplay between genetics, environment, and evolution of a population across time [27,28]. In genomic studies, the comparison of data from *herbaria* and modern samples can provide valuable information for optimizing the management of plant genetic resources [29]. Cozzolino et al. (2007) [30] evaluated changes in the genetic diversity of the *Anacamptis palustris* population because of habitat transformation in Italy over time. To do so, authors sequenced and compared a plastid region between *herbarium* specimens collected prior to the Second World War and modern samples. Thus, this approach could be very useful as a complement to large genotyping and phenotyping programs of modern accessions aimed at understanding the environmental adaptation and the evolutionary dynamics of genetic diversity, such as the INCRESE project [31], which is focused on food legumes such as chickpeas [32], the common bean [33], lupin [34], and lentil [35]. However, from the work carried out by Cozzolino et al. (2007) [30], some possible limitations of complementing modern collections with *herbarium* collections have arisen due to the limited sample size (i.e., the number of specimens) and potential methodological biases in collecting procedures. For instance, only a few individuals could have been collected from each collection site, which may not capture the full extent of genetic diversity within a population. Despite these limitations, *herbaria* can be considered a valuable source for detecting variations in the distribution and frequency of haplotypes and alleles.

1.3. Challenges in the Use of Herbaria and Ancient Samples for Genomics Studies

The advent of NGS platforms has opened unprecedented possibilities for studying genome-level diversity from *herbaria* and ancient samples. However, for this purpose, the successful implementation of NGS may be challenging due to the unique nature and the issues associated with the manipulation of ancient samples, such as damage and fragmentation of the aDNA. The development of robust protocols for aDNA purification, amplification, and sequence data analysis is crucial for obtaining reliable results [18]. The aDNA from *herbaria* and ancient samples is often degraded and fragmented, and it is often challenging to purify fragments longer than 500 bp [36–38]; thus, the optimization of extraction protocols can represent a crucial bottleneck in obtaining high-quality aDNA suitable for sequencing. Additionally, the limited amount of aDNA that can be rescued from ancient specimens poses the risk of contamination between samples and from modern

DNA while handling and processing [39,40]. Moreover, aDNA in *herbarium*-preserved samples can be susceptible to biochemical postmortem damage [41]. Indeed, over time, the aDNA molecules can undergo various forms of degradation, including fragmentation and chemical modifications; these damages pose additional challenges when the original DNA sequence must be assembled and in the analysis of sequences. The biochemical postmortem damage in *herbarium* samples can result from different factors, such as light exposure, not-controlled moisture, temperature fluctuations, and the presence of reactive chemicals. These factors can lead to DNA strand breaks, base modifications, and crosslinking of DNA molecules, which can affect the quality and integrity of the assembled sequence [42]. Efficient protocols have been developed to overcome these issues, especially regarding aDNA purification. Here, we report the most relevant protocols for aDNA purification from *herbaria* specimens (Table 1). Furthermore, the choice of the bioinformatics pipeline and of the strategy for the analysis of NGS data from *herbaria* and ancient samples is particularly relevant, especially in the case of low-quality and fragmented aDNA. In the present review, we want to give insight into the best strategies for the extraction, purification, amplification, and sequencing of *herbarium* aDNA, which can be exploited in genetic diversity and evolutionary studies of plant species as well as in the reconstruction of phylogenetic and taxonomic relationships.



Figure 1. *Herbaria* illustrations. (**a**) The oldest illustration of a European common bean plant (*Phaseolus vulgaris*) from "*Di Historias Stirpium*" (Fuchs, 1542); "Courtesy of Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh, PA". (**b**) *Phaseolus vulgaris* specimen from the Muséum National d'Histoire Naturelle (MNHN) (France), Paris, Collection: Vascular plants (P), Specimen P02872197, dated 1833. The label notes include information about the collector, the collection site, and the date. (**c**) *Phaseolus vulgaris* specimen from the *herbarium* of Ulisse Aldrovandi; "Courtesy of Alma Mater Studiorum University of Bologna—University Museum System—Herbarium and Botanical Garden".

Table 1. Summary of relevant aDNA purification protocols from *herbaria* specimens. The source of plant material, timing of the sampling before storage, success with the tested samples, procedures for the evaluation of aDNA quality, and the reference are provided for each of the tested protocols.

DNA Extraction/Purification Protocol	Source of Plant Material	Timing of Sampling	Suitable for Extracting *Herbarium* DNA	Quality Evaluation Approach	Reference
CTAB, according to the protocol of Doyle and Doyle (1990)	*Juncus* and *Luzula genera (Juncaceae)*	*	Yes	PCR amplification	[43]
CTAB + pre-wash with a sorbitol-containing buffer	*Lafoensia* spp.	N.A.	Yes	PCR amplification	[44]

material is a reliable preservation method, preventing degradation and better preserving aDNA when compared to alternative practices, such as microwaving, boiling, or immersion in chemical solutions [37]. Moreover, it has been observed that well-preserved older materials tend to yield higher quantities of purified aDNA of better quality when compared to younger materials that have been subjected to unfavorable storage conditions [38,43,54]. As a general consideration, we also have to take into account that the initial condition of the tissues at the sampling stage, before storage, inevitably impacts the quality and the quantity of the purified aDNA [55]. Overall, the purification from *herbarium* specimens results in a lower yield of DNA when compared to the utilization of fresh plant tissues [4,56]. In their pioneering work, Rogers and Bendich (1985) [45] employed a modified CTAB (cetyltrimethylammonium bromide) nucleic acid extraction protocol to purify DNA from a small number of mummified seeds, *herbarium* specimens, and fresh plant tissues. Although this extraction method confirmed a higher yield when applied to fresh tissue, it has also proven to be effective for the extraction and purification of aDNA from ancient plant materials. A relevant step during aDNA purification is the grinding of the plant material obtained from the *herbarium* specimen. One of the most adopted practices relies on bead-based homogenization equipment, such as a mixer mill or bead mill. As noted by Drábková (2014) [57], this approach can reduce aDNA loss and the risk of contamination compared to alternative methods, such as grinding in mortars with liquid nitrogen. Additionally, the use of a bead-based homogenization approach enables the simultaneous processing of multiple samples, ensuring consistent and homogenous results [43]. It is important to remember that best practices must always be adopted in the laboratory to prevent contamination since the use of bead-based grinding methods might introduce a risk of contamination, which can particularly impact the analysis of aDNA. Over the past two decades, numerous aDNA extraction protocols have been developed, tested, and compared to optimize the quality and quantity of purified aDNA from *herbaria* specimens, for which a report is provided in Table 1. Drabkova et al. (2002) [43] tested seven DNA extraction protocols in *herbarium* specimens from *Juncus* and *Luzula* genera (*Juncaceae*) species collected during the twentieth century, including fresh tissue materials as a control. They evaluated quality parameters, such as the aDNA A260/280 and A230/280 absorbance ratios, and estimated the quantity of aDNA through spectrometry. From their results, they concluded that the use of a DNeasy Plant Kit (Qiagen), with some modifications, was the best approach for the aDNA purification. In detail, the authors found that both optimal homogenization of the plant tissue and extension of the precipitation time were crucial steps. Modifications to the manufacturer's protocol included a longer cell lysis time (i.e., 30 min), an increased volume of AP1 buffer (i.e., 450 µL), a reduced volume of elution buffer (i.e., 50 µL) that increased aDNA concentration, and an extended elution time (i.e., 10 min). They also pointed out the protocols that gave the worst results, particularly the phenol extraction due to the potential contamination of aDNA with phenol, which can negatively affect downstream reactions, including amplification and sequencing, and, similarly, the CsCl gradients protocol, as aDNA could be lost within the gradient due to the limited amount of tissue [43].

Preserving the integrity of museum collections, which prioritizes maintaining specimens in their original state [58], poses a significant challenge when extracting DNA from ancient materials. To address this issue, efficient DNA extraction protocols that use minimal tissue have been established for historical specimens. Shepherd (2017) [59] introduced a non-invasive approach to sampling material directly from *herbaria* by using a Staedtler "Mars Plastic" eraser that minimized damage to the specimens while obtaining the necessary amount of tissue for DNA extraction. Sugita et al. (2020) [48], taking as an example a previously published protocol for DNA extraction in arthropods, established a plant-suitable non-disruptive protocol, particularly useful in species with small (i.e., <25 mm^2) and fragile leaves [60,61]. Nevertheless, despite the importance of non-disruptive methods, CTAB and modified DNeasy Plant Mini Kit (QIAGEN) protocols are the most frequently adopted methods for aDNA extraction [57]. Among the reasons for this, aDNA-specific protocols are generally more expensive, more time-consuming, and require specific proto-

cols and facilities to avoid contamination. As stated above, storage conditions for *herbarium* specimens (e.g., fluctuating temperatures or high humidity) and the species from which the tissue is obtained can influence the efficiency of the aDNA extraction [62]; thus, such factors must be taken into account when choosing the aDNA extraction protocol. With this regard, Marinček et al. (2022) [63] compared the standard Qiagen DNeasy Plant Mini Kit and a specific dithiothreitol (PTB)-dithiothreitol (DTT) protocol, previously developed by Gutaker et al. (2017) [49], with the aim of evaluating the efficiency of these protocols on specimens from the genera *Xanthium L.* and *Salix L.* They showed a lower efficiency of the Qiagen DNeasy Plant Mini Kit when working with older specimens (i.e., collected before 1900). Thus, a researcher might want to consider the PTB-DTT protocol as an alternative for aDNA extraction in older and more challenging specimens, for example, due to the higher concentration of secondary metabolites in the tissue. Indeed, several protocols have been developed, specifically taking into account the issues related to the presence of secondary metabolites that can inhibit PCR reactions. Ribeiro and Lovato (2007) [51] tested five DNA extraction protocols on fresh and *herbarium* leaves of various species belonging to the genus *Dalbergia*, known to contain significant amounts of secondary metabolites potentially interfering with DNA amplification. Based on their results, the most efficient protocol was developed by Jobes et al. (1995) [64]. This protocol uses three key reagents: polyvinylpyrrolidone (PVP), which binds phenolic compounds, sodium chloride with a high molar concentration, which inhibits the co-precipitation of polysaccharides and DNA and, in turn, enhances the solubility of polysaccharides in ethanol, and, finally, lithium chloride, which is useful to selectively precipitate and subsequently remove RNA. Krinitsina et al. (2015) [52] proposed a cost-effective DNA extraction method specifically designed for *herbarium* specimens that involves the utilization of AMPure XP magnetic beads diluted in a buffer containing PEG 8000. The authors suggest that the use of magnetic beads can reduce the concentration of PCR inhibitors in the aDNA. Hofreiter (2012) [65] adapted a protocol that had been originally designed for isolating aDNA from human hair samples [66] to ancient plant specimens. In particular, the protocol includes two key steps: a phenol–chloroform extraction followed by a silica spin column purification. Phenol–chloroform extraction has proven to be effective for those samples containing low amounts of DNA and that may contain polymerase chain reaction (PCR) inhibitors, affecting the enzymatic activity of the Taq polymerase [38,67]; although, one may consider minimizing the use of phenol-chloroform due to the environmental hazards associated with its use. However, the use of phenol–chloroform for DNA extraction can introduce contamination issues associated with phenol–chloroform residues that can significantly impact downstream processes, particularly DNA quantification and PCR, potentially compromising data accuracy [43].

Overall, selecting the most appropriate aDNA purification protocols involves careful consideration of factors such as preservation conditions and the specific goals of the research. By taking these factors into account, researchers can optimize their DNA purification methods and enhance the chances of obtaining high-quality aDNA for subsequent analysis (Figure 2).

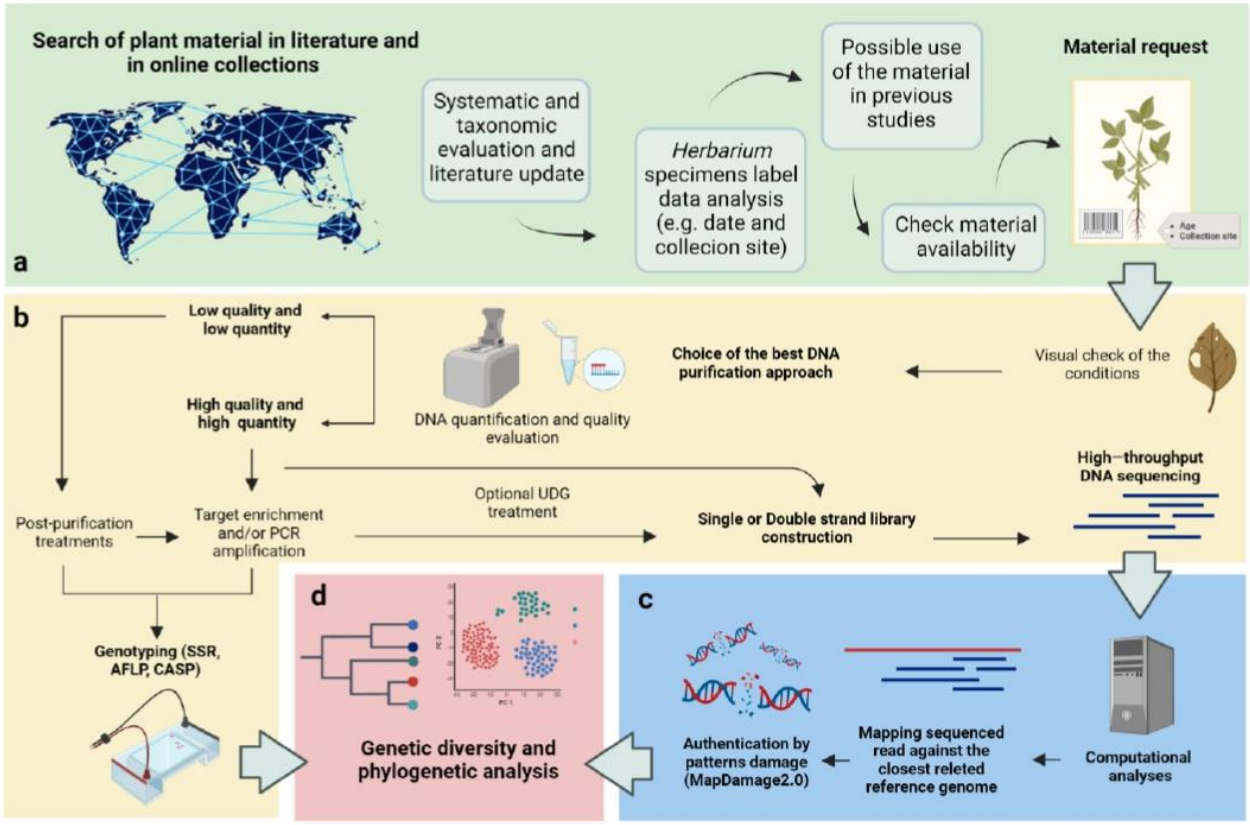


Figure 2. Workflow for the exploitation of *herbarium* specimens for genomic studies. (**a**) A preliminary investigation allows for gaining a comprehensive understanding of the available information on *herbarium* specimens. The use of web tools, such as web *herbaria*, academic databases, and online libraries, allows for finding reliable and up-to-date sources and conducting proper research before making any material requests; (**b**) the identification and selection of the best protocol—aDNA extraction, purification, or genotyping—to obtain molecular information from a specimen should take into account the characteristics of the material, available resources, and research objectives; (**c**) bioinformatics approaches are needed for data processing; (**d**) maximizing the quality of sequencing data is of great importance, as it significantly enhances the chances of reconstructing the phylogeny, understanding the demographic history of a species, and identifying selection signatures in response to natural or human-driven selection. Created using BioRender.com.

3. DNA Amplification by PCR from *Herbarium* Specimens

Since its development, PCR has become a routine and indispensable technique in several protocols for molecular biology studies that can be extended to the analysis of ancient samples. Several studies reported on the successful amplification of aDNA from plant *herbarium* material collected over a wide range of ages, resulting in variable lengths of amplified fragments [38,46,68]. Staats et al. (2011) [37] also reported that there are no significant differences in PCR amplification efficiency among plastid, mitochondrial, and nuclear aDNA from *herbarium* material, similar to what can be generally observed in amplification from plastid, mitochondrial, and nuclear DNA obtained from fresh tissue. As mentioned, DNA extraction from *herbarium* material can often yield low quantities of fragmented genetic material. However, the quality and purity, rather than the quantity, of the aDNA template mostly affect PCR amplification efficiency [38]. Quality control procedures, including rigorous purification steps and measures for minimizing and identifying contaminations, are crucial to ensuring the accuracy and reliability of the PCR amplification. As stated above, the presence of secondary metabolites in plant cells from fresh

tissues, and thus in *herbarium* samples, can have a negative impact on the amplification efficiency. To address this issue, various extraction protocols have been developed with the aim of reducing the concentration of these metabolites [47,51,52]. In order to mitigate the effects of PCR inhibitors, as a post-aDNA purification step, a potential solution is diluting DNA extracts. However, this approach is not always feasible for *herbarium* samples due to the typically low yield and concentration of the eluted aDNA obtained from these specimens [69]. An alternative approach to improving the quality and quantity of the aDNA template available for PCR amplification in the presence of potential inhibitors has been suggested by Samarakoon et al. (2013) [69]; it involves the use of a reagent called TBT-PAR, which contains trehalose, bovine serum albumin (BSA), and polysorbate-20 (Tween-20). Multiple studies investigated the effects of using high concentrations of BSA on PCR amplification efficiency when working with poor-quality template aDNA. Overall, these studies agreed that a higher concentration of BSA can have a strong positive impact on PCR efficiency [51,57]. In particular, Särkinen et al. (2012) [38] suggested the utilization of a high concentration of BSA when routinely amplifying aDNA from *herbarium* specimens. Several factors can indeed affect PCR amplification efficiency, and among these factors are the purity and specific type of polymerase and the purity of buffers and reagents used. It is crucial to consider these factors to ensure reliable and consistent PCR results, especially in amplifying aDNA [70]. A suitable and cost-effective alternative to mitigate the effects of potential PCR inhibitors is the adoption of a polymerase enzyme that is less susceptible to their presence [71]. Särkinen et al. (2012) [38] suggest that the utilization of a polymerase without proofreading activity may perform better than one with proofreading activity. Moreover, it is essential to investigate and optimize the performance of different polymerases to improve the amplification efficiency for aDNA. Furthermore, Särkinen et al. (2012) [38] reported a significant negative correlation between the length of the aDNA template fragments from *herbarium* specimens and PCR efficiency, that is, a strong reduction in the success rate of the amplification when the template regions have a greater length than 100 pb. Thus, this suggests that shorter fragments can be easily amplified, which is an interesting aspect considering the fragmented nature of the aDNA template usually obtained from these specimens. However, in the case of extreme aDNA fragmentation, amplification of certain loci may be challenging. It is worth noting that mitochondrial and chloroplast aDNA can also be successfully retrieved from severely degraded samples, being present in high copy numbers within a cell. Moreover, there is a higher chance that some mtDNA and cpDNA fragments are long enough for amplification, even when nuclear DNA is highly fragmented [72]. Another relevant aspect is the instability of nucleic acids during long-term preservation, which can lead to the formation of various postmortem damages, affecting the quantity and quality of template aDNA for the amplification; such damages can also introduce potential errors that make the reconstruction of the aDNA sequence challenging. An example of postmortem damage is depurination, which can result in the formation of single- or double-strand breaks in the DNA molecule. Additionally, crosslinking reactions can occur, limiting or even preventing DNA amplification altogether [73]. Thus, it is crucial to identify the best methodologies and strategies to take into account the occurrence of postmortem damage when working with *herbarium* samples in order to mitigate their effects on quality and amplification efficiency. Indeed, oxidative and hydrolytic modifications of bases can lead to the formation of miscoding lesions, such as the deamination of cytosine to uracil. Thus, these lesions can lead to the incorporation of incorrect bases during the amplification process [37]. Certain oxidative damage can also create lesions that block the polymerase enzyme and promote the generation of chimeric sequences through 'jumping PCR' [41,74].To address the PCR artifact caused by misincorporation, Uracil-N-glycosylase (UNG) can be employed prior to amplification to remove deaminated cytosines [41,75]. UNG creates an abasic site that is subsequently hydrolyzed through β-elimination, resulting in a strand break [41,76]. However, it is important to carefully consider the use of UNG treatments since it inevitably leads to a reduction in the aDNA sequence length, which is often highly fragmented [77]. In addition to the sensitivity

of PCR amplification, the low template aDNA concentration, and the fragmented nature of potential target loci in the aDNA, the presence of exogenous DNA contamination is also a common challenge in aDNA manipulation. Indeed, exogenous DNA might have been introduced during the collection and conservation of the plant material, by the presence of microorganisms or pathogens [39], or by handling samples during molecular processes. Kistler et al., 2020 [40] emphasized the importance of manipulating *herbarium* specimens in a physically isolated laboratory during all the steps. Nevertheless, DNA contaminants can still be introduced into the experimental workflow through various means, as summarized in Figure 3, including the use of contaminated reagents or samples as well as the presence of residual DNA and PCR amplification products from previous experiments [78].

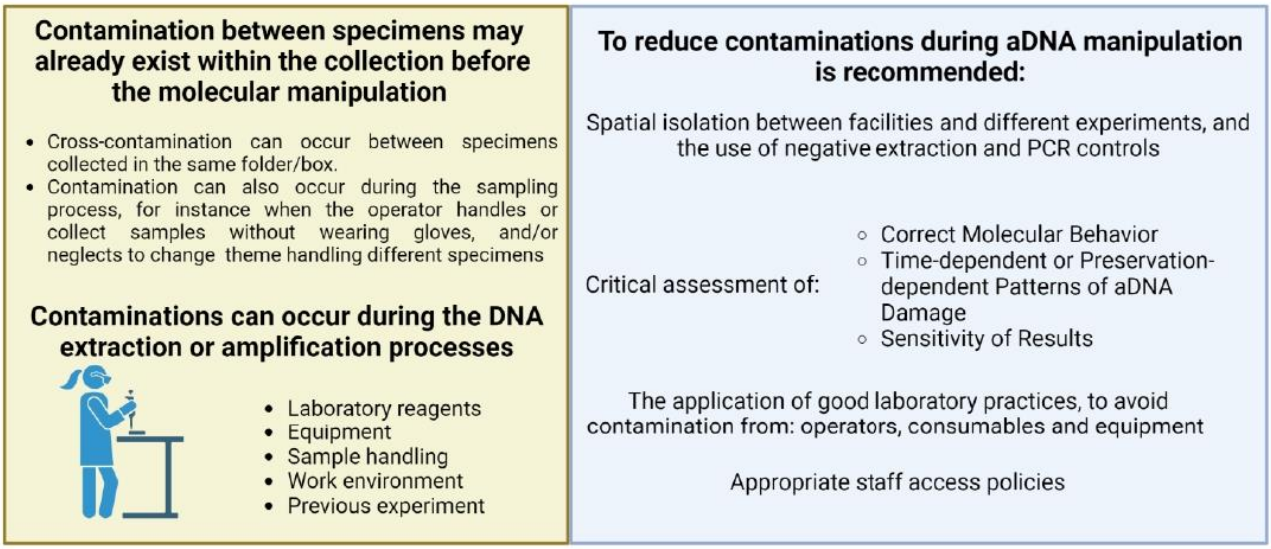


Figure 3. Contamination represents a challenge to acquire endogenous aDNA. Contamination sources and guidelines for minimizing negative impact during molecular manipulation of *herbarium* samples. Created using BioRender.com.

To minimize the risk of contamination, Knapp et al. (2011) [79] developed a set of guidelines for establishing a molecular laboratory exclusively dedicated to studies on ancient specimens. The guidelines recommend conducting the different steps of aDNA extraction and amplification in dedicated rooms and/or hoods, expanding the concept of spatial isolation within the same laboratory. Furthermore, the authors highlight the importance of implementing a limited access policy, granting facility access only to qualified personnel who are aware of the risks of contamination.

4. Next-Generation Sequencing (NGS) and Genotyping on *Herbarium* Specimens to Disentangle Relevant Aspects of the Evolutionary History of a Species

Several studies investigated the genetic diversity preserved in herbarium collections, providing insights into relevant aspects of plant species' evolutionary histories, including adaptation processes and the effects of environmental changes over time. Several approaches have been employed to genotype herbarium specimens [80–82]. Malenica et al. (2011) [83] successfully genotyped a 90-year-old Tribidrag grapevine herbarium specimen by utilizing a set of nine microsatellite markers (SSRs) and a whole-genome amplification (WGA) procedure. Other amplification-based genotyping approaches have proven to be suitable for molecular analysis of herbarium samples despite the limitations in the aDNA amplification from such materials, as described above [46,72]. As an example, while AFLP fingerprinting can be utilized to analyze herbarium material, the fragmented nature of the purified aDNA can introduce a bias when utilizing such an approach since AFLP is

based on the presence and distribution of restriction sites. In such cases, Lambertini et al. (2008) [84] suggested using only those markers that can also be detected in fresh tissue and, in parallel, increasing the number of primer combinations to ensure the amplification of an adequate number of polymorphic fragments. By studying chloroplast and nuclear microsatellite diversity in 57 historical herbarium specimens, Roullier et al. (2013) [85] were able to trace relevant events in the evolutionary history of the sweet potato over time, such as the effect of the domestication process, migration patterns, and genetic interactions. By comparing the genetic data from historical herbarium collections with that of modern samples, researchers can assess changes in the level of genetic diversity and identify potential genetic bottlenecks or shifts that occurred during domestication and cultivation. The emergence of next-generation sequencing (NGS) led to a revolutionary change in our ability to obtain multi-locus genetic data from natural historical collections [72]. An approach that has significantly impacted this field of research is the Sequencing by Synthesis (SBS) strategy, which has been developed to enable cost-effective shotgun sequencing of whole genomes [86]. High-throughput SBS technologies involve the preparation of sequencing libraries, which entails attaching artificial DNA segments known as adapters to both ends of those template fragments characterized by a specific range of fragments [81]. This strategy allows for efficient and accurate sequencing of the aDNA fragments. When working with aDNA, differently from the SBS procedure for library preparation in modern samples, the fragmentation step that comes before adapter ligation can be skipped [81]. The adapters used in sequencing library preparation can serve multiple purposes. Indeed, they enable the priming of both whole-genome shotgun sequencing or the enrichment of specific genomic regions of interest through hybridization capture techniques [81]. These approaches allow for the characterization of various types of DNA, such as organellar DNA or nuclear loci, as well as the detection of a vast number of single-nucleotide polymorphisms (SNPs) distributed throughout the nuclear genome. Generally, library preparation protocols for *herbarium* DNA templates do not require significant modifications compared to standard ones. Most of these protocols were originally developed for Illumina sequencing-based platforms and can be broadly categorized into two main types: the single-stranded library (ss-library) and the double-stranded library (ds-library) construction methods. Briefly, ss-library preparation starts with heat denaturation of DNA followed by the attachment of a biotinylated adapter oligonucleotide to the 3′ ends. A second adapter is then blunt end-ligated to complete the library preparation, which is further amplified through PCR [87]. Whereas, ds-library preparation can be further categorized into Blunt-End and Y-adapter methods, both involving the end-repairing of the DNA fragments and ligation of double-strand adapters but differing in the type of adapter used. Bennett et al. (2014) [88] tested all the different methodologies for library preparation, starting from aDNA faunal and human remains, finding that the Y-adapter method led to the formation of adapter–dimer artifacts, while the ss-library approach allowed for an increased proportion of shorter endogenous molecules incorporated into the libraries. Their outcome has been further confirmed in recent research, suggesting that the ss-library preparation approach is better suited to the aDNA features compared to double-stranded library protocols [41], exhibiting greater sensitivity to degraded and poorly preserved ancient samples, such as *herbarium* specimens [89]. In protocols that require the amplification of libraries by PCR, the presence of artifacts caused by misincorporation can impact sequencing accuracy and efficiency. To mitigate the effects of postmortem damage, UDG enzyme-based protocols (see previous section) can be employed during library preparation [90]. One of the key advantages of using UDG is that it helps eliminate misincorporations (C to T and G to A) in the recovered sequences, improving the mapping efficiency. This is particularly useful when the reference sequence belongs to a species that is distantly related to the query sequence, which reduces the number of possible gaps and mismatches in the alignment [89]. However, the presence of uracil and other types of postmortem damage can also serve as an indicator to assess the presence of modern contamination in aDNA samples [91]. Damage profiles can be inferred using computational methods such as mapDamage 2.0, which allows for distinguishing

ancient molecules from modern contaminants [92,93]. Indeed, the aDNA damage pattern is a crucial factor that can significantly impact aDNA analysis and the authentication of ancient sequences obtained from *herbarium* samples. Another strategy for addressing errors in sequence reconstruction resulting from misincorporations in aDNA is to focus downstream analysis on transversions rather than transitions since the latter are more susceptible to postmortem damage. Trucchi et al. (2021) [94] applied this strategy in their paleontological study of ancient bean seeds from different archaeological sites in Argentina. They analyzed the temporal dynamics of genetic variation and selection during the domestication process of the common bean (*Phaseolus vulgaris* L.) in the southern Andes by comparing WGS data from a panel of 15 ancient beans dating between 2500 and 600 years ago and modern wild and domesticated common bean accessions (from both Mesoamerican and Andean gene pools). The work of Trucchi et al. (2021) [94] clearly demonstrated how, by considering transversions, researchers can significantly reduce the impact of postmortem damage on sequence analysis and improve the accuracy of genetic variation inference in aDNA studies.

5. Conclusions

The employment of *herbaria* in paleogenomics can offer an opportunity for investigating the genetic and phenotypic diversities of ancient samples compared to modern ones, being available plant genetic resources that are largely unexplored. This requires optimized strategies for *herbaria* handling and for the analysis of the aDNA, including the employment of advanced next-generation sequencing approaches and computational methods to authenticate and analyze aDNA data and access an unexplored source of genetic information. However, the quality of attainable genetic data can be strongly affected by the protocols used for aDNA extraction, amplification, and sequencing due to several factors, such as the preservation state of the specimens and the aDNA. Consequently, an in-depth understanding of the strengths and limitations of available protocols for aDNA exploitation becomes crucial. Here, we reviewed the most recommended strategies and approaches for handling *herbaria* specimens and aDNA. Moreover, we discussed the relevance of integrating available high-quality ancient sequence data from *herbarium* specimens with modern sequences to perform population genetics and genomics studies and inferences on the evolution of a species.

Author Contributions: S.P., V.D.V., A.P. and R.P. conceived the structure and wrote and contributed to the drafting of the manuscript. S.P., V.D.V., A.P., M.A., G.F., L.N., E.B. (Elena Bitocchi), E.B. (Elisa Bellucci), T.G., L.G.P., K.S., M.T., K.N. and R.P. contributed to the editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the INCREASE project (European Union's Horizon 2020 research and innovation programme under grant agreement No. 862862) and by the Italian Government (MIUR) Grant n. 20177RL4KL, Project Progetti di Ricerca di Interesse Nazionale (PRIN PARDOM) 2017.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bebber, D.P.; Carine, M.A.; Wood, J.R.I.; Wortley, A.H.; Harris, D.J.; Prance, G.T.; Davidse, G.; Paige, J.; Pennington, T.D.; Robson, N.K.B.; et al. Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 22169–22171. [CrossRef] [PubMed]
2. Zunic, L.; Skrbo, A.; Dobraca, A. Historical Contribution of Pharmaceutics to Botany and Pharmacognosy Development. *Mater. Socio Medica* **2017**, *29*, 291–300. [CrossRef] [PubMed]
3. Willis, C.G.; Ellwood, E.R.; Primack, R.B.; Davis, C.C.; Pearson, K.D.; Gallinat, A.S.; Yost, J.M.; Nelson, G.; Mazer, S.J.; Rossington, N.L.; et al. Old Plants, New Tricks: Phenological Research Using Herbarium Specimens. *Trends Ecol. Evol.* **2017**, *32*, 531–546. [CrossRef] [PubMed]

4. Taylor, J.W.; Swann, E.C. DNA from Herbarium Specimens. In *Ancient DNA: Recovery and Analysis of Genetic Material from Paleontological, Archaeological, Museum, Medical, and Forensic Specimens*; Herrmann, B., Hummel, S., Eds.; Springer: New York, NY, USA, 1994; pp. 166–181. [CrossRef]
5. Heberling, J.M. Herbaria as Big Data Sources of Plant Traits. *Int. J. Plant Sci.* **2022**, *183*, 87–118. [CrossRef]
6. Von Engelhardt, D. *Luca Ghini (1490–1556): Il Padre Fondatore Della Botanica Moderna Nel Contesto dei Rapporti Scientifici Europei del Sedicesimo Secolo*; Museo: New York, NY, USA, 2011; Volume 27.
7. Bellorini, C. *The World of Plants in Renaissance Tuscany*; Routledge: London, UK, 2016. [CrossRef]
8. Cristofolini, G.; Mossetti, U. Pre-Linnean herbaria in Bologna: Some Newly Discovered Collections from the Time of Ulisse Aldrovandi. 1993. Available online: https://www.researchgate.net/publication/260179625 (accessed on 25 August 2023).
9. Nepi, C. La "slegatura" dell'erbario di Andrea Cesalpino (1525–1603). *Museol. Sci. Nuova Ser.* **2007**, *1*, 50–54.
10. Signorini, M.A. Sulle Piante Dipinte Dal Bachiacca Nello Scrittoio Di Cosimo I A Palazzo Vecchi. 1993. Available online: https://www.jstor.org/stable/27654359 (accessed on 25 August 2023).
11. Stefanaki, A.; Porck, H.; Grimaldi, I.M.; Thurn, N.; Pugliano, V.; Kardinaal, A.; Salemink, J.; Thijsse, G.; Chavannes-Mazel, C.; Kwakkel, E.; et al. Breaking the silence of the 500-year-old smiling garden of everlasting flowers: The En Tibi book herbarium. *PLoS ONE* **2019**, *14*, e0217779. [CrossRef]
12. Stearn, W.T. A New Photographic Record of the Linnaean Herbarium. *Taxon* **1961**, *10*, 16–19. [CrossRef]
13. Müller-Wille, S. Linnaeus' herbarium cabinet: A piece of furniture and its function. *Endeavour* **2006**, *30*, 60–64. [CrossRef]
14. Thiers, B. The World's Herbaria 2020: A Summary Report Based on Data from Index Herbariorum. 2021. Available online: http://sweetgum.nybg.org/science/ih/ (accessed on 25 August 2023).
15. Thiers, B.M.; Tulig, M.C.; Watson, K.A. Digitization of The New York Botanical Garden Herbarium. *Brittonia* **2016**, *68*, 324–333. [CrossRef]
16. Heberling, J.M.; Prather, L.A.; Tonsor, S.J. The Changing Uses of Herbarium Data in an Era of Global Change: An Overview Using Automated Content Analysis. *BioScience* **2019**, *69*, 812–822. [CrossRef]
17. Nelson, G.; Paul, D.; Riccardi, G.; Mast, A. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* **2012**, *209*, 19–45. [CrossRef] [PubMed]
18. Pont, C.; Wagner, S.; Kremer, A.; Orlando, L.; Plomion, C.; Salse, J. Paleogenomics: Reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.* **2019**, *20*, 29. [CrossRef] [PubMed]
19. Saville, A.C.; Martin, M.D.; Ristaino, J.B. Historic Historic late blight outbreaks caused by a widespread dominant lineage of *Phytophthora infestans* (Mont.) de Bary. *PLoS ONE* **2016**, *11*, e0168381. [CrossRef] [PubMed]
20. Yoshida, K.; Schuenemann, V.J.; Cano, L.M.; Pais, M.; Mishra, B.; Sharma, R.; Lanz, C.; Martin, F.N.; Kamoun, S.; Krause, J.; et al. The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife* **2013**, *2*, e00731. [CrossRef]
21. Muñoz-Rodríguez, P.; Carruthers, T.; Wood, J.R.I.; Williams, B.R.M.; Weitemier, K.; Kronmiller, B.; Goodwin, Z.; Sumadijaya, A.; Anglin, N.L.; Filer, D.; et al. A taxonomic monograph of Ipomoea integrated across phylogenetic scales. *Nat. Plants* **2019**, *5*, 1136–1144. [CrossRef]
22. Bellucci, E.; Benazzo, A.; Xu, C.; Bitocchi, E.; Rodriguez, M.; Alseekh, S.; Di Vittori, V.; Gioia, T.; Neumann, K.; Cortinovis, G.; et al. Selection and adaptive introgression guided the complex evolutionary history of the European common bean. *Nat. Commun.* **2023**, *14*, 1908. [CrossRef]
23. Cortinovis, G.; Di Vittori, V.; Bellucci, E.; Bitocchi, E.; Papa, R. Adaptation to novel environments during crop diversification. *Curr. Opin. Plant Biol.* **2020**, *56*, 203–217. [CrossRef]
24. Damerval, C.; Ben Othman, W.; Manicacci, D.; Jabbour, F. Distribution area of the two floral morphs of *Nigella damascena* L. (Ranunculaceae): A diachronic study using herbarium specimens collected in France. *Bot. Lett.* **2018**, *165*, 396–403. [CrossRef]
25. Davis, C.C.; Willis, C.G.; Connolly, B.; Kelly, C.; Ellison, A.M. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *Am. J. Bot.* **2015**, *102*, 1599–1609. [CrossRef]
26. Myers, J.R.; Formiga, A.K.; Janick, J. Iconography of Beans and Related Legumes Following the Columbian Exchange. *Front. Plant Sci.* **2022**, *13*, 851029. [CrossRef]
27. Bakker, F.T. Herbarium Genomics: Plant Archival DNA Explored. In *Paleogenomics. Population Genomics*; Springer: Cham, Switzerland, 2018; pp. 205–224. [CrossRef]
28. Gutaker, R.M.; Burbano, H.A. Reinforcing plant evolutionary genomics using ancient DNA. *Curr. Opin. Plant Biol.* **2017**, *36*, 38–45. [CrossRef] [PubMed]
29. Rosche, C.; Baasch, A.; Runge, K.; Brade, P.; Träger, S.; Parisod, C.; Hensen, I. Tracking population genetic signatures of local extinction with herbarium specimens. *Ann. Bot.* **2022**, *129*, 857–868. [CrossRef] [PubMed]
30. Cozzolino, S.; Cafasso, D.; Pellegrino, G.; Musacchio, A.; Widmer, A. Genetic variation in time and space: The use of herbarium specimens to reconstruct patterns of genetic variation in the endangered orchid *Anacamptis palustris*. *Conserv. Genet.* **2006**, *8*, 629–639. [CrossRef]
31. Bellucci, E.; Mario Aguilar, O.; Alseekh, S.; Bett, K.; Brezeanu, C.; Cook, D.; De la Rosa, L.; Delledonne, M.; Dostatny, D.F.; Ferreira, J.J.; et al. The INCREASE project: Intelligent Collections of food-legume genetic resources for European agrofood systems. *Plant J.* **2021**, *108*, 646–660. [CrossRef]

32. Rocchetti, L.; Gioia, T.; Logozzo, G.; Brezeanu, C.; Pereira, L.G.; De la Rosa, L.; Marzario, S.; Pieri, A.; Fernie, A.R.; Alseekh, S.; et al. Towards the development, maintenance and standardized phenotypic characterization of single-seed-descent genetic re-sources for chickpea. *Curr. Protoc.* **2022**, *2*, e371. [CrossRef]
33. Cortinovis, G.; Oppermann, M.; Neumann, K.; Graner, A.; Gioia, T.; Marsella, M.; Alseekh, S.; Fernie, A.R.; Papa, R.; Bellucci, E.; et al. Towards the Development, Maintenance, and Standardized Phenotypic Characterization of Single-Seed-Descent Genetic Resources for Common Bean. *Curr. Protoc.* **2021**, *1*, e133. [CrossRef]
34. Kroc, M.; Tomaszewska, M.; Czepiel, K.; Bitocchi, E.; Oppermann, M.; Neumann, K.; Guasch, L.; Bellucci, E.; Alseekh, S.; Graner, A.; et al. Towards Development, Maintenance, and Standardized Phenotypic Characterization of Single-Seed-Descent Genetic Resources for Lupins. *Curr. Protoc.* **2021**, *1*, e191. [CrossRef]
35. Guerra-García, A.; Gioia, T.; von Wettberg, E.; Logozzo, G.; Papa, R.; Bitocchi, E.; Bett, K.E. Intelligent characterization of lentil genetic resources: Evolutionary history, genetic diversity of germplasm, and the need for well-represented collections. *Curr. Protoc.* **2021**, *1*, e134. [CrossRef] [PubMed]
36. Pääbo, S.; Poinar, H.; Serre, D.; Jaenicke-Després, V.; Hebler, J.; Rohland, N.; Kuch, M.; Krause, J.; Vigilant, L.; Hofreiter, M. Genetic Analyses from Ancient DNA. *Annu. Rev. Genet.* **2004**, *38*, 645–679. [CrossRef]
37. Staats, M.; Cuenca, A.; Richardson, J.E.; Ginkel, R.V.-V.; Petersen, G.; Seberg, O.; Bakker, F.T. DNA Damage in Plant Herbarium Tissue. *PLoS ONE* **2011**, *6*, e28448. [CrossRef]
38. Särkinen, T.; Staats, M.; Richardson, J.E.; Cowan, R.S.; Bakker, F.T. How to Open the Treasure Chest? Optimising DNA Extraction from Herbarium Specimens. *PLoS ONE* **2012**, *7*, e43808. [CrossRef] [PubMed]
39. Bieker, V.C.; Barreiro, F.S.; Rasmussen, J.A.; Brunier, M.; Wales, N.; Martin, M.D. Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. *Mol. Ecol. Resour.* **2020**, *20*, 1206–1219. [CrossRef] [PubMed]
40. Kistler, L.; Bieker, V.C.; Martin, M.D.; Pedersen, M.W.; Madrigal, J.R.; Wales, N. Ancient Plant Genomics in Archaeology, Herbaria, and the Environment. *Annu. Rev. Plant Biol.* **2020**, *71*, 605–629. [CrossRef] [PubMed]
41. Orlando, L.; Allaby, R.; Skoglund, P.; Der Sarkissian, C.; Stockhammer, P.W.; Ávila-Arcos, M.C.; Fu, Q.; Krause, J.; Willerslev, E.; Stone, A.C.; et al. Ancient DNA analysis. *Nat. Rev. Methods Prim.* **2021**, *1*, 14. [CrossRef]
42. Weiß, C.L.; Schuenemann, V.J.; Devos, J.; Shirsekar, G.; Reiter, E.; Gould, B.A.; Stinchcombe, J.R.; Krause, J.; Burbano, H.A. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *R. Soc. Open Sci.* **2016**, *3*, 160239. [CrossRef]
43. Drábková, L.; Kirschner, J.; Vlček, Č. Comparison of Seven DNA Extraction and Amplification Protocols in Historical Herbarium Specimens of Juncaceae. *Plant Mol. Biol. Rep.* **2002**, *20*, 161–175. [CrossRef]
44. Inglis, P.W.; Pappas, M.D.C.R.; Resende, L.V.; Grattapaglia, D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS ONE* **2018**, *13*, e0206085. [CrossRef]
45. Rogers, S.O.; Bendich, A.J. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol. Biol.* **1985**, *5*, 69–76. [CrossRef]
46. Cota-Sánchez, J.H.; Remarchuk, K.; Ubayasena, K. Ready-to-use DNA extracted with a CTAB method adapted for herbarium specimens and mucilaginous plant tissue. *Plant Mol. Biol. Rep.* **2006**, *24*, 161–167. [CrossRef]
47. Riahi, M.; Zarre, S.; Maassoumi, A.; Attar, F.; Osaloo, S.K. Methodology An inexpensive and rapid method for extracting papilionoid genomic DNA from herbarium specimens. *Evolution* **2010**, *9*, 1334–1342. [CrossRef]
48. Sugita, N.; Ebihara, A.; Hosoya, T.; Jinbo, U.; Kaneko, S.; Kurosawa, T.; Nakae, M.; Yukawa, T. Non-destructive DNA extraction from herbarium specimens: A method particularly suitable for plants with small and fragile leaves. *J. Plant Res.* **2019**, *133*, 133–141. [CrossRef] [PubMed]
49. Gutaker, R.M.; Reiter, E.; Furtwängler, A.; Schuenemann, V.J.; Burbano, H.A. Extraction of ultrashort DNA molecules from herbarium specimens. *BioTechniques* **2017**, *62*, 76–79. [CrossRef] [PubMed]
50. Wales, N.; Kistler, L. Extraction of Ancient DNA from Plant Remains. In *Ancient DNA: Methods and Protocols*; Shapiro, B., Barlow, A., Heintzman, P., Hofreiter, M., Paijmans, J., Soares, A., Eds.; Methods in Molecular Biology; Humana Press: New York, NY, USA, 2019; Volume 1963. [CrossRef]
51. Ribeiro, R.; Lovato, M. DNA extraction in fresh and herbarium specimens of Dalbergia Comparative analysis of different DNA extraction protocols in fresh and herbarium specimens of the genus Dalbergia. *Genet. Mol. Res.* **2007**, *6*, 173–187.
52. Krinitsina, A.A.; Sizova, T.V.; Zaika, M.A.; Speranskaya, A.S.; Sukhorukov, A.P. A rapid and cost-effective method for DNA extraction from archival herbarium specimens. *Biochemistry* **2015**, *80*, 1478–1484. [CrossRef] [PubMed]
53. Wang, W. A primer to the use of herbarium specimens in plant phylogenetics. *Bot. Lett.* **2018**, *165*, 404–408. [CrossRef]
54. Doyle, J.J.; Dickson, E.E. Preservation of Plant Samples for DNA Restriction Endonuclease Analysis. *Taxon* **1987**, *36*, 715–722. [CrossRef]
55. Jankowiak, K.; Buczkowska, K.; Szweykowska-Kulinska, Z. Successful extraction of DNA from 100-year-old herbarium specimens of the liverwort *Bazzania trilobata*. *Taxon* **2005**, *54*, 335–336. [CrossRef]
56. Savolainen, V.; Cuénoud, P.; Spichiger, R.; Martinez, M.D.; Crèvecoeur, M.; Manen, J.F. Systematics and Evolution The use of herbarium specimens in DNA phylogenetics: Evaluation and improvement. *Plant Syst. Evol.* **1995**, *197*, 87–98. [CrossRef]
57. Drábková, L.Z. DNA Extraction from Herbarium Specimens. *Methods Mol. Biol.* **2013**, *1115*, 69–84. [CrossRef]

58. Freedman, J.; Dorp, V.; Brace, L. Destructive Sampling Natural Science Collections: An Overview for Museum Professionals and Researchers. 2018. Available online: http://www.natsca.org/article/2440 (accessed on 25 August 2023).
59. Shepherd, L.D. A non-destructive DNA sampling technique for herbarium specimens. *PLoS ONE* **2017**, *12*, e0183555. [CrossRef]
60. Gilbert, M.T.P.; Moore, W.; Melchior, L.; Worobey, M. DNA extraction from dry museum beetles without conferring external morphological damage. *PLoS ONE* **2007**, *2*, e272. [CrossRef] [PubMed]
61. Santos, D.; Ribeiro, G.C.; Cabral, A.D.; Sperança, M.A. A non-destructive enzymatic method to extract DNA from arthropod specimens: Implications for morphological and molecular studies. *PLoS ONE* **2018**, *13*, e0192200. [CrossRef] [PubMed]
62. Zeng, C.-X.; Hollingsworth, P.M.; Yang, J.; He, Z.-S.; Zhang, Z.-R.; Li, D.-Z.; Yang, J.-B. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods* **2018**, *14*, 43. [CrossRef] [PubMed]
63. Marinček, P.; Wagner, N.D.; Tomasello, S. Ancient DNA extraction methods for herbarium specimens: When is it worth the effort? *Appl. Plant Sci.* **2022**, *10*, e11477. [CrossRef]
64. Jobes, D.V.; Hurley, D.L.; Thien, L.B. Plant DNA Isolation: A Method to Efficiently Remove Polyphenolics, Polysaccharides, and RNA. *Taxon* **1995**, *44*, 379–386. [CrossRef]
65. Hofreiter, M. *Ancient DNA: Methods and Protocols*; Shapiro, B., Barlow, A., Heintzman, P.D., Hofreiter, M., Paijmans, J.L.A., Soares, A.E.R., Eds.; Humana Press Incorporated: Totowa, NJ, USA, 2012; Volume 840.
66. Thomas, M.; Gilbert, P.; Wilson, A.S.; Bunce, M.; Hansen, A.J.; Willerslev, E.; Shapiro, B.; Higham, T.F.G.; Richards, M.P.; O'connell, T.C.; et al. Ancient mito-chondrial DNA from hair. *Curr. Biol.* **2004**, *14*, R463–R464.
67. Khanuja, S.P.; Shasany, A.K.; Darokar, M.; Kumar, S. Rapid Isolation of DNA from Dry and Fresh Samples of Plants Producing Large Amounts of Secondary Metabolites and Essential Oils. *Plant Mol. Biol. Rep.* **1999**, *17*, 74. [CrossRef]
68. Lister, D.L.; Bower, M.A.; Howe, C.J.; Jones, M.K. Extraction and amplification of nuclear DNA from herbarium specimens of emmer wheat: A method for assessing DNA preservation by maximum amplicon length recovery. *Taxon* **2008**, *57*, 254–258. [CrossRef]
69. Samarakoon, T.; Wang, S.Y.; Alford, M.H. Enhancing PCR Amplification of DNA from Recalcitrant Plant Specimens Using a Trehalose-Based Additive. *Appl. Plant Sci.* **2013**, *1*, 1200236. [CrossRef]
70. Telle, S.; Thines, M. Amplification of cox2 (~620 bp) from 2 mg of up to 129 Years Old Herbarium Specimens, Comparing 19 Extraction Methods and 15 Polymerases. *PLoS ONE* **2008**, *3*, e3584. [CrossRef]
71. Monroe, C.; Grier, C.; Kemp, B.M. Evaluating the efficacy of various thermo-stable polymerases against co-extracted PCR inhibitors in ancient DNA samples. *Forensic Sci. Int.* **2013**, *228*, 142–153. [CrossRef] [PubMed]
72. Burrell, A.S.; Disotell, T.R.; Bergey, C.M. The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* **2015**, *79*, 35–44. [CrossRef] [PubMed]
73. Binladen, J.; Willerslev, E. Why study ancient DNA damage? *J. Nord. Archaeol. Sci.* **2010**, *17*, a012567.
74. Pääbo, S.; Irwin, D.M.; Wilson, A.C. DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.* **1990**, *265*, 4718–4721. [CrossRef]
75. Willerslev, E.; Cooper, A. Ancient DNA. *Proc. R. Soc. B Boil. Sci.* **2004**, *272*, 3–16. [CrossRef]
76. Hofreiter, M. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **2001**, *29*, 4793–4799. [CrossRef]
77. Pruvost, M.; Grange, T.; Geigl, E.-M. Minimizing DNA contamination by using UNG-coupled quantitative real-time PCR on degraded DNA samples: Application to ancient DNA studies. *BioTechniques* **2005**, *38*, 569–575. [CrossRef] [PubMed]
78. Champlot, S.; Berthelot, C.; Pruvost, M.; Bennett, E.A.; Grange, T.; Geigl, E.-M. An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. *PLoS ONE* **2010**, *5*, e13042. [CrossRef]
79. Knapp, M.; Clarke, A.C.; Horsburgh, K.A.; Matisoo-Smith, E.A. Setting the stage—Building and working in an ancient DNA laboratory. *Ann. Anat.—Anat. Anz.* **2011**, *194*, 3–6. [CrossRef]
80. Beck, J.B.; Semple, J.C. Next-Generation Sampling: Pairing Genomics with Herbarium Specimens Provides Species-Level Signal in *Solidago* (Asteraceae). *Appl. Plant Sci.* **2015**, *3*, 1500014. [CrossRef]
81. Hart, M.L.; Forrest, L.L.; Nicholls, J.A.; Kidner, C.A. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* **2016**, *65*, 1081–1092. [CrossRef]
82. Rowe, K.C.; Singhal, S.; Macmanes, M.D.; Ayroles, J.F.; Morelli, T.L.; Rubidge, E.M.; Bi, K.; Moritz, C.C. Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* **2011**, *11*, 1082–1092. [CrossRef]
83. Malenica, N.; Šimon, S.; Besendorfer, V.; Maletić, E.; Kontić, J.K.; Pejić, I. Whole genome amplification and microsatellite genotyping of herbarium DNA revealed the identity of an ancient grapevine cultivar. *Sci. Nat.* **2011**, *98*, 763–772. [CrossRef] [PubMed]
84. Lambertini, C.; Frydenberg, J.; Gustafsson, M.H.G.; Brix, H. Herbarium specimens as a source of DNA for AFLP fingerprinting of Phragmites (Poaceae): Possibilities and limitations. *Plant Syst. Evol.* **2008**, *272*, 223–231. [CrossRef]
85. Roullier, C.; Benoit, L.; Mckey, D.B.; Lebot, V. Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 2205–2210. [CrossRef] [PubMed]
86. Liu, L.; Li, Y.; Li, S.; Hu, N.; He, Y.; Pong, R.; Lin, D.; Lu, L.; Law, M. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012**, *2012*, 251364. [CrossRef] [PubMed]
87. Gansauge, M.-T.; Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **2013**, *8*, 737–748. [CrossRef] [PubMed]

88. Bennett, E.A.; Massilani, D.; Lizzo, G.; Daligault, J.; Geigl, E.-M.; Grange, T. Library construction for ancient genomics: Single strand or double strand? *BioTechniques* **2014**, *56*, 289–300. [CrossRef]
89. Psonis, N.; Vassou, D.; Kafetzopoulos, D. Testing a series of modifications on genomic library preparation methods for ancient or degraded DNA. *Anal. Biochem.* **2021**, *623*, 114193. [CrossRef] [PubMed]
90. Briggs, A.W.; Heyn, P. Preparation of Next-Generation Sequencing Libraries from Damaged DNA. *Methods Mol. Biol.* **2011**, *840*, 143–154. [CrossRef]
91. Prüfer, K.; Stenzel, U.; Hofreiter, M.; Pääbo, S.; Kelso, J.; Green, R.E. Open Access METHOD Computational challenges in the analysis of ancient DNA. *Genome Biol.* **2010**, *11*. [CrossRef] [PubMed]
92. Ginolhac, A.; Rasmussen, M.; Gilbert, M.T.P.; Willerslev, E.; Orlando, L. mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics* **2011**, *27*, 2153–2155. [CrossRef] [PubMed]
93. Jónsson, H.; Ginolhac, A.; Schubert, M.; Johnson, P.L.F.; Orlando, L. mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **2013**, *29*, 1682–1684. [CrossRef] [PubMed]
94. Trucchi, E.; Benazzo, A.; Lari, M.; Iob, A.; Vai, S.; Nanni, L.; Bellucci, E.; Bitocchi, E.; Raffini, F.; Xu, C.; et al. Ancient genomes reveal early Andean farmers selected common beans while preserving diversity. *Nat. Plants* **2021**, *7*, 123–128. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

General discussions

This thesis work reflects the research lines carried out during the years of PhD, and the efforts made to investigate the genetic diversity of crop legumes through innovative methods and approaches. Specifically, two distinct approaches were implemented and evaluated. Both approaches enhance the capacity to investigate genetic diversity, making information more accessible in types of samples that present different challenges. The first approach, fully discussed in Chapter II was based on the application of a molecular innovation capable of expanding the applicability of species with large and highly repetitive genomes in studies involving extensive collections. The other examined in Chapter III is focused on implementing studies that explore genetic diversity using *herbarium* specimens, to enhance the temporal perspective of genetic diversification. In particular, with the first approach, we investigate genetic diversity in lentils. To achieve this, a CRISPR-Cas9-based strategy for genomic complexity reduction was tested. This method reduces the abundance of repetitive sequences during the library preparation phase, facilitating the alignment and improving the analysis of non-repetitive regions in such challenging genomes. The approach was applied to characterize a landrace population compound of 80 genotypes from Castelluccio di Norcia, comparing them with 25 accessions from the INCREASE project. The application of this approach led to several considerations. From a methodological perspective, the genotypic characterization resulted in low-coverage data, with an average of approximately 20% missing data per individual. This led to the development of a specific variant filtering strategy, which optimizes data quality while considering downstream analyses. Various considerations can be made regarding the sequencing approach. Firstly, following the depletion of part of the repetitive sequences, we observed, as expected, an increase in the alignment of non-repetitive sequences, which showed a higher average coverage, as previously discussed. In Chapter II, the application of this approach to our dataset enabled us to clarify conditions related to alignment problems that arise as a direct consequence of regions affected by duplications. Consequently, this artifact becomes more pronounced

in highly repetitive genomes, making it even more evident. Through our work, we developed a model capable of identifying the sites where the misalignment occurs, laying the basis for the development of future filtering approaches. The implications of this are not limited to a general improvement in data quality but also extend to the possibility of more accurate assessments of heterozygosity levels, which in some cases may be biased due to alignment artifacts caused by this phenomenon. The sequencing data resulted in an average coverage low, and the levels of missingness were consequently relatively high. This led to several considerations and methodological approaches that addressed this condition. Among these, it has emerged that a genotype-likelihoods-based variant calling approach is the optimal strategy, particularly for future applications of this genotyping method (Sand Korneliussen et al., 2013). Despite the potential for further improvements, data analysis has confirmed the suitability of this approach for this type of study. Furthermore, it has demonstrated the potential for broader application to other species with similar genomic characteristics, such as pea and faba bean. This sequencing method is already being utilized in various ongoing projects involving these species, addressing multiple research questions. Beyond the methodological aspect related to genotyping in Chapter II, various analytical strategies were applied, particularly aimed at identifying genetic signs of adaptation in the Castelluccio di Norcia population. The strategies applied and developed were critically evaluated, assessing both the challenges and strengths of the analysis. This led to the development of a robust strategy, based on the specific characteristics of the produced data that resulted in the identification of concrete signals of selection within our population. Chapter III differently sets out to evaluate the possibilities and opportunities, but also the challenges related to implementation in studies assessing genetic diversity using herbarium samples. In particular, beyond the nature of these samples, a series of considerations regarding the molecular approaches necessary for their implementation is also presented. These considerations led to a greater awareness of the potential enclosed in this material. Moreover, they enabled the application of these strategies and precautions. Consequently, although in a preliminary phase, this study is ongoing and is leading to the effective inclusion of herbarium samples in studies on genetic diversity in crop legume species.

General Conclusion

In the evolving agricultural scenarios, the study of genetic diversity emerges as a critical aspect for understanding and addressing the challenges posed by climate change. Landraces are cultivated through generations by farmer selection. They incorporate the diversity resulting from adaptive processes over time, preserving genetic characteristics that have been shaped by the interactions between local environments, cultural practices, and agricultural requirements. In the same way, samples preserved over time can conserve genetic diversity from environments that no longer exist and provide insights into how genetic diversity has acquired its current shape. Through our studies, we have expanded the possibilities for studying genetic diversity by exploring challenging samples for various reasons: in lentils due to their genome characteristics, and in *herbarium* specimens based on their long-term preservation methods. We can conclude that these studies, despite bringing incremental advancements in genetic diversity characterization, represent an initial step and a solid foundation for multiple research lines to be developed in the future.

Bibliography

Abi-Ghanem, R., Carpenter-Boggs, L., & Smith, J. L. (2011). Cultivar effects on nitrogen fixation in peas and lentils. *Biology and Fertility of Soils*, *47*(1), 115–120. https://doi.org/10.1007/s00374-010-0492-6

Acito, M., Fatigoni, C., Villarini, M., & Moretti, M. (2023). Effect of Cooking and Domestic Storage on the Antioxidant Activity of Lenticchia di Castelluccio di Norcia, an Italian PGI Lentil Landrace. *International Journal of Environmental Research and Public Health*, *20*(3). https://doi.org/10.3390/ijerph20032585

Arunkumar, R., Ness, R. W., Wright, S. I., & Barrett, S. C. H. (2015). The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics*, *199*(3), 817–829. https://doi.org/10.1534/genetics.114.172809

Bakker, F. T. (2018). *Herbarium Genomics: Plant Archival DNA Explored* (pp. 205–224). https://doi.org/10.1007/13836_2018_40

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J. L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5473–5478. https://doi.org/10.1073/pnas.1418631112

Bellucci, E., Benazzo, A., Xu, C., Bitocchi, E., Rodriguez, M., Alseekh, S., Di Vittori, V., Gioia, T., Neumann, K., Cortinovis, G., Frascarelli, G., Murube, E., Trucchi, E., Nanni, L., Ariani, A., Logozzo, G., Shin, J. H., Liu, C., Jiang, L., … Papa, R. (2023). Selection and adaptive introgression guided the complex evolutionary history of the European common bean. *Nature Communications*. https://doi.org/10.1038/s41467-023-37332-z

Bellucci, E., Mario Aguilar, O., Alseekh, S., Bett, K., Brezeanu, C., Cook, D., De la Rosa, L., Delledonne, M., Dostatny, D. F., Ferreira, J. J., Geffroy, V., Ghitarrini, S., Kroc, M., Kumar Agrawal, S., Logozzo, G., Marino, M., Mary-Huard, T., McClean, P., Meglič, V., … Papa, R. (2021). The INCREASE project: Intelligent Collections of food-legume genetic resources for European agrofood systems. *The Plant Journal*, *108*(3), 646–660. https://doi.org/10.1111/TPJ.15472

Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M. L., Gioia, T., Santo, D., Nanni, L., Attene, G., & Papa, R. (2017). Beans (Phaseolus ssp.) as a model for

understanding crop evolution. In *Frontiers in Plant Science* (Vol. 8). Frontiers Media S.A. https://doi.org/10.3389/fpls.2017.00722

Bonin, A., Bellemain, E., Eidesen, P. B., Pompanon, F., Brochmann, C., & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. In *Molecular Ecology* (Vol. 13, Issue 11, pp. 3261–3273). https://doi.org/10.1111/j.1365-294X.2004.02346.x

Booker, T. R., Yeaman, S., & Whitlock, M. C. (2021). Global adaptation complicates the interpretation of genome scans for local adaptation. In *Evolution Letters* (Vol. 5, Issue 1, pp. 4–15). Oxford University Press. https://doi.org/10.1002/evl3.208

Broccanello, C., Bellin, D., DalCorso, G., Furini, A., & Taranto, F. (2023). Genetic approaches to exploit landraces for improvement of Triticum turgidum ssp. durum in the age of climate change. In *Frontiers in Plant Science* (Vol. 14). Frontiers Media S.A. https://doi.org/10.3389/fpls.2023.1101271

Brown, W. L. (1983). *Genetic Diversity and Genetic Vulnerability-An Appraisal 1*.

Burns, M., Starrett, J., Derkarabetian, S., Richart, C. H., Cabrero, A., & Hedin, M. (2017). Comparative performance of double-digest RAD sequencing across divergent arachnid lineages. *Molecular Ecology Resources*, *17*(3), 418–430. https://doi.org/10.1111/1755-0998.12575

Calleja-Cabrera, J., Boter, M., Oñate-Sánchez, L., & Pernas, M. (2020). Root Growth Adaptation to Climate Change in Crops. In *Frontiers in Plant Science* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fpls.2020.00544

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barret, K., Blanco, G., Cheung, W. W. L., Connors, S. L., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., … Ha, M. (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland*. (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge-Vorsatz, C. Xiao, N. Yassaa, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, … Y. Park, Eds.). https://doi.org/10.59327/IPCC/AR6-9789291691647.001

Caracuta, V., Vardi, J., Paz, Y., & Boaretto, E. (2017). Farming legumes in the pre-pottery Neolithic: New discoveries from the site of Ahihud (Israel). *PLoS ONE*, *12*(5). https://doi.org/10.1371/journal.pone.0177859

Ceccobelli, S., Ciancaleoni, S., Lancioni, H., Veronesi, F., Albertini, E., & Rosellini, D. (2019). Genetic distinctiveness of a Protected Geographic Indication lentil landrace from the Umbria region, Italy, over 20 years. *Genetic Resources and Crop Evolution*, *66*(7), 1483–1493. https://doi.org/10.1007/s10722-019-00799-1

Chahota, R. K., & Sharma, S. K. (1993). STUDIES ON GENETIC VARIABILITY AND COMPONENT ANALYSIS IN MACROSPERMA AND MICROSPERMA LENTILS. In *J. Genet* (Vol. 53, Issue 4). www.IndianJournals.com

Chapman, S. C., Chakraborty, S., Dreccer, M. F., & Howden, S. M. (2012). Plant adaptation to climate changeopportunities and priorities in breeding. *Crop and Pasture Science*, *63*(3), 251–268. https://doi.org/10.1071/CP11303

Chung, E., Mi Kim, K., ee Eun Heo, J., Cho, C.-W., -woo Lee, S., & ai-Heon Lee, J. (2009). Molecular Characterization of Mungbean Peroxisomal Alanine Glyoxylate Aminotransferase Gene Induced by Low Temperature Stress. In *GENES & GENOMICS* (Vol. 3, Issue 09).

Condón, F., Gustus, C., Rasmusson, D. C., & Smith, K. P. (2008). Effect of advanced cycle breeding on genetic diversity in barley breeding germplasm. *Crop Science*, *48*(3), 1027–1036. https://doi.org/10.2135/cropsci2007.07.0415

Cooper, J. W., Wilson, M. H., Derks, M. F. L., Smit, S., Kunert, K. J., Cullis, C., & Foyer, C. H. (2017). Enhancing faba bean (Vicia faba L.) genome resources. *Journal of Experimental Botany*, *68*(8), 1941–1953. https://doi.org/10.1093/jxb/erx117

Cortés, A. J., López-Hernández, F., & Blair, M. W. (2022). Genome–Environment Associations, an Innovative Tool for Studying Heritable Evolutionary Adaptation in Orphan Crops and Wild Relatives. In *Frontiers in Genetics* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fgene.2022.910386

Cortinovis, G., Di Vittori, V., Bellucci, E., Bitocchi, E., & Papa, R. (2020). Adaptation to novel environments during crop diversification. In *Current Opinion in Plant Biology* (Vol. 56, pp. 203–217). Elsevier Ltd. https://doi.org/10.1016/j.pbi.2019.12.011

Crawford, J. E., & Lazzaro, B. P. (2012). Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Frontiers in Genetics*, *3*(APR). https://doi.org/10.3389/fgene.2012.00066

Dallaire, X., Bouchard, R., Henault, P., Ulmo-Diaz, G., Normandeau, E., Merot, C., Bernatchez, L., & Moore, J. S. (2023). Widespread Deviant Patterns of Heterozygosity in Whole-Genome Sequencing Due to Autopolyploidy, Repeated Elements, and Duplication. *Genome Biology and Evolution*, *15*(12). https://doi.org/10.1093/gbe/evad229

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011a). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/BIOINFORMATICS/BTR330

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

Denham, T., Barton, H., Castillo, C., Crowther, A., Dotte-Sarout, E., Florin, S. A., Pritchard, J., Barron, A., Zhang, Y., & Fuller, D. Q. (2020). The domestication syndrome in vegetatively propagated field crops. *Annals of Botany*, *125*(4), 581–597. https://doi.org/10.1093/aob/mcz212

Di Vittori, V., Bellucci, E., Bitocchi, E., Rau, D., Rodriguez, M., Murgia, M. L., Nanni, L., Attene, G., & Papa, R. (2017). *Domestication and Crop History* (pp. 21–55). https://doi.org/10.1007/978-3-319-63526-2_2

Di Vittori, V., Gioia, T., Rodriguez, M., Bellucci, E., Bitocchi, E., Nanni, L., Attene, G., Rau, D., & Papa, R. (2019). Convergent evolution of the seed shattering trait. In *Genes* (Vol. 10, Issue 1). MDPI AG. https://doi.org/10.3390/genes10010068

Dissanayake, R., Braich, S., Cogan, N. O. I., Smith, K., & Kaur, S. (2020). Characterization of Genetic and Allelic Diversity Amongst Cultivated and Wild Lentil Accessions for Germplasm Enhancement. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00546

Drabkova, L., Kirschner, J., & Vl~k, C. (2002). Comparison of Seven DNA Extraction and Amplification Protocols in Historical Herbarium Specimens of Juncaceae. In *International Society for Plant Molecular Biology* (Vol. 20).

Drábková, L. Z. (2014). DNA extraction from herbarium specimens. *Methods in Molecular Biology*, *1115*, 69–84. https://doi.org/10.1007/978-1-62703-767-9_4

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. B. (2016). Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*, *33*(4), 1082–1093. https://doi.org/10.1093/molbev/msv334

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0019379

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0721-2

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

European Commission. (1997). *Commission Regulation (EC) No 1065/97 of 12 June 1997 supplementing the Annex to Regulation (EC) No 1107/96 on the registration of geographical indications and designations of origin under the procedure laid down in Article 17 of Council Regulation (EEC) No 2081/92 (Text with EEA relevance)*.

FAOSTAT. (2024).

Fatichenti F. (2019). *LA RITERRITORIALIZZAZIONE DELLA MONTAGNA DELL'UMBRIA SUD-ORIENTALE*. CISGE – Centro Italiano per gli Studi Storico-Geografici.

Felsenstein, J. (1974). *THE EVOLUTIONARY ADVANTAGE OF RECOMBINATION'*. https://academic.oup.com/genetics/article/78/2/737/5991150

Ferguson, M. E., Maxted, N., Slageren, M. V., & Robertson, L. D. (2000). A re-assessment of the taxonomy of Lens Mill. (Leguminosae, Papilionoideae, Vicieae). *Bot. J. Linn. Soc.*, *133*, 41–59.

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929. https://doi.org/10.1111/2041-210X.12382

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Fuller, D. Q. (2007). Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the old world. In *Annals of Botany* (Vol. 100, Issue 5, pp. 903–924). https://doi.org/10.1093/aob/mcm048

Galludo, M., Canals, J., Pineda-Cirera, L., Esteve, C., Rosselló, M., Balanyà, J., Arenas, C., & Mestres, F. (2018). Climatic adaptation of chromosomal inversions in Drosophila subobscura. *Genetica*, *146*(4–5), 433–441. https://doi.org/10.1007/s10709-018-0035-x

Gamble, T., Coryell, J., Ezaz, T., Lynch, J., Scantlebury, D. P., & Zarkower, D. (2015). Restriction site-associated DNA sequencing (RAD-seq) reveals an extraordinary number of transitions among gecko sex-determining systems. *Molecular Biology and Evolution*, *32*(5), 1296–1309. https://doi.org/10.1093/molbev/msv023

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

Gao, L., Kantar, M. B., Moxley, D., Ortiz-Barrientos, D., & Rieseberg, L. H. (2023). Crop adaptation to climate change: An evolutionary perspective. In *Molecular Plant* (Vol. 16, Issue 10, pp. 1518–1546). Cell Press. https://doi.org/10.1016/j.molp.2023.07.011

Gepts, P. (2006). Plant genetic resources conservation and utilization: The accomplishments and future of a societal insurance policy. *Crop Science*, *46*(5), 2278–2292. https://doi.org/10.2135/cropsci2006.03.0169gas

Gorkovskiy, A., & Verstrepen, K. J. (2021). The role of structural variation in adaptation and evolution of yeast and other fungi. In *Genes* (Vol. 12, Issue 5). MDPI AG. https://doi.org/10.3390/genes12050699

Goron, T. L., & Raizada, M. N. (2015). Genetic diversity and genomic resources available for the small millet crops to accelerate a New Green Revolution. *Frontiers in Plant Science*, *6*(MAR). https://doi.org/10.3389/fpls.2015.00157

Guerra-Garcia, A., Haile, T., Ogutcen, E., Bett, K. E., & von Wettberg, E. J. (2022). An evolutionary look into the history of lentil reveals unexpected diversity. *Evolutionary Applications*, *15*(8), 1313–1325. https://doi.org/10.1111/eva.13467

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., & Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics*, *37*(5), 549–554. https://doi.org/10.1038/ng1547

Gupta, A. K. (2004). Origin of agriculture and domestication of plants and animals linked to early Holocene climate amelioration. In *Current Science* (Vol. 87, Issue 1). https://www.jstor.org/stable/24107979

Hajjar, R., Jarvis, D. I., & Gemmill-Herren, B. (2008). The utility of crop genetic diversity in maintaining ecosystem services. In *Agriculture, Ecosystems and Environment* (Vol. 123, Issue 4, pp. 261–270). https://doi.org/10.1016/j.agee.2007.08.003

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., Depamphilis, C. W., Tiffin, P., & Rieseberg, L. H. (2021). *Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree*. *118*, 2102914118. https://doi.org/10.1073/pnas.2102914118/-/DCSupplemental

Hansen, J., & Renfrew, J. (1978). Palaeolithic–Neolithic seed remains at Franchthi Cave, Greece. *Nature*, *271*, 349–352. https://doi.org/10.1038/271349a0

Harlan, J. R. (1975). *Crops & man*. American Society of Agronomy, Crop Science Society of America.

Hoban, S., Bruford, M., D'Urban Jackson, J., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Grueber, C. E., Hunter, M. E., Jaffe, R., … Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation*, *248*. https://doi.org/10.1016/j.biocon.2020.108654

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., Cregan, P. B., & Tanksley, S. D. (2006). *Impacts of genetic bottlenecks on soybean genome diversity*. www.pnas.orgcgidoi10.1073pnas.0604379103

Jarpa-Parra, M. (2018). Lentil protein: a review of functional properties and food application. An overview of lentil protein functionality. In *International Journal of Food Science and Technology* (Vol. 53, Issue 4, pp. 892–903). Blackwell Publishing Ltd. https://doi.org/10.1111/ijfs.13685

Joshi, M., Timilsena, Y., & Adhikari, B. (2017). Global production, processing and utilization of lentil: A review. In *Journal of Integrative Agriculture* (Vol. 16, Issue 12, pp. 2898–2913). Chinese Academy of Agricultural Sciences. https://doi.org/10.1016/S2095-3119(17)61793-3

Kang, Y., Khan, S., & Ma, X. (2009). Climate change impacts on crop yield, crop water productivity and food security - A review. In *Progress in Natural Science* (Vol. 19, Issue 12, pp. 1665–1674). Science Press. https://doi.org/10.1016/j.pnsc.2009.08.001

Kapun, M., Fabian, D. K., Goudet, J., & Flatt, T. (2016). Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster. *Molecular Biology and Evolution*, *33*(5), 1317–1336. https://doi.org/10.1093/molbev/msw016

Kapun, M., Mitchell, E. D., Kawecki, T. J., Schmidt, P., & Flatt, T. (2023). An Ancestral Balanced Inversion Polymorphism Confers Global Adaptation. *Molecular Biology and Evolution*, *40*(6). https://doi.org/10.1093/molbev/msad118

Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., de Haan, S., Engels, J. M. M., Guarino, L., Hoban, S., Mercer, K. L., Miller, A. J., Nabhan, G. P., Perales, H. R., Richards, C., Riggins, C., & Thormann, I. (2022). Crop genetic erosion: understanding and responding to loss of crop diversity. In *New Phytologist* (Vol. 233, Issue 1, pp. 84–118). John Wiley and Sons Inc. https://doi.org/10.1111/nph.17733

Khoury, C., Laliberté, B., & Guarino, L. (2010). Trends in ex situ conservation of plant genetic resources: A review of global crop and regional conservation strategies. In *Genetic Resources and Crop Evolution* (Vol. 57, Issue 4, pp. 625–639). https://doi.org/10.1007/s10722-010-9534-z

Kislev, M. E., Nadel, D., & Carmi, I. (1992). Epipalaeolithic (19,000 BP) cereal and fruit diet at Ohalo II, Sea of Galilee, Israel. In *Review of Palaeobotany and Palynology* (Vol. 73).

Kislev ME. (1985). Early Neolithic Horsebean from Yiftah'el, Israel. *Science*, *228*, 319–320. https://doi. org/10.1126/science.228.4697.319

Kockum, I., Huang, J., & Stridh, P. (2023). Overview of Genotyping Technologies and Methods. *Current Protocols*, *3*(4). https://doi.org/10.1002/cpz1.727

Koul, P. M., Sharma, V., Rana, M., Chahota, R. K., Kumar, S., & Sharma, T. R. (2017). Analysis of genetic structure and interrelationships in lentil species using morphological and SSR markers. *3 Biotech*, *7*(1). https://doi.org/10.1007/s13205-017-0683-z

Kreplak, J., Madoui, M. A., Cápal, P., Novák, P., Labadie, K., Aubert, G., Bayer, P. E., Gali, K. K., Syme, R. A., Main, D., Klein, A., Bérard, A., Vrbová, I., Fournier, C., d'Agata, L., Belser, C., Berrabah, W., Toegelová, H., Milec, Z., … Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. *Nature Genetics*, *51*(9), 1411–1422. https://doi.org/10.1038/s41588-019-0480-1

Kumar, R., Sharma, V., Suresh, S., Ramrao, D. P., Veershetty, A., Kumar, S., Priscilla, K., Hangargi, B. S., Narasanna, R., Pandey, M. K., Naik, G. R., Thomas, S., & Kumar, A. (2021). Understanding Omics Driven Plant Improvement and de novo Crop Domestication: Some Examples. In *Frontiers in Genetics* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fgene.2021.637141

Ladizinsky, G. (1993). LENTIL DOMESTICATION" ON THE QUALITY OF EVIDENCE AND ARGUMENTS. *Economic Botany*, *47*, 60–64.

Laurentin, H. (2009). Data analysis for molecular characterization of plant genetic resources. In *Genetic Resources and Crop Evolution* (Vol. 56, Issue 2, pp. 277–292). https://doi.org/10.1007/s10722-008-9397-8

Lee, C. yu, Yen, H. Y., Zhong, A. W., & Gao, H. (2021). Resolving misalignment interference for NGS-based clinical diagnostics. *Human Genetics*, *140*(3), 477–492. https://doi.org/10.1007/s00439-020-02216-5

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289–304. https://doi.org/10.1534/genetics.118.301747

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/gr.078212.108

Li, Z. M., Zheng, X. M., & Ge, S. (2011). Genetic diversity and domestication history of African rice (Oryza glaberrima) as inferred from multiple gene sequences. *Theoretical and Applied Genetics*, *123*(1), 21–31. https://doi.org/10.1007/s00122-011-1563-2

Liber, M., Duarte, I., Maia, A. T., & Oliveira, H. R. (2021). The History of Lentil (Lens culinaris subsp. culinaris) Domestication and Spread as Revealed by Genotyping-by-Sequencing of Wild and Landrace Accessions. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.628439

Liu, Q., Zhou, Y., Morrell, P. L., Gaut, B. S., & Ge, S. (2017). Deleterious variants in Asian Rice and the potential cost of domestication. *Molecular Biology and Evolution*, *34*(4), 908–924. https://doi.org/10.1093/molbev/msw296

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S., & Wu, C. I. (2006). The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends in Genetics*, *22*(3), 126–131. https://doi.org/10.1016/j.tig.2006.01.004

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. In *Nature Reviews Genetics* (Vol. 17, Issue 11, pp. 704–714). Nature Publishing Group. https://doi.org/10.1038/nrg.2016.104

M. Sidari, C. Santonoceto, U. Anattasi, G. Pretini, & A. Muscolo. (2008). Variation in Four Genotypes of Lentil under NaCl-Salinity Stress. *American Journal of Agricultural and Biological Sciences*, *3*, 410–416.

Martins, H., Caye, K., Luu, K., Blum, M. G. B., & François, O. (2016). *Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics*. https://doi.org/10.1101/054585

Mercer, K. L., & Perales, H. R. (2010). Evolutionary response of landraces to climate change in centers of crop diversity. *Evolutionary Applications*, *3*(5–6), 480–493. https://doi.org/10.1111/j.1752-4571.2010.00137.x

Moyers, B. T., Morrell, P. L., & McKay, J. K. (2018). Genetic costs of domestication and improvement. *Journal of Heredity*, *109*(2), 103–116. https://doi.org/10.1093/jhered/esx069

Muscolo, A., Sidari, M., Anastasi, U., Santonoceto, C., & Maggio, A. (2014). Effect of PEG-induced drought stress on seed germination of four lentil genotypes. *Journal of Plant Interactions*, *9*(1), 354–363. https://doi.org/10.1080/17429145.2013.835880

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. In *Nature Reviews Methods Primers* (Vol. 1, Issue 1). Springer Nature. https://doi.org/10.1038/s43586-020-00011-0

Özkan, H., Willcox, G., Graner, A., Salamini, F., & Kilian, B. (2011). Geographic distribution and domestication of wild emmer wheat (Triticum dicoccoides). In *Genetic Resources and Crop Evolution* (Vol. 58, Issue 1, pp. 11–53). Kluwer Academic Publishers. https://doi.org/10.1007/s10722-010-9581-5

Palacios, L. T., Guindón, F., Maglia, F., & Bermejo, C. (2023). Identification of Macrosperma and Microsperma Lentil (Lens culinaris Medik) Genotypes with Superior Nutritional, Industrial, and Productive Traits to Get Transgressive Segregants. *ACS Agricultural Science and Technology*, *3*(8), 683–693. https://doi.org/10.1021/acsagscitech.3c00131

Parween, S., Nawaz, K., Roy, R., Pole, A. K., Venkata Suresh, B., Misra, G., Jain, M., Yadav, G., Parida, S. K., Tyagi, A. K., Bhatia, S., & Chattopadhyay, D. (2015). An advanced draft genome assembly of a Desi type chickpea (Cicer arietinum L.). *Scientific Reports*, *5*. https://doi.org/10.1038/srep12806

Pavan, S., Bardaro, N., Fanelli, V., Marcotrigiano, A. R., Mangini, G., Taranto, F., Catalano, D., Montemurro, C., De Giovanni, C., Lotti, C., & Ricciardi, L. (2019). Genotyping by Sequencing of Cultivated Lentil (Lens culinaris Medik.) Highlights Population Structure in the Mediterranean Gene Pool Associated With Geographic Patterns and Phenotypic Variables. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00872

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C., & Salse, J. (2019). Paleogenomics: Reconstruction of plant evolutionary trajectories from modern and ancient DNA. In *Genome Biology* (Vol. 20, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1627-1

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Purugganan, M. D., & Fuller, D. Q. (2009). The nature of selection during plant domestication. In *Nature* (Vol. 457, Issue 7231, pp. 843–848). https://doi.org/10.1038/nature07895

Ramsay, L., Koh, C. S., Kagale, S., Gao, D., Kaur, S., Haile, T., Gela, T. S., Chen, L.-A., Cao, Z., Konkin, D. J., Toegelová, H., Doležel, J., Rosen, B. D., Stonehouse, R., Humann, J. L., Main, D., Coyne, C. J., Mcgee, R. J., Cook, D. R., … Bett, K. E. (2021). Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. *BioRxiv Preprint*. https://doi.org/10.1101/2021.07.23.453237

Rane, R. V., Rako, L., Kapun, M., Lee, S. F., & Hoffmann, A. A. (2015). Genomic evidence for role of inversion 3RP of Drosophila melanogaster in facilitating climate change adaptation. *Molecular Ecology*, *24*(10), 2423–2432. https://doi.org/10.1111/mec.13161

Rao, V. R., & Hodgkin, T. (2002). Genetic diversity and conservation and utilization of plant genetic resources. In *Plant Cell, Tissue and Organ Culture* (Vol. 68).

Rawal, V., & Navarro, D. K. (2019). *The Global Economy of Pulses* (V. Rawal & D. K. , Navarro, Eds.; 1st ed.). FAO.

Raza, A., Razzaq, A., Mehmood, S. S., Zou, X., Zhang, X., Lv, Y., & Xu, J. (2019). Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. In *Plants* (Vol. 8, Issue 2). MDPI AG. https://doi.org/10.3390/plants8020034

Reif, J. C., Zhang, P., Dreisigacker, S., Warburton, M. L., Van Ginkel, M., Hoisington, D., Bohn, M., & Melchinger, A. E. (2005). Wheat genetic diversity trends during domestication and breeding. *Theoretical and Applied Genetics*, *110*(5), 859–864. https://doi.org/10.1007/s00122-004-1881-8

Rogers L., D. (2004). Genetic Erosion - No longer just an agricultural issue. *Native Plants Journal*, *5*(2). https://doi.org/DOI: https://doi.org/10.2979/NPJ.2004.5.2.112

Rosche, C., Baasch, A., Runge, K., Brade, P., Träger, S., Parisod, C., & Hensen, I. (2022). Tracking population genetic signatures of local extinction with herbarium specimens. *Annals of Botany*, *129*(7), 857–868. https://doi.org/10.1093/aob/mcac061

Rossato, M., Marcolungo, L., De Antoni, L., Lopatriello, G., Bellucci, E., Cortinovis, G., Frascarelli, G., Nanni, L., Bitocchi, E., Di Vittori, V., Vincenzi, L., Lucchini, F., Bett, K. E., Ramsay, L., Konkin, D. J., Delledonne, M., & Papa, R. (2023a). CRISPR-Cas9-based repeat depletion for highthroughput genotyping of complex plant genomes. *Genome Research*, *33*(5), 787–797. https://doi.org/10.1101/gr.277628.122

Rossato, M., Marcolungo, L., De Antoni, L., Lopatriello, G., Bellucci, E., Cortinovis, G., Frascarelli, G., Nanni, L., Bitocchi, E., Di Vittori, V., Vincenzi, L., Lucchini, F., Bett, K.

E., Ramsay, L., Konkin, D. J., Delledonne, M., & Papa, R. (2023b). CRISPR-Cas9-based repeat depletion for highthroughput genotyping of complex plant genomes. *Genome Research*, *33*(5), 787–797. https://doi.org/10.1101/gr.277628.122

Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). *Plant domestication, a unique opportunity to identify the genetic basis of adaptation*. www.nasonline.org/adaptationandcomplexdesign.

Salgotra, R. K., & Chauhan, B. S. (2023). Genetic Diversity, Conservation, and Utilization of Plant Genetic Resources. In *Genes* (Vol. 14, Issue 1). MDPI. https://doi.org/10.3390/genes14010174

Sanchez, D., Sadoun, S. Ben, Mary-Huard, T., Allier, A., Moreau, L., & Charcosse, A. (2023). Improving the use of plant genetic resources to sustain breeding programs' efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(14). https://doi.org/10.1073/pnas.2205780119

Sand Korneliussen, T., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). *Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data*. http://www.biomedcentral.com/1471-2105/14/289

Särkinen, T., Staats, M., Richardson, J. E., Cowan, R. S., & Bakker, F. T. (2012). How to Open the Treasure Chest? Optimising DNA Extraction from Herbarium Specimens. *PLoS ONE*, *7*(8). https://doi.org/10.1371/journal.pone.0043808

Schaefer, H., Hechenleitner, P., Santos-Guerra, A., De Sequeira, M. M., Pennington, R. T., Kenicer, G., & Carine, M. A. (2012). Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evolutionary Biology*, *12*(1). https://doi.org/10.1186/1471-2148-12-250

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., … Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, *46*(7), 707–713. https://doi.org/10.1038/ng.3008

Scippa, G. S., Rocco, M., Ialicicco, M., Trupiano, D., Viscosi, V., Di Michele, M., Arena, S., Chiatante, D., & Scaloni, A. (2010). The proteome of lentil (Lens culinaris Medik.) seeds: Discriminating between landraces. *Electrophoresis*, *31*(3), 497–506. https://doi.org/10.1002/elps.200900459

Smale, M. (1997). *The Green Revolution and Wheat Genetic Diversity: Some Unfounded Assumptions* (Vol. 25, Issue 8).

Sonnante, G., Hammer, K., & Pignone, D. (2009a). From the cradle of agriculture a handful of lentils: History of domestication. In *Rendiconti Lincei* (Vol. 20, Issue 1, pp. 21–37). Springer-Verlag Italia s.r.l. https://doi.org/10.1007/s12210-009-0002-7

Sonnante, G., Hammer, K., & Pignone, D. (2009b). From the cradle of agriculture a handful of lentils: History of domestication. In *Rendiconti Lincei* (Vol. 20, Issue 1, pp. 21–37). Springer-Verlag Italia s.r.l. https://doi.org/10.1007/s12210-009-0002-7

Sonnante, G., & Pignone, D. (2001). Assessment of genetic variation in a collection of lentil using molecular tools. In *Euphytica* (Vol. 120).

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., Gentzbittel, L., Childs, K. L., Yandell, M., Gundlach, H., Mayer, K. F., Schwartz, D. C., & Town, C. D. (2014). *An improved genome release (version Mt4.0) for the model legume Medicago truncatula*. http://www.biomedcentral.com/1471-2164/15/312

Taylor, W., & Swan, E. C. (1994a). *DNA from Herbarium Specimens. In: Herrmann, B., Hummel, S. (eds) Ancient DNA*. https://doi.org/10.1007/978-1-4612-4318-2_11

Taylor, W., & Swan, E. C. (1994b). *Dried Samples: Soft Tissues 11 DNA from Herbarium Specimens*.

Thiers, B. M., Tulig, M. C., & Watson, K. A. (2016). Digitization of The New York Botanical Garden Herbarium. *Brittonia*, *68*(3), 324–333. https://doi.org/10.1007/s12228-016-9423-7

Thomas, M., Dawson, J. C., Goldringer, I., & Bonneuil, C. (2011). Seed exchanges, a key to analyze crop diversity dynamics in farmer-led on-farm conservation. In *Genetic Resources and Crop Evolution* (Vol. 58, Issue 3, pp. 321–338). Springer Science and Business Media B.V. https://doi.org/10.1007/s10722-011-9662-0

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. (2017). AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, *45*(W1), W122–W129. https://doi.org/10.1093/nar/gkx382

Torricelli, R., Silveri, D. D., Ferradini, N., Venora, G., Veronesi, F., & Russi, L. (2012). Characterization of the lentil landrace Santo Stefano di Sessanio from Abruzzo, Italy. *Genetic Resources and Crop Evolution*, *59*(2), 261–276. https://doi.org/10.1007/s10722-011-9682-9

Tosh, S. M., & Yada, S. (2010). Dietary fibres in pulse seeds and fractions: Characterization, functional attributes, and applications. In *Food Research International* (Vol. 43, Issue 2, pp. 450–460). https://doi.org/10.1016/j.foodres.2009.09.005

Trucchi, E., Benazzo, A., Lari, M., Iob, A., Vai, S., Nanni, L., Bellucci, E., Bitocchi, E., Raffini, F., Xu, C., Jackson, S. A., Lema, V., Babot, P., Oliszewski, N., Gil, A., Neme, G., Michieli, C. T., De Lorenzi, M., Calcagnile, L., … Bertorelle, G. (2021). Ancient genomes reveal early Andean farmers selected common beans while preserving diversity. *Nature Plants*, *7*(2), 123–128. https://doi.org/10.1038/s41477-021-00848-7

Van De Wouw, M., Kik, C., Van Hintum, T., Van Treuren, R., & Visser, B. (2010). Genetic erosion in crops: Concept, research results and challenges. *Plant Genetic Resources: Characterisation and Utilisation*, *8*(1), 1–15. https://doi.org/10.1017/S1479262109990062

Van Oss, H., Aron, · Y, & Ladizinsky, · G. (1997). Chloroplast DNA variation and evolution in the genus Lens Mill. In *Theor Appl Genet* (Vol. 94). Springer-Verlag.

Verza, M., Ceccacci, A., Frigo, G., Mulazzani, L., & Chatzinikolaou, P. (2024). Legumes on the Rise: The Impact of Sustainability Attributes on Market Prices. *Sustainability (Switzerland)* , *16*(7). https://doi.org/10.3390/su16072644

Viana, J. P. G., Fang, Y., Avalos, A., Song, Q., Nelson, R., & Hudson, M. E. (2022). Impact of multiple selective breeding programs on genetic diversity in soybean germplasm. *Theoretical and Applied Genetics*, *135*(5), 1591–1602. https://doi.org/10.1007/s00122-022-04056-5

Von Engelhardt, D. (2011). *LUCA GHINI (1490-1556) IL PADRE FONDATORE DELLA BOTANICA MODERNA NEL CONTESTO DEI RAPPORTI SCIENTIFICI EUROPEI DEL SEDICESIMO SECOLO (1)* (Vol. 27).

Wade, M. J., & Kalisz, S. (1990). THE CAUSES OF NATURAL SELECTION. In *Evolution* (Vol. 44, Issue 8). https://academic.oup.com/evolut/article/44/8/1947/6869674

Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Faller, M., Bomblies, K., Lukens, L., & Doebley, J. F. (2005). The origin of the naked grains of maize. *Nature*, *436*(7051), 714–719. https://doi.org/10.1038/nature03863

Wang, W. (2018a). A primer to the use of herbarium specimens in plant phylogenetics. *Botany Letters*, *165*(3–4), 404–408. https://doi.org/10.1080/23818107.2018.1438311

Wang, W. (2018b). A primer to the use of herbarium specimens in plant phylogenetics. *Botany Letters*, *165*(3–4), 404–408. https://doi.org/10.1080/23818107.2018.1438311

Warne, T., Ahmed, S., Byker Shanks, C., & Miller, P. (2019). Sustainability Dimensions of a North American Lentil System in a Changing World. *Frontiers in Sustainable Food Systems*, *3*. https://doi.org/10.3389/fsufs.2019.00088

Weisdorf, J. L. (2003). *03-41 From Foraging to Farming: Explaining the Neolithic Revolution From Foraging to Farming: Explaining the Neolithic Revolution **. http://www.econ.ku.dk

Williams, R. C. (1989). Restriction fragment length polymorphism (RFLP). *American Journal of Physical Anthropology*, *32*(10 S), 159–184. https://doi.org/10.1002/ajpa.1330320508

Willis, C. G., Ellwood, E. R., Primack, R. B., Davis, C. C., Pearson, K. D., Gallinat, A. S., Yost, J. M., Nelson, G., Mazer, S. J., Rossington, N. L., Sparks, T. H., & Soltis, P. S. (2017). Old Plants, New Tricks: Phenological Research Using Herbarium Specimens. In *Trends in Ecology and Evolution* (Vol. 32, Issue 7, pp. 531–546). Elsevier Ltd. https://doi.org/10.1016/j.tree.2017.03.015

Wong, M. M. L., Gujaria-Verma, N., Ramsay, L., Yuan, H. Y., Caron, C., Diapari, M., Vandenberg, A., & Bett, K. E. (2015). Classification and characterization of species within the genus lens using genotyping-by-sequencing (GBS). *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0122025

Wright, K. I. (1994). Society for American Archaeology Ground-Stone Tools and Hunter-Gatherer Subsistence in Southwest Asia: Implications for the Transition to Farming. In *Source: American Antiquity* (Vol. 59, Issue 2).

Zaccardelli, M., Lupo, F., Piergiovanni, A. R., Laghetti, G., Sonnante, G., Daminati, M. G., Sparvoli, F., & Lioi, L. (2012). Characterization of Italian lentil (Lens culinaris Medik.) germplasm by agronomic traits, biochemical and molecular markers. *Genetic Resources and Crop Evolution*, *59*(5), 727–738. https://doi.org/10.1007/s10722-011-9714-5

Zeven, A. C. (1998). Landraces: A review of definitions and classifications. In *Euphytica* (Vol. 104).

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., & Yang, T. L. (2019). PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, *35*(10), 1786–1788. https://doi.org/10.1093/bioinformatics/bty875

Zhu, Y., Chen, H., Fan, J., Wang, Y., Li, Y., Chen, J., Fan, J., Yang, S., Hu, L., Leung, H., Mew, T. W., Teng, P. S., Wang, Z., & Mundt, C. C. (2000). Genetic diversity and disease control in rice. *Nature*, *406*, 718–722.

Zohary, D. (1972). *The Wild Progenitor and the Place of Origin of the Cultivated Lentil: Lens culinaris* (Vol. 26, Issue 4).