PRO-GRACE

PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

DELIVERABLE 3.2

Demonstration of DNA barcoding, reduced representation sequencing/resequencing and cytogenomic methods and services

This deliverable has been submitted and is currently pending approval by the European Commission.

Call identifier: HORIZON-INFRA-2022-DEV-01-01
PRO-GRACE

Grant agreement no: 101094738

Promoting a plant genetic resource community for Europe

Deliverable No. 3.2

Demonstration of DNA barcoding, reduced representation sequencing/resequencing and cytogenomic methods and services

Contractual delivery date:
32

Actual delivery date:
32

Responsible partner:
UNITO

Contributing partners:
(Partners' short names; ENEA, INRAE, UEB)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094738.

Grant agreement no.	Horizon Europe – 101094738
Project full title	PRO-GRACE – Promoting a plant genetic resource community for Europe

Deliverable number	**D3.2**
Deliverable title	**Demonstration of DNA barcoding, reduced representation sequencing/resequencing and cytogenomic methods and services**
Type	R
Dissemination level	PU
Work package number	3
Author(s)	Lorenzo Barchi, Giuseppe Aprea, Maria Tiziana Sirangelo, Paola Ferrante, Luciana Gaccione, Vèronique Lefebvre, Giovanni Giuliano, Jaroslav Doležel, Jan Bartoš, Jana Čížková, Jan Šafář
Keywords	DNA barcoding, reduced representation sequencing, SPET, Pangenome, resequencing, cytogenomics flow cytometry, molecular cytogenetics

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094738**.**

Contents

Executive summary

This document presents a detailed overview of the refinement, integration, and practical demonstration of genomic methodologies, including DNA barcoding, reduced representation sequencing (such as Genotyping-by-Sequencing, GBS, and Restriction-site Associated DNA sequencing, RAD-Seq), whole-genome resequencing (WGRS), and cytogenomic approaches. These techniques are changing our capacity to analyse biological diversity, characterize genome structure and function, and apply this knowledge across multiple domains.

The application of these methods has significantly enhanced the accuracy and resolution of species identification, population structure analysis, and functional genomic studies. DNA barcoding, for instance, enables rapid and reliable classification of organisms through short, standardized genetic markers. Meanwhile, reduced representation and whole-genome sequencing approaches provide deep insights into genomic variability, enabling targeted investigations of traits of ecological, evolutionary, or agronomic relevance. Cytogenomic techniques complement these methods by offering chromosomal-level perspectives on genome organization and gene regulation.

Notably, the benefits of these advances are not limited to the "research" domain. In agriculture, these genomic tools are being leveraged to support crop breeding programmes aimed at improving resistance to pathogens, pests, and climate-related stressors. In conservation biology, they are instrumental in identifying and monitoring endangered species, guiding restoration programmes, and ensuring the genetic integrity of regenerated populations. Moreover, these techniques are increasingly contributing to areas such as environmental monitoring, biosecurity, etc.

This document reports the current state-of-the-art in genomic methodologies and documents the refinement processes and demonstration activities carried out within the project framework. It highlights the methodologies' adaptability and relevance to key priorities, such as biodiversity protection, sustainable agriculture, and climate resilience. Finally, it provides a guide to help genebanks, researchers and breeding companies to choose the best genomic methodology according to their needs.

Sequencing methodologies

Overview of DNA Sequencing Methods

DNA sequencing methods have revolutionized our understanding of genetics and molecular biology by allowing researchers to decipher the exact sequence of nucleotides in DNA molecules. Over the past few decades, these methods have undergone significant transformation, moving from labour-intensive and time-consuming processes to highly automated and rapid techniques. This evolution has enabled the scientific community to explore complex biological questions with unprecedented depth and precision. The Sanger sequencing method laid the groundwork for genetic sequencing, while the development of second- and third-generation technologies has

enabled high-throughput sequencing at lower costs. Today, DNA sequencing is a cornerstone of research in genomics, evolutionary biology, and medical diagnostics. Here we provide an in-depth overview of the historical development, technical principles, and applications of various DNA sequencing methods, focusing on the advancements that have led to the adoption of next-generation sequencing (NGS) techniques.

First-Generation Sequencing

The first-generation sequencing methods, primarily represented by the Sanger method, were pivotal in the initial sequencing of the human genome. Developed in the 1970s, the Sanger method (**Figure 1**) uses the principle of chain termination to generate DNA fragments of varying lengths (Sanger *et al.*, 1977). These fragments are then separated based on size using gel electrophoresis, and the sequence is read by analysing the pattern of the separated fragments. The method's accuracy and reliability made it the gold standard for many years, especially in smaller-scale projects.

In parallel another first-generation method, the Maxam-Gilbert sequencing technique (Maxam and Gilbert, 1977), was released. However, it was not as widely adopted due to its complexity and the use of hazardous chemicals such as dimethyl sulfate and hydrazine (**Figure 2**). Maxam-Gilbert sequencing relies on chemical modification of DNA and subsequent cleavage at specific bases, followed by separation using electrophoresis. Although the Sanger method is still used for certain applications today, such as validating sequences obtained from newer technologies, its limitations in terms of scalability and cost have led to the development of more advanced methods. Indeed, these first-generation techniques were not suitable for large-scale genome projects due to their high cost and time-intensive nature.

Second-Generation Sequencing

The advent of second-generation sequencing, also known as next-generation sequencing (NGS), marked a significant turning point in genomic research. These methods are characterized by their ability to perform massively parallel sequencing, in which millions

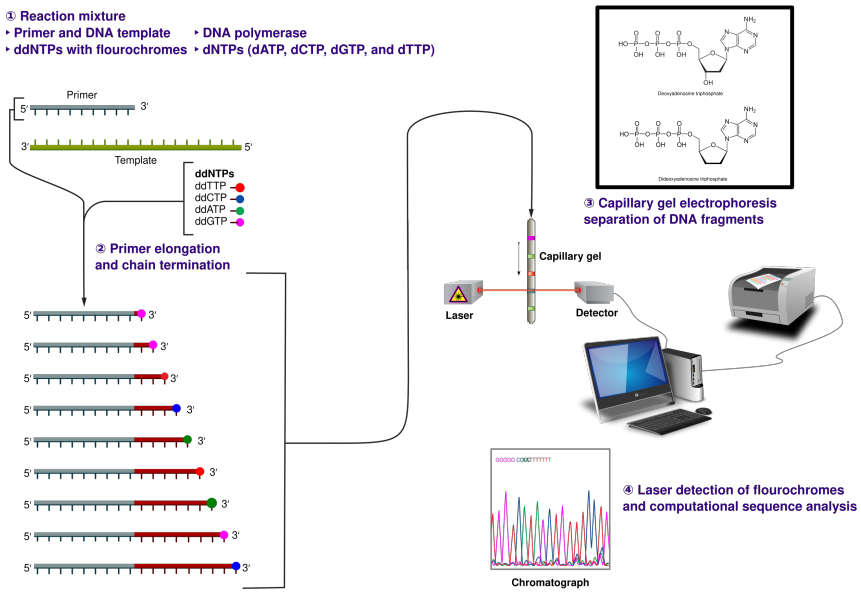

Figure 1: The Sanger sequencing approach

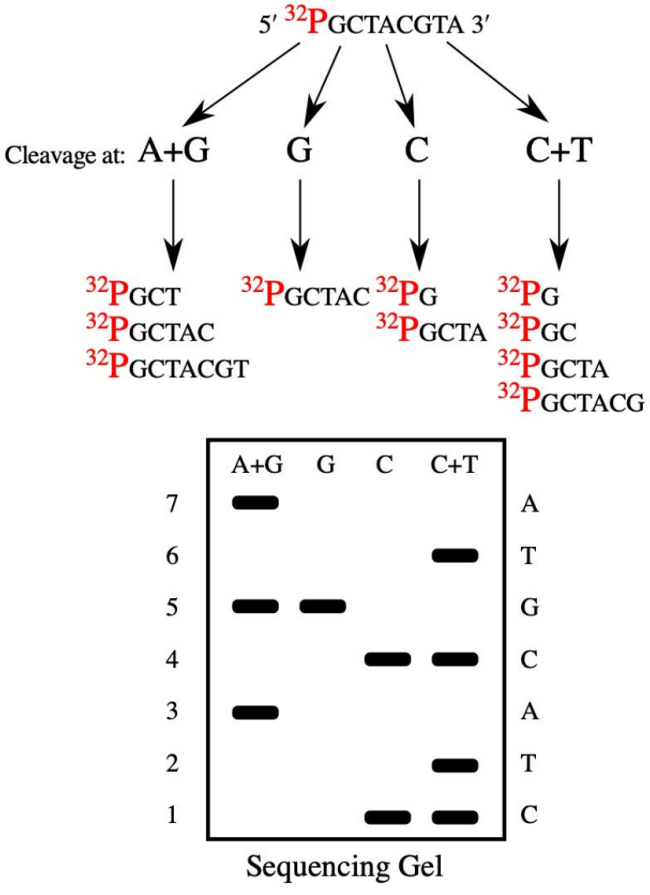

Figure 2: The Maxam-Gilbert sequencing approach

of DNA fragments can be sequenced simultaneously. The distinctive trait of the second-generation sequencing is the short read length, barely encompassing the 500 bp. This breakthrough has enabled researchers to undertake large-scale projects, such as the sequencing of entire genomes and transcriptomes, at a fraction of the time and cost required by first-generation methods. The second-generation sequencing platforms originally include the Illumina sequencing, Roche 454 pyrosequencing, and the Ion Torrent technologies. These platforms are based on different approaches to achieve high-throughput sequencing. For instance, the Illumina platform uses sequencing-by-synthesis, where fluorescently labelled nucleotides are incorporated into the growing DNA strand, and the emitted fluorescence is captured to determine the sequence of bases. One of the key features of second-generation methods is the amplification of DNA fragments, either through bridge amplification or emulsion PCR, which ensures that there is sufficient signal strength for accurate detection (**Figure 3**).

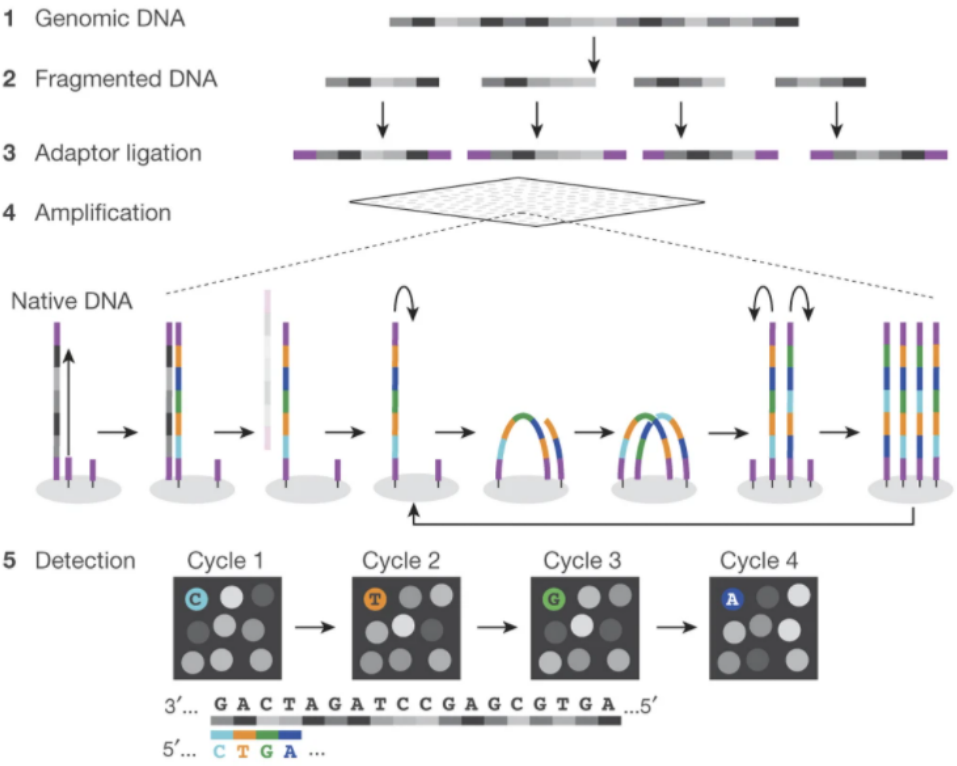

Figure 3: Illumina sequencing (form Hackl *et al.*, 2022)

More recently additional technologies were released, namely the DNBSEQ technology (MGI), the Aviti sequencer Element bioscience), the Onso platform (Pacbio), as well as the latest Illumina technology, all focused in increasing the overall quality of sequencing reads (having a quality score up to Q40). Overall, despite the short read lengths, the reduced cost and increased throughput (up to 8TB using the Illumina Novaseq X) have made NGS the method of choice for many research applications.

Third-Generation Sequencing

Third-generation sequencing (TGS) technologies have further enhanced the capabilities of sequencing by enabling the direct sequencing of single molecules of DNA or RNA. This advancement eliminates the need for PCR amplification, thereby reducing bias and preserving the original sequence context. Two prominent third-generation platforms are Pacific Biosciences' (PacBio) Single Molecule Real-Time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing. SMRT sequencing uses zero-mode waveguides (ZMWs) to observe the incorporation of nucleotides in real-time **(Figure 4**). This technique allows the sequencing of long fragments, providing a more comprehensive view of the genome, including regions that are difficult to sequence with second-generation methods. Oxford Nanopore sequencing, on the

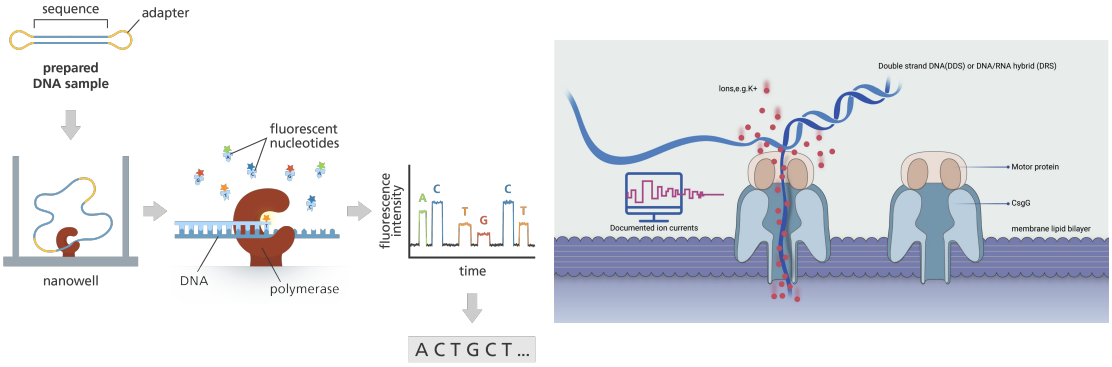

Figure 4: Pacbio SMRT (left, from yourgenome, 2017) and ONT sequencing technologies (right, from Xie *et al.*, 2021).

other hand, measures changes in electrical conductivity as single-stranded DNA molecules pass through nanopores embedded in a membrane. This approach can generate extremely long reads, making it ideal for studying structural variations and complex genomic rearrangements. The real-time capabilities and long-read potential of third-generation sequencing are particularly valuable in applications such as de novo genome assembly, epigenetic profiling, and the characterization of isoforms. However, third-generation methods are still maturing and still face challenges related to error rates and data analysis complexity. Ongoing advancements in chemistry improved and are still improving the overall quality of sequencing reads, while progress in bioinformatics will help to further expand the utility of third-generation sequencing technologies.

Applications and Future Directions in DNA Sequencing

Advances in DNA sequencing technologies opened a new era of plant genomics, driving significant progress in understanding genetic diversity, adaptation, and the domestication of crops. With the advent of more efficient and cost-effective sequencing methods, researchers are now able to explore plant genomes at unprecedented depths, uncovering critical insights that pave the way for improving crop resilience and productivity.

Whole-genome sequencing (WGS)

The primary distinction between whole-genome sequencing (WGS) and other next-generation sequencing (NGS) methodologies lies in the absence of sequence capture and the significantly

larger volume of data generated. Historically, the cost of WGS was prohibitive, but advances in second and third generation sequencing and optimized chemistries have reduced its price.

The high-quality assembly of a genome sequence is a critical foundation for understanding the biology of an organism, as well as the genetic variation within a species. Approximately two decades ago, *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) became the first plant genome to be sequenced, using first-generation capillary sequencing technology. This milestone marked the beginning of a surge in sequenced plant genomes, driven by technological advancements. First-generation automated DNA sequencing instruments could only sequence thousands of base pairs daily. In contrast, current technologies can process billions of base pairs at a fraction of the cost. Despite these advancements, sequence assembly remains a significant challenge, requiring more effort than the sequencing process itself.

The advent of second-generation sequencing, or short-read Next-Generation Sequencing (NGS) technologies, dramatically increased the availability of plant genome sequences and reduced costs. However, the assemblies produced were highly fragmented, with numerous contigs. This limitation arose from the short-read lengths and the complexity of plant genomes, which often include extensive repetitive sequences from transposable elements that are difficult to resolve in *de novo* assembly (Alkan *et al.*, 2011). Unlike vertebrate genome assemblies (Gnerre *et al.*, 2011), plant assemblies frequently consist of isolated gene "islands" interspersed with repetitive elements, complicating comprehensive assembly.

Third-generation sequencing (TGS) technologies have revolutionized genome assembly, offering high accuracy for *de novo* sequencing (Pareek *et al.*, 2011). Key platforms such as Pacific Biosciences (PacBio), Oxford Nanopore Technologies (ONT), and BioNano Genomics provide significant improvements over earlier technologies. TGS platforms generate long-read sequences with improved consensus accuracy, reduced G+C bias, and simultaneous epigenetic profiling (Nakano *et al.*, 2017). Compared to NGS, TGS offers longer average read lengths (up to 10,000 bp or more), faster sequencing (from days to hours), and reduced bias due to direct DNA sequencing without polymerase chain reaction amplification.

Repetitive regions remain the primary obstacle to high-quality genome assembly. While short reads from second-generation sequencing are insufficient to resolve these regions, long reads from TGS platforms significantly improve assembly quality by spanning repetitive sequences (Berlin *et al.*, 2015). *De novo* long-read genome assembly typically involves raw read mapping, error correction, and assembly using overlap–layout–consensus (OLC) algorithms. Error correction methods include self-correction, which aligns long reads to form consensus sequences, and hybrid correction, which uses high-quality short reads for alignment. These techniques enhance assembly quality but cannot eliminate errors, necessitating polishing steps (Sohn and Nam, 2018).

Improved sequencing accessibility has expanded focus from agriculturally significant crops to neglected species (as crops wild relatives, CWR) and non-model organisms, fostering biodiversity

research and particularly orphan crops with untapped potential (Siadjeu *et al.*, 2020). Furthermore, the integration of Hi-C and optical mapping data facilitates the generation of pseudochromosomes, bridging gaps in assembly while maintaining cost efficiency.

Going beyond: the Pangenome concept

Although a single reference genome from a 'typical' individual of a given species has served as a roadmap for the species over long time-periods, scientists soon realized that a single reference genome does not capture the genetic diversity of a species and is inadequate for many purposes (Ballouz *et al.*, 2019), leading to the concept of a pan-genome, namely a collection of all the DNA sequences that occur in a species.

The pangenome concept encapsulates the entirety of genetic information, including all genes, within a specific group of individuals, such as a population, species, or higher taxonomic classification. A pangenome consists of a small subset of essential "core" genes and numerous "accessory" genes, which exhibit varying levels of dispensability (Tettelin *et al.*, 2005). Since a single genome assembly cannot capture the full genetic repertoire of a species, the pangenome provides a more accurate representation of genomic diversity. In plants, accessory genes are often associated with responses to biotic and abiotic stresses (Bayer *et al.*, 2020).

Early genome sequencing projects aimed to generate a reference genome that would benefit studies not only of a specific species but also of related taxa. Variations across cultivars or species were typically explored using short-read resequencing and alignment to the reference genome. Such approaches have been applied to pangenome studies in *Arabidopsis thaliana* (Alonso-Blanco *et al.*, 2016), rice (Lv *et al.*, 2020), eggplant (Barchi *et al.*, 2021) and grapevine (Liang *et al.*, 2019). However, short-read sequencing projects face limitations, including challenges in resolving large insertions and identifying variants within repetitive or heterozygous regions (Cameron *et al.*, 2019; Schilbert *et al.*, 2020).

Long-read sequencing technologies address these limitations by enabling the identification of structural variants that are undetectable with short reads (Chawla *et al.*, 2021), especially in complex genomic regions (Olson *et al.*, 2022). This technology has enabled the construction of pan-genomes for important crops like soybean (Liu *et al.*, 2020), wheat (Walkowiak *et al.*, 2020; Bayer *et al.*, 2022), tomato (Zhou *et al.*, 2022) and rice (Qin *et al.*, 2021), highlighting the presence-absence variations that are critical for understanding genetic diversity and improving breeding programs. For example, in soybean, the identification of structural variants has elucidated gene loss during domestication, providing insights into the genetic basis of traits such as protein content and stress tolerance. Similarly, long-read sequencing has been applied to banana (Rijzaani *et al.*, 2022) and sorghum (Ruperao *et al.*, 2021), advancing research on these less-studied crops by constructing high-quality genome assemblies.

Recently, the concept of graph pangenomes has also emerged. Representing pangenomes as graphs enables the efficient study of genetic diversity across populations by collapsing redundant sequences into a compact structure that serves as a stable coordinate system for annotation,

alignment, and variant calling (Rakocevic *et al.*, 2019). In this framework, nodes represent sequences, edges represent connections between them, and polymorphic sites create branching paths. Each path represents a real or potential haploid genome, and the total number of paths increases with additional samples, eventually reflecting the species' maximum haplotype diversity. This graph-based approach allows for evolutionary and breeding studies by comparing theoretical haplotype diversity with observed diversity, potentially revealing genetic constraints or diversity losses. Features like node and edge counts, branching patterns, and shortest paths provide insights into selective sweeps, sequence convergence, recombination, and the core genome fraction. While sampling bias remains a challenge, graph representations can mitigate its effects by preserving paths absent in the observed population. This method offers a powerful tool to interpret genetic variation and diversity within and across species.

Initial crop genome sequencing efforts employed long-read assemblies to explore genetic variation within species, as exemplified by studies in rice (Qin *et al.*, 2021), wheat (Jiao *et al.*, 2024), barley (Jayakodi *et al.*, 2024), and tomato (Zhou *et al.*, 2022). These analyses identified extensive structural variants, such as translocations, insertions, deletions, inversions, and chromosome fusions. Accessory genes with significant phenotypic impacts—ranging from ecotype differentiation to stress tolerance and seed weight—were also uncovered.

While pangenome studies primarily focus on cultivated species, examining wild relatives can provide additional insights due to their broader repertoire of accessory genes (Bayer *et al.*, 2020). For instance, pathogen resistance genes from wild species could be integrated into crops via breeding programs. Long-read sequencing technologies are pivotal in advancing the field of plant pangenomics, enabling the exploration of the whole genomic diversity within and across species.

Case study: the eggplant pangenome graph

To explore the genetic diversity of *Solanum melongena* and its related species while avoiding bias toward any specific reference genome, a reference-free pangenome graph, was constructed using PGGB (Garrison *et al.*, 2023) and 40 chromosome-level genome assemblies, representing *S. melongena* and its closely related species (*S. insanum*, *S. incanum*) based on geographic distribution and breeding potential. The variation graph had a total length of 2.94 Gbp—more than twice the size of our previously published reference-based pangenome (1.21 Gbp) (Barchi *et al.*, 2021). It consisted of 141.8 million nodes, 194.7 million edges, and an average node degree of 2.75. Overall, a total of 83,147 structural variants (SVs) were identified. Mapping of short reads from the G2p-sol core collection of 368 accessions to this graph clarified the domestication history and population structure. Indeed, PCA and ML tree analyses revealed clear genetic separation

between *S. melongena*, *S. insanum*, and *S. incanum*, with intermediate clustering indicating potential introgression events (**Figure 5**).

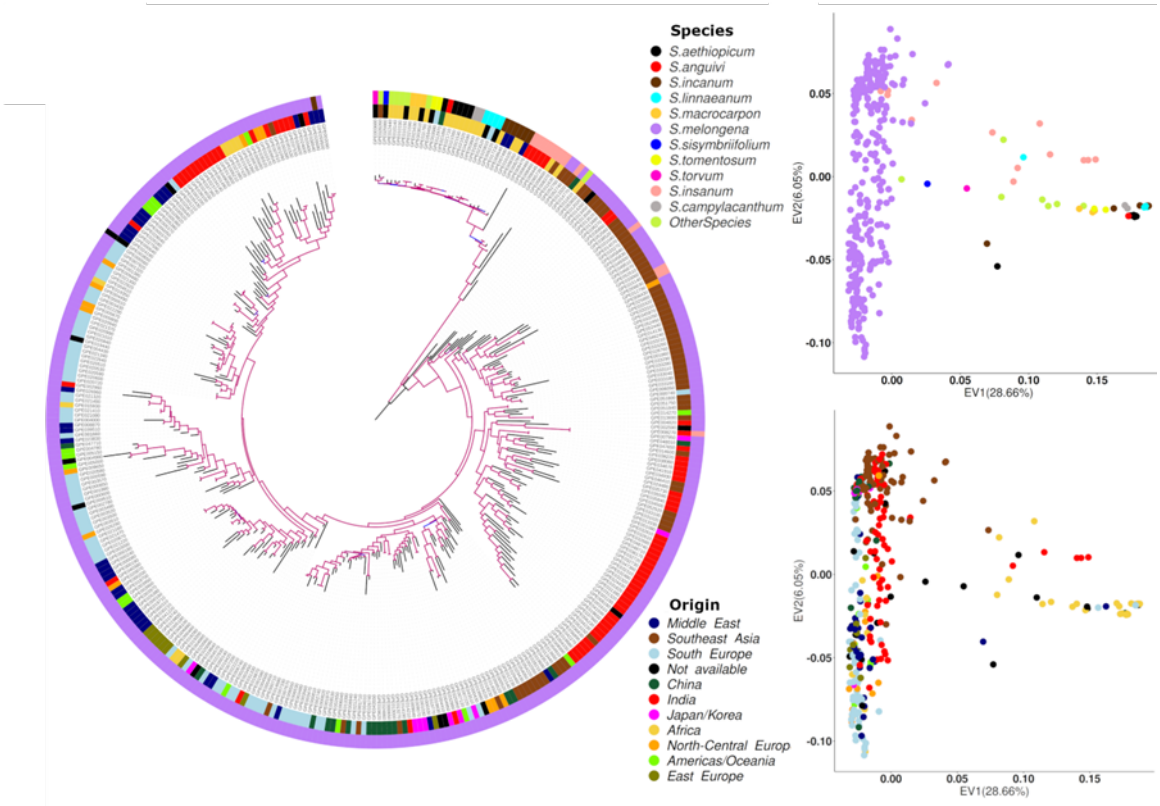

Figure 5: Phylogeny and population structure of *S. melongena* and wild relatives.

A second pangenome graph was built with Minigraph-Cactus (Hickey *et al.*, 2023), using only the 33 *S. melongena* accessions, with the GPE001970 genome sequence as the reference, to facilitate alignment of the reads of the 321 *S. melongena* accessions of the core collection available from G2P-sol project and polymorphism identification. This second graph was then used to carry out Pan-GWA studies, significantly enhancing the capacity to identify QTLs (**Figure 6**).

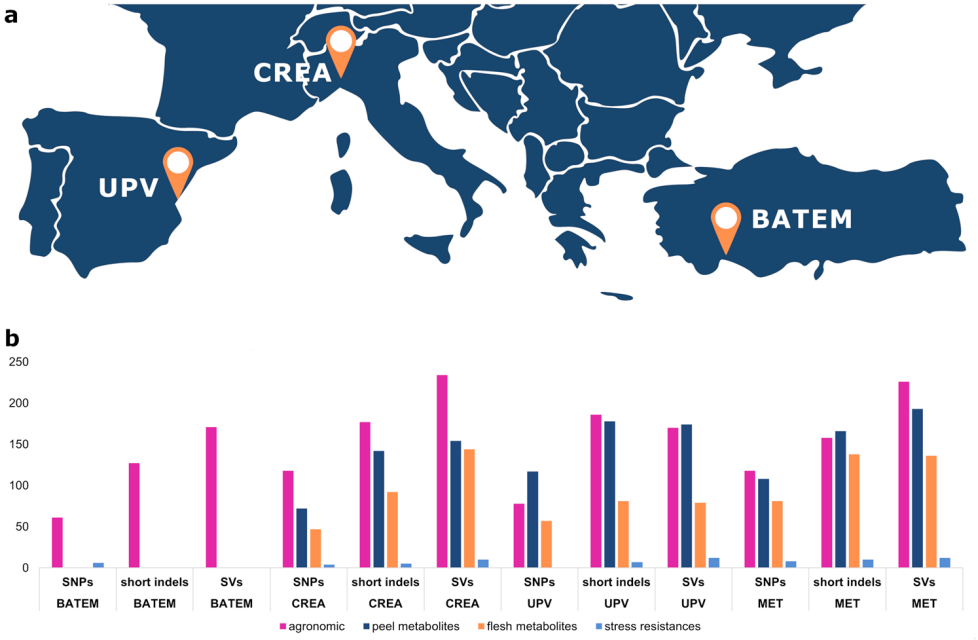

Figure 6: The eggplant pan-phenome and Pan-GWA

Whole-genome resequencing (WGR)

Whole-genome resequencing involves sequencing the entire genome of species with known reference genomes, allowing for the identification of diverse genetic variants such as single nucleotide polymorphisms (SNPs), insertions and deletions (InDels), and structural variations (SVs).

Whole-genome resequencing necessitates a robust computational infrastructure to ensure rapid and reliable data processing. The challenge posed by WGR data volume becomes apparent when compared to large gene panels or exomes. While gene panel and exome analyses generate approximately 0.15GB and 5GB of raw data, respectively, WGS generates data at least an order of magnitude higher. The corresponding variant files (.vcf) for gene panels or exomes are approximately 7E-05GB and 0.04GB, respectively, whereas WGR variant files are close to 1GB per individual, representing an increase in data size of 13,000- and 24-fold, respectively.

Three steps are essential in WGR (and RRS (see below)) analysis: i) mapping of sequencing reads against a reference genome, using tools such as bwa-mem (Li, 2013); ii) variant calling, and iii) interpretation. The development of standardized, end-to-end variant calling workflows was pioneered by the open-source Genome Analysis Toolkit (GATK) (DePristo *et al.*, 2011). However, several commercial hardware-accelerated solutions, such as Illumina DRAGEN™, Sentieon and Parabricks (Franke and Crowgey, 2020), as well as prediction-based approaches (Poplin *et al.*, 2018) are also available. None of these solutions are plug-and-play, and institutions performing large-scale WGR analyses must be prepared to contribute to pipeline development and maintenance to ensure a safe, reliable, and up-to-date analytical environment.

Large-scale resequencing projects, such as the 1001 Genomes Project for *Arabidopsis thaliana* (Alonso-Blanco *et al.*, 2016) and the 3000 Rice Genomes Project (W., Wang *et al.*, 2018), have provided invaluable resources for plant breeders. These studies revealed genes involved in domestication sweeps, which reduce genetic diversity in cultivated crops but confer advantageous traits such as non-shattering seeds and increased yield. For instance, genes like PROG1 in rice (Huang *et al.*, 2012), associated with tiller angle, and *TtBtr1* in wheat, linked to reduced shattering, have been identified as key targets for breeding (Cheng *et al.*, 2019; Zhou *et al.*, 2020). Additional analyses have been pivotal in uncovering genes associated with critical agronomic traits, such as drought tolerance, nutrient efficiency, salt tolerance (Zheng *et al.*, 2020) and disease resistance (Y., Wang *et al.*, 2018; Ahn *et al.*, 2018).

The application of WGR in crop genomics extends beyond the identification of genetic variants. It is also being used to study gene expression and epigenetic modifications, which play critical roles in the regulation of important traits. By integrating WGR with transcriptomic and epigenomic data, researchers can gain a more comprehensive understanding of the molecular mechanisms underlying trait expression and inheritance. This knowledge is essential for the development of crops that are not only high-yielding but also robust and/or resilient to environmental stresses and adaptable to different growing conditions.

Whole-genome resequencing has also played a critical role in understanding the evolutionary history of domesticated species as well as to reveal differences between populations' genetic diversity and genetic structure based on the obtained genome-wide variation data. By comparing the genomes of wild and domesticated populations, researchers can trace the genetic changes that occurred during domestication and identify the genes that were selected for desirable traits (T., Wei *et al.*, 2021; Zhao *et al.*, 2019). Furthermore, WGR can be exploited to study the genetic diversity, as well as population structure within the same species (as an example see (Barchi *et al.*, 2021)), since the genetic diversity is the basis and core of biodiversity and the fundamental guarantee of the evolutionary potential of species (Ma *et al.*, 2022).

Resequencing efforts have significantly contributed to understanding how plants adapt to changing environments. Genes associated with temperature tolerance, precipitation responses, and altitude adaptation have been identified through genome-environment association studies. For instance, in sunflower, the *HEAT-INTOLERANT1* gene has been linked to heat stress tolerance (Todesco *et al.*, 2020), while genomic loci associated with drought resistance have been discovered in crops such as fonio millet and soybean (Abrouk *et al.*, 2020; Wang *et al.*, 2022). Adaptive traits such as flowering time have also been extensively studied. Genes like *FLOWERING LOCUS C* (*FLC*) and *TERMINAL FLOWER 1* (*TFL1*) have been shown to play crucial roles in enabling plants to synchronize flowering with favourable environmental conditions, ensuring reproductive success. Such findings are particularly important in the context of climate change, where breeding crops robust or resilient to extreme weather is a global priority.

Despite remarkable progress, several challenges hamper the full potential of sequencing technologies in plant genomics. Data accessibility remains a significant issue, with many genomic datasets locked behind institutional barriers or published without proper repository links. This lack of openness limits the opportunities for collaborative research and secondary analyses. Addressing this issue requires stronger policies mandating data sharing and the creation of robust, permanent repositories. Looking ahead, integrating sequencing data with advanced computational tools, such as machine learning and *in silico* breeding platforms, holds immense potential. Systems like RiceNavi (X., Wei *et al.*, 2021) have already demonstrated the ability to optimize breeding strategies, reducing the time required for crop improvement. Expanding such approaches to other crops will be essential for addressing food security challenges in a growing global population.

A case study: resequencing the G2P-SOL pepper core collection

As a demonstration activity within the PRO-GRACE project, we resequenced the G2P-SOL core collection of *Capsicum* spp., consisting of 423 accessions representing the genetic variability of a panel of 10,083 accessions (McLeod *et al.*, 2023). In particular, it contains 393 *C. annuum* accessions and 32 accessions from other cultivated species; these include 16 *C. chinense*, 4 *C. frutescens*, 7 *C. baccatum*, 1 *C. chacoense*, 1 *C. praetermissum*, and one unclassified accession. This collection underwent resequencing using the MGI platform at 20X coverage depth. Subsequently, the raw reads obtained from the sequencing step were aligned to the *C. annuum cultivar* Zhangshugang (Liu *et al.*, 2023) pepper reference genome using the BWA-MEM tool (Vasimuddin *et al.*, 2019). Following this, SNP and small indel calling was conducted using GATK v4.3.0 (DePristo *et al.*, 2011) following the software best practices in June 2024 for germline short variant discovery. Starting from more than 300 million unfiltered variants, a final set of about 31.33M SNPs were retained, after applying filters based on depth of coverage (DP=10), no more than 5% of missing data at SNP level and retaining only biallelic sites using bcftools (Danecek *et al.*, 2021).

Analysis of the post-filtering marker set revealed a mean minor allele frequency (MAF) of 20,7% and an average density of 10,36 markers/Kb. At the marker level, the dataset exhibited 2% missing data and a 2,1% rate of heterozygosity, the accessions displayed on average 2% missing data and a heterozygosity rate of 7,5%.

A principal component analysis (PCA) was conducted on the panel of 423 Capsicum accessions to evaluate the genetic diversity characterizing the core collection **(Figure 7**). This analytical approach provided insights into the genetic structure and variability present within the core collection, revealing evident intraspecific differentiation and distinct clustering among the different species.

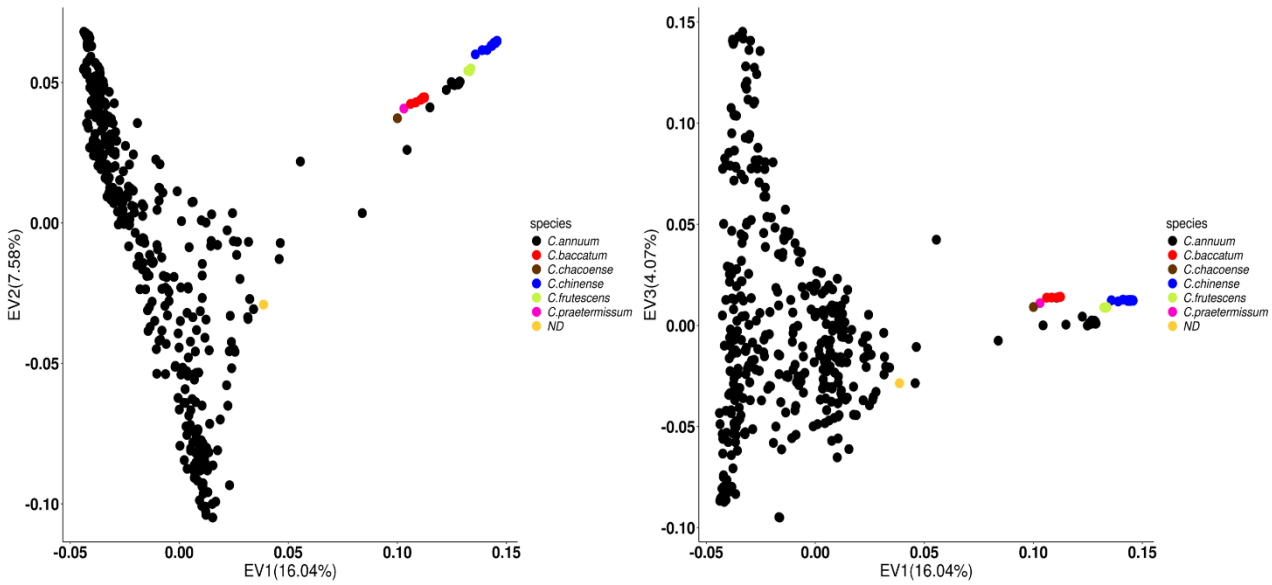

Figure 7: Principal component analysis of genetic diversity in Capsicum spp. The scatter plots illustrate the distribution of 423 accessions along the first three principal components, using 662.455 single nucleotide polymorphism (SNP) markers, which were selected after pruning for linkage disequilibrium and a minor allele frequency (MAF) > 0.05

The initial core collection underwent a series of refinement steps. First, the collection was narrowed to 393 accessions specifically belonging to *C. annuum*. Subsequently, accessions exhibiting excessive heterozygosity, defined as those deviating by more than two standard deviations from the mean individual heterozygosity, were excluded. Further filtration removed accessions with over 20% missing markers. Additionally, accessions displaying multiple phenotypes in field trials or lacking phenotypic data were eliminated. This comprehensive selection process yielded a final set of 362 accessions, which will serve as the foundation for genome-wide association studies (GWAS).

The final marker set composed by 17.4M SNPs, exhibited a mean density of 5,75 markers per Kb, but not homogeneously distributed across the different chromosomes (**Figure 8**); with an average minor allele frequency (MAF) of 21%. Heterozygosity rates were observed at 1.36%, while missing data at the marker level accounted for 1.3%, the accessions displayed on average 1.3% missing data and a heterozygosity rate of 3.4%.

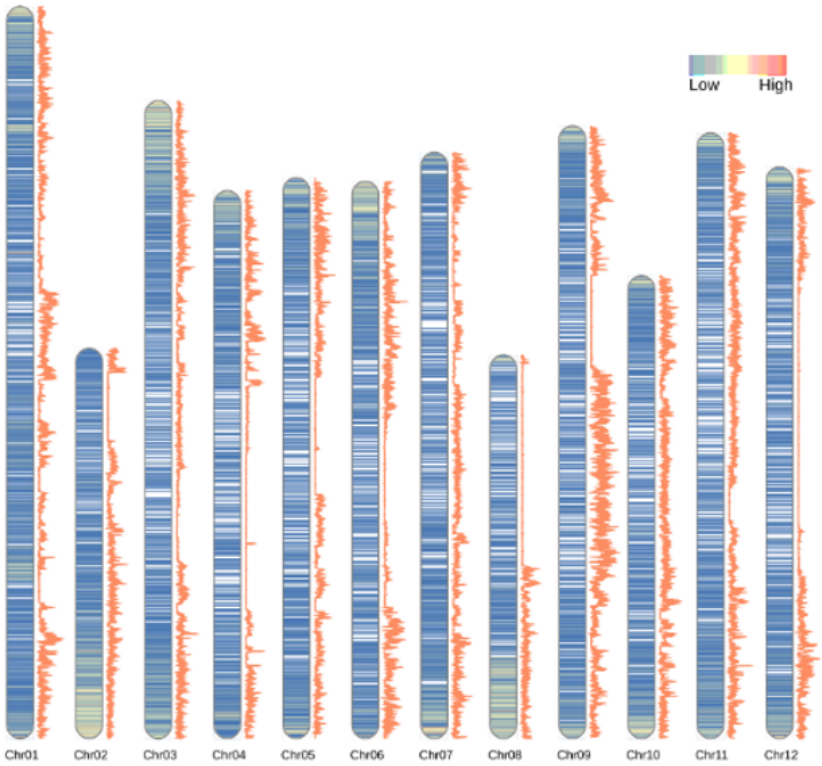

Figure 8: Marker distribution along the 12 chromosomes in pepper core collection resequencing project

The genetic diversity characterizing the core collection composed of 362 *C. annuum* accessions, which will be utilized for genome-wide association studies (GWAS), was assessed through principal component analysis (PCA) and the construction of a Maximum Likelihood (ML) phylogenetic tree using the IQ-TREE2 (Minh *et al.*, 2020). The PCA results, displayed in **Figure 9**, indicate that there is no clear evidence of strong population stratification within the *C. annuum* core collection.

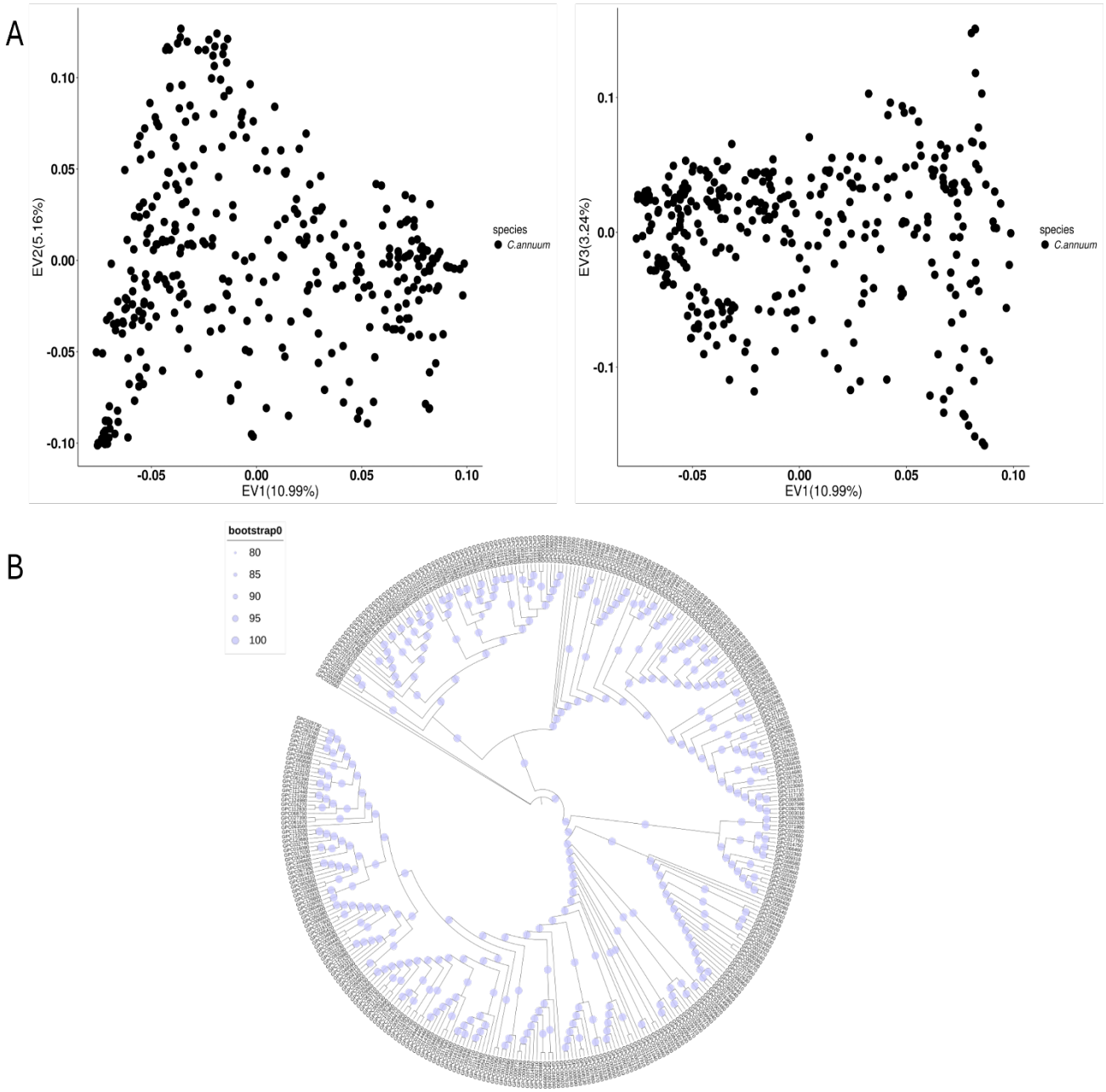

Figure 9: A) Distribution of the 362 *C. annuum* accessions along the first three principal coordinates. B) Maximum likelihood tree representing the distribution of C. annuum accessions.

Reduced representation sequencing (RRS)

Reduced representation sequencing (RRS) is a genomic technique that focuses on sequencing a subset of the genome, allowing for the efficient analysis of genetic variation across populations. Methods such as Genotyping by Sequencing (GBS; (Elshire *et al.*, 2011)) and Restriction site Associated DNA sequencing (RAD-Seq; (Baird *et al.*, 2008)) are commonly used in RRS to identify genetic markers linked to traits of interest. These methods have been widely applied in crop improvement, animal breeding, and conservation genetics, providing insights into the genetic basis of adaptation and selection.

Genotyping by sequencing (GBS)

Next-generation sequencing (NGS) technologies have significantly reduced the cost of DNA sequencing, enabling genotyping-by-sequencing (GBS) to be applied to species with high genetic diversity and large genomes. GBS is a highly multiplexed and streamlined method for constructing reduced-representation libraries on the Illumina NGS platform. This approach generates extensive single nucleotide polymorphism (SNP) datasets for genetic analysis and genotyping (Beissinger *et al.*, 2013). The GBS methodology offers several advantages, including low cost, minimal sample handling, fewer PCR and purification steps, the elimination of size fractionation and reference sequence requirements, efficient barcoding, and scalability (Davey *et al.*, 2011) (**Figure 10**). Consequently, GBS is becoming a pivotal tool for genomics-assisted breeding in diverse plant species. When coupled with genome-independent imputation, GBS facilitates efficient genetic map construction in pseudo-testcross progenies (Ward et al., 2013). The technique's streamlined library preparation process is particularly suited for high-throughput studies involving large sample. A two-enzyme GBS protocol (*PstI/MspI*) has been developed, offering enhanced complexity reduction and uniform sequencing libraries compared to the initial *ApeKI*-based protocol. This method has been successfully implemented in wheat and barley (Poland *et al.*, 2012). Moreover, Sonah et al. (2013) introduced a modified protocol incorporating selective amplification, increasing SNP yield and coverage depth while reducing per-sample costs.

GBS is an invaluable platform for plant breeding applications, from single-gene marker studies to comprehensive genomic profiling. It enables genome-wide association studies (GWAS), diversity analyses, linkage mapping, molecular marker discovery, and genomic selection (GS). The method is robust across species, allowing simultaneous SNP discovery and genotyping without prior genomic knowledge (Poland and Rife, 2012; NARUM *et al.*, 2013).

GBS has enabled cost-effective, large-scale genome sequencing in different species. For example, 5,000 maize recombinant inbred lines have been resequenced, identifying 1.4 million SNPs and 200,000 indels (Gore *et al.*, 2009). More recently, GBS has also facilitated genomic resource development in tomato (Olivieri *et al.*, 2020) and potato (Olivieri *et al.*, 2020). Furthermore, it was used to study the diversity in pepper (Ortega-Albero *et al.*, 2024) as well as to identify QTLs related to fruit traits (McLeod *et al.*, 2023). In pepper, Tripodi et al. (Tripodi *et al.*, 2021) used GBS markers to depict the demography history of the species, as well as to detected up to 1,618 duplicate accessions within and between genebanks, showing that taxonomic ambiguity and misclassification often involve interspecific hybrids that are difficult to classify morphologically.

While GBS offers numerous advantages, challenges remain, particularly in aligning true alleles in polyploid species. Computational methods for data analysis and bioinformatics pipelines are critical for maximizing GBS's potential. Enzyme-based genome complexity reduction may exclude certain regions due to mutations at restriction sites, but this limitation is shared with other methods using similar strategies.

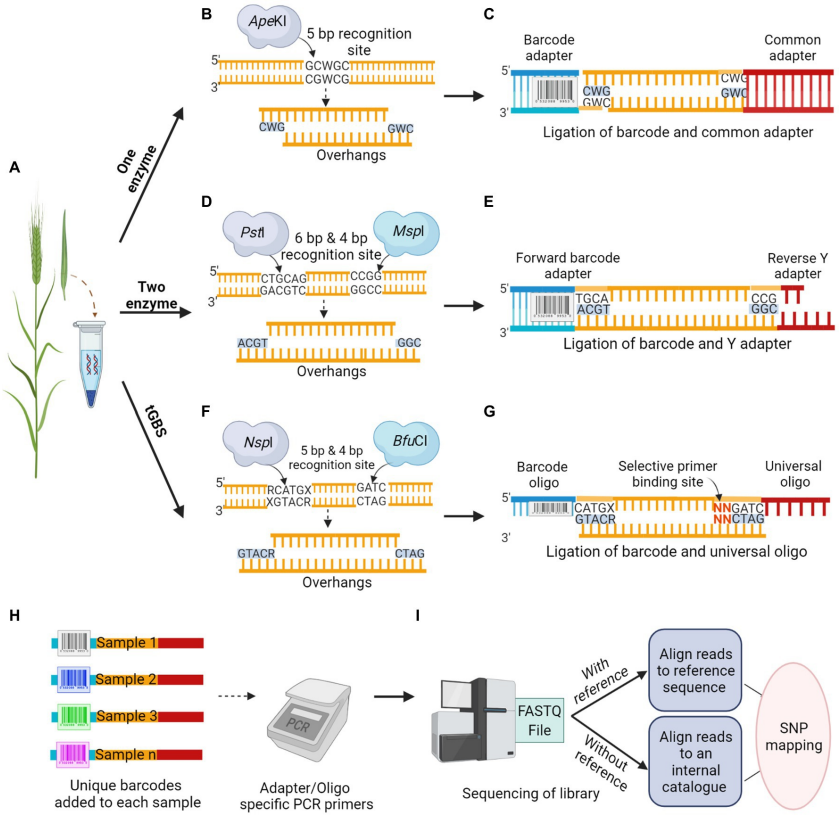

Figure 10: Schematic overview of Restriction Enzyme based GBS methodology (from Rajendran *et al.*, 2022).

Restriction-site Associated DNA sequencing (RAD-Seq)

RAD-Seq is an advanced NGS method that has significantly influenced phylogeography and phylogenetics. Like other related techniques, RAD sequencing involves digesting genomic DNA with a restriction enzyme, followed by size selection of the resulting fragments using an agarose gel, which are then sequenced using next-generation sequencing (NGS) platforms. The double digest restriction-site associated DNA (ddRAD-seq) is a variation on the RAD sequencing protocol, which is used for SNP discovery and genotyping. In this variation, the fragment shearing is replaced with a second restriction digestion to improve the tunability and accuracy of the size-selection step (**Figure 11**). The protocol also includes a second index to allow combinatorial indexing. The RAD-Seq procedure might retain more alleles due to the single enzyme digestion and size selection process, while ddRAD-Seq shows a reduced complexity, improving the reproducibility of and accuracy of variant calling and genotyping.

The distinguishing feature of RAD sequencing is its precise control over the fragments produced by the restriction digest, coupled with ultra-deep sequencing across a large number of individuals (Baird *et al.*, 2008). This level of precision makes it one of the most reproducible methods among restriction digest-based approaches.

The generated NGS reads are analysed to identify single nucleotide polymorphisms (SNPs) located immediately adjacent to common restriction digest sites. RAD sequencing has been successfully applied to marker development (Miller *et al.*, 2007; Barchi *et al.*, 2011), as well as constructing linkage maps (Barchi *et al.*, 2012; Jiang *et al.*, 2022) and phylogenetic studies (Emerson *et al.*, 2010; Acquadro *et al.*, 2017).

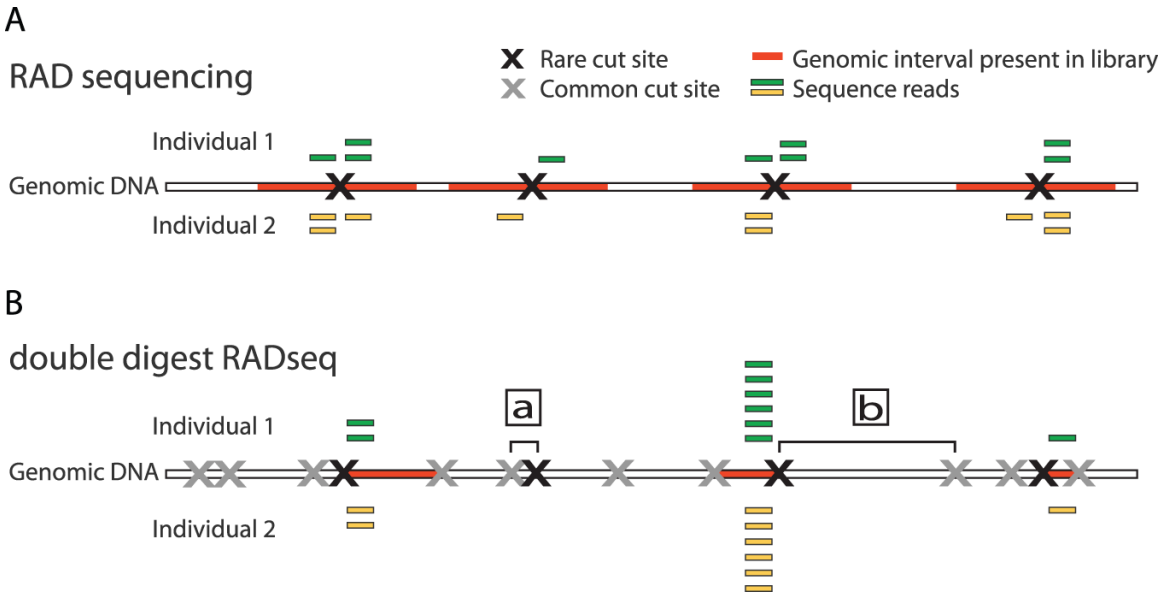

Figure 11: RADseq and ddRADseq outline (from Peterson *et al.*, 2012)

Target capture

Target DNA enrichment, in combination with high-throughput sequencing technologies, represents a highly effective and cost-efficient strategy for the large-scale investigation and characterization of genomic loci. Although whole-genome sequencing is technically feasible, it remains cost-prohibitive for large-scale population studies. For many research objectives, obtaining a moderate number of loci (hundreds to thousands) across many individuals is more appropriate than sequencing entire genomes from a few individuals (Lemmon and Lemmon, 2013). This sequence enrichment is generally achieved by using two major approaches, namely PCR-based amplicon and hybrid capture-based methodologies

PCR-based target enrichment has demonstrated significant utility in next-generation sequencing (NGS) workflows, particularly in scenarios involving limited or degraded nucleic acid material. As amplification is the central mechanism in this approach, it is especially well-suited for clinical specimens with low DNA yield or compromised quality. Given that NGS often involves multiplexed analysis of numerous genomic regions, PCR-based enrichment requires the simultaneous use of hundreds to thousands of primers under harmonized reaction conditions. To achieve uniform

amplification across all regions of interest, careful primer design, optimization of primer concentrations, and fine-tuning of thermal cycling parameters are essential. Several computational tools have been developed to aid in primer design (as an example see (Kechin *et al.*, 2020)), and various commercial platforms (e.g., ThermoFisher Scientific, Qiagen, Fluidigm, and Integrated DNA Technologies) offer predesigned or customizable primer panels. While predesigned panels provide validated and ready-to-use solutions, custom panels may require empirical optimization to ensure consistent and efficient enrichment across all targets. Innovations such as droplet digital PCR (ddPCR) have also enhanced enrichment strategies. In this technique, PCR reactions are compartmentalized into millions of nanoliter-scale droplets, each functioning as an individual microreactor. This format supports extensive multiplexing while reducing primer interference, resulting in highly uniform amplification of targeted sequences (Tewhey *et al.*, 2009). Although ddPCR requires dedicated instrumentation, its capacity for high-throughput and precise enrichment makes it a powerful tool in NGS workflows. Microfluidic-based enrichment platforms represent another compartmentalization strategy, enabling parallel processing of multiple reactions in miniaturized volumes. This technology offers reduced reagent consumption, automation, and high-throughput capability via nanofluidic chips. Post-amplification, enriched amplicons can be recovered and processed for sequencing (Murphy *et al.*, 2020).

The hybrid capture-based approach encompasses a suite of molecular techniques that selectively increase the representation of specific DNA regions within next-generation sequencing (NGS) libraries through the application of oligonucleotide probes, commonly referred to as “baits”, which are employed either in solution (Lemmon *et al.*, 2012) or immobilized on arrays (Albert *et al.*, 2007). These baits are designed based on their high nucleotide sequence similarity to the genomic regions of interest, enabling specific hybridization to target sequences within DNA samples and thereby facilitating their enrichment. The terminology used to describe this method varies depending on the genomic regions targeted. For instance, when exons or coding DNA sequences are enriched, the approach is often referred to as “exome capture” or “gene capture” (Ng *et al.*, 2009). When the flanking regions of (ultra)conserved elements are targeted, the method is known as “anchored hybrid enrichment” (Bejerano *et al.*, 2004), while “hyRAD” refers to the enrichment of specific RAD loci (Suchan *et al.*, 2016).

Methodological advances have facilitated the capture of more divergent loci using a given bait set, thereby significantly broadening the applicability of the approach, although enrichment efficiency generally declines as the sequence divergence between bait and target increases (Paijmans *et al.*, 2016). Consequently, designing bait sets that are effective across a broad phylogenetic spectrum remains a considerable challenge. For example, a bait may show high sequence similarity and consequently high enrichment efficiency for species within one clade, but may be significantly less effective or entirely ineffective for enriching homologous sequences in more distantly related clades. To address this, it may be necessary to design multiple baits per locus to ensure more uniform enrichment across all target taxa. To this purpose, a common strategy involves designing baits from a reference genome and optimizing hybridization

parameters to compensate for sequence divergence (Li *et al.*, 2013). Alternatively, conserved genomic regions across taxa can serve as anchors to enrich adjacent, more variable sequences (Lemmon *et al.*, 2012). However, this method is limited to loci near conserved regions. To address these limitations, Hugall et al. (2016) utilized transcriptomes to reconstruct phylogenies and infer ancestral sequences for clade-specific bait design, minimizing sequence divergence. This highlights the promise of target enrichment in museomics and biodiversity studies, particularly where PCR-based methods fail due to DNA degradation.

Single primer enrichment technology (SPET)

SPET is a customizable solution for targeted sequencing at an affordable price. SPET requires a priori genomic or transcriptomic information and identification of SNPs for probe design. A SPET probe is around 40-bases long and is designed adjacent to a region containing a sequence variant, thus enabling detection of both the target SNP and the discovery of additional *de novo* or off-target SNPs) surrounding the target one (**Figure 12**). The relatively high sequence conservation of exons should facilitate the hybridization of SPET probes designed on these regions across different related species and thus increase the chances to identify novel SNPs, especially if the region downstream of the probe falls in less conserved regions, such as introns and untranslated regions (UTR). In addition, SPET provides the ability of multiplexing thousands of samples in a single sequencing run, which can be genotyped with 5 to 10 of thousands of probes, and with a good coverage at target sites. Finally, thanks to the sequencing of the genomic regions around the target SNPs, SPET allows the discovery of thousands of novel SNPs not originally included in the panel.

SPET genotyping has been firstly applied in plants for evaluating its feasibility in *Zea mays* L. and to *Populus nigra* L (Scaglione *et al.*, 2019). Then two panels composed of 400 tomato and 422 eggplant accessions, comprising both domesticated material and wild relatives, were genotyped with this methodology. Overall, a total of 12,002 and 30,731 high-confidence SNPs, respectively, which comprised both target and novel SNPs were generated and used for studying the genetic relationships among the accessions. More recently, SPET was used for building linkage maps (Toppino *et al.*, 2020), as well as for demography and GWA studies (Gardoce *et al.*, 2023; Barchi *et al.*, 2023; Tripodi *et al.*, 2023) in different species. In particular in eggplant, SPET markers allowed the identification of misclassified and putatively duplicated accessions, facilitating genebank management (Barchi *et al.*, 2023).

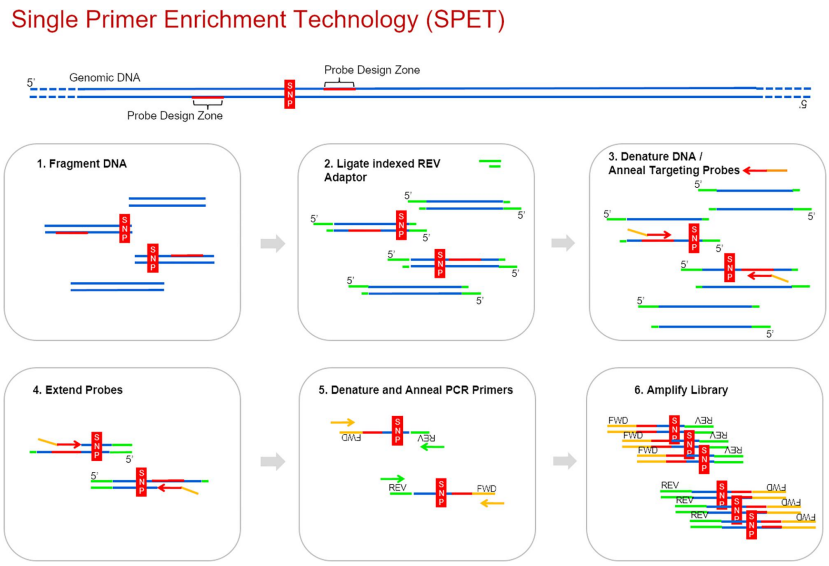

Figure 12: The six main steps of the SPET workflow. Probes can be designed up or downstream the identified SNP (from Barchi *et al.*, 2019).

Case study: Design and application of a 12K SPET tomato genotyping panel

To further demonstrate the utility of targeted genotyping in the assessment of plant genetic resources, a 12K SPET (Single Primer Enrichment Technology) panel was developed and applied to multiple tomato populations within the EU funded HARNESSTOM project (http://harnesstom.eu/en/index.html). The SPET panel was obtained from a selection process starting from approximately 0.5 million polymorphic SNPs derived from prior genotyping arrays (G2P-SOL 5K, SOLCAP 7K, Agriplex 1K), resequencing data, and partner-selected *loci* associated with traits of agronomic interest.

The selection process for the 12 K panel design included the following steps:

- Preliminary filtering of SNPs not mapped to chromosome 0, with minor allele frequency (MAF) ≥ 0.05 and missing data ≤ 5%, as well as inclusion of SNPs directly provided by partners;
- Selection of one representative marker per gene from the G2P-SOL, SOLCAP, and Agriplex panels, supplemented by partner-provided SNPs located on genes not covered by those markers;
- Completion of the panel up to 12,000 markers was achieved by randomly selecting additional SNPs from the resequencing dataset, ensuring that they targeted genes not already covered by previous selections.

The panel was used to genotype over 1,500 tomato samples across diverse genetic materials, including intraspecific biparental populations, MAGIC populations, landrace collections, prebreeding materials, and interspecific backcross lines (BILs). Variant calling was performed to evaluate both target SNPs and additional polymorphisms found in the flanking regions captured by the SPET technology.

Across the tested populations, the number of polymorphic markers ranged from ~4,000 (in low diversity biparental populations) to nearly 8,000 in more heterogeneous materials such as MAGIC lines and interspecific BILs. Remarkably, when considering both target and non-target SNPs, the proportion of polymorphic markers consistently exceeded 92% across all populations, reaching nearly complete coverage in most cases (**Figure 13**).

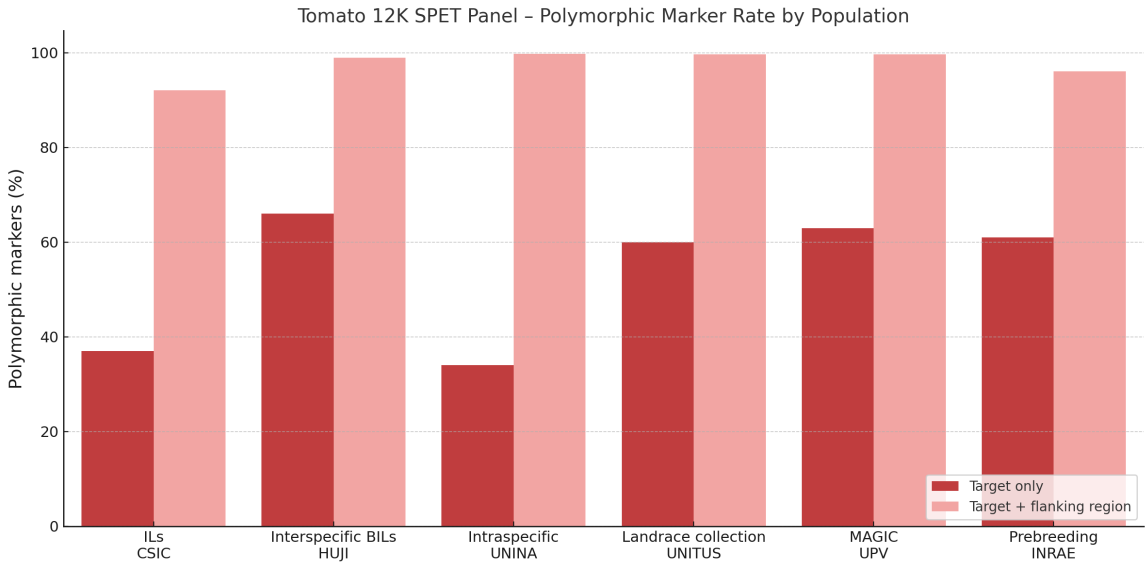

Figure 13: Percentage of polymorphic markers observed using the 12K SPET panel in different tomato populations.

This case study highlights the versatility of SPET genotyping for reduced-representation analysis in tomato. The technology enables both high-resolution diversity analysis and supports downstream applications such as GWAS, QTL mapping, and duplicate accession detection in genebank collections. The 12K panel developed for HARNESSTOM is compatible with previous datasets and provides dense coverage of the tomato gene space, making it a powerful tool for future applications in genetic resource management.

Kompetitive Allele Specific PCR (KASP)

KASP is a simple, rapid, and cost-effective method that enables high-precision biallelic characterization of single nucleotide polymorphisms (SNPs), as well as insertions and deletions at specific loci (**Figure 14**). The KASP assay employs fluorescence resonance energy transfer (FRET) to generate allele-specific signals, utilizing two fluorescently labelled cassettes for the detection of a single bi-allelic SNP. The process begins with an initial round of PCR in which one of the allele-specific primers matches the target SNP, leading to amplification of the target region in conjunction with a common reverse primer. As the PCR progresses, the allele-specific primer is incorporated into the newly synthesized DNA template. During the initial stages of the reaction, fluorophore-labelled oligonucleotides remain bound to their complementary quencher-labelled oligonucleotides, preventing the generation of a fluorescent signal. However, as PCR continues, the fluorophore-labelled oligonucleotide corresponding to the amplified allele becomes

incorporated into the DNA template and is no longer quenched. This results in the generation of a fluorescence signal specific to the amplified allele, enabling precise allele discrimination.

KASP methodology has been used for various purposes, including development of markers for biotic resistance (Zhang *et al.*, 2023), carried out GWA (Tian *et al.*, 2024) and diversity analysis (Bhattacharjee *et al.*, 2024; Xing *et al.*, 2024).

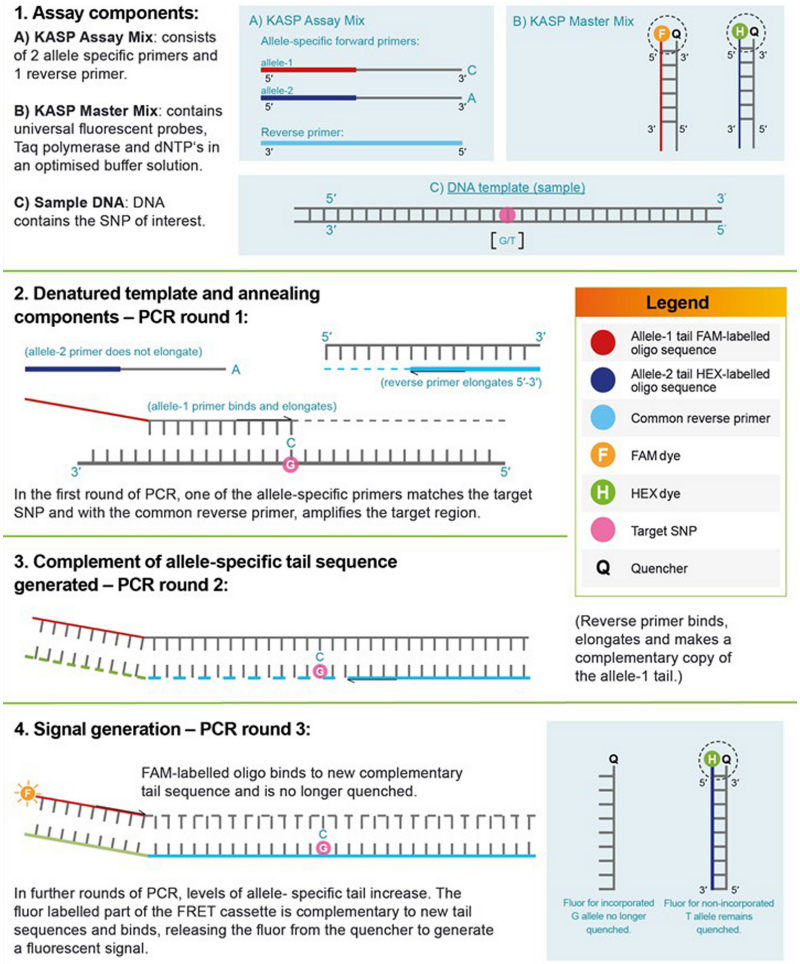

Figure 14: A schematic approach to the KASP mechanism of action (from Alvarez-Fernandez *et al.*, 2021)

DNA Barcoding

The success of DNA barcoding in species identification relies on its ability to use short, standardized regions of DNA that exhibit sufficient variability to distinguish between species. The choice of barcode region is critical, as it must be highly conserved within species but sufficiently variable between species. In animals, the mitochondrial cytochrome oxidase subunit 1 (COI) gene has proven to be highly effective for this purpose, but in plants, the situation is more complex due to the slower mutation rates in chloroplast DNA. As a result, researchers have explored various loci, including *matK*, *rbcL*, and *trnH-psbA* (Hollingsworth *et al.*, 2009), to find the most suitable markers for different plant groups.

Moreover, the ability to distinguish species using DNA barcoding has far-reaching implications for ecological studies and environmental monitoring. For instance, in marine biology, DNA barcoding is used to monitor fish populations and detect illegal fishing activities. In forestry, it helps in the identification of tree species, which is crucial for the management and conservation of forest ecosystems. These applications underscore the versatility of DNA barcoding as a tool for addressing a wide range of biological and environmental challenges.

Principles of DNA Barcoding

DNA barcoding is a method that uses a short genetic sequence from a standardized region of the genome to identify species. This technique has become a fundamental tool in taxonomy, ecology, and conservation. The approach involves the amplification and sequencing of specific DNA regions, such as the mitochondrial cytochrome c oxidase I (COI) gene in animals, which provides a unique identifier for each species. In plants, however, the search for a universal barcode has been more challenging, as no single locus has been found to be universally applicable across all plant taxa.

The development of next-generation sequencing technologies has been a game-changer for plant DNA barcoding. These technologies enable the simultaneous sequencing of multiple loci or even entire genomes, providing a much more comprehensive dataset for species identification. The concept of 'super barcodes,' which involves the use of entire chloroplast genomes (Li *et al.*, 2015), has emerged as a powerful tool for resolving complex taxonomic relationships. By analysing larger portions of the genome, researchers can achieve higher resolution in distinguishing closely related species, which is particularly important in groups with low genetic divergence (Parks *et al.*, 2009).

In addition to its use in species identification, DNA barcoding has been applied in various other fields. For example, in ecological research, DNA barcoding allows for the rapid assessment of biodiversity in different ecosystems. By analysing environmental DNA (eDNA) samples, researchers can detect the presence of species without the need for physical specimens, making it possible to monitor elusive or rare species more effectively. In conservation, DNA barcoding is used to combat the illegal trade of endangered species by providing a reliable method for species identification in confiscated wildlife products.

The establishment of global databases, such as the Barcode of Life Datasystem (BOLD), has been crucial in supporting these applications by providing a repository of barcode sequences for comparison and identification.

Cytogenomics

Cytogenomics integrates both the chromosomal and molecular aspects of genetics. It employs techniques from cytogenetics and genomics to examine the structure and function of genomes in relation to the cellular components that house genetic material, particularly chromosomes (Liehr, 2021). By utilizing methods such as chromosome imaging and genomic sequencing as well as a variety of genomic resources, researchers can identify genetic variations and uncover their implications for plant performance.

Cytogenomics provides a detailed understanding of how the genome is organized and how chromosomal changes can affect gene expression and phenotypic traits. Cytogenomics has been particularly important in the study of polyploidy, a condition where organisms have more than two sets of chromosomes, which is common in plants. Understanding the cytogenomic basis of polyploidy has implications for plant breeding and evolution, as polyploid species often exhibit greater genetic diversity and adaptability (Heslop-Harrison *et al.*, 2023).

Cytogenomics is not only crucial for understanding genome organization but also for studying the dynamics of genome evolution. Chromosomal rearrangements, such as inversions, translocations, and duplications, can have profound effects on gene expression and phenotype. By mapping these chromosomal changes, cytogenomics provides insights into the evolutionary processes that shape genomes and drive species diversity (Mayrose and Lysak, 2021). This knowledge is essential for understanding the genetic basis of adaptation and speciation, particularly in plants, where polyploidy and other chromosomal changes are common.

Advanced techniques like ATAC-Seq and ChIP-Seq have revolutionized the field of cytogenomics by providing high-resolution maps of chromatin accessibility and protein-DNA interactions across the genome. ATAC-Seq, for instance, identifies regions of the genome that are accessible to regulatory proteins, shedding light on the epigenetic mechanisms that control gene expression. ChIP-Seq, on the other hand, is used to identify the binding sites of transcription factors and other regulatory proteins, allowing researchers to understand how these factors influence gene expression and contribute to cellular function

The development of advanced cytogenomic techniques, such as single-cell sequencing and high-resolution chromosome conformation capture (Hi-C), has revolutionized our understanding of genome organization at the cellular level. These techniques allow researchers to study how chromatin structure and nuclear architecture influence gene expression and genome stability. For example, Hi-C has revealed the three-dimensional organization of the genome, showing how genes and regulatory elements are spatially arranged within the nucleus to facilitate or limit gene expression. These insights are crucial for understanding how changes in genome organization can lead to diseases, such as cancer, and for developing targeted therapies.

Introduction to Cytogenomics

Cytogenomics combines the study of chromosomes (cytogenetics) with genomic analysis (genomics) to explore the organization and function of the genome at a cellular level. This field has advanced significantly with the advent of techniques like ATAC-Seq (Assay for Transposase-Accessible Chromatin using sequencing) and ChIP-Seq (Chromatin Immunoprecipitation Sequencing), which allow for the mapping of regulatory elements and chromatin accessibility across the genome. These techniques provide insights into how the genome is organized within the nucleus and how this organization affects gene expression and cellular function.

The integration of cytogenomic data with resequencing efforts is particularly valuable in the study of complex traits, where multiple genetic and epigenetic factors interact to influence the phenotype. By combining WGRS with data from techniques like ATAC-Seq and ChIP-Seq, researchers can identify regulatory elements that are affected by genetic variants. This approach is especially useful in crop breeding, where understanding the regulatory mechanisms underlying important traits can lead to more precise and effective breeding strategies. For instance, the identification of regulatory variants associated with yield or stress response can guide the selection of breeding lines that are more likely to perform well under different environmental conditions.

Techniques in Cytogenomics

Flow cytometry

Flow cytometric analysis of nuclear DNA content is a high-throughput method for determining ploidy, defined as the number of basic chromosome sets (x), which is a fundamental genetic characteristic of any plant accession (Vrána *et al.*, 2014). The assay (**Figure 15**) begins with the preparation of a suspension of cell nuclei, which can be quickly and easily obtained by chopping a small amount of leaf tissue with a sharp razor blade. The nuclei are then stained with a fluorescent dye that specifically binds to DNA. This suspension is passed through a flow cytometer, which uses a laser to illuminate the nuclei as they flow in single file. The intensity of the emitted fluorescence correlates with the amount of DNA, allowing for quantification. The resulting data is plotted on a histogram, where the x-axis represents fluorescence intensity (indicating DNA content) and the y-axis indicates the number of nuclei.

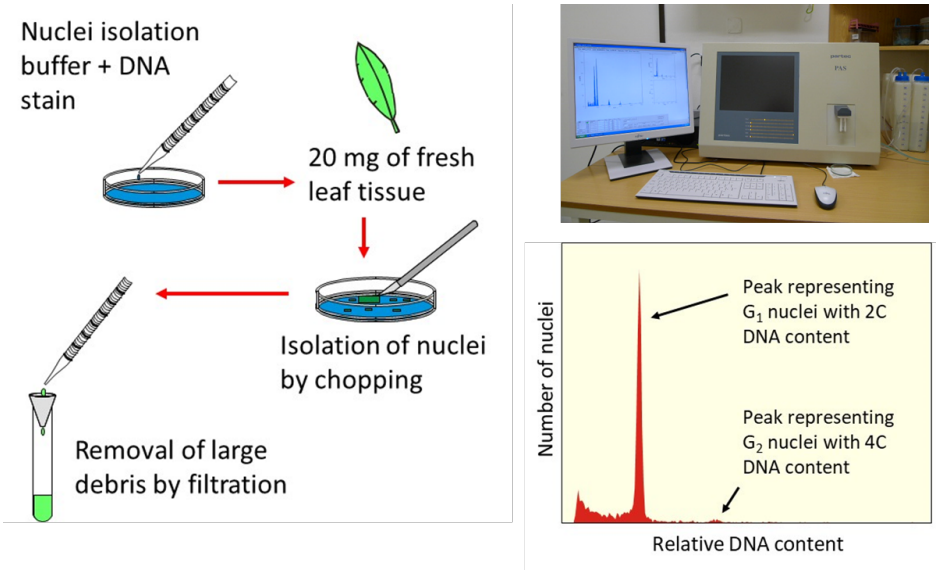

Figure 15: Flow cytometric estimation of nuclear DNA content.

Ploidy is determined by comparing the positions of G1-phase nuclei from the unknown sample to those of a reference standard (belonging to the same species) with known ploidy. This method is widely used in plant systematics, population biology, and plant breeding, as it provides a quick and accurate assessment of ploidy levels in individual accessions (Loureiro *et al.*, 2023). For example, the method was instrumental to characterize ploidy of all accessions of banana (*Musa* spp.) maintained in global gene bank (International Musa Germplasm Transit Centre in Leuven, Belgium) (Christelová *et al.*, 2017), (**Figure 16**). Under certain conditions, flow cytometry can also be used to identify interspecific hybrids and even aneuploids (Roux *et al.*, 2003).

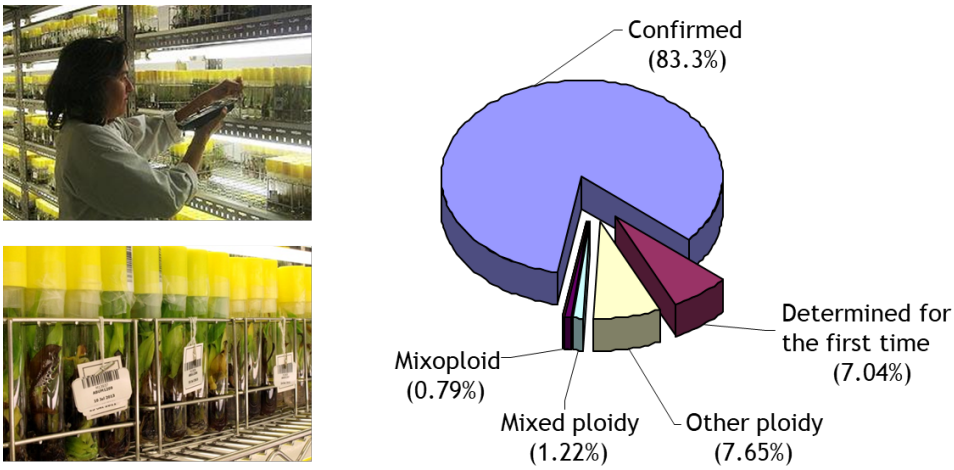

Figure 16: Characterization of ploidy of banana (*Musa* spp.) accessions maintained in global gene bank (Bioversity International Musa Germplasm Transit Centre, Leuven, Belgium).

Flow cytometry can also be used to estimate genome size in absolute units, such as picograms of DNA or megabase pairs (Mbp) (Doležel and Bartoš, 2005). The sample preparation is similar to that used for ploidy determination, but the fluorochromes selected for staining nuclear DNA—such as ethidium bromide or propidium iodide—should not exhibit a preference for AT or GC base pairs. To enhance accuracy, the nuclei of both the unknown sample and the reference standard are isolated, stained, and analysed simultaneously, enabling internal standardization (Doležel *et al.*, 1994), (**Figure 17**). Genome size is determined by comparing the positions of G1-phase nuclei from the unknown sample to those of a reference standard with a known genome size. A set of reference standards is available from the Centre of Plant Structural and Functional Genomics of the Institute of Experimental Botany, Olomouc (Czech Republic), facilitating comparison of genome size estimations obtained in different laboratories (Doležel *et al.*, 2007).

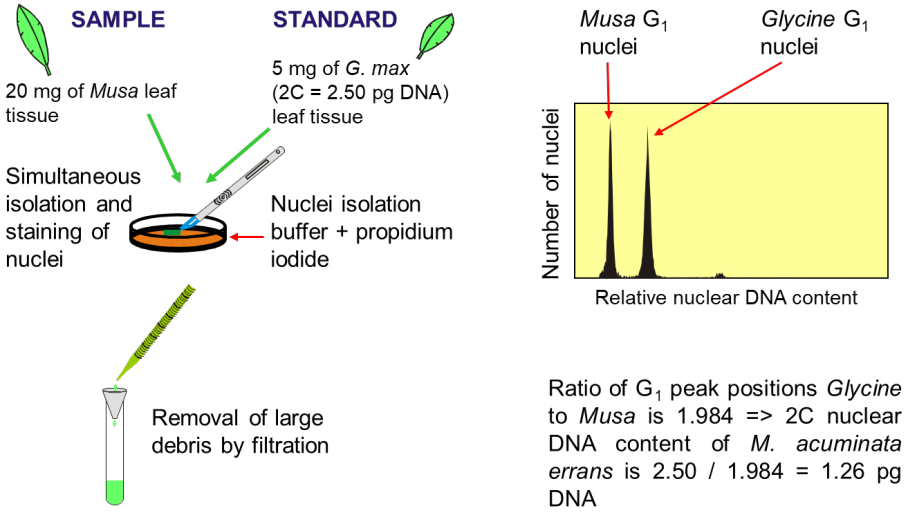

Figure 17: Flow cytometric estimation of nuclear genome size of banana (*Musa* spp.) in absolute units.

Molecular cytogenetics

Molecular cytogenetics integrates traditional cytogenetics—focused on the study of chromosomes and their structures—with molecular biology techniques to analyse and understand genetic material at the molecular level. This field enhances our understanding of the relationship between chromosome structure and genetic function, primarily through the use of fluorescence *in situ* hybridization (FISH) with fluorescently labelled DNA probes (Jiang, 2019). In this process, fluorescent signals are observed using fluorescence microscopy. Depending on the type of DNA probes used, FISH can identify individual chromosomes within a karyotype, distinguish chromosome arms, pinpoint specific chromosomal regions, or even locate single genes based on unique labelling patterns (**Figure 18**).

Initially, a variety of probes derived from repetitive DNA sequences limited the capabilities of FISH. However, advancements in methodology now allow for the identification of individual genes, and the development of oligonucleotide painting has further enhanced our ability to visualize specific chromosomes or regions (Harun *et al.*, 2023). This method employs short

synthetic DNA sequences labelled with fluorescent dyes and, among other, has been used to reveal striking variation in chromosome structure within the accessions maintained in banana (*Musa*) global gene bank (International Musa Germplasm Transit Centre in Leuven, Belgium) (Beránková *et al.*, 2024).

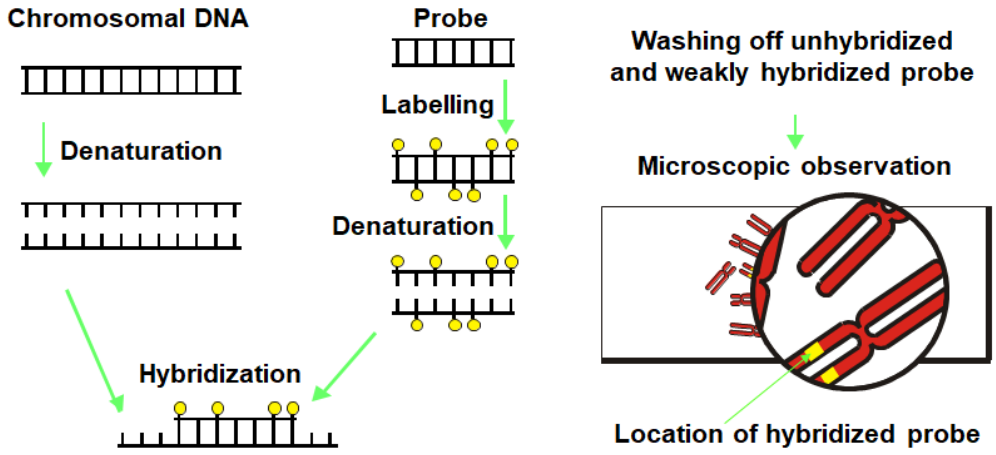

Figure 18. Principle of fluorescence *in situ* hybridization (FISH).

Genomic *in situ* hybridization (GISH) is a modification of FISH, utilizing labelled genomic DNA as probes. This technique enables researchers to distinguish chromosomes from different parents or genomes in interspecific or intergeneric hybrids and allopolyploids (Silva and Souza, 2013). Thus, Kopecký *et al.* (2006; 2017) used GISH to characterize genetic diversity in *Festulolium* (*Festuca* x *Lolium*), determine genetic constitution and identify homoeologous recombinations.

In summary, the applications of molecular cytogenetics are extensive, including the identification of chromosomal structures and their arrangements, as well as detecting changes such as deletions, duplications, or translocations. It has also been instrumental in locating specific genes on chromosomes, analysing genomic stability, and exploring evolutionary relationships between species. Other important methods of cytogenomics include ATAC-Seq and ChIP-seq. While ATAC-Seq identifies regions of open chromatin, which are accessible to transcription factors and other regulatory proteins, ChIP-Seq is used to map the binding sites of these proteins on the DNA. Together, these methods offer a detailed view of the regulatory landscape of the genome, helping researchers understand the complex interactions that govern gene expression. The integration of cytogenomic data with genomic and transcriptomic data provides a comprehensive understanding of genome function.

Case study: Characterization of banana accessions stored in the global Musa gene bank

To demonstrate the usefulness of cytogenomics to characterize plant genetic resources, a set of 125 new banana accessions introduced to the *Musa* gene bank was selected. A global *Musa* gene bank, the International Musa Germplasm Transit Centre (ITC) in Leuven, Belgium, plays a vital role in safeguarding the genetic diversity of one of the world's most important staple and cash crops. Bananas are highly vulnerable to pests, diseases, and climate change, largely because most commercial varieties are clonally propagated and genetically uniform. By conserving a wide range of wild species and traditional cultivars, the gene bank provides essential genetic resources for breeding resilient, high-yielding, and climate-adapted bananas. This invaluable reservoir ensures that future generations can continue to depend on bananas for food security, livelihoods, and economic stability worldwide.

Bananas (*Musa* spp.) and their two related genera, *Ensete* and *Musella*, belong to the family Musaceae. The primary centre of diversity is located in Southeast Asia. Most edible banana cultivars originated from interspecific or intersubspecific crosses between *M. balbisiana* (B genome, 2n = 2x = 22) and several subspecies of *M. acuminata* (A genome, 2n = 2x = 22). Other edible bananas derive from different *Musa* species related to *M. textilis* (2n = 2x = 20) or *M. schizocarpa* (2n = 2x = 22). Hybridization between the A and B genomes gave rise to edible diploid and triploid cultivars, which are almost completely sterile and highly susceptible to many diseases and disorders. The sterility of edible clones, combined with their unclear genetics and evolutionary history, complicates breeding efforts. Therefore, it is essential to characterize the genetic diversity and variability of both wild diploid and edible banana accessions stored in the *Musa* gene bank, which provides individual accessions to breeders.

The International Musa Germplasm Transit Centre currently holds more than 1,700 accessions of edible and wild banana species. Originally, the accessions introduced to the gene bank have been classified primarily using morpho-taxonomic markers. However, this approach was not found reliable, as a study of Christelová *et al.*(2017) revealed mislabelling of 22 accessions. These findings highlighted the need for more precise characterization of all newly added material. Consequently, each new accession is now analysed for ploidy and genotyped using DNA markers. Ploidy is determined using flow cytometry, a fast and accurate method that does not require actively dividing (mitotic) cells. All measurements are performed together with a consistent internal standard, allowing direct comparison with previously analysed accessions. Genotyping is conducted using a defined set of 19 microsatellite (SSR) markers (Christelová *et al.*, 2011; Christelová *et al.*, 2017). The combination of ploidy estimation and SSR genotyping provides detailed insight into the genetic background of each accession. This enables accurate identification of species, subspecies, and groups or types of edible banana clones, thereby preventing misclassification and delivering essential information for all gene bank users.

In specific cases, for gene bank accessions used in breeding programs, flow cytometric analysis and SSR genotyping are complemented by karyotype analysis using FISH with chromosome-

specific painting probes. This approach provides detailed information on chromosome organization and allows the detection of large chromosomal translocations (Šimoníková *et al.*, 2019; Šimoníková *et al.*, 2020; Beránková *et al.*, 2024).

Due to the seed sterility of most edible banana clones and the low germination rates observed in some wild species, the International Musa Germplasm Transit Centre maintains all accessions as *in vitro* plants. These are kept under slow-growth conditions and periodically transferred to fresh media to ensure long-term preservation. For genetic characterization of newly introduced accessions, samples from five randomly selected plants were analysed for each accession. This approach proved useful for detecting accessions containing multiple genotypes, most likely due to human error during the introduction into culture.

The genetic characterization began with ploidy estimation using flow cytometry following a well-established protocol (Doležel *et al.*, 2007). This technique is particularly advantageous for *in vitro*-grown plantlets, which are small and provide only a limited amount of leaf tissue—yet sufficient to prepare high-quality samples. Crude suspensions of isolated nuclei were supplemented with chicken red blood cell (CRBC) nuclei, serving as an internal reference standard (Roux *et al.*, 2003; Christelová *et al.*, 2017). The advantage of this standardization is that CRBC nuclei can be prepared in large quantities, allowing all *Musa* samples to be analysed against the same reference. Ploidy was then determined based on the DNA peak ratios of *Musa* and CRBC nuclei (**Figure 19 A,B**).

Of the 125 accessions analysed (625 plantlets in total), 16 had a ploidy level different from that expected based on morpho-taxonomic identification. In 10 accessions, ploidy was estimated for the first time. In four accessions, at least one of the five plantlets displayed mixoploidy (**Figure 19 C,D**).

Following the flow cytometric analysis, the new accessions were genotyped using a set of microsatellite (SSR) markers as described in Christelová *et al.* (2011; 2017). Briefly, genomic DNA was extracted from leaf tissue and used for PCR amplification and fragment analysis. A total of 19 SSR loci were amplified with locus-specific primers modified by 5´-M13 tails to enable the use of a universal fluorescently labelled primer. Four distinct fluorophores were employed for primer labelling, allowing subsequent multiplexing of PCR reactions. To improve allele binning accuracy, three independent PCR reactions were performed for each sample.

PCR products were purified and used for capillary electrophoresis. Optimized amounts of amplification products were combined with an internal size standard and loaded onto an ABI 3730xl DNA Analyzer equipped with a 96-capillary array. To reduce costs and increase throughput, samples were multiplexed in the second and third electrophoretic runs. Up to four-fold multiplexing was achieved by combining PCR products labelled with different fluorescent dyes into a single injection.

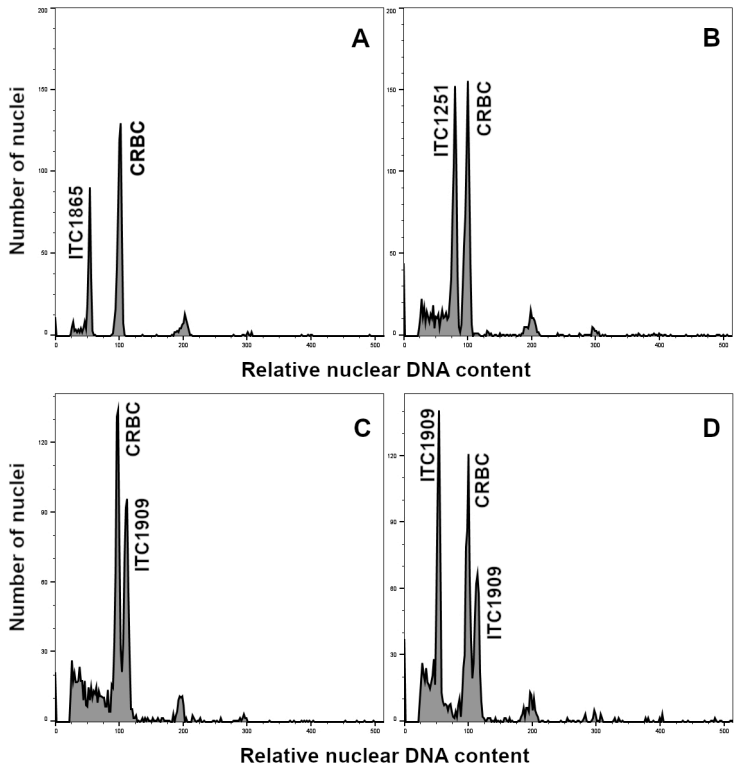

Figure 19: Histograms of relative nuclear DNA content obtained after simultaneous flow cytometric analysis of DAPI-stained nuclei isolated from fresh leaf tissues of *Musa* and chicken red blood cell nuclei (CRBC), which served as an internal reference standard. G1 peaks of CRBC were positioned on channel 100, peaks appearing on channels 200 and 300 correspond to doublets and triplets of CRBC nuclei. Ploidy of *Musa* accessions was determined based on the ratio of G1 peak positions (*Musa* : CRBC), knowing that in diploid, triploid and tetraploid plants, the ratio is ~ 0.5, 0.75 and 1 respectively. (A) Diploid accession *M. acuminata* (ITC1865); the ratio of G1 peak means was 0.53. (B) Triploid accession Vietnam no.5 (ITC1251); the ratio of G1 peak means was 0.77. (C) Tetraploid *M. acuminata* x *M. schizocarpa* cultivar (ITC1909); the ratio of G1 peak means was 1.18. (D) Mixoploid plant of *M. acuminata* x *M. schizocarpa* cultivar (ITC1909) with diploid and tetraploid nuclei; the ratio of G1 peak means was 0.59 and 1.18 respectively.

Alleles were automatically called using GeneMarker software and manually verified. The SSR profiles of the newly analysed accessions were integrated into the binary SSR dataset (Christelová *et al.*, 2017) and analysed jointly. Genetic similarity matrices were calculated using Nei's genetic distance coefficient and dendrogram was constructed based on the results of UPGMA analysis implemented in DARwin software v6.0.021 (Perrier and Jacquemoud-Collet, 2006) and visualized in FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/) (**Figure 20**). This enabled taxonomic classification of previously unidentified accessions (**Table 1**).

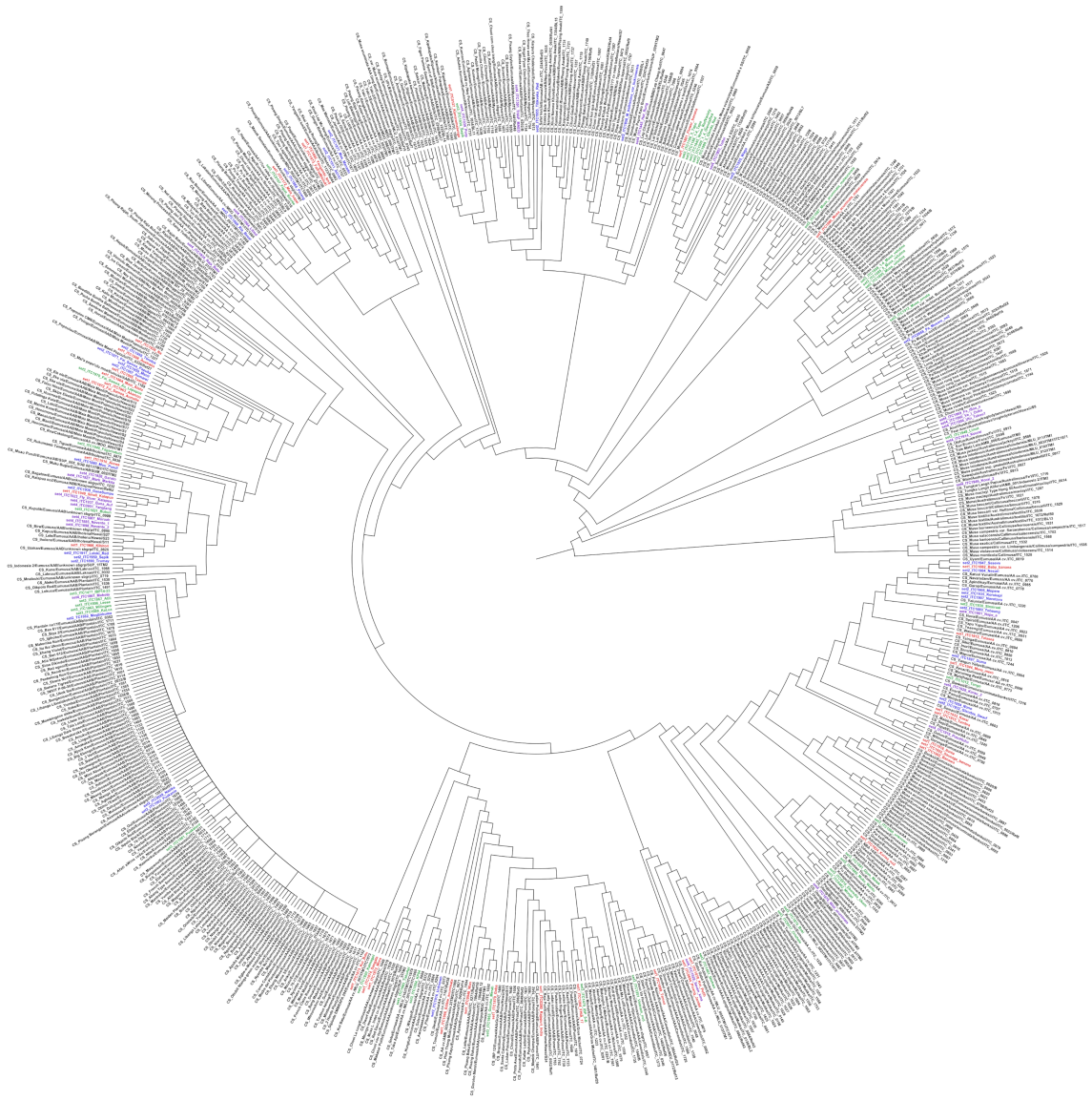

Figure 20. UPGMA dendrogram constructed based on SSR data of 125 newly analysed accessions (in colours) and dataset obtained by Christelová *et al.* (2017).

Table 1. Summary of flow cytometry and SSR genotyping results for 125 new *Musa* accessions introduced to the International Musa Germplasm Transit Centre (ITC)

ITC code	Accession name	Expected ploidy based on classification	Ploidy estimated by flow cytometry	SSR clustering
0028	Nazika	3x	3x	AAB Plantains
0107	Figue Sucrée	2x	2x	AA cv. Sucrier (AA cv. ISEA 2)
0127	Kamaramasenge	3x	3x	AAB Silk
0280	Rajapuri India	3x	3x	AAB Pome
0462	Monjet	2x	4x	AA cv. banksii *sensu lato*
0516	Eberedia Ukom	3x	2x	AA cv. banksii *sensu lato*
0517	Orishele	3x	2x	AA cv. banksii *sensu lato*
0668	Pa (Musore) no.2	2x	2x	M. acuminata
0694	Pisang Jambe	4x	3x	AA cv. IndonTriPh
0901	Tangtang	3x	3x	closely related to AAB/Iholena
0948	Wan	2x	3x	AAA Rio/Ambon
1007	Moruah	-	3x	closely related to AAB/Iholena
1012	Tango	2x	2x	AA cv. banksii derivatives
1172	Mai'a hapai	2x	2x	AA cv. Sucrier (AA cv. ISEA 2)
1188	Tapo	2x	2x; one mixoploid plant	AS cv.
1251	Vietnam no.5	3x	3x	AAA Cavendish
1264	FHIA-17	4x	4x	AAA Gros Michel – connected
1301	THAP MAEO	3x	3x	clusters together with AAB/Mysore
1445	Hot Rung	2x	2x	clusters together with BB/M. balbisiana
1461	Ntebwa	3x	3x	AAA Lujugira/Mutika
1471	Zanzebar	3x	3x	clusters together with Cavendish
1477	IBP 14-23	4x	4x	AAB Plantains
1479	IBP 5-B	3x	3x	AAA Gros Michel
1850	Musa balbisiana var. liukiuensis	2x	2x	M. balbisiana
1860	Muku Wahtu	2x	2x	AA cv. IndonTriNG
1861	Maboto	3x	3x	clusters together with AAB/Plantain.
1862	Mogbokuma	3x	3x	AAB Plantains
1863	Wilingwa	3x	3x	AAB Plantains
1864	Libod	2x	2x	clusters together with AA cv.
1865	Tako Api	2x	2x	AA cv. IndonTriNG
1866	Edible acuminata	2x	2x	AA cv. P. Tongat/AA cv. IndonTriNG
1867	Atili	3x	Mixoploid plants	AAB Plantains
1868	Pinang	3x	3x	AA cv. ISEA 2
1869	Roa Besar	2x	2x	AA cv. ISEA 2/IndonTriNG
1870	Boki	2x	2x	AA cv. banksii *sensu lato*
1872	Mora	-	3x	subclade with close connection to P. Jari Buaya and AAA Lujugira/Mutika
1873	Koi Batu	2x	2x	Indonesian AA cv.
1874	Sangate	-	3x	subclade with close connection to P. Jari Buaya and AAA Lujugira/Mutika
1876	Wild acuminata	2x	2x	clusters together with AA/banksii (Hawain accessions)
1877	Mas Manado	2x	2x	ISEA 1 / connected to AAA Ibota
1878	Koi Putih	2x	2x	Indonesian AA cv.
1879	Goba	2x	3x	AAA Rio/Ambon

1880	Mu'u Pundi	2x	2x	plantain-linked
1881	Flower banana	-	2x	AS cv.
1882	Baby banana	2x	2x	AA cv. banksii derivatives
1883	Tobaung	2x	2x	AA cv. banksii (derivates and *sensu lato*)
1884	KP 04	2x	2x	AA cv. IndonTriNG
1885	Navente 1	3x	3x	closely related to AAB/Iholena
1886	Musa acuminata ssp. malaccensis	2x	2x	M. acuminata ssp. malaccensis
1887	Musa acuminata malaccencis	2x	2x	M. acuminata ssp. malaccensis
1888	Musa velutina	2x	2x; one mixoploid plant	Rhodochlamys/M. velutina
1890	Trumay	2x	3x	AAB Iholena (plantain linked)
1891	Djum Metek	2x	2x	AA cv. IndonTriNG
1892	Yangambi KM 5	3x	3x	AAA Ibota
1895	Musa velutina	2x	2x	Rhodochlamys/M. velutina
1896	Leese	3x	3x	AAB Plantains
1897	Duma	2x	2x	AA cv. banksii derivatives
1898	Navente 2	3x	3x	closely related to AAB/Iholena
1899	Tamoa	-	3x	AAA Cavendish
1900	unknown	2x	3x	connected to AA cv. banksii *sensu lato/M. ac.* ssp. *banksii*
1901	Nape"e	2x	2x	clusters together with AA cv.
1902	Banawa	2x	3x	AA cv. banksii (derivates and *sensu lato*)
1903	Toitoi	-	3x	clusters together with AA X SS hybrids
1904	Nesuri	2x	2x	AA cv. banksii derivatives
1905	Mopere	2x	2x	AA cv. banksii derivatives
1906	Kibirori	3x	3x	AAB Iholena (plantain linked)
1907	Navotavu	2x	2x	AA cv. banksii derivatives
1908	Australia	3x	3x	AAB Maia Maoli/Popoulu
1909	Glenda Red	-	4x; four mixoploid plants	AS cv.
1910	Arawa	2x	3x	AAB plantain-linked
1911	A0 157	3x	3x	AAA Ibota
1912	Musa ornata	2x	2x	*M. ornata*
1913	Talasea	2x	2x	AA cv. banksii derivatives
1914	Kourai	2x	2x	clusters together with M. troglodytarum
1915	Tambra	2x	2x	AA cv. banksii *sensu lato*
1916	Unknown	2x	2x	closest accession: ITC0373_Uwati/AA cv.
1917	Laloki Red	3x	3x	AAB Iholena (plantain linked)
1918	Itonia	2x	2x	AA cv. banksii *sensu lato*
1919	PosoHuhu	2x	2x	clusters together with AA cv.
1920	Kalmagol	4x	3x	subclade with close connection to P. Jari Buaya and AAA Lujugira/Mutika
1921	Bubun	3x	3x	AAB Iholena (plantain linked)
1922	Sausage banana	-	2x	AA cv. banksii (derivates and *sensu lato*)
1923	Fly River Kalapua	3x	3x	closely retabed to AAB/Kalapua
1924	Kurisa No. 2	2x	2x	AA cv. banksii *sensu lato*
1925	Waga	2x	3x	closest accession: ITC0299_Guyod/AAcv.

1926	Garoto	3x	3x	closely related to AAB/Kalapua
1927	Mark Markila	3x	3x	closely related to AAB/Kalapua
1928	Karau 2	2x	2x	clusters together with AA cv.
1929	Seven kina	2x	3x	closely related to Lujugira/Mutika
1930	Sinsiruai	2x	2x	AA cv. banksii derivatives
1932	Glenda's Dwarf	2x	2x	AA cv. banksii (derivates and *sensu lato*)
1933	Kaesi	2x	2x	AA cv. banksii *sensu lato*
1934	Sepik	3x	4x	AAB Silk
1935	Korukapi	3x	3x	AA cv. banksii derivatives
1936	N/A	3x	3x	AAB Maia Maoli/Popoulu
1937	Gana Auf	3x	3x	closely retabed to AAB/Kalapua
1938	GanaSumpu	3x	3x	AAB plantain-linked (Kalapua)
1939	Porp	4x	4x	clusters together with AAB/Silk
1940	Fagamutum	3x	3x	AAB Iholena (plantain linked)
1944	Mero mero	2x	2x	AA cv. banksii derivatives
1945	Nono (Red Variety)	-	4x	AS cv.
1946	Korai 2	-	2x	clusters together with Fe'i
1947	Seseve	2x	2x	AA cv. banksii derivatives
1948	Limot	2x	2x	Australimusa/Fe'i
1949	Small Kalapua	3x	3x	AAB plantain-linked (Kalapua)
1950	Sepik Red	3x	3x	AAB Iholena (plantain linked)
1953	Goroho Merah	3x	3x	AAA Rio/Ambon
1954	Nono	3x	2x	AA cv. banksii (derivates and *sensu lato*)
1955	Titikaveka Red	3x	3x	AAB Mysore
1956	Utu Tekou1	2x	2x	clusters with M. troglodytarum
1957	Aumarei	3x	3x	AAB Maia Maoli/Popoulu
1958	Torotea	3x	3x	AAB Maia Maoli/Popoulu
1959	Mani'i	3x	3x	AAB Maia Maoli/Popoulu
1960	Ve"I Ooka	2x	2x	clusters with M. troglodytarum
1961	Puakatoro	3x	3x	AAB Plantains
1962	Puakanio	3x	3x	AAB Plantains
1963	Akamou	3x	3x	AAB Maia Maoli/Popoulu
1964	Mao'i Atetu	3x	3x	AAB Maia Maoli/Popoulu
1965	Ta'anga	3x	3x	AAB Maia Maoli/Popoulu
1968	KaLua	3x	3x	AAB Plantains
1969	Fa"iSoa"a	2x	2x	clusters with M. troglodytarum
1970	Fa''i Samoa Au Lapopoa	3x	3x	AAB Maia Maoli/Popoulu
1971	Fa'i Samoa Pupuka	3x	3x	AAB Maia Maoli/Popoulu
1972	Fa'i Samoa Aumalie	3x	3x	AAB Maia Maoli/Popoulu
1973	Fa'iFiaMisiluki	2x	3x	AAA Ibota

In several cases, the classification of newly introduced accessions based on SSR genotyping did not match the genomic constitution information available at the time of their introduction into the gene bank. To resolve such discrepancies, karyotype analysis was employed. For instance, SSR genotyping of the accession 'Zanzebar' (ITC 1471)—originally described as a triploid clone with an AAB genome belonging to the Plantain subgroup—placed it instead among accessions of the Cavendish subgroup (genome AAA) (**Figure 20)**. Because individual *Musa* species, subspecies, and edible clones differ in the presence of large chromosome translocations, we applied oligo painting FISH to verify the genome organization of this accession, following the protocol of Šimoníková *et al.* (2019).

Briefly, sets of 20,000 oligomers (45 nt) covering individual chromosome arms were synthesized and labelled either directly with CY5 or Texas Red fluorochromes or indirectly via digoxigenin or biotin. Mitotic metaphase chromosome spreads were prepared from actively growing root tips and hybridized with the oligonucleotide probes. Digoxigenin- and biotin-labelled probes were detected using anti-digoxigenin–FITC and streptavidin–Cy3, respectively. Images of metaphase spreads were captured with a fluorescence microscope, and final image processing and idiogram construction were performed in Adobe Photoshop. The analysis revealed a chromosome organization typical of the Cavendish subgroup (**Figure 21**), thus confirming the SSR genotyping results.

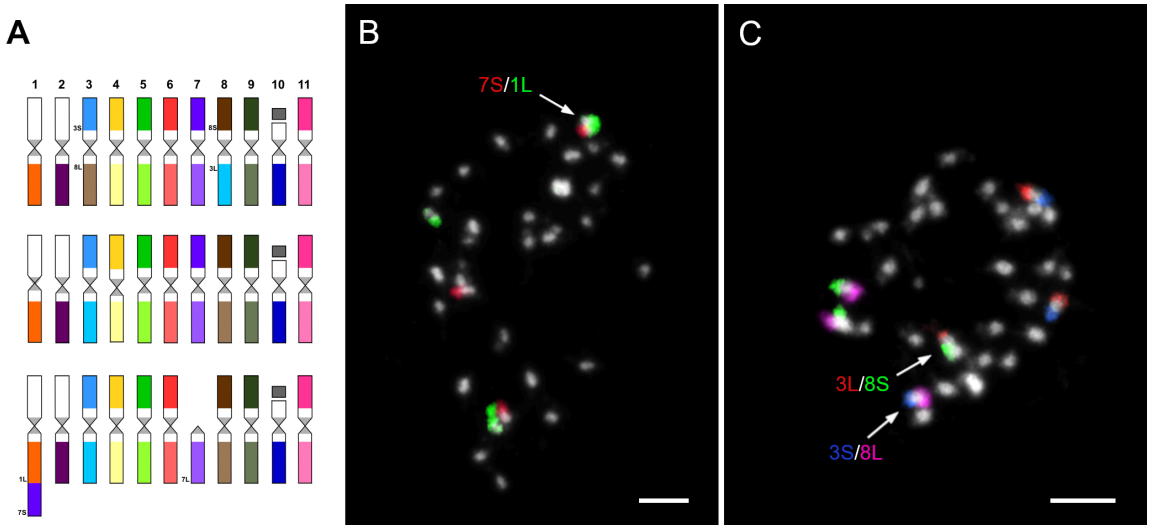

Figure 21. Chromosome translocations identified by oligo-painting FISH on mitotic metaphase chromosomes of the triploid accession 'Zanzebar' (ITC1471) (A) Ideogram of 'Zanzebar' with translocations specific for the Cavendish subgroup. (B) Oligo-painting FISH with probes specific for short arm of chromosome 7 (red) and long arm of chromosome 1 (green). (C) Oligo-painting FISH with probes specific for short and long arm of chromosome 3 (blue and red respectively) and short and long arm of chromosome 8 (green and magenta respectively). Chromosomes were counterstained with DAPI (light grey pseudo-colour). Arrows point to chromosomes with translocation. Bars = 5 µm.

Conclusions

As we look to the future, the integration of DNA barcoding, reduced representation sequencing, and cytogenomic methods will continue to drive innovation in genomic research. These technologies offer unprecedented opportunities to explore the complexities of the genome, from understanding species diversity to uncovering the genetic basis of important traits (Christelová *et al.*, 2017). The ongoing refinement of these methodologies will undoubtedly lead to new discoveries and applications, particularly in areas such as agriculture, where the need for sustainable and resilient crops is more pressing than ever. By leveraging the power of genomics, we can address some of the most critical challenges facing our world today, from food security to biodiversity conservation.

The integration of DNA barcoding, reduced representation sequencing, and cytogenomic methods represents a significant advancement in our ability to understand and manipulate the genome. These tools have not only enhanced our capacity to classify and identify species but have also provided deep insights into the genetic basis of important traits in plants, animals, and other organisms. As these technologies continue to evolve, they will undoubtedly play a critical role in addressing some of the most pressing challenges facing our world today, from food security to biodiversity conservation. The future of genomic research lies in the continued refinement and integration of these methodologies, which will allow us to explore the complexities of the genome with unprecedented detail and accuracy.

Integrating cytogenomic data with resequencing efforts enhances our understanding of genome organization and function. For instance, combining WGRS with ATAC-Seq or ChIP-Seq can reveal how genetic variants affect chromatin structure and gene regulation. This integrated approach is particularly valuable in identifying regulatory variants that contribute to complex traits in crops, ultimately guiding more effective breeding strategies.

The refinement and demonstration of DNA barcoding, reduced representation sequencing, and cytogenomic methods represent significant advancements in genomic research. These technologies have not only enhanced our ability to identify and classify species but have also provided deep insights into the genetic basis of important traits in plants and other organisms. The future of these fields lies in the integration of various genomic techniques, which will allow for a more comprehensive understanding of genome function and evolution, ultimately contributing to advancements in agriculture, conservation, and biodiversity research.

Deviations

None.

References

Abrouk, M., Ahmed, H.I., Cubry, P., et al. (2020) Fonio millet genome unlocks African orphan crop diversity for agriculture in a changing climate. *Nat Commun*, **11**, 4488.

Acquadro, A., Barchi, L., Gramazio, P., Portis, E., Vilanova, S., Comino, C., Plazas, M., Prohens, J. and Lanteri, S. (2017) Coding SNPs analysis highlights genetic relationships and evolution pattern in eggplant complexes M. Singh, ed. , **12**, e0180774.

Ahn, Y.-K., Manivannan, A., Karna, S., Jun, T.-H., Yang, E.-Y., Choi, S., Kim, J.-H., Kim, D.-S. and Lee, E.-S. (2018) Whole Genome Resequencing of Capsicum baccatum and Capsicum annuum to Discover Single Nucleotide Polymorphism Related to Powdery Mildew Resistance. *Scientific Reports 2018 8:1*, **8**, 5188.

Albert, T.J., Molla, M.N., Muzny, D.M., et al. (2007) Direct selection of human genomic loci by microarray hybridization. *Nat Methods*, **4**, 903–905.

Alkan, C., Sajjadian, S. and Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nat Methods*, **8**, 61–65.

Alonso-Blanco, C., Andrade, J., Becker, C., et al. (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, **166**, 481–491.

Alvarez-Fernandez, A., Bernal, M.J., Fradejas, I., Martin Ramírez, A., Md Yusuf, N.A., Lanza, M., Hisam, S., Pérez de Ayala, A. and Rubio, J.M. (2021) KASP: a genotyping method to rapid identification of resistance in Plasmodium falciparum. *Malaria Journal*, **20**, 16.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.

Ballouz, S., Dobin, A. and Gillis, J.A. (2019) Is it time to change the reference genome? *Genome Biol*, **20**, 159.

Barchi, L., Acquadro, A., Alonso, D., et al. (2019) Single Primer Enrichment Technology (SPET) for High-Throughput Genotyping in Tomato and Eggplant Germplasm. *Front. Plant Sci.*, **10**. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01005/full [Accessed December 12, 2020].

Barchi, L., Aprea, G., Rabanus-Wallace, M.T., et al. (2023) Analysis of >3400 worldwide eggplant accessions reveals two independent domestication events and multiple migration-diversification routes. *The Plant Journal*, **116**, 1667–1680.

Barchi, L., Lanteri, S., Portis, E., et al. (2012) A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. G. Bonaventure, ed. *PloS one*, **7**, e43740.

Barchi, L., Lanteri, S., Portis, E., et al. (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, **12**.

Barchi, L., Rabanus-Wallace, M.T., Prohens, J., et al. (2021) Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *The Plant Journal*, **107**, 579–596.

Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J. and Edwards, D. (2020) Plant pan-genomes are the new reference. *Nat. Plants*, **6**, 914–920.

Bayer, P.E., Petereit, J., Durant, É., et al. (2022) Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. *The Plant Genome*, **15**, e20221.

Beissinger, T.M., Hirsch, C.N., Sekhon, R.S., et al. (2013) Marker Density and Read Depth for Genotyping Populations Using Genotyping-by-Sequencing. *Genetics*, **193**, 1073–1081.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved Elements in the Human Genome. *Science*, **304**, 1321–1325.

Beránková, D., Čížková, J., Majzlíková, G., Doležalová, A., Mduma, H., Brown, A., Swennen, R. and Hřibová, E. (2024) Striking variation in chromosome structure within Musa acuminata subspecies, diploid cultivars, and F1 diploid hybrids. *Front. Plant Sci.*, **15**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1387055/full [Accessed July 1, 2025].

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M. and Phillippy, A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*, **33**, 623–630.

Bhattacharjee, R., Luseni, M.M., Ametefe, K., Agre, P.A., Kumar, P.L. and Grenville-Briggs, L.J. (2024) Genetic Diversity and Population Structure of Cacao (Theobroma cacao L.) Germplasm from Sierra Leone and Togo Based on KASP–SNP Genotyping. *Agronomy*, **14**, 2458.

Cameron, D.L., Di Stefano, L. and Papenfuss, A.T. (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*, **10**, 3240.

Chawla, H.S., Lee, H., Gabur, I., et al. (2021) Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnology Journal*, **19**, 240–250.

Cheng, H., Liu, J., Wen, J., et al. (2019) Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biol*, **20**, 136.

Christelová, P., De Langhe, E., Hřibová, E., et al. (2017) Molecular and cytological characterization of the global Musa germplasm collection provides insights into the treasure of banana diversity. *Biodivers Conserv*, **26**, 801–824.

Christelová, P., Valárik, M., Hřibová, E., Van den houwe, I., Channelière, S., Roux, N. and Doležel, J. (2011) A platform for efficient genotyping in Musa using microsatellite markers. *AoB PLANTS*, **2011**, plr024.

Danecek, P., Bonfield, J.K., Liddle, J., et al. (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.

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

DePristo, M.A., Banks, E., Poplin, R., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dolezel, J. and Bartos, J. (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Ann Bot*, **95**, 99–110.

Doležel, J., Greilhuber, J. and Suda, J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nat Protoc*, **2**, 2233–2244.

Doležel, J., Lucretti ,Sergio and Schubert, I. and (1994) Plant Chromosome Analysis and Sorting by Flow Cytometry. *Critical Reviews in Plant Sciences*, **13**, 275–309.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, 1–10.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E. and Holzapfel, C.M. (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, **107**, 16196–16200.

Franke, K.R. and Crowgey, E.L. (2020) Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. *Genomics Inform*, **18**, e10.

Gardoce, R.R., Manohar, A.N.C., Mendoza, J.-V.S., Tejano, M.S., Nocum, J.D.L., Lachica, G.C., Gueco, L.S., Cueva, F.M.D. and Lantican, D.V. (2023) A novel SNP panel developed for targeted genotyping-by-sequencing (GBS) reveals genetic diversity and population structure of Musa spp. germplasm collection. *Mol Genet Genomics*, **298**, 857–869.

Garrison, E., Guarracino, A., Heumos, S., et al. (2023) Building pangenome graphs. , 2023.04.05.535718. Available at: https://www.biorxiv.org/content/10.1101/2023.04.05.535718v1 [Accessed January 24, 2024].

Gnerre, S., MacCallum, I., Przybylski, D., et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, **108**, 1513–1518.

Gore, M.A., Chia, J.M., Elshire, R.J., et al. (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115–1117.

Hackl, S.T., Harbig, T.A. and Nieselt, K. (2022) Technical report on best practices for hybrid and long read de novo assembly of bacterial genomes utilizing Illumina and Oxford Nanopore Technologies reads. , 2022.10.25.513682. Available at: https://www.biorxiv.org/content/10.1101/2022.10.25.513682v1 [Accessed July 21, 2025].

Harun, A., Liu, H., Song, S., Asghar, S., Wen, X., Fang, Z. and Chen, C. (2023) Oligonucleotide Fluorescence In Situ Hybridization: An Efficient Chromosome Painting Method in Plants. *Plants*, **12**, 2816.

Heslop-Harrison, J.S. (Pat), Schwarzacher, T. and Liu, Q. (2023) Polyploidy: its consequences and enabling role in plant diversification and evolution. *Annals of Botany*, **131**, 1–10.

Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Marschall, T., Li, H. and Paten, B. (2023) Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol*, 1–11.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., et al. (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12794–12797.

Huang, X., Kurata, N., Wei, X., et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.

Hugall, A.F., O'Hara, T.D., Hunjan, S., Nilsen, R. and Moussalli, A. (2016) An Exon-Capture System for the Entire Class Ophiuroidea. *Molecular Biology and Evolution*, **33**, 281–294.

Jayakodi, M., Lu, Q., Pidon, H., et al. (2024) Structural variation in the pangenome of wild and domesticated barley. *Nature*, 1–9.

Jiang, J. (2019) Fluorescence in situ hybridization in plants: recent developments and future applications. *Chromosome Res*, **27**, 153–165.

Jiang, X., Yang, T., Zhang, F., et al. (2022) RAD-Seq-Based High-Density Linkage Maps Construction and Quantitative Trait Loci Mapping of Flowering Time Trait in Alfalfa (Medicago sativa L.). *Front. Plant Sci.*, **13**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.899681/full [Accessed December 9, 2024].

Jiao, C., Xie, X., Hao, C., et al. (2024) Pan-genome bridges wheat structural variations with habitat and breeding. *Nature*, 1–10.

Kechin, A., Borobova, V., Boyarskikh, U., Khrapov, E., Subbotin, S. and Filipenko, M. (2020) NGS-PrimerPlex: High-throughput primer design for multiplex polymerase chain reactions. *PLOS Computational Biology*, **16**, e1008468.

Kopecký, D., Loureiro, J., Zwierzykowski, Z., Ghesquière, M. and Doležel, J. (2006) Genome constitution and evolution in Lolium × Festuca hybrid cultivars (Festulolium). *Theor Appl Genet*, **113**, 731–742.

Kopecký, D., Šimoníková, D., Ghesquière, M. and Doležel, J. (2017) Stability of Genome Composition and Recombination between Homoeologous Chromosomes in Festulolium (Festuca × Lolium) Cultivars. *Cytogenetic and Genome Research*, **151**, 106–114.

Lemmon, A.R., Emme, S.A. and Lemmon, E.M. (2012) Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Systematic Biology*, **61**, 727–744.

Lemmon, E.M. and Lemmon, A.R. (2013) High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 99–121.

Li, C., Hofreiter ,Michael, Straube ,Nicolas, Corrigan ,Shannon and Naylor, G.J.P. and (2013) Capturing Protein-Coding Genes Across Highly Divergent Species. *BioTechniques*, **54**, 321–326.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997.

Li, X., Yang, Y., Henry, R.J., Rossetto, M., Wang, Y. and Chen, S. (2015) Plant DNA barcoding: from gene to genome. *Biological Reviews*, **90**, 157–166.

Liang, Z., Duan, S., Sheng, J., et al. (2019) Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nat Commun*, **10**, 1190.

Liehr, T. (2021) Chapter 1 - A definition for cytogenomics - Which also may be called chromosomics. In T. Liehr, ed. *Cytogenomics*. Academic Press, pp. 1–7. Available at: https://www.sciencedirect.com/science/article/pii/B9780128235799000011 [Accessed July 1, 2025].

Liu, F., Zhao, J., Sun, H., et al. (2023) Genomes of cultivated and wild Capsicum species provide insights into pepper domestication and population differentiation. *Nat Commun*, **14**, 5487.

Liu, Y., Du, H., Li, P., et al. (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell*, **182**, 162-176.e13.

Loureiro, J., Čertner, M., Lučanová, M., Sliwinska, E., Kolář, F., Doležel, J., Garcia, S., Castro, S. and Galbraith, D.W. (2023) The Use of Flow Cytometry for Estimating Genome Sizes and DNA Ploidy Levels in Plants. In T. Heitkam and S. Garcia, eds. *Plant Cytogenetics and Cytogenomics: Methods and Protocols*. New York, NY: Springer US, pp. 25–64. Available at: https://doi.org/10.1007/978-1-0716-3226-0_2 [Accessed July 1, 2025].

Lv, Q., Li, W., Sun, Z., et al. (2020) Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. *Nat Commun*, **11**, 4778.

Ma, D., Lai, Z., Ding, Q., Zhang, K., Chang, K., Li, S., Zhao, Z. and Zhong, F. (2022) Identification, Characterization and Function of Orphan Genes Among the Current Cucurbitaceae Genomes. *Front. Plant Sci.*, **13**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.872137/full [Accessed December 9, 2024].

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, **74**, 560–564.

Mayrose, I. and Lysak, M.A. (2021) The Evolution of Chromosome Numbers: Mechanistic Models and Experimental Approaches. *Genome Biology and Evolution*, **13**, evaa220.

McLeod, L., Barchi, L., Tumino, G., et al. (2023) Multi-environment association study highlights candidate genes for robust agronomic quantitative trait loci in a novel worldwide Capsicum core collection. *The Plant Journal*, **n/a**. Available at: https://doi.org/10.1111/tpj.16425 [Accessed August 31, 2023].

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A. and Johnson, E.A. (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Haeseler, A. von and Lanfear, R. (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*. Available at: https://doi.org/10.1093/molbev/msaa015.

Murphy, T.W., Hsieh, Y.-P., Zhu, B., Naler, L.B. and Lu, C. (2020) Microfluidic Platform for Next-Generation Sequencing Library Preparation with Low-Input Samples. *Anal. Chem.*, **92**, 2519–2526.

Nakano, K., Shiroma, A., Shimoji, M., et al. (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Human Cell*, **30**, 149–161.

NARUM, S.R., BUERKLE, C.A., DAVEY, J.W., MILLER, M.R. and HOHENLOHE, P.A. (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol*, **22**, 2841–2847.

Ng, S.B., Turner, E.H., Robertson, P.D., et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

Olivieri, F., Calafiore, R., Francesca, S., Schettini, C., Chiaiese, P., Rigano, M.M. and Barone, A. (2020) High-Throughput Genotyping of Resilient Tomato Landraces to Detect Candidate Genes Involved in the Response to High Temperatures. *Genes (Basel)*, **11**, 626.

Olson, N.D., Wagner, J., McDaniel, J., et al. (2022) PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics*, **2**. Available at: https://www.cell.com/cell-genomics/abstract/S2666-979X(22)00058-1 [Accessed December 9, 2024].

Ortega-Albero, N., Barchi, L., Fita, A., Díaz, M., Martínez, F., Luna-Prohens, J.-M. and Rodríguez-Burruezo, A. (2024) Genetic diversity, population structure, and phylogeny of insular Spanish pepper landraces (Capsicum annuum L.) through phenotyping and genotyping-by-sequencing. *Front. Plant Sci.*, **15**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1435427/full [Accessed November 29, 2024].

Paijmans, J.L.A., Fickel, J., Courtiol, A., Hofreiter, M. and Förster, D.W. (2016) Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources*, **16**, 42–55.

Pareek, C.S., Smoczynski, R. and Tretyn, A. (2011) Sequencing technologies and genome sequencing. *J Appl Genetics*, **52**, 413–435.

Parks, M., Cronn, R. and Liston, A. (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol*, **7**, 84.

Perrier, X. and Jacquemoud-Collet, J.P. (2006) DARwin software. Available at: https://darwin.cirad.fr/.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and 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**, e37135.

Poland, J.A., Brown, P.J., Sorrells, M.E. and Jannink, J.-L. (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE*, **7**, e32253.

Poland, J.A. and Rife, T.W. (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome Journal*, **5**, 92.

Poplin, R., Chang, P.-C., Alexander, D., et al. (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*, **36**, 983–987.

Qin, P., Lu, H., Du, H., et al. (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542-3558.e16.

Rajendran, N.R., Qureshi, N. and Pourkheirandish, M. (2022) Genotyping by Sequencing Advancements in Barley. *Front. Plant Sci.*, **13**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.931423/full [Accessed July 21, 2025].

Rakocevic, G., Semenyuk, V., Lee, W.-P., et al. (2019) Fast and accurate genomic analyses using genome graphs. *Nat Genet*, **51**, 354–362.

Rijzaani, H., Bayer, P.E., Rouard, M., Doležel, J., Batley, J. and Edwards, D. (2022) The pangenome of banana highlights differences between genera and genomes. *The Plant Genome*, **15**, e20100.

Roux, N., Toloza, A., Radecki, Z., Zapata-Arias, F.J. and Dolezel, J. (2003) Rapid detection of aneuploidy in Musa using flow cytometry. *Plant Cell Rep*, **21**, 483–490.

Ruperao, P., Thirunavukkarasu, N., Gandham, P., et al. (2021) Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain. *Front. Plant Sci.*, **12**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.666342/full [Accessed December 4, 2024].

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**, 5463–5467.

Scaglione, D., Pinosio, S., Marroni, F., et al. (2019) Single primer enrichment technology as a tool for massive genotyping: a benchmark on black poplar and maize. *Annals of Botany*. Available at: https://academic.oup.com/aob/advance-article/doi/10.1093/aob/mcz054/5424191.

Schilbert, H.M., Rempel, A. and Pucker, B. (2020) Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. *Plants*, **9**, 439.

Siadjeu, C., Pucker, B., Viehöver, P., Albach, D.C. and Weisshaar, B. (2020) High Contiguity de novo Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read Sequencing. *Genes*, **11**, 274.

Silva, G.S. and Souza, M.M. (2013) Genomic in situ hybridization in plants. *Genet Mol Res*, **12**, 2953–2965.

Šimoníková, D., Němečková, A., Čížková, J., Brown, A., Swennen, R., Doležel, J. and Hřibová, E. (2020) Chromosome Painting in Cultivated Bananas and Their Wild Relatives (Musa spp.) Reveals Differences in Chromosome Structure. *Int J Mol Sci*, **21**, 7915.

Šimoníková, D., Němečková, A., Karafiátová, M., Uwimana, B., Swennen, R., Doležel, J. and Hřibová, E. (2019) Chromosome Painting Facilitates Anchoring Reference Genome Sequence to Chromosomes In Situ and Integrated Karyotyping in Banana (Musa Spp.). *Front. Plant Sci.*, **10**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2019.01503/full [Accessed August 14, 2025].

Sohn, J. and Nam, J.-W. (2018) The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, **19**, 23–40.

Sonah, H., Bastien, M., Iquira, E., et al. (2013) An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLOS ONE*, **8**, e54603.

Suchan, T., Pitteloud, C., Gerasimova, N.S., Kostikova, A., Schmid, S., Arrigo, N., Pajkovic, M., Ronikier, M. and Alvarez, N. (2016) Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLOS ONE*, **11**, e0151651.

Tettelin, H., Masignani, V., Cieslewicz, M.J., et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13950–13955.

Tewhey, R., Warner, J.B., Nakano, M., et al. (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol*, **27**, 1025–1031.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Tian, Y., Liu, P., Zhang, X., et al. (2024) Genome-wide association study and KASP marker development for starch quality traits in wheat. *The Plant Genome*, **n/a**, e20514.

Todesco, M., Owens, G.L., Bercovich, N., et al. (2020) Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, **584**, 602–607.

Toppino, L., Barchi, L., Mercati, F., et al. (2020) A New Intra-Specific and High-Resolution Genetic Map of Eggplant Based on a RIL Population, and Location of QTLs Related to Plant Anthocyanin Pigmentation and Seed Vigour. *Genes*, **11**, 745.

Tripodi, P., Beretta, M., Peltier, D., et al. (2023) Development and application of Single Primer Enrichment Technology (SPET) SNP assay for population genomics analysis and candidate gene discovery in lettuce. *Front. Plant Sci.*, **14**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1252777/full [Accessed December 9, 2024].

Tripodi, P., Rabanus-Wallace, M.T., Barchi, L., et al. (2021) Global range expansion history of pepper (Capsicum spp.) revealed by over 10,000 genebank accessions. *PNAS*, **118**. Available at: https://www.pnas.org/content/118/34/e2104315118 [Accessed August 30, 2021].

Vasimuddin, Md., Misra, S., Li, H. and Aluru, S. (2019) Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. pp. 314–324. Available at: https://ieeexplore.ieee.org/document/8820962 [Accessed February 21, 2024].

Vrána, J., Cápal, P., Bednářová, M. and Doležel, J. (2014) Flow Cytometry in Plant Research: A Success Story. In P. Nick and Z. Opatrny, eds. *Applied Plant Cell Biology: Cellular Tools and Approaches for Plant Biotechnology*. Berlin, Heidelberg: Springer, pp. 395–430. Available at: https://doi.org/10.1007/978-3-642-41787-0_13 [Accessed July 1, 2025].

Walkowiak, S., Gao, L., Monat, C., et al. (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **588**, 277–283.

Wang, J., Hu, Z., Liao, X., et al. (2022) Whole-genome resequencing reveals signature of local adaptation and divergence in wild soybean. *Evolutionary Applications*, **15**, 1820–1833.

Wang, W., Mauleon, R., Hu, Z., et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wang, Y., Jiang, J., Zhao, L., Zhou, R., Yu, W. and Zhao, T. (2018) Application of Whole Genome Resequencing in Mapping of a Tomato Yellow Leaf Curl Virus Resistance Gene. *Sci Rep*, **8**, 9592.

Wei, T., Treuren, R. van, Liu, Xinjiang, et al. (2021) Whole-genome resequencing of 445 Lactuca accessions reveals the domestication history of cultivated lettuce. *Nat Genet*, **53**, 752–760.

Wei, X., Qiu, J., Yong, K., et al. (2021) A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat Genet*, **53**, 243–253.

Xie, S., Leung, A.W.-S., Zheng, Z., Zhang, D., Xiao, C., Luo, R., Luo, M. and Zhang, S. (2021) Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era. *The Innovation*, **2**, 100153.

Xing, X., Hu, T., Wang, Y., et al. (2024) Construction of SNP fingerprints and genetic diversity analysis of radish (Raphanus sativus L.). *Front. Plant Sci.*, **15**. Available at: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1329890/full [Accessed December 9, 2024].

yourgenome (2017) yourgenome. Available at: http://www.yourgenome.org/copyright-information/.

Zhang, Z., Cao, Y., Wang, Y., et al. (2023) Development and validation of KASP markers for resistance to Phytophthora capsici in Capsicum annuum L. *Mol Breeding*, **43**, 20.

Zhao, G., Lian, Q., Zhang, Z., et al. (2019) A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat Genet*, **51**, 1607–1615.

Zheng, C., Zhou, J., Zhang, F., Yin, J., Zhou, G., Li, Y., Chen, F. and Xie, X. (2020) OsABAR1, a novel GRAM domain-containing protein, confers drought and salt tolerance *via* an ABA-dependent pathway in rice. *Plant Physiology and Biochemistry*, **152**, 138–146.

Zhou, Y., Zhang, Zhiyang, Bao, Z., et al. (2022) Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, **606**, 527–534.

Zhou, Y., Zhao, X., Li, Y., et al. (2020) Triticum population sequencing provides insights into wheat adaptation. *Nat Genet*, **52**, 1412–1422.