

# **A pan-generic marker panel for apples to enable genetic research and breeding across *Malus* species**

Aafreen Sakina<sup>1</sup>, Richard Tegtmeier<sup>1</sup>, Hana Feulner<sup>1</sup>, David Hickok<sup>1</sup>, Anze Švara<sup>1</sup>, Patrick Cho<sup>1</sup>, Matthew Clark<sup>2</sup>, Andrzej Kilian<sup>3</sup>, Jeffrey Glaubitz<sup>4</sup>, Qi Sun<sup>4</sup>, Awais Khan<sup>\*1</sup>

<sup>1</sup>Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Geneva, NY, USA

<sup>2</sup>Department of Horticultural Science, University of Minnesota, Saint Paul, MN 55108, USA

<sup>3</sup>Diversity Arrays Technology Pty Ltd., University of Canberra, Bruce, Australian Capital Territory, Australia

<sup>4</sup>Biotechnology Resource Center, Bioinformatics Facility, Cornell University, Ithaca, NY 14853, USA.

\*Corresponding author: ([awais.khan@cornell.edu](mailto:awais.khan@cornell.edu))

## **Abstract**

Wild *Malus* species harbor untapped genetic diversity to advance apple breeding, particularly for disease resistance and stress tolerance. However, existing marker panels, developed mainly using *Malus domestica* accessions, introduce ascertainment bias and limit detecting rare variants in wild species. We developed and validated a medium-density and cost-effective pan-generic 3K apple DArTag panel optimized to capture genome-wide variation across the *Malus* genus. The panel was constructed using conserved, syntenic, and collinear genomic blocks identified within the core genome of 13 *Malus* accessions for cross-species transferability. The panel was validated across three bi-parental mapping populations totaling 593 progeny. Across these populations, 2,461–3,234 SNP markers were polymorphic and 1,482–2,620 were informative. Each population contained over 900 multiallelic micro-haplotype loci, with several hundred loci exhibiting three or four distinct haplotypes. Markers were uniformly distributed across all 17 chromosomes, each containing between 60–230 informative SNPs. The panel was further evaluated on 174 diverse germplasm accessions from 20 *Malus* species. It exhibited strong cross-species transferability, exceptionally low rates of missing data (<0.5%), and clear genetic differentiation between wild and domesticated accessions. Genome-wide association studies (GWAS) identified a major locus on chromosome 4 significantly linked to fruit length, weight, and width in addition to trait-specific associations on chromosomes 1, 6, 9, and 11. The cost-effectiveness of genotyping per sample (<\$15), combined with these results, underscore the panel's broad utility for quantitative trait locus (QTL) mapping in bi-parental and diverse populations, marker-assisted selection in the breeding programs, and genetic diversity analysis across the *Malus* genus.

**Keywords:** DArTag, pan-generic markers, population structure, diversity panel, phylogenetic analysis.

## 44 Introduction

45 Apple (*Malus domestica*) is one of the most important fruit crops grown in temperate  
46 regions worldwide (Duan et al. 2017). Apple breeding programs can greatly benefit  
47 from leveraging the extensive natural genetic diversity found in *Malus spp.* (Khan et  
48 al. 2021; Cornille et al. 2012). The wide genetic diversity found across *Malus spp.* is  
49 largely attributed to their predominantly outcrossing nature and the relatively weak  
50 domestication bottleneck (Dan et al. 2015). By broadening the genetic base beyond  
51 the limited number of elite cultivars currently in use, breeding programs can harness  
52 this diversity to enhance apple breeding (Migicovsky et al. 2021). Given the high  
53 heterozygosity, gametophytic self-incompatibility, prolonged juvenile phase, and  
54 linkage drag of unfavorable traits, a genomics-informed breeding approach can  
55 greatly improve apple breeding (Sabety et al. 2024).

56 Molecular markers are indispensable for genomics-assisted breeding and have been  
57 employed to develop genetic maps and advance the study of apple genetics (Longi et  
58 al. 2013). Previously, conventional molecular markers like Restriction Fragment  
59 Length Polymorphisms (RFLPs), Amplified Fragment Length Polymorphisms  
60 (AFLPs), Sequence Characterized Amplified Regions (SCARs), Simple Sequence  
61 Repeats (SSRs) and Diversity Arrays Technology (DArT) have been used to develop  
62 several saturated genetic maps (Hemmat et al. 1994; Conner et al. 1997; Maliepaard,  
63 1998; Liebhart 2003; Fernández-fernández et al. 2008; Schouten et al. 2012). These  
64 maps have facilitated the genomic localization of economically important traits  
65 through many studies focused on novel QTL identification and marker-assisted  
66 selection (MAS) (Maric et al. 2010). However, these marker platforms are less suited  
67 for the fine dissection of functional genetic variations mainly due to their low density  
68 across the genome (Zhang et al. 2012).

69 Recent advances in genomic technologies and the availability of high-quality  
70 reference genomes in apples have revolutionized the study of genetic variations in  
71 apples (Chu et al. 2025, Khan et al. 2022; Sun et al. 2020; Švara et al. 2024a; Sabety  
72 et al. 2024). The development of SNP arrays such as Infinium® IRSC 8K and 20K  
73 apple SNP arrays, Expressed sequence tags (EST) based apple SNP genotyping  
74 platform, 50K SNP array and Affymetrix Axiom® Apple 480K SNP array has  
75 enabled the simultaneous genotyping of thousands of SNPs across the apple genome,  
76 and they have been widely used in the development of dense SNP-based linkage maps  
77 (Antanaviciute et al. 2012; Clark et al. 2014; Rymenants et al. 2020; Khan et al. 2012),  
78 QTL mapping (Falginella et al. 2015), pedigree-based analysis, association mapping  
79 (Kumar et al. 2013), and genomic selection (Kumar et al. 2012). However, the  
80 development of these arrays primarily relied on discovery panels consisting mainly of  
81 *M. domestica* scion cultivars (Rasheed et al. 2017). For example, the 8K SNP array  
82 was developed using 27 cultivars re-sequenced at low coverage, the 20K array was  
83 based on 14 major founders of European breeding programs, and the 480K array  
84 utilized data from 67 re-sequenced accessions (Chagne et al. 2012; Bianco et al. 2014;  
85 Bianco et al. 2016). This focus on domesticated apples could introduce ascertainment  
86 bias by failing to capture the rare and unique genetic variants present in wild *Malus*  
87 species (Peace et al. 2019). The main drawbacks of SNP arrays include a bias toward  
88 variants present in the accessions of the discovery panel used for the development of  
89 the array and under-representation of alleles with extreme allele frequencies (Giebel  
90 et al. 2021). Subsequently, the use of the Infinium 20K SNP array for genotyping 178

91 *Malus sylvestris* accessions revealed that 20% of the markers exhibited more than 10%  
 92 missing values, 30% had a minor allele frequency (MAF) below 5%, and 6% were  
 93 entirely monomorphic (Buiteveld et al. 2021). Wild apple species, which are  
 94 reservoirs of novel alleles for disease resistance, abiotic stress tolerance, fruit quality,  
 95 and rootstock traits, may harbor important genetic diversity that is not represented in  
 96 domesticated varieties (Peace and Norelli 2009; Fazio et al. 2009, 2014; Duan et  
 97 al. 2017; Norelli et al. 2017). Consequently, the SNP arrays may miss valuable  
 98 genetic variation present in wild species, which can hinder efforts to breed apple  
 99 cultivars with better traits. This underscores the importance of incorporating diverse  
 100 reference panels including both domesticated and wild accessions to enable  
 101 comprehensive and accurate genetic analyses (Yang et al., 2016).

102 The availability of high-quality draft genomes of domesticated apple and its  
 103 wild relatives has provided a reference framework for developing novel genomic  
 104 resources ([Malus all species | GDR \(rosaceae.org\)](#); Chu et al. 2025; Sabety et al. 2024;  
 105 Švara et al. 2024a; Sun et al. 2020;). Advancements in next-generation sequencing  
 106 (NGS) technologies offer the development of low-cost, high-throughput, and mid-  
 107 density genetic marker screening platforms based on targeted genotyping-by-  
 108 sequencing (GBS) (Endelman et al. 2024; Sabety et al. 2024). One such example of a  
 109 targeted GBS marker platform is “DARtag” marker assays offered by Diversity  
 110 Arrays Technology Ltd. (DART; Canberra, Australia) (Hardigan et al. 2023;  
 111 Endelman et al. 2024). DARtag marker panels utilize molecular inversion probe (MIP)  
 112 genotyping to efficiently analyze large-scale single nucleotide polymorphisms (SNPs)  
 113 and small indels. The custom-designed molecular probes amplify target regions  
 114 containing SNPs, which are then sequenced using NGS. The sequenced amplicons are  
 115 processed using DART’s proprietary pipeline, enabling precise and comprehensive  
 116 identification of SNPs (Sansaloni et al. 2011; Hardigan et al. 2023). Several DARtag  
 117 panels have been developed for different crops such, as 2.5K and 4K DARtag panels  
 118 for potato (*Solanum tuberosum* L.) (Endelman et al. 2024), 3K and 5K DARtag panels  
 119 for cultivated strawberry (*F. × ananassa*) (Hardigan et al. 2023), 3K marker panel in  
 120 cultivated blueberry (*Vaccinium spp.*), 3K marker panels in alfalfa (Zhao et al. 2023),  
 121 3K DARtag v.1 panel in cranberry (*Vaccinium macrocarpon*), 3K DARtag panel v.1  
 122 in cucumber (*Cucumis sativus*), 3K DARtag panel v.2 for lettuce (*Lactuca sativa*),  
 123 DARtag panel v.1 in pecan (*Carya illinoensis*), 3K DARtag panel v.1 for *Ipomoea*  
 124 *batata* ([https://breedinginsight.org/breeding-solutions/open-source-dartag-marker-](https://breedinginsight.org/breeding-solutions/open-source-dartag-marker-panels)  
 125 [panels](https://breedinginsight.org/breeding-solutions/open-source-dartag-marker-panels)). Twelve DARtag panels are also publicly accessible via CGIAR’s EiB  
 126 platform ([https://excellenceinbreeding.org/toolbox/services/mid-density-genotyping-](https://excellenceinbreeding.org/toolbox/services/mid-density-genotyping-service)  
 127 [service](https://excellenceinbreeding.org/toolbox/services/mid-density-genotyping-service)) to support genomic selection in breeding programs of CGIAR and its national  
 128 partners.

129 Here, we report the development of a cost-effective 3K DARtag marker panel  
 130 for *Malus*, designed using a diverse reference panel comprising both domesticated  
 131 cultivars and wild accessions. The panel was validated in multiple bi-parental  
 132 mapping populations and a *Malus* diversity panel. This high-throughput, transferable  
 133 marker system enables robust QTL mapping, GWAS, linkage map construction, and  
 134 population genomics analyses across diverse genetic backgrounds.

## 135 **Material and Methods**

### 136 **1. Development of DARtag marker panel**

138

## 139 **1.1 Data preparation**

140 Publicly available whole-genome assemblies of 13 *Malus* accessions, representing  
 141 seven wild *Malus* species and five domesticated apples, were used for the  
 142 development of the DArTag marker panel. Twelve assemblies were downloaded from  
 143 the Genome Database for Rosaceae (GDR; <https://www.rosaceae.org>), while one, *M.*  
 144 *orientalis* (Khan et al. unpublished), was obtained separately. The *M. domestica*  
 145 assemblies included ‘Golden Delicious’ (GDDH13v1.1; Daccord et al. 2017),  
 146 ‘Antonovka’ (Švara et al. 2023), ‘Gala’ (Sun et al. 2020), ‘Hanfu’ (Qin et al. 2023),  
 147 and ‘Honeycrisp’ (Khan et al. 2022). The wild *Malus* species represented were *M.*  
 148 *ioensis* (Švara et al. 2024a), *M. orientalis* (Khan et al. unpublished), *M. prunifolia* (Li  
 149 et al. 2022), *M. sieversii* (Sun et al. 2020), *M. sylvestris* (Sun et al. 2020), *M.*  
 150 *coronaria* (Švara et al. 2024a), *M. fusca* (Mansfeld et al. 2023), and *M. baccata* (Chen  
 151 et al. 2019). These genomes were used for the core-genome identification and  
 152 subsequent development of the DArTag marker panel.

## 153 **1.2 Syntenic core-genome construction**

154 The assemblies were aligned to the reference genome GDDH13v1.1 (Daccord et al.  
 155 2017) using Minimap2 (version 2.17) (Li et al. 2018) with the -x asm10 alignment  
 156 option to construct the syntenic core-genome. The parameter was employed to ensure  
 157 a global alignment score greater than 400 and over 90% sequence identity with the  
 158 reference genome for cross-species whole-genome alignment. The resulting BAM  
 159 files were converted into BED format using a custom script to facilitate further  
 160 analysis. Coverage depth per site was calculated using bedtools genomecov (Aaron et  
 161 al. 2010), and coverage plots were generated by counting sites within sliding windows  
 162 of 1 Mb. The data was curated and merged using Bedtools (Aaron et al. 2010) to  
 163 refine the final output.

164

## 165 **1.3 Genus-wide variant calling**

166

167 Illumina short-read sequences were obtained for 461 *Malus* accessions, as described  
 168 in Liao et al. (2021). Sequences were downloaded from the Genome Sequence  
 169 Archive (GSA) database (Genome Sequence Archive - CNGB-NGDC) under  
 170 BioProject PRJCA004568, using the “fastq-dump” command from the SRA-Toolkit  
 171 software (version 3.1.1). The sequence reads were aligned to the GDDH13 genome  
 172 (Daccord et al. 2017) using BWA-MEM (v7.17) (Li and Durbin, 2009) with default  
 173 parameters. The resulting alignments were then converted to BAM format using  
 174 SAMtools (v1.16.1) (Li et al. 2009). Variants were called using the BCFtools package  
 175 (v1.16) (Li et al. 2011), which produced SNP predictions from the GDDH13  
 176 alignments for all 461 *Malus* accessions. SNPs within the core genomic regions were  
 177 filtered from the whole-genome sequencing data using the BCFtools package (Li et al.  
 178 2011). The SNPs were filtered into two categories: (1) high variance between wild  
 179 and domestic apples, which are informative for hybrid populations; and (2) low  $F_{st}$   
 180 and high MAF between wild and domestic apples, which are informative for general  
 181 apple populations. BCFtools (Li et al. 2011) was used to filter WGS high  $F_{st}$  SNPs by  
 182 applying a series of criteria; biallelic variants were retained using the -m2 -M2 option;  
 183 variants with more than 5% missing data were excluded across all samples using  
 184  $F_{MISSING} > 0.05$ . The data was filtered to ensure that no more than one individual  
 185 in the wild population had missing data at a given site. Finally, the  $F_{st}$  values between

186 wild and domesticated populations were calculated, and SNPs with a minimum  $F_{st}$  of  
 187 0.8% were retained to identify regions with significant population differentiation (min  
 188 Q30; min MAF 1%; bi-allelic; maxMiss 5%; maxMiss wild 1; max  $F_{st}$  WildDom0.8).  
 189 To identify WGS low $F_{st}$  high MAF SNPs filter criteria were set to min Q30; min  
 190 MAF 40 %; bi allelic; maxMiss 5%; maxMiss wild 1; max  $F_{st}$  WildDom0.1. In  
 191 addition, SNPs from the previously known 8K and 50K apple SNP arrays were  
 192 filtered using criteria of a minimum quality score of Q30 and a MAF of at least 1% to  
 193 ensure representation of variable SNPs within the core genome from these arrays. 128  
 194 QTL-linked SNPs from IRSCOA v1.0 apple SNP array were marked as “required”  
 195 (Chagne et al. 2019). A total of 18,880 markers were submitted to DArT for QC, and  
 196 104 QTL-linked SNPs from IRSCOA v1.0 apple SNP array linked to known QTLs,  
 197 which passed the design phase, were marked as “required” (Supplementary Table 1)  
 198 (Chagne et al. 2019). The DArT algorithm selected SNPs in 54 base regions with high  
 199 variant density for amplicon inserts and minimal flanking variant density for PCR  
 200 primers. 9,134 markers (65.6%) passed DArT silico test for designing suitable  
 201 flanking primers. Final marker selection was done among the “pass” SNPs for even  
 202 distribution across genome-space and gene density. Ultimately, 3100 “genomic  
 203 blocks” were selected for the final DArTag design. These regions are referred to as  
 204 DArTag markers or amplicons, and each represents a specific genomic block targeted  
 205 for amplification. For each of these 3,100 markers, a pair of primers (oligos) was  
 206 designed to amplify a short DNA segment. Within each amplified region, multiple  
 207 SNPs and micro-haplotypes are present. These SNPs and micro-haplotypes are the  
 208 actual genetic variants used for genotyping, while the DArTag marker refers to the  
 209 entire amplified region that contains them. This distinction is important throughout  
 210 the manuscript; DArTag markers refer to the amplified genomic segments (not  
 211 individual SNPs), each of which can harbor several informative variants used for  
 212 downstream analyses.

## 213 214 **2. Validation of DArTag marker panel**

### 215 **2.1 Plant material**

216 The marker panel was tested on three bi-parental mapping populations: Honeycrisp ×  
 217 Gala (n = 235), Gala × Demir (n = 179), and Gala × Chisel Jersey (n = 186), as well  
 218 as on a diversity panel comprising 174 accessions representing a broad range of  
 219 *Malus* species. These included *Malus angustifolia* (n= 10), *M. arnoldiana* (n=2), *M.*  
 220 *baccata* (n=33), *M. coronaria* (n=5), *M. domestica* (n=5), *M. ‘Evereste’* (n=1), *M.*  
 221 *florentina* (n=1), *M. floribunda* (n=9), *M. hupehensis* (n=5), *M. hybrid* (n=2), *M.*  
 222 *ioensis* (n=7), *M. micromalus* (n=12), *M. orientalis* (n=10), *M. fusca* (n=23), *M.*  
 223 *prunifolia* (n=18), *M. × robusta* (n=10), *M. sargentii* (n=11), *M. sieversii* (n=2), *M.*  
 224 *sylvestris* (n=5), and *M. zumi* (n=3). Leaf tissue for DNA extraction was collected  
 225 from actively growing shoots, placed in 96-well plates, and shipped to Diversity  
 226 Arrays Technology Pty Ltd (DArT, Canberra, Australia), where high-quality genomic  
 227 DNA was isolated following their standard protocol (Jaccoud et al. 2001).

### 228 **2.2 Genotyping and allele calling pipeline**

229 The Apple 3K DArTag marker panel is designed to call SNP variants from 54 bp  
 230 amplicons, each strategically developed to include two or more SNPs within the target  
 231 sequence. This design enables the identification of both individual SNPs and micro-  
 232 haplotypes, which are contiguous sequence blocks within a single read containing

multiple polymorphisms, including SNPs and small indels. The proprietary DArT genotyping platform provided a single SNP call per locus across the 3K marker panel, despite multiple SNPs within some amplicons (referred to as DArT markers onwards). To capture other variants and micro-haplotypes, raw amplicon reads were obtained from DArT and analyzed using a robust pipeline optimized for identifying multiple SNPs and micro-haplotypes within these amplicons ([https://bitbucket.org/cornell\\_bioinformatics/amplicon/src/master/](https://bitbucket.org/cornell_bioinformatics/amplicon/src/master/)). The repository was downloaded from GIT ([git clone https://bitbucket.org/cornell\\_bioinformatics/amplicon.git](https://bitbucket.org/cornell_bioinformatics/amplicon.git)) (Zou et al. 2020). The pipeline de-multiplexes amplicons based on PCR primer sequences and subsequently calls novel alleles from FASTQ files by splitting reads using primer information. It identifies haplotypes across the population and optionally performs PCR error correction. Haplotype variants were obtained for each marker, and the results are compiled into a hap\_genotype file, which is essentially a matrix with haplotype, genotype, and read count per allele information, and a haplotype allele fasta file which has haplotype sequence information for each allele. The haplotype file is converted to a haplotype VCF file using to\_lep\_map.pl script ([https://bitbucket.org/cornell\\_bioinformatics/amplicon/src/master/](https://bitbucket.org/cornell_bioinformatics/amplicon/src/master/)), converts the haplotype genotype data into a haplotype VCF format. This script records four haplotype alleles per marker in each bi-parental mapping population as the pseudo standard nucleotide codes (A, C, G, T). Multiple variants within each amplicon were called by aligning the haplotype sequences generated by hap\_genotype and the HaplotypeAllele.fasta file to the indexed reference genome GDDH13 (Daccord et al. 2017). The reference genome was indexed using bwa index. The resulting SNP VCF files were filtered using bcftools to retain sites with missing data < 30% and MAF > 0.01. The SNPs and haplotypes identified are referred to as “new pipeline SNP and micro-haplotype markers” in this paper.

## 2.3 Bi-parental mapping population

The marker panel was tested with three bi-parental mapping populations: Honeycrisp × Gala (n = 235), Gala × Demir (n = 179), and Gala × Chisel Jersey (n = 186). Principal component analysis (PCA) was performed to validate the robustness of the panel in assessing genetic variation among parental replicates and F<sub>1</sub> individuals. PCA was performed separately using SNP and micro-haplotype data from three different bi-parental mapping populations. The VCF file was imported into R version 4.3.2 using the vcfR package (Knaus and Grunwald, 2017) and converted into a genlight object with the vcfR2genlight function (Knaus and Grunwald, 2017). PCA was conducted on the genlight object using the glPca function from the adegenet package, retaining the first 10 principal components (PCs) (Jombart, 2008). The PCA results were visualized using ggplot2 (Wickham, 2016), with parents and F<sub>1</sub>s colored distinctly.

## 2.4 SNP filtering and informative marker selection

SNP VCF files were first filtered to retain only those variants with consistent genotypes across biological replicates of both parents. Specifically, markers were retained if each parent exhibited identical genotype calls across all replicates and had no missing genotype values. Among the consistent markers, monomorphic loci were identified as those where both parents shared the same homozygous genotype and

subsequently removed. The remaining polymorphic markers were then screened for informativeness. Markers were retained if they met one of the following conditions: 1. One parent was homozygous, and the other was heterozygous, 2. both parents were heterozygous. The resulting informative marker dataset was used to plot the chromosomal distribution of informative markers across the 17 chromosomes. Marker counts per chromosome were calculated and visualized using a bar plot to evaluate genome-wide coverage of polymorphic markers.

## 2.5 Diversity population

The marker panel was evaluated on a diverse *Malus* population comprising 174 unique genotypes and 247 total samples, including technical replicates to ensure genotyping consistency. This population represented a broad spectrum of *Malus* diversity, encompassing 20 species with varying sample sizes (Supplementary Table 2).

## 2.6 Genetic properties of populations

### 2.6.1 Principal component analysis

PCA was conducted using a total of 15,885 SNP markers (from the new pipeline), which were previously filtered to remove insertions and deletions (INDELs) using TASSEL v5 (Bradbury et al. 2007). The resulting VCF file was imported into R (version 4.3.2) using the vcfR package (Knaus and Grunwald, 2017). SNP data were read using the read.vcfR function and subsequently converted into a genlight object via the vcfR2genlight function from the adegenet package (Jombart 2008) for multivariate analysis. PCA was performed using the glPca function (Jombart 2008), retaining the first two PCs for downstream visualization and interpretation. The PCA results were visualized using the ggplot2 package (Wickham, 2016). Samples were color-coded by species, with distinct species-specific palettes generated using the randomcoloR packages (Ammar 2006).

### 2.6.2 Phylogenetic analysis

To investigate genetic relationships among the *Malus* accessions, a phylogenetic tree was constructed using SNP data (from the new pipeline) in R (version 4.3.2). 16,002 SNPs were used, and missing data were handled by excluding loci with over 50% missing values. The cleaned matrix was converted into a numeric matrix and subsequently into a genlight object using the dartR package (Gruber et al. 2018), with ploidy set to 2. Monomorphic and all-missing loci were filtered out using gl.filter.monomorphs and gl.filter.allna. A neighbor-joining (NJ) tree was computed based on bitwise genetic distances using the bitwise.dist function from poppr (Kamvar et al. 2015), and the NJ algorithm implemented in the ape package (Paradis et al. 2004). Bootstrap support values were estimated using 1,000 replicates with a custom bootstrap function that resamples SNP columns and reconstructs NJ trees. Node support values were assigned using the boot.phylo function. The final NJ tree with bootstrap support was visualized in R and saved in Newick format using write.tree from the ape package (Paradis et al. 2004). The Newick file was then imported into the Interactive Tree of Life (iTOL) platform (<https://itol.embl.de/>) for enhanced

327 visualization, including taxon-specific color coding and annotations (Letunic and  
328 Bork, 2021).

### 329 **2.6.3 Population structure analysis**

330 The population genetic structure of *Malus* accessions was assessed using the model-  
331 based Bayesian clustering algorithm implemented in STRUCTURE v2.3.4 (Pritchard  
332 et al. 2000). This method uses a Markov Chain Monte Carlo (MCMC) approach to  
333 estimate the proportion of each individual's genome derived from  $K$  ancestral  
334 populations, aiming to minimize deviations from Hardy–Weinberg equilibrium and  
335 linkage disequilibrium within inferred clusters. Analyses were conducted using the  
336 admixture model with correlated allele frequencies. Five independent replicates were  
337 run for each value of  $K$  (ranging from 2 to 12), each with a burn-in period of 10,000  
338 iterations followed by 50,000 MCMC iterations to ensure convergence and model  
339 stability. The results were processed using CLUMPP v1.1.2 to align cluster  
340 assignments across replicate runs and identify consistent clustering patterns  
341 (Jakobsson and Rosenberg 2007). The aligned outputs were also visualized using  
342 DISTRUCT v1.1 to display individual admixture proportions (Rosenberg,  
343 2004). Final plots and summary statistics were generated using the Pophelper R  
344 package (Francis, 2017), facilitating visualization and interpretation of population  
345 structure across varying  $K$  values (Francis et al. 2017).

346

### 347 **2.7 LD decay**

348

349 Linkage disequilibrium (LD) decay across the genome was analyzed using PLINK  
350 v1.9 (Chang et al. 2015). SNPs were filtered to retain variants with a MAF  $\geq 0.01$ , a  
351 missing genotype rate per SNP  $\leq 10\%$ , and a missing genotype rate per individual  $\leq$   
352  $50\%$ . LD decay was computed using pairwise squared correlation coefficients ( $r^2$ )  
353 within a 50 kb sliding window across all chromosomes. The following PLINK options  
354 were used to define the LD window: `--ld-window 99999`: no limit on the number of  
355 variant pairs considered, `--ld-window-kb 50`: restricts comparisons to variant pairs  
356 within 50 kilobases, `--ld-window-r2 0`: includes all SNP pairs regardless of their  $r^2$   
357 value. The final dataset was then converted to PLINK binary format (.bed, .bim, .fam)  
358 using `--make-bed` for downstream analyses. The LD decay pattern was visualized in R  
359 (v4.3.2) using the ggplot2 package (Wickham, 2016). The absolute genomic distance  
360 between SNP pairs was computed, and LOESS smoothing was applied to estimate the  
361 decay of LD ( $r^2$ ) with increasing physical distance. The half-LD decay distance was  
362 calculated as the physical distance at which the fitted LOESS curve dropped to half  
363 the maximum observed  $r^2$  value. Multiple visualizations were generated using raw and  
364 smoothed LD values to represent decay patterns at different resolutions (base pairs  
365 and kilobases).

### 366 **2.8 Marker distribution and coverage**

367 The SNP marker distribution across chromosomes was evaluated using the .bim file  
368 generated by PLINK v1.9 (Chang et al. 2015). SNP positions were binned into 1 Mb  
369 windows, and the number of markers per bin was summarized using dplyr and tidyr  
370 (Wickham et al. 2023; Wickham et al. 2024). SNP coverage was displayed as point  
371 plots, showing exact marker locations and densities across chromosomes. A

372 chromosome-wide summary of LD was performed by grouping pairwise SNP  
373 comparisons based on chromosome identity.

## 374 **2.9 Genome wide association study**

375  
376 Phenotypic data for fruit size traits were obtained from the publicly accessible USDA  
377 National Plant Germplasm System (NPGS) *via* the GRIN-Global database  
378 (<https://npgsweb.ars-grin.gov/gringlobal/search?q=APPLE>). Three quantitative fruit  
379 traits were analyzed; fruit length, fruit width, and fruit weight for 169 accessions.  
380 Measurements were recorded as part of standard germplasm evaluation protocols.  
381 Fruit length and *fruit width* were recorded as the average of 10 mature fruits collected  
382 from a vigorous tree, measured in millimeters (mm). Fruit weight was expressed as  
383 the mean weight (in grams) of 10 mature fruits at physiological maturity. The SNP  
384 VCF file was filtered to retain only SNPs by removing indels using bcftools (Li et al.  
385 2011). GWAS was conducted for fruit quality traits using the General Linear Model  
386 (GLM) and the Mixed Linear Model (MLM), both accounting for population  
387 structure using the first five PCs and Q matrix. The MLM additionally accounted for  
388 kinship by incorporating (K), a kinship matrix, accounting for relatedness. Analysis  
389 was performed using the Genomic Association and Prediction Integrated Tool  
390 (GAPIT) package version 3.5.0 (Wang et al. 2021).

391

## 392 **Results**

393

### 394 **1. Development of DArTag marker panel**

395

#### 396 **1.1 Core genome construction**

397 Thirteen *Malus* genomes representing phylogenetically and geographically distinct  
398 species were selected to construct the core genome. These assemblies collectively  
399 span the genetic diversity across the *Malus* genus and form the foundation for  
400 identifying conserved genomic regions in the core genome. While the ideal approach  
401 would involve designing markers exclusively from genomic regions shared across all  
402 13 genomes, this was not feasible. In practice, only 2.1% of the reference genome,  
403 primarily encompassing 80.1% of annotated protein-coding gene regions, was  
404 conserved across all assemblies. A hierarchical strategy was employed to address  
405 potential gaps in marker coverage. Markers were first placed in regions present in  $\geq 12$   
406 genomes, ensuring a high-confidence representation of conserved loci. Subsequently,  
407 remaining gaps were filled using regions found in  $\geq 8$  genomes, maximizing genomic  
408 coverage while maintaining cross-genome relevance. This approach resulted in  
409 marker coverage of approximately 35 Mb out of the 625 Mb reference genome.  
410 Among the selected genomes, Hanfu and Honeycrisp exhibited the highest genome  
411 coverage, each with 99.50%, followed closely by *M. sieversii* (99.30%), Gala  
412 (99.20%), and *M. sylvestris* (98.50%). Antonovka also showed high coverage at  
413 98.40%. Wild North American species such as *M. fusca* (95.80%), *M. orientalis*  
414 (95.10%), and *M. coronaria* (94.00%) demonstrated moderately high coverage, while  
415 *M. ioensis* had a slightly lower value at 92.60%. *M. baccata* and *M. prunifolia* showed  
416 the lowest genome coverage among the assemblies, with 89.60% and 84.60%,  
417 respectively.

418

#### 419 **1.2 Reference genome coverage across multiple assemblies**

To evaluate the extent of conserved genomic regions across *Malus* assemblies, we calculated reference genome coverage based on presence across 12, 11, and 10 assemblies, respectively (Table 1). When considering regions conserved in  $\geq 12$  assemblies, only 2.1% of the 625 Mb reference genome was covered, corresponding to 13.2 Mb of sequence. These regions, however, captured 80.1% of annotated protein-coding gene space, indicating a strong representation of functionally important loci. Chromosome-level coverage ranged from 0.7% (Chr01) to 2.7% (Chr07 and Chr10), with the highest average marker density observed on Chr04 (377 bp) and Chr14 (397 bp) and no coverage on Chr00 (unplaced contig). Expanding the inclusion threshold to regions conserved in 11 assemblies resulted in 5.6-fold greater coverage, spanning 34.8 Mb or 5.6% of the reference genome. This set included 76.5% of protein-coding gene regions, suggesting a modest trade-off between broader genomic coverage and gene content. Per-chromosome coverage increased substantially from 2.9% (Chr01) to 6.8% (Chr15). Fragment spacing also increased slightly, averaging between 431 bp (Chr00) and 522 bp (Chr13). Further relaxing the threshold to include regions present in  $\geq 10$  assemblies yielded the most extensive coverage: 55.0 Mb, representing 8.8% of the reference genome. However, the proportion of gene space captured dropped to 69.1%, reflecting the inclusion of more variable, non-genic regions. Chromosome-level coverage reached up to 12.7% (Chr15), with average fragment sizes ranging from 534 bp (Chr01) to 628 bp (Chr06) (Figure 1).

**Table 2.** Chromosome-wise reference genome coverage by conserved regions across 12 *Malus* genome assemblies. This table summarizes the extent of conserved genomic regions mapped onto each chromosome of the reference genome. “Length” refers to the total physical length (in base pairs) of each chromosome. “Coverage percentage (12 genomes),” “Coverage percentage (11 genomes),” and “Coverage percentage (10 genomes)” indicate the proportion of each chromosome covered by regions conserved across all 12 genomes, at least 11 genomes, and at least 10 genomes, respectively. “Average fragment” values represent the average length (in base pairs) of continuous conserved regions detected at each coverage threshold.

| Chr   | Length (bp) | Coverage percentage |            |            | Average fragment |            |            |
|-------|-------------|---------------------|------------|------------|------------------|------------|------------|
|       |             | 12 genomes          | 11 genomes | 10 genomes | 12 genomes       | 11 genomes | 10 genomes |
| Chr00 | 52728359    | 0.0%                | 0.0%       | 0.0%       | 0                | 431        | 564        |
| Chr01 | 32625452    | 0.7%                | 2.9%       | 6.8%       | 367              | 434        | 534        |
| Chr02 | 37577729    | 1.2%                | 3.1%       | 6.7%       | 369              | 480        | 552        |
| Chr03 | 37524076    | 1.6%                | 4.8%       | 9.4%       | 370              | 471        | 570        |
| Chr04 | 32301874    | 2.6%                | 6.1%       | 11.2%      | 377              | 490        | 607        |
| Chr05 | 47952461    | 1.8%                | 4.8%       | 8.9%       | 340              | 463        | 590        |
| Chr06 | 37137259    | 2.1%                | 5.3%       | 9.2%       | 369              | 501        | 628        |
| Chr07 | 36691129    | 2.7%                | 6.3%       | 11.1%      | 371              | 496        | 591        |
| Chr08 | 31609270    | 1.5%                | 5.0%       | 9.5%       | 357              | 451        | 574        |
| Chr09 | 37604908    | 2.2%                | 5.9%       | 10.6%      | 359              | 476        | 602        |
| Chr10 | 41762413    | 2.7%                | 6.4%       | 11.5%      | 354              | 489        | 598        |

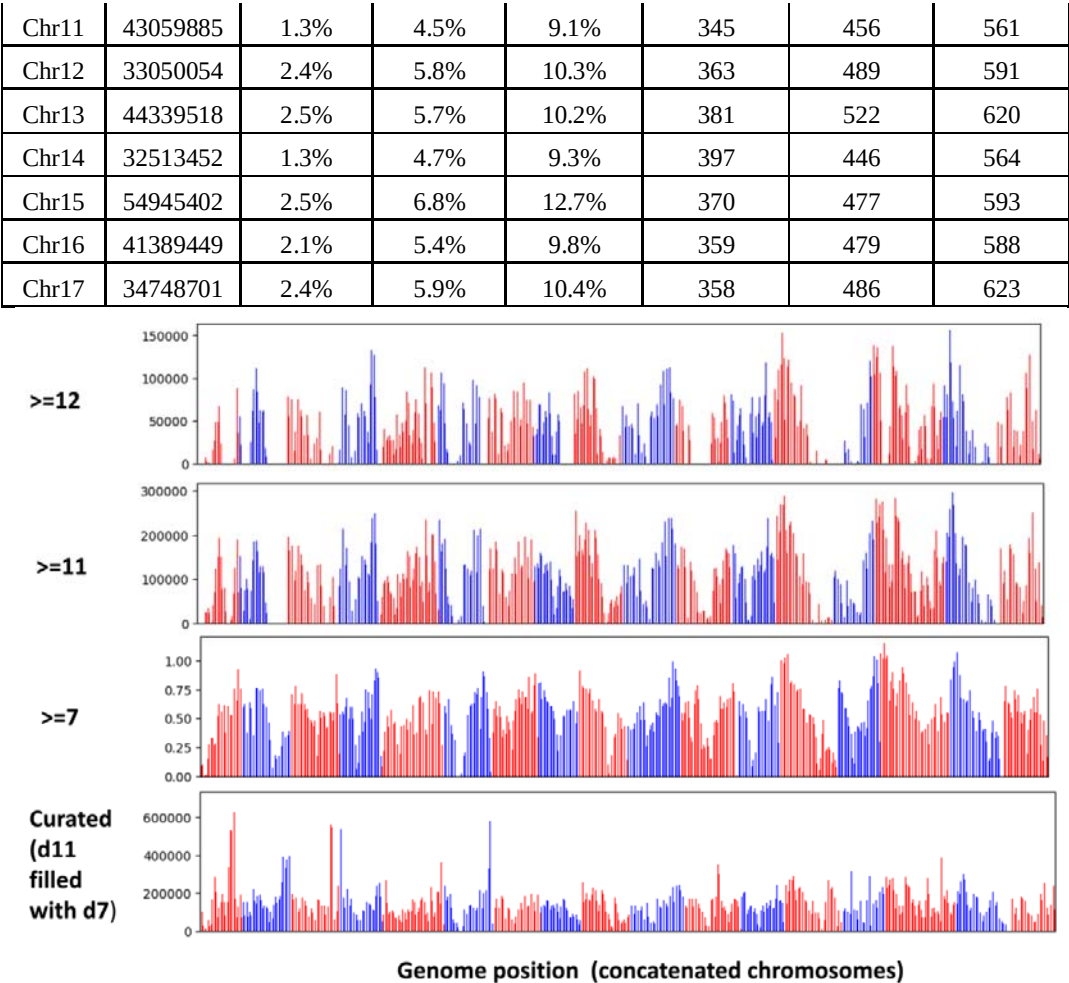


**Figure 1.** Genomic distribution of conserved regions across multiple *Malus* genome assemblies used to design DArTag marker panel. Bar plots show the genomic distribution of regions conserved across *Malus* assemblies using three thresholds: regions present in  $\geq 12$  assemblies,  $\geq 11$  assemblies, and  $\geq 7$  assemblies, respectively. The bottom panel represents a curated marker panel, where regions present in  $\geq 11$  assemblies were supplemented with regions from  $\geq 7$  assemblies to fill gaps. Red and blue bar denote alternating chromosomes for visual clarity.

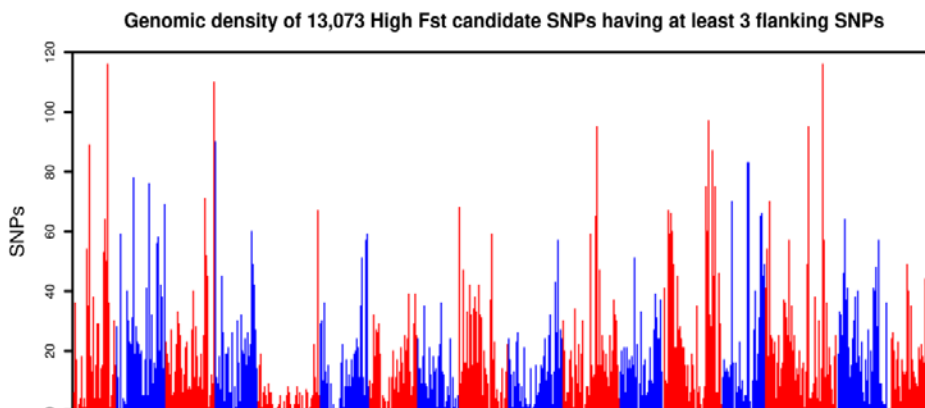
### 1.3 Marker filtering and DArTag panel design

The DArTag marker panel was developed through a two-phase marker selection process designed to optimize genomic coverage. In phase one, the DArT algorithm was employed to identify SNP-rich 54 bp regions suitable for amplicon design, while minimizing flanking sequence variation to ensure primer efficiency. A total of 13,919 candidate SNP loci were submitted for *in silico* evaluation. Additionally, 104 SNPs from the IRSCOA v1.0 apple SNP array, each linked to known QTLs, were prioritized for inclusion based on their association with key disease resistance and

470 fruit quality traits (Chagné et al. 2019). These include major resistance loci for apple  
471 scab (*Rvi6*, *Rvi11*, *Rvi2*, *Rvi4*, *Rvi15/Vr2*, *Rvi3*, *Rvi12*), fire blight (*RLP1*, *MR5*, *FBE*),  
472 woolly apple aphid (*Ermis*), and powdery mildew (*Pl2*). SNPs related to fruit quality  
473 traits represented by associated genes include fructose content (*LG1Fru*), esters  
474 (*MdAAT1*), acidity (*Acidity-LG8*, *Ma1*), red skin color (*MYB10*), fruit firmness  
475 (*MdPG1*, *LG15-BB*), vitamin C (*GGP3*), ethylene (*MdACO1*, *MdACS1*), bitter pit  
476 (*LAR/Ma*, *Bp13*), polyphenols and bitter pit (*LAR1*), and additional traits such as type  
477 2 red flesh (*MYB110*) and maturity time (*Maturity-LG3*) were also included (Chagné  
478 et al. 2019). Of all submitted markers, 9,134 (65.6%) passed the *in silico* design filters.  
479 In phase two, marker selection was refined to ensure even representation across the  
480 genome based on gene density. From the passing markers, 3,100 genomic blocks,  
481 each representing approximately equal numbers of genes, were selected to ensure  
482 balanced coverage. All 104 QTL-linked SNPs were retained in the final panel.

483 Table 2 summarizes the filtering statistics from different SNP sources used in panel  
484 development. Of particular interest were the 13,073 high  $F_{st}$  SNPs and 3,105 low  $F_{st}$ ,  
485 high MAF SNPs, which had sufficient flanking SNP support ( $\geq 3$  and  $\geq 2$  flanking  
486 SNPs, respectively) and passed conserved primer design checks. These sets are  
487 visualized in Figure 2, displaying their genomic density across the *Malus* reference  
488 genome. The total number of SNPs in the final design is described in Table 3.

**a.**



**b.**

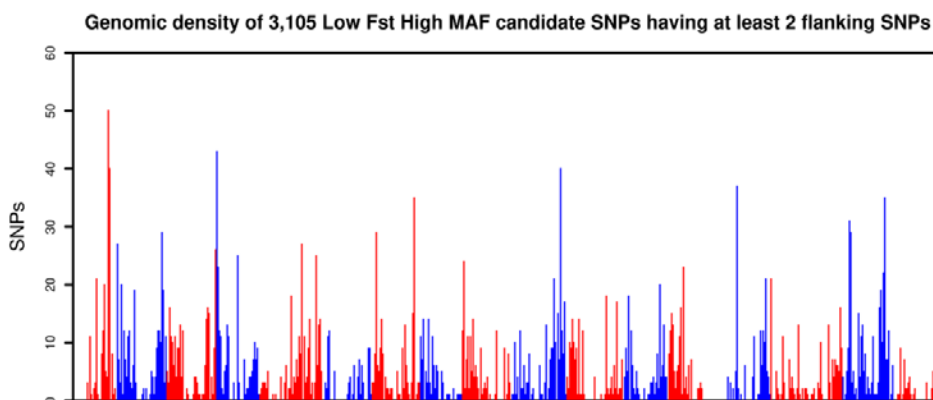


**Figure 2.** Genomic distribution of candidate SNPs across the apple genome a) Genomic density of 13,073 high-  $F_{st}$  candidate SNPs, each flanked by at least three neighboring SNPs, indicating genomic regions with strong population differentiation. b) Genomic density of 3,105 low-  $F_{st}$ , high-MAF candidate SNPs, each supported by at least two flanking SNPs, selected for their broad allele representation across diverse apple accessions. The x-axis represents genomic positions (in megabases, Mb), concatenated across all chromosomes of the apple reference genome. The y-axis shows the number of candidate SNPs within each genomic window, reflecting local marker density and distribution. Alternating colors along the x-axis indicate transitions between chromosomes, allowing chromosome boundaries to be visually distinguished across the concatenated genome.

**Table 2.** Summary of SNP marker selection and filtering across different sources for DArTag panel development during phase 1. The table outlines the number of SNPs at each stage of the filtering pipeline across five sources: whole-genome sequencing (WGS) high-  $F_{st}$  and low-  $F_{st}$ , high-MAF candidates, the 9K and 50K apple arrays, and a curated set of 128 QTL-linked SNPs.

| SNP Source           | WGS High $F_{st}$ | WGS Low $F_{st}$ HighMAF | 9K    | 50K    | 128 QTL-linked |
|----------------------|-------------------|--------------------------|-------|--------|----------------|
| Total original SNPs  | 82,310,931        | 82,310,931               | 7,399 | 52,638 | 114            |
| Pass Filter          | 60, 291           | 14,596                   | 1,056 | 6,057  | 104            |
| minFlankSNPs         | 3                 | 2                        | 2     | 2      | 0              |
| conservedPrimers     | 13,073            | 3,105                    | 361   | 2,238  | NA             |
| ReadyForDArT         | 13,073            | 3,105                    | 361   | 2,238  | 104            |
| Total ready for DArT | 18,880            |                          |       |        |                |

**Table 3.** Classification of SNPs included in the final DArTag panel design. This table summarizes the composition of the final SNP set ( $n = 3,100$ ), categorized by source and selection criteria. The panel integrates 104 QTL-linked SNPs without filtering, SNPs from legacy apple arrays (9K and 50K) located in core genomic regions, and SNPs chosen based on population differentiation and allele frequency metrics.

| SNPs in final design | Category                                                       | Category description                |
|----------------------|----------------------------------------------------------------|-------------------------------------|
| 104                  | 128 Apple QTL SNPs (No Filter)                                 | QTL SNPs                            |
| 945                  | Apple 50K in core genomic regions                              | 50K SNP Chip                        |
| 271                  | Apple 9K in core genomic regions                               | 9K SNP Chip                         |
| 1406                 | High variation between wild species and <i>Malus domestica</i> | Between wild & domesticated species |
| 378                  | Low $F_{st}$ DomVsWild HighMAF                                 | Generally good with high MAF        |

## 2. Validation of DArTag marker panel

### 2.1 Bi-parental $F_1$ mapping populations

The newly developed DArTag marker panel was validated on the Honeycrisp  $\times$  Gala bi-parental population, consisting of 235 individuals (Figure 3). A total of 3,320 SNPs obtained from the new SNP pipeline were analyzed. Among these, 3,137 were polymorphic and 2,841 were found to be informative, while 183 markers showed genotype inconsistencies between parental replicates. The genomic distribution of

SNPs revealed even coverage across the 17 apple chromosomes. Chromosome-wise marker distribution of informative markers ranged from 112 markers on Chr13 to 230 markers on Chr15, with the remaining chromosomes containing between 115 and 212 markers. Among 3,097 haplotype markers, 829 loci were identified as multiallelic. Specifically, 612 loci exhibited three haplotypes, and 217 loci displayed four distinct haplotypes, indicating a high level of allelic diversity. The DArTag markers included 3,099 SNPs, with 13 markers differing between replicates. Of these, 1,676 SNPs were polymorphic and 1,482 were informative.

The Chisel Jersey  $\times$  Gala bi-parental population ( $n = 186$ ) was genotyped using the SNP panel comprising 3,097 markers (Figure 3). A total of 3,099 DArTag markers were initially analyzed across 186 progeny. Of these, 11 markers showed genotype discrepancies between technical parental replicates, and 1,824 markers were polymorphic, with 1,560 classified as informative. Using the refined set of 3,097 markers from the new pipeline, 171 were found to differ between parental replicates. The refined dataset yielded 2,461 polymorphic and 1,941 informative markers. A total of 1,513 informative SNPs from this population were successfully placed across the 17 apple chromosomes. Chromosome 7 contained the highest number of markers (132), while Chromosome 1 had the fewest (60). Marker counts for the remaining chromosomes ranged from 60 to 130. Among these, 904 loci were identified as multiallelic and subsequently excluded from PCA. Haplotype-based analysis revealed that, of the multiallelic loci, 632 loci exhibited three distinct micro-haplotypes, while 272 loci had four micro-haplotypes, indicating high allelic diversity in the panel. Notably, one parental sample (Chisel Jersey\_002) appeared to cluster incorrectly with progeny, suggesting potential sample mislabeling or contamination.

The Demir  $\times$  Gala bi-parental population was evaluated using both the pipeline developed by Diversity Arrays Technology (DArT) and new SNP pipelines (Figure 3). The newly developed SNP panel included 3,514 markers, of which seven loci exhibited more than two alleles and were excluded from further biallelic analysis. The population consisted of 172 individuals. Haplotype analysis using a set of 3,097 markers revealed 940 multiallelic loci. Of these, 680 loci exhibited three distinct haplotypes, and 260 loci exhibited four haplotypes, reflecting substantial haplotypic diversity in the panel. In the DArT panel pipeline, 3,099 SNP markers were used on a population of 185 individuals. Among these, 30 markers showed discrepancies between replicates, 1,726 were polymorphic, and 1,336 were informative. The updated pipeline showed significant improvements: among the 3,514 markers evaluated, only 111 exhibited genotype differences between technical replicates. Furthermore, 3,234 markers were polymorphic and 2,620 were informative. A total of 2,407 informative markers were successfully mapped across the 17 apple chromosomes. Chromosome 15 had the highest number of markers (214), while Chromosome 13 had the fewest (94). The remaining chromosomes had marker counts ranging from 102 to 173.

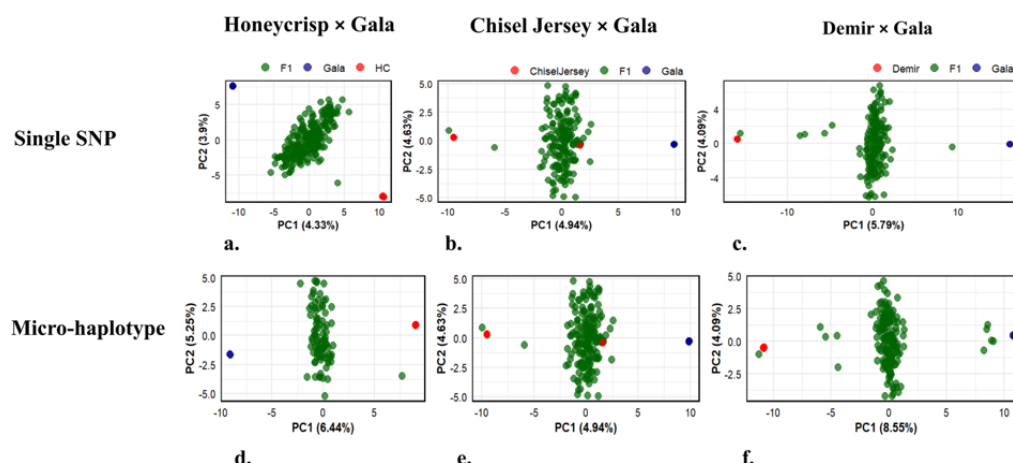


**Figure 3.** Principal component analysis of three bi-parental mapping apple populations based on new pipeline SNP (upper panel) and micro-haplotype markers (lower panel) (a and d) Honeycrisp × Gala population (n = 235), (b and e) Chisel Jersey × Gala population (n = 186), and (c and f) Demir × Gala population (n = 172). Each dot represents an individual. F<sub>1</sub> progeny are shown in green, while parental genotypes are colored separately: red for Honeycrisp (a, d), Chisel Jersey (b, c), and Demir (e, f); and blue for Gala.

## 2.2 Genetic properties of diversity panel

### 2.2.1 Principal component analysis

PCA was performed on SNP genotype data from multiple *Malus* species to assess genetic relationships. The first two PCs, PC1 and PC2, explained 23.23% and 17.37% of the total genetic variance, respectively, and were used to visualize genetic clustering among accessions (Figure 4). Species-level grouping was supported by 95% confidence ellipses, which delineate the distribution of each species in PC space. North American species *M. angustifolia*, *M. ioensis*, and *M. coronaria* formed a tight cluster, distinct from *M. fusca*, indicating divergent genetic backgrounds. Interestingly, two accessions of *M. fusca* clustered apart from the main *M. fusca* group, while *M. florentina* was positioned closely to *M. fusca*, suggesting potential genetic relatedness. One accession of *M. ioensis* also clustered separately, highlighting intraspecific divergence. European and Central Asian wild species *M. sieversii*, *M. orientalis*, and *M. sylvestris* clustered near *M. domestica*, reflecting their known contributions to the domesticated apple gene pool. Several accessions of *M. robusta* and *M. sargentii* occupied intermediate positions or showed partial overlap with *M. baccata*, potentially indicating shared ancestry or historical admixture events. Overall, the clustering pattern underscores clear genetic differentiation between North American and Eurasian wild *Malus* species and supports evidence of divergence and gene flow among wild and cultivated lineages (Figure 4).

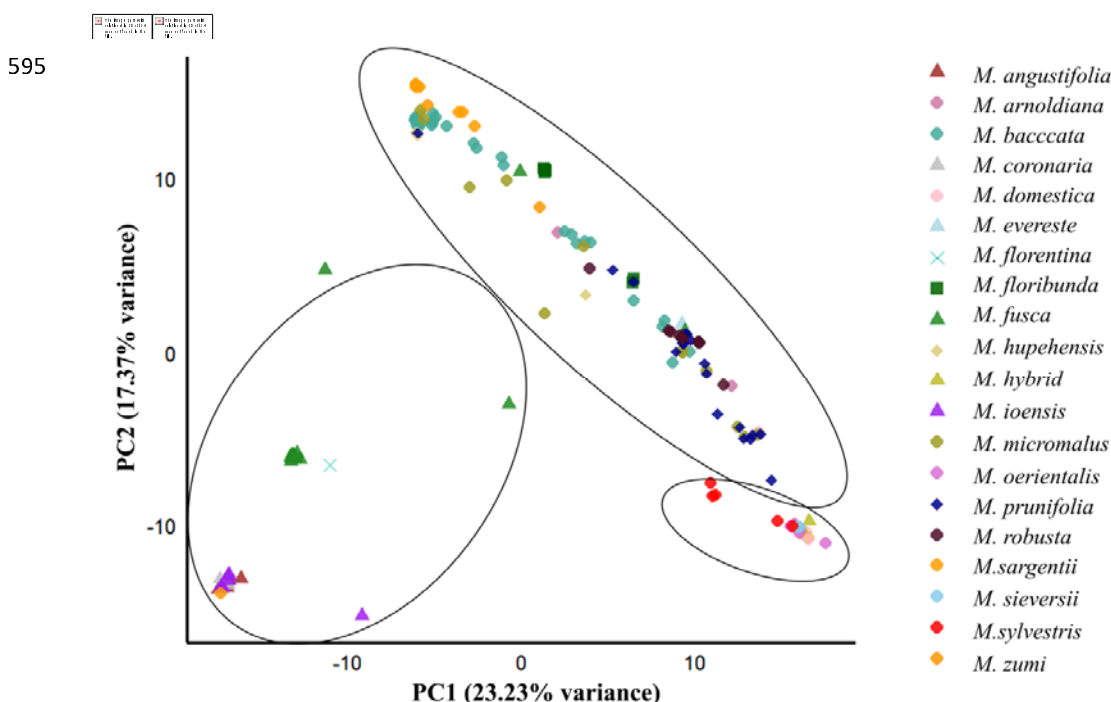


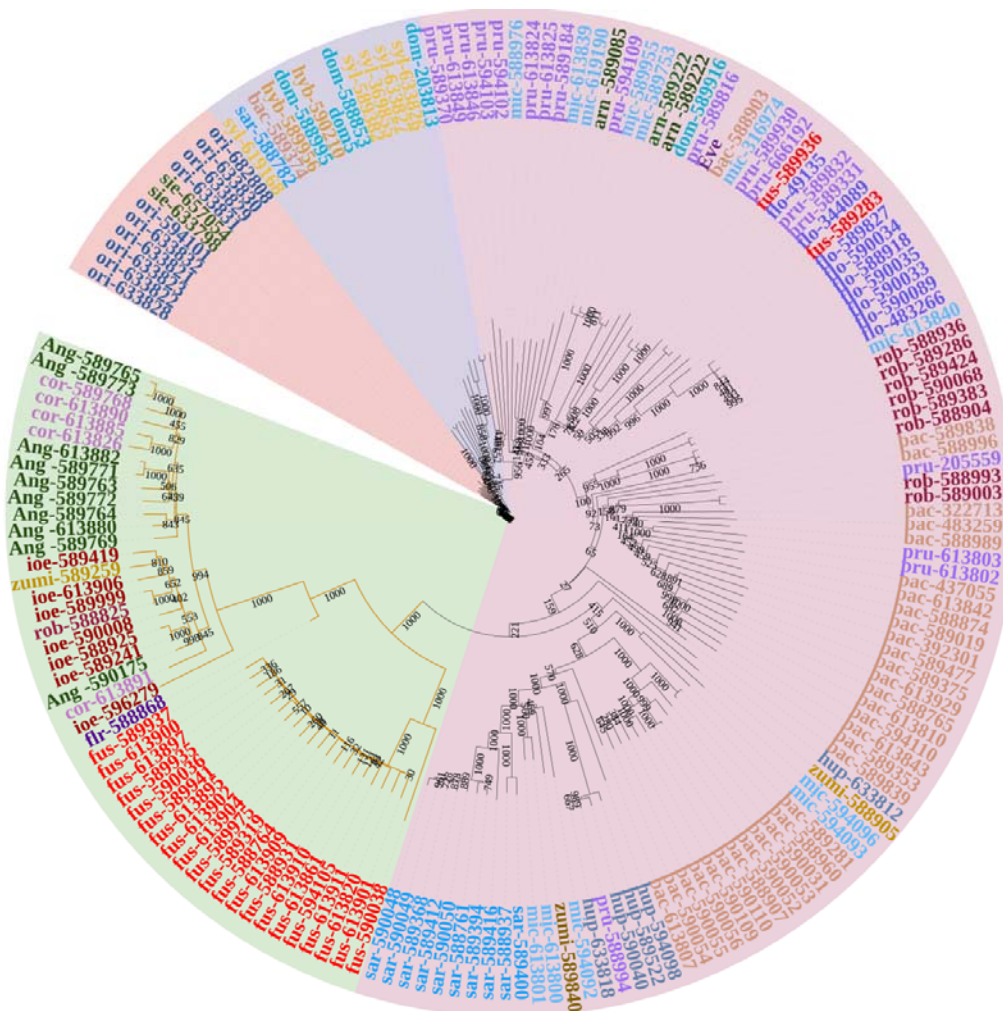
**Figure 4.** PCA of diverse *Malus* species based on genome-wide SNP data. Each point represents an individual accession, colored and shaped by species identity. The first two principal components (PC1 and PC2) explain 23.235 and 17.37% of the total genetic variance, respectively, and capture the major axes of genetic variation. Distinct clustering is observed for North American (*M. angustifolia*, *M. ioensis*, *M. coronaria*, *M. fusca*), Eurasian species (*M. sieversii*, *M. orientalis*, *M. sylvestris*) and East Asian *Malus* species. To improve visual distinction among taxa, selected species including *M. angustifolia*, *M. fusca*, *M. coronaria*, and *M. ioensis* were assigned unique point shapes (triangles), while remaining species were displayed using standard plotting symbols such as circles or squares. Solid outlines represent 95% confidence ellipses for each species, indicating the expected distribution of accessions around the species centroid in PC space.

## 2.2.2 Phylogenetic analysis

A maximum likelihood phylogenetic tree was constructed to infer the evolutionary relationships among diverse *Malus* accessions (Figure 5). The tree revealed distinct clades corresponding to species-level groupings, with strong bootstrap support for major nodes. North American species such as *M. angustifolia*, *M. ioensis*, and *M. coronaria* formed a well-supported monophyletic clade, indicating a shared genetic background. *M. fusca* also formed a distinct group, though some accessions appeared slightly divergent within the clade. *M. florentina* clustered near *M. fusca*, consistent with PCA results. Eurasian wild species, *M. sieversii*, *M. sylvestris*, and *M. orientalis* were grouped closely with *M. domestica*, reflecting their documented contribution to the domesticated apple genome. Several *M. baccata* accessions were interspersed with *M. robusta* and *M. sargentii*, suggesting admixture. The clustering pattern supports clear phylogenetic separation between North American and Eurasian wild species,

while also highlighting areas of potential introgression or shared ancestry, particularly among East Asian species and cultivated apples.

625



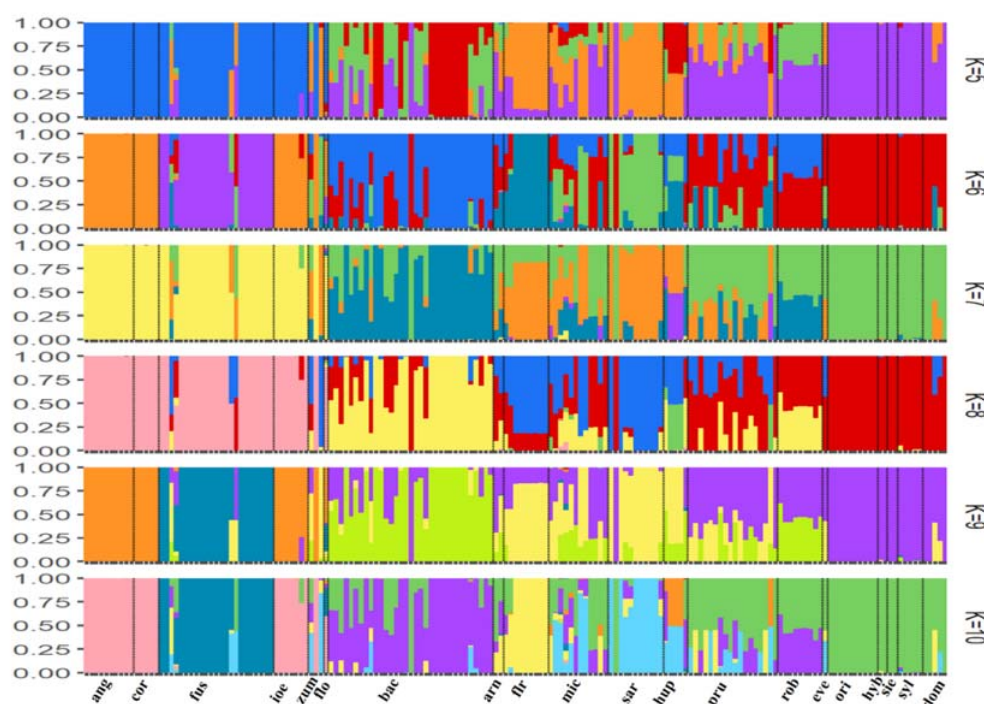
626

**Figure 5:** Maximum likelihood phylogenetic tree of diverse *Malus* accessions based on genome-wide SNP data, with accessions color-coded by species: North American species (green), East Asian species (lavender), followed by *Malus sylvestris* from Europe and domesticated *M. domestica* (light blue), and finally *M. sieversii* and *M. orientalis* from Central Asia (peach); bootstrap support values are shown at major nodes. Abbreviations: Ang = *M. angustifolia*, cor = *M. coronaria*, fus = *M. fusca*, ioe = *M. ioensis*, zumi = *M. zumi*, flo = *M. florentina*, bac = *M. baccata*, arn = *M. arnoldiana*, flr = *M. floribunda*, mic = *M. micromalus*, sar = *M. sargentii*, hup = *M. hupehensis*, pru = *M. prunifolia*, rob = *M. robusta*, eve = *M. evereste*, ori = *M. orientalis*, hyb = *M. hyb*, sie = *M. sieversii*, syl = *M. sylvestris*, dom = *M. domestica* and Dom5 = T1190.

## 2.4 Population structure analysis

638

Structure analysis, supported by the Evanno  $\Delta K$  method (Supplementary figure 1), revealed nine distinct populations best represent the ancestry of the *Malus* accessions under study. *M. ioensis*, *M. coronaria*, and *M. angustifolia* share a common ancestral background, indicating close genetic relatedness among these North American species. *M. fusca* exhibits a distinct ancestral background, separate from other North American *Malus* species. Two *M. fusca* accessions that did not cluster with the main *fusca* group, did not show the same ancestral background as other *fusca* accessions, instead they, show evidence of admixture with *M. sargentii* and *M. sieverssii*. *M. baccata* represents a separate ancestral population, although many of its accessions exhibit admixture, likely resulting from gene flow with other wild *Malus* species. *M. floribunda* also shows signs of admixture, particularly from other East Asian *Malus* species (Figure 6).

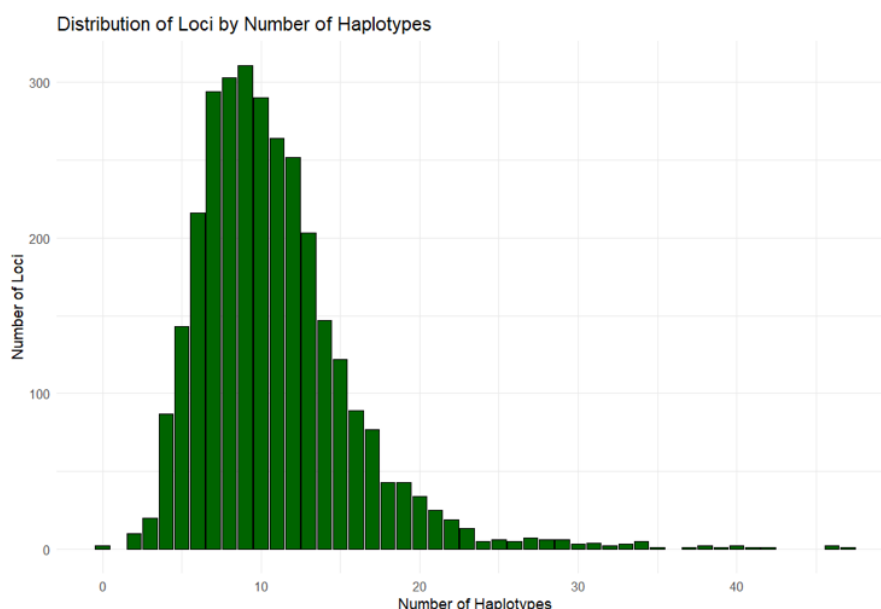


651

**Figure 6:** Population structure analysis of *Malus* accessions inferred using STRUCTURE for  $K = 5$  to  $K = 10$ . Each vertical bar represents an individual accession, and the colors indicate estimated membership proportions in inferred ancestral populations. Species codes are shown along the x-axis. As  $K$  increases, substructure within species becomes more apparent. Abbreviations: ang = *M. angustifolia*, cor = *M. coronaria*, fus = *M. fusca*, ioe = *M. ioensis*, zum = *M. zumi*, flo = *M. florentina*, bac = *M. baccata*, arn = *M. arnoldiana*, flr = *M. floribunda*, mic = *M. micromalus*, sar = *M. sargentii*, hup = *M. hupehensis*, pru = *M. prunifolia*, rob = *M. robusta*, eve = *M. evereste*, ori = *M. orientalis*, hyb = *M. hyb*, sic = *M. sieverssii*, syl = *M. sylvestris* and dom = *M. domestica*.

Across the *Malus* diversity population, micro-haplotype analysis revealed substantial variation in the number of haplotypes per locus. Most loci exhibited moderate haplotype diversity, with the number of distinct haplotypes per locus ranging from 1 to over 40. The distribution was right-skewed, with a peak around 9 haplotypes,

666 indicating that the majority of loci (over 300) contained 8–10 haplotypes. A smaller  
667 proportion of loci showed high haplotype richness (>20 haplotypes), suggesting  
668 regions of elevated allelic diversity or recombination. These patterns reflect a high  
669 level of genetic diversity within the panel and support the utility of multi-allelic loci  
670 for resolving fine-scale population structure and evolutionary relationships in *Malus*  
671 (Figure 7).



672

673 **Figure 7:** Distribution of loci by number of haplotypes in the *Malus* diversity panel,  
674 showing the frequency of loci with varying levels of haplotype diversity across  
675 accessions; the x-axis represents the number of distinct haplotypes per locus, while  
676 the y-axis indicates the number of loci exhibiting each haplotype count.

## 677 2.5 LD decay analysis and GWAS

678 There is a sharp decline in LD over short distances, with  $r^2$  values dropping rapidly  
679 within the first few kilobases. The half-decay point occurs at approximately 8,000 bp  
680 (Figure 8). Genome-wide association studies (GWAS) were conducted to identify  
681 SNPs associated with fruit length, weight, and width in an apple diversity panel. All  
682 three traits exhibited right-skewed distributions (Supplementary figure 2), with most  
683 accessions clustering in the lower value ranges and a smaller number of individuals  
684 showing substantially larger fruit dimensions (Supplementary Table 3). The genome-  
685 wide association analysis revealed a polygenic basis for fruit size traits, with distinct  
686 but partially overlapping loci contributing to variation in fruit length, weight, and  
687 width. For fruit length, most accessions measured between 10 and 30 mm, with a few  
688 exceeding 60 mm, and significant associations were identified on chromosomes 1, 4,  
689 and 9, where several markers surpassed the Bonferroni threshold ( $-\log_{10}(p) = 5$ ),  
690 along with moderate associations on chromosomes 5, 6, and 11. The strongest GLM  
691 associations for fruit length were observed at *S01\_31069925*, *S04\_30479303*, and  
692 *S09\_5414418*, with  $-\log_{10}(p)$  values of 8.15, 9.58, and 7.67, respectively. In  
693 addition to these major loci on chromosomes 1, 4, and 9, several SNPs on  
694 chromosomes 5, 6, 11, and 15—such as *S06\_31198297*, *S11\_41050955*, and

695 *S15\_31363354*—showed moderate significance. Fruit weight was skewed toward  
696 smaller sizes, with most accessions weighing less than 150 g and relatively few  
697 exceeding 200 g. Genome-wide association analysis revealed a highly significant  
698 locus on chromosome 4 at *S04\_30479303*, with a peak  $-\log_{10}(p)$  of 11.86 in the  
699 GLM model and 10.48 in the MLM model, indicating a strong and consistent  
700 association across models. Additional significant associations were observed at  
701 *S09\_1227048* ( $-\log_{10}(p) = 7.13$  GLM; 5.86 MLM), *S10\_38529651* ( $-\log_{10}(p) =$   
702 6.91 GLM; 6.69 MLM), *S11\_41163261* ( $-\log_{10}(p) = 6.83$  GLM; 6.16 MLM), and  
703 *S12\_31337187* ( $-\log_{10}(p) = 6.27$  GLM; 5.69 MLM). Additional moderate  
704 associations included *S05\_16541510* ( $-\log_{10}(p) = 5.88$ ), *S06\_31198297* ( $-\log_{10}(p)$   
705  $= 5.97$ ), and *S01\_31810145* ( $-\log_{10}(p) = 5.58$  GLM; 5.55 MLM). The identification  
706 of overlapping loci across models supports a polygenic basis for fruit weight, with  
707 several loci displaying robust and replicable signals. Among the top associated loci,  
708 *S01\_31810145*, *S04\_30479303*, *S06\_31198297*, *S09\_1227048*, *S11\_41163261*, and  
709 *S12\_31337187* were common to both fruit weight and fruit length, indicating shared  
710 genetic regulation of these traits. Similarly, fruit width was concentrated between 10  
711 and 30 mm, with some individuals exceeding 75 mm. The strongest association was  
712 detected on chromosome 4 at *S04\_30479303* ( $-\log_{10}(p) = 8.97$ ), followed by a  
713 moderate association on chromosome 6 at *S06\_31198297* ( $-\log_{10}(p) = 5.95$ ). Both  
714 loci were also associated with fruit length and weight, highlighting a shared yet  
715 complex genetic architecture underlying fruit size traits in *Malus* (Figure 9)  
716 (Supplementary Table 4). Overall, these GWAS results highlight overlapping as well  
717 as trait-specific genomic regions associated with apple fruit morphology. The  
718 identified loci provide promising targets for further validation and marker-assisted  
719 breeding to improve fruit size and shape.

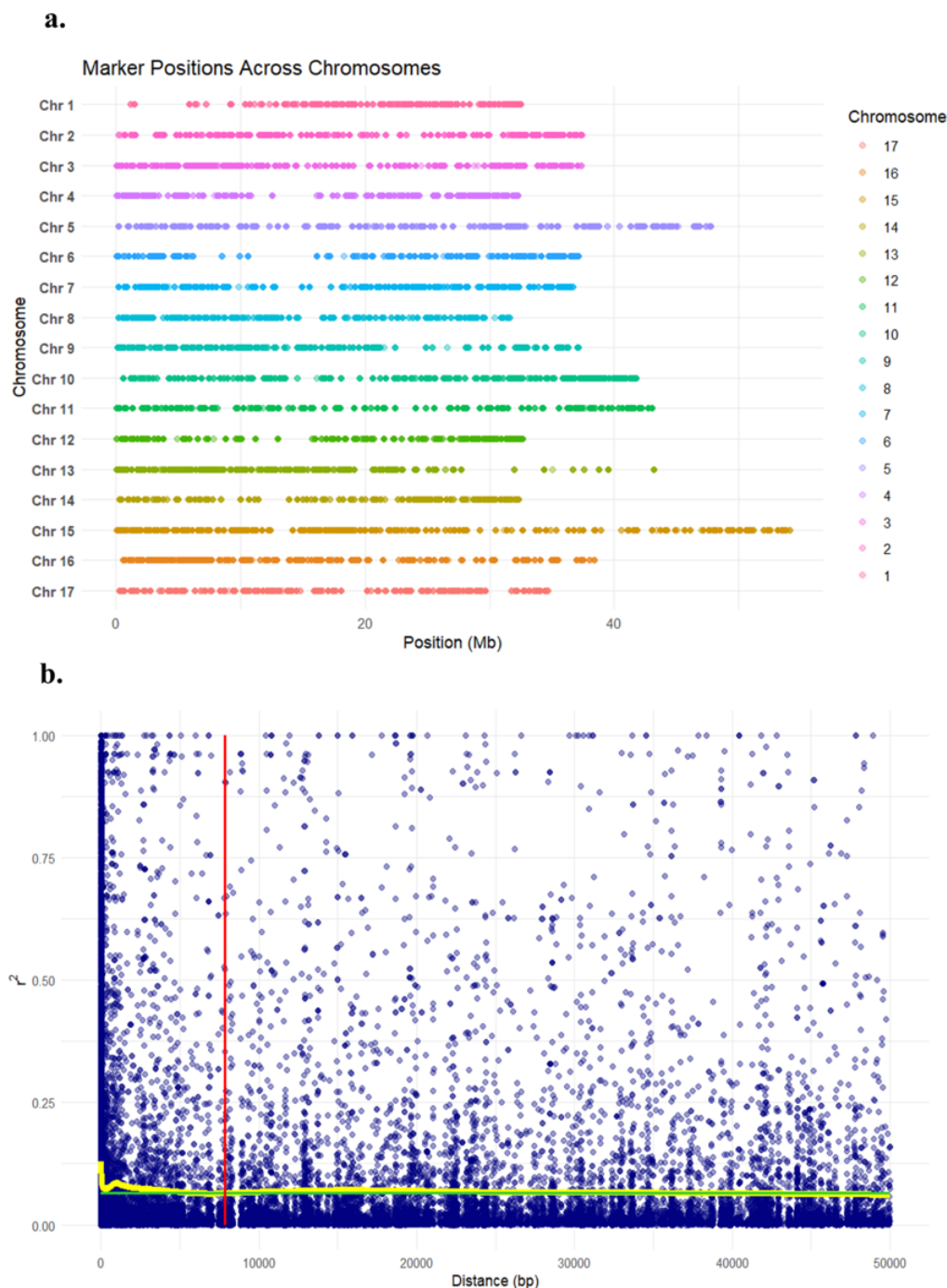


**Figure 8: Genomic distribution of markers and linkage disequilibrium decay in *Malus*** **a.** Distribution of genetic markers across the 17 *Malus domestica* chromosomes. Each dot represents the physical position of a marker, color-coded by chromosome. Markers are evenly distributed across the genome, with selection based on gene density to ensure uniform coverage **b.** Linkage disequilibrium (LD) decay plot showing pairwise  $r^2$  values against physical distance (bp). The yellow LOESS curve represents the average LD decay, with  $r^2$  declining with increasing inter-marker distance. A red vertical line indicates the approximate LD decay threshold used to inform marker spacing in downstream analyses.

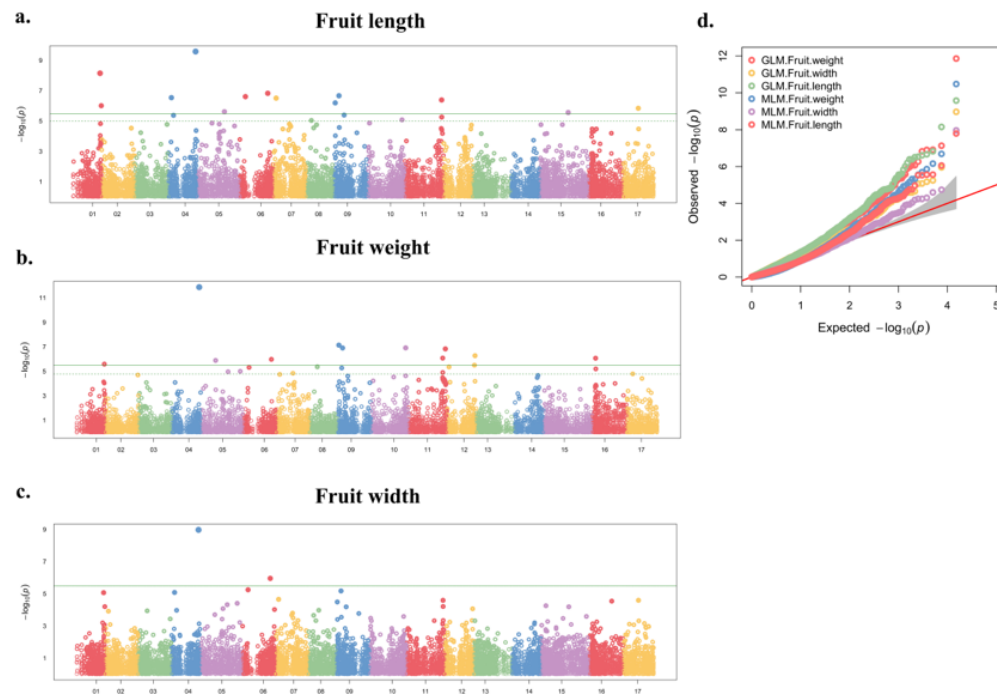


**Figure 9: Genome-wide association analysis for fruit length, weight and width in *Malus*.** Manhattan plots (a–c) display the  $-\log_{10}(p)$  values of SNP–trait associations across all 17 chromosomes for: (a) fruit length, (b) fruit weight, and (c) fruit width. The x-axis represents the physical positions of SNPs ordered by chromosome, and the y-axis shows the statistical significance of association ( $-\log_{10}(p)$ ). Each point represents a single SNP, color-coded by chromosome. The horizontal green dashed line indicates the genome-wide significance threshold ( $-\log_{10}(p) = 5$ ). (d) Quantile–quantile (Q–Q) plot comparing observed vs. expected  $-\log_{10}(p)$  values under the null hypothesis for each trait analyzed using both GLM and MLM models. Trait–model combinations are color-coded as indicated in the legend. Deviation from the diagonal red line indicates potential true associations.

## Discussion

Wild relatives of the domesticated apple (*Malus domestica*) represent a valuable reservoir of genetic diversity, particularly for traits related to disease resistance and tolerance to abiotic stresses. This underutilized genepool hold immense potential for accelerating the development of novel cultivars with improved resilience through targeted breeding strategies (Sabety et al. 2024). Here, we present a cost-effective medium density pan-generic SNP and micro-haplotype marker panel optimized to capture genome-wide variation across the *Malus* genus. The marker panel was developed based on genome-wide conserved, syntenic, and collinear genomic blocks identified within the core genome of 13 diverse *Malus* species, thereby enabling cross-species marker transferability across the genus.

*Malus* is a highly heterozygous genus, shaped by extensive gene recombination and natural mutations throughout its evolution. Understanding the genetic architecture underlying valuable traits is essential for advancing fruit tree breeding (Khan et al. 2021). Identifying the molecular mechanisms that drive trait variation is particularly important to meet the increasing demand for high-quality, and sustainably produced pome fruits (Chagne et al. 2012). One of the major challenges in breeding disease resistant apples today is elucidating the complex relationship between phenotypic diversity and underlying genomic variation (Khan et al. 2024; Liu et al. 2022; Švara et al. 2024b). This marker panel was developed and validated to capture genomic variation, including SNPs, INDELs, and micro-haplotypes across the diverse *Malus* genus. While previous studies assessed genetic diversity in wild apple germplasm using a limited number of SSR markers (Zhang et al. 2012; Gao et al. 2015; Wang et al. 2024; Gao et al. 2021), most available SNP arrays were developed using discovery panels composed primarily of *M. domestica* cultivars (Chagné et al. 2012; Bianco et al. 2014, 2016). This domesticated apple-focused nature of the SNP arrays introduces ascertainment bias, resulting in poor representation of rare alleles from wild *Malus* species (Peace et al. 2019; Giebel et al. 2021). Subsequently, the use of the Infinium 20K SNP array for genotyping 178 *Malus sylvestris* accessions revealed that 20% of the markers exhibited more than 10% missing values, 30% had a minor allele frequency (MAF) below 5%, and 6% were entirely monomorphic (Buiteveld et al. 2021). In contrast, the DArTag panel we developed has 16,000 universal genetic markers showing less than 0.5% missing data, demonstrating cross-species transferability and applicability across diverse *Malus* accessions.

DArT-based marker panels have demonstrated superior performance over GBS approaches in highly heterozygous and polyploid crops such as alfalfa, strawberry, and sweetpotato, offering increased genome coverage, reduced missing data, improved genotyping accuracy, and enhanced cross-species transferability (Zhao et al. 2023; Zhao et al. 2024; Jose et al. 2015). DArTag technology addresses this need by designing oligonucleotides that target known SNPs and small INDELs (<50 bp), along with their flanking regions. The resulting amplicons, 54 bp or 81 bp are sequenced using next-generation sequencing (NGS), allowing for the detection of multiple alleles per locus (Zhao et al. 2023 & 2024). Unlike traditional SNP arrays or KASP assays, which typically detect only two alleles per site, DArT enables identification of short multiallelic haploblocks (Zou et al. 2020; Jose et al. 2015). In this study, the DArT markers exhibited an average of 9-10 alleles per locus across the *Malus* diversity panel, highlighting their high resolution and informativeness. Similar results have been observed in sweetpotato, where 183 marker loci revealed more than 10 distinct micro-haplotypes, reflecting extensive genetic diversity. Compared to biallelic SNPs, these multiallelic haplotype markers provide greater information content, simplify haplotype phasing, and improve inference of ancestry. The even genomic coverage and reduced ascertainment bias associated with DArTag marker panel makes it particularly valuable for genetic studies in highly heterozygous systems, as well as for applications in population genetics and ecological genomics (Zhao et al. 2023; Zhao et al. 2024).

This pan-generic marker panel offers high-resolution, gene-informed marker distribution, suitable for large-scale genotyping, QTL validation, and genome-wide association studies across diverse *Malus* populations. To ensure uniform genome-wide coverage, marker selection was refined based on gene density resulting in the

retention of 3,100 gene-informed genomic blocks including 104 SNPs linked to previously identified important QTLs. This design strategy provides balanced representation of both coding and regulatory regions, enhancing the utility of the panel for functional and evolutionary genomics. Notably, markers exhibiting high allele frequency variance between wild and domesticated apples were prioritized, making them particularly informative for studying hybrid populations and introgression patterns. Additionally, markers with low  $F_{st}$  but high MAF between wild and cultivated groups offer valuable resolution for capturing shared genetic diversity across *Malus*, increasing their applicability in broad population genomic analyses. The panel enables fine-scale marker-trait association discovery, dissection of the genetic architecture of complex traits, and the study of both neutral and adaptive genetic variation. To validate the performance of the panel, it was applied to three bi-parental mapping populations and a diverse *Malus* germplasm panel. Across these datasets, more than 1,500 SNP and haplotype markers were identified as informative, underscoring the panel's robustness and versatility. PCA further demonstrated its discriminatory power, with  $F_2$  individuals clearly segregating between both parents. This pattern confirms the panel's high resolution and reliability in capturing inherited genetic variation and population structure. Similar results have been found in DArT panels developed for potato (*Solanum tuberosum* L.) (Endelman et al. 2024), strawberry (*F. × ananassa*) (Hardigan et al. 2023), and alfalfa (Zhao et al. 2023). The 3K DArTag panel for apple offers substantial practical benefits, particularly in terms of scalability and cost-efficiency. With a per-sample cost of less than \$15, it provides an affordable solution for high-throughput genotyping in large-scale studies. This makes it especially suitable for bi-parental mapping populations, where it enables the construction of high-resolution linkage maps while simultaneously allowing the detection of additional loci associated with important agronomic traits. Its high informativeness, genome-wide coverage, and low ascertainment bias make it very attractive for genetic studies in complex, heterozygous systems such as *Malus*.

The pan-generic marker panel was validated on a broad *Malus* diversity panel, revealing clear phylogenetic structure and supporting the panel's resolution in detecting both broad-scale divergence and fine-scale variation within and between species. Phylogenetic clustering identified three major clades. Clade I comprised of North American *Malus* species, including *M. ioensis*, *M. coronaria*, *M. angustifolia*, and *M. fusca*. Within this clade, *M. angustifolia* and *M. coronaria* formed a close sister group, while *M. ioensis* clustered as a more distinct subgroup. Interestingly, one accession of *M. coronaria* (PI 613891) collected from Ontario and a *M. angustifolia* accession PI 590175, collected from New York clustered with *M. ioensis*, whereas the remaining *M. coronaria* accessions grouped with *M. angustifolia*. These clustering patterns may reflect admixture zones, historical gene flow within North American *Malus*, or possible misclassification. *Malus fusca*, native to the Pacific Northwest and commonly referred to as the Oregon or Pacific crabapple, formed a genetically distinct group, in agreement with recent pan-genome studies that place *M. fusca* as divergent from other North American species (Svara, 2024a; Li et al. 2024). Notably, two accessions of *M. fusca* (PI 589941 and PI 590038) from California and one from an unspecified U.S. location clustered apart from the main *M. fusca* group, suggesting mis-labelling or potential admixture.

A second major cluster consisted of East Asian *Malus* species, including *M. arnoldiana*, *M. baccata*, *M. 'Evereste'*, *M. floribunda*, *M. hupehensis*, *M. micromalus*,

850 *M. prunifolia*, *M. × robusta*, *M. sargentii*, and *M. zumi*. A third clade comprised  
851 domesticated apples (*M. domestica*) and its progenitor species including *M. sieversii*,  
852 *M. sylvestris*, and *M. orientalis*. This grouping reflects known evolutionary histories,  
853 where *M. domestica* originated in Central Asia from *M. sieversii*, with introgression  
854 from *M. sylvestris* and *M. orientalis* during the process of domestication and  
855 westward dispersal (Duan et al. 2017).

856 These findings underscore the capacity of this pan-generic marker panel to detect  
857 genetic relationships across the *Malus* genus. The clear separation of major  
858 geographic groups, detection of intra-specific variation, and identification of potential  
859 admixture or mislabeling events highlight its utility for taxonomy, germplasm  
860 curation, evolutionary biology, and pre-breeding applications. Moreover, the panel's  
861 effectiveness in resolving complex polyploid and hybrid structures provides a strong  
862 foundation for its use in trait mapping, genomic selection, and conservation strategies  
863 across wild and cultivated *Malus* populations.

864 We identified multiple loci significantly ( $-\log_{10}(p) \geq 5$ ) associated with fruit length,  
865 weight, and width, highlighting the polygenic and partially overlapping genetic  
866 architecture of fruit size traits. The most robust associations were observed on  
867 chromosomes 1, 4, 5, 6, 9, 11, 12, and 15, with *S04\_30479303* and *S06\_31198297*  
868 consistently detected across all traits. The detection of major loci on chromosomes 4,  
869 5, 6, 9, 11, and 15 is broadly consistent with the results reported by Devoghalaere et al.  
870 (2012), who identified six QTLs for fruit weight on LGs 5, 8, 11, 15, 16, and 17 using  
871 bi-parental populations derived from 'Royal Gala' × 'Braeburn' and 'Starkrimson' ×  
872 'Granny Smith'. Their study reported conserved QTLs on LGs 8 and 15, along with  
873 epistatic interactions among loci on LGs 8, 15, and 17. In line with these findings,  
874 significant associations in the current study were also observed on chromosomes 5, 11,  
875 and 16, with *S05\_16541510* and *S16\_3804813* influencing fruit weight, and  
876 *S11\_41050955* and *S11\_41163261* associated with both fruit weight and length. The  
877 repeated detection of signals on chromosome 15 across independent studies highlights  
878 its potential role as a key genomic region involved in the regulation of fruit size in  
879 apple. Moreover, our findings are in strong agreement with the genome-wide QTL  
880 analysis conducted by Chang et al. (2014), who used a 'Jonathan' × 'Golden  
881 Delicious' population to identify multiple QTLs for fruit size, shape, length, and  
882 diameter across at least 11 linkage groups. Their results highlighted LGs 4, 5, 8, 11,  
883 12, 15, and 17 as recurrent hotspots for fruit-related traits. Notably, QTLs for both  
884 fruit length and diameter were mapped to LGs 4 and 11. Our study similarly detected  
885 significant ( $-\log_{10}(p) \geq 5$ ) associations for fruit size traits on chromosomes 4, 5, and  
886 11, and we additionally identified *S04\_30479303* on chromosome 4 as a major shared  
887 locus across all three traits, supporting the hypothesis that this region harbors genes of  
888 broad regulatory influence.

889 Overall, our study confirms the highly polygenic and partially overlapping genetic  
890 architecture underlying fruit morphology in apple. Importantly, the ability of the  
891 DArTag panel to detect both major and minor associations across a diverse panel  
892 confirms its effectiveness not only for diversity analysis but also for high-resolution  
893 GWAS applications in apple.

## 894 Conclusion

We report the successful development of a pan-generic 3K DArTag marker panel tailored for comprehensive genomic applications in *Malus* species. This cost-effective, high-throughput, transferable marker system enables robust QTL mapping, GWAS, linkage map construction, and population genomics analyses across diverse genetic backgrounds. The panel's broad cross-species utility and genome-wide coverage provide a valuable resource for dissecting trait architecture, facilitating marker-assisted selection, and accelerating breeding efforts in apples and its wild relatives.

## **Declarations**

### **Availability of data**

All data used in the manuscript is cited or included in the manuscript.

### **Competing Interest**

The authors declare that they have no competing interests.

### **Ethics approval and consent to participate**

Not applicable

### **Consent for publication**

Not applicable

### **Funding**

This research was financially supported by the USDA-NIFA Specialty Crop Research Initiative (SCRI) grant 1023572 (subaward RC111414A).

### **Acknowledgments**

We acknowledge Della Cobb-Smith for her help to collect leaf samples for DNA extraction and genotyping.

### **Author Contributions**

A.Kh conceived the study, acquired the funding, and supervised the research. R.T, H.F, A.Š, and P.C, collected the leaf samples for DNA extraction. A.S. performed bioinformatic analysis, conducted genetic analysis, and wrote the original draft of the manuscript. J.G and Q.S, performed bioinformatic analysis for the oligo design. A.S, A.Ki and A.Kh, revised the manuscript. All authors have read and approved the manuscript.

930 **References:**

- 931 Ammar R (2016) randomcoloR: Generate Attractive Random Colors. R package  
932 version 1.1.0.1
- 933 Antanaviciute L, Fernández-Fernández F, Jansen J, Banchi E, Evans KM et al (2012)  
934 Development of a dense SNP-based linkage map of an apple rootstock progeny using  
935 the Malus Infinium whole genome genotyping array. BMC Genomics 13:203
- 936 Bianco L, Cestaro A, Linsmith G, Muranty H, Denancé C et al (2016) Development  
937 and validation of the Axiom® Apple480K SNP genotyping array. Plant J 86:62–74
- 938 Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S et al (2014) Development and  
939 validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping  
940 array for apple (*Malus × domestica* Borkh). PLoS One 9:e110377
- 941 Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007)  
942 TASSEL: software for association mapping of complex traits in diverse samples.  
943 *Bioinformatics* 23:2633–2635
- 944 Buiteveld J, Koehorst-van Putten HJ, Kodde L et al (2021) Advanced genebank  
945 management of genetic resources of European wild apple, *Malus sylvestris*, using  
946 genome-wide SNP array data. Tree Genet Genomes 17:32
- 947 Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B et al (2012) Genome-  
948 wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS  
949 One 7:e31745
- 950 Chagné D, Vanderzande S, Kirk C, Profitt N, Weskett R, Gardiner SE, Peace CP,  
951 Volz RK, Bassil NV (2019) Validation of SNP markers for fruit quality and disease  
952 resistance loci in apple (*Malus × domestica* Borkh.) using the OpenArray platform.  
953 Hortic Res 6:30
- 954 Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-  
955 generation PLINK: rising to the challenge of larger and richer datasets. GigaScience  
956 4:7
- 957 Chang Y, Sun R, Sun H, Zhao Y, Han Y, Chen D, Wang Y, Zhang X, Han Z (2014)  
958 Mapping of quantitative trait loci corroborates independent genetic control of apple  
959 size and shape. Sci Hortic 174:126–132
- 960 Chen X, Li S, Zhang D, Han M, Jin X, Zhao C, Wang S, Xing L, Ma J, Ji J, An N  
961 (2019) Sequencing of a wild apple (*Malus baccata*) genome unravels the differences  
962 between cultivated and wild apple species regarding disease resistance and cold  
963 tolerance. G3 Genes Genom Genet 9:2051–2060
- 964 Clark MD, Schmitz CA, Rosyara UR et al (2014) A consensus ‘Honeycrisp’ apple  
965 (*Malus × domestica*) genetic linkage map from three full-sib progeny populations.  
966 Tree Genet Genomes 10:627–639

967 Conner P, Brown S, Weeden N (1997) Randomly amplified polymorphic DNA-based  
968 genetic linkage maps of three apple cultivars. J Am Soc Hortic Sci 122:350–359

969 Cornille A, Gladieux P, Smulders MJM, Roldán-Ruiz I, Laurens F et al (2012) New  
970 insight into the history of domesticated apple: secondary contribution of the European  
971 wild apple to the genome of cultivated varieties. PLoS Genet 8:e1002703

972 Daccord N, Celton JM, Linsmith G et al (2017) High-quality de novo assembly of the  
973 apple genome and methylome dynamics of early fruit development. Nat Genet  
974 49:1099–1106

975 Dan, C., Sestras, A.F., Bozdog, C., and Sestras, R.E. 2015. Investigation of wild  
976 species potential to increase genetic diversity useful for apple breeding. Genetika  
977 47:993–1011

978 Devoghalaere F, Doucen T, Guitton B, Keeling J, Payne T, Ling TJ, Davy M, Hilario  
979 E, Chagné D, David K (2012) A genomics approach to understanding the role of  
980 auxin in apple (*Malus × domestica*) fruit size control. BMC Plant Biol 12:7

981 Duan N, Bai Y, Sun H, Wang N, Ma Y et al (2017) Genome re-sequencing reveals the  
982 history of apple and supports a two-stage model for fruit enlargement. Nat Commun  
983 8:249

984 Endelman JB, Kante M, Lindqvist-Kreuzer H, Kilian A, Shannon LM et al (2024)  
985 Targeted genotyping-by-sequencing of potato and data analysis with R/polyBreedR.  
986 Plant Genome e20484

987 Falginella L, Cipriani G, Monte C et al (2015) A major QTL controlling apple skin  
988 russetting maps on the linkage group 12 of ‘Renetta Grigia di Torriana’. BMC Plant  
989 Biol 15:150

990 Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using  
991 multilocus genotype data: linked loci and correlated allele frequencies. Genetics  
992 164:1567–1587

993 Fazio G, Aldwinckle HS, Volk GM, Richards CM, Janisiewicz W, Forsline PL (2009)  
994 Progress in evaluating *Malus sieversii* for disease resistance and horticultural traits.  
995 Acta Hort 814:59–66

996 Fazio G, Chao CT, Forsline P, Richards C, Volk G (2014) Tree and root architecture  
997 of *Malus sieversii* seedlings for rootstock breeding. Acta Hort 1058:585–594

998 Fernández-Fernández F, Evans KM, Clarke JB, Govan CL, James CM et al (2008)  
999 Development of an STS map of an interspecific progeny of *Malus*. Tree Genet  
1000 Genomes 4:469–479

1001 Francis RM (2017) pophelper: an R package and web app to analyse and visualize  
1002 population structure. Mol Ecol Resour 17:27–32

- 1003 Gao Y, Liu F, Wang K et al (2015) Genetic diversity of *Malus* cultivars and wild  
1004 relatives in the Chinese National Repository of Apple Germplasm Resources. Tree  
1005 Genet Genomes 11:106
- 1006 Gao Y, Wang D, Wang K, Cong P, Li L, Piao J (2021) Analysis of genetic diversity  
1007 and structure across a wide range of germplasm reveals genetic relationships among  
1008 seventeen species of *Malus* Mill. native to China. J Integr Agric 20:3186–3198
- 1009 Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H (2021) How array  
1010 design creates SNP ascertainment bias. PLoS One 16:e0245178
- 1011 Gruber B, Unmack PJ, Berry OF, Georges A (2018) dartr: An R package to facilitate  
1012 analysis of SNP data generated from reduced representation genome sequencing. Mol  
1013 Ecol Resour 18:691–699
- 1014 Hardigan, M.A., Feldmann, M.J., Carling, J., Zhu, A., Kilian, A., Famula, R.A., Cole,  
1015 G.S., and Knapp, S.J. 2023. A medium-density genotyping platform for cultivated  
1016 strawberry using DArTag technology. Plant Genome 16:e20399
- 1017 Hemmat M, Weeden NF, Manganaris AG, Lawson DM (1994) Molecular marker  
1018 linkage map for apple. J Hered 85:4–11
- 1019 Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity Arrays: A solid state  
1020 technology for sequence information independent genotyping. Nucleic Acids Res  
1021 29:e25
- 1022 Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation  
1023 program for dealing with label switching and multimodality in analysis of population  
1024 structure. Bioinformatics 23:1801–1806
- 1025 Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic  
1026 markers. Bioinformatics 24:1403–1405
- 1027 Kamvar ZN, Brooks JC, Grünwald NJ (2015) Novel R tools for analysis of genome-  
1028 wide population genetic data with emphasis on clonality. Front Genet 6:208
- 1029 Khan A, Carey SB, Serrano A, Zhang H, Hargarten H, Hale H, Harkess A, Honaas L  
1030 (2022) A phased, chromosome-scale genome of 'Honeycrisp' apple (*Malus domestica*).  
1031 GigaByte 2022:gigabyte69
- 1032 Khan A, Gutierrez B, Chao CT, Singh J (2021) Origin of the domesticated apples. In:  
1033 Korban SS (ed) The Apple Genome. Compend Plant Genomes. Springer, Cham.
- 1034 Khan A, Švara A, Wang N (2024) Comparing apples and oranges: Advances in  
1035 disease resistance breeding of woody perennial fruit crops. Annu Rev Phytopathol  
1036 62:263–287
- 1037 Khan MA, Han Y, Zhao YF, Korban SS (2012) A high-throughput apple SNP  
1038 genotyping platform using the GoldenGate™ assay. Gene 494:196–201

1039 Knaus BJ, Grünwald NJ (2017) vcfr: a package to manipulate and visualize variant  
1040 call format data in R. *Mol Ecol Resour* 17:44–53

1041 Kostick SA, Luby JJ (2022) Apple fruit size QTLs on chromosomes 8 and 16  
1042 characterized in ‘Honeycrisp’-derived germplasm. *Agronomy* 12:1279

1043 Kumar S, Chagné D, Bink MC, Volz RK, Whitworth C, Carlisle C (2012) Genomic  
1044 selection for fruit quality traits in apple (*Malus × domestica* Borkh.). *PLoS One*  
1045 7:e36674

1046 Kumar S, Garrick DJ, Bink MC, Whitworth C, Chagné D, Volz RK (2013) Novel  
1047 genomic approaches unravel genetic architecture of complex traits in apple. *BMC*  
1048 *Genomics* 14:393

1049 Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for  
1050 phylogenetic tree display and annotation. *Nucleic Acids Res* 49:W293–W296.

1051 Li H (2011) A statistical framework for SNP calling, mutation discovery, association  
1052 mapping and population genetical parameter estimation from sequencing data.  
1053 *Bioinformatics* 27:2987–2993

1054 Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*  
1055 34:3094–3100

1056 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler  
1057 Transform. *Bioinformatics* 25:1754–1760

1058 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,  
1059 Durbin R; 1000 Genomes Project Data Processing Subgroup (2009) The Sequence  
1060 Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079

1061 Li W, Chu C, Zhang T, Hou N, Guo J, Niu C, Li C, Liu S, Xu J, Xie Y, Zhang D,  
1062 Shen X, Lu L, Geng D, Chen P, Jiang L, Wang L, Li H, Malnoy M, Deng C, Zou Y,  
1063 Li C, Zhan X, Dong Y, Notaguchi M, Ma F, Xu Q, Guan Q (2025) Pan-genome  
1064 analysis reveals the evolution and diversity of *Malus*. *Nat Genet* 57:1274–1286

1065 Li Z, Wang L, He J, Li X, Hou N, Guo J, Niu C, Li C, Liu S, Xu J, Xie Y, Zhang D,  
1066 Shen X, Lu L, Geng D, Chen P, Jiang L, Wang L, Li H, Malnoy M, Deng C, Zou Y,  
1067 Li C, Zhan X, Dong Y, Notaguchi M, Ma F, Xu Q, Guan Q (2022) Chromosome-  
1068 scale reference genome provides insights into the genetic origin and grafting-mediated  
1069 stress tolerance of *Malus prunifolia*. *Plant Biotechnol J* 20:1015–1017

1070 Liebhard R, Koller B, Gianfranceschi L, Gessler C (2003) Creating a saturated  
1071 reference map for the apple (*Malus × domestica* Borkh.) genome. *Theor Appl Genet*  
1072 106:1479–1508

1073 Liu W, Chen Z, Jiang S, Wang Y, Fang H, Zhang Z, Chen X, Wang N (2022)  
1074 Research progress on genetic basis of fruit quality traits in apple (*Malus × domestica*).  
1075 *Front Plant Sci* 13:918202

1076 Longhi S, Cappellin L, Guerra W et al (2013) Validation of a functional molecular  
1077 marker suitable for marker-assisted breeding for fruit texture in apple (*Malus ×*  
1078 *domestica* Borkh). *Mol Breed* 32:841–852

1079 Maliepaard C, Alston HF, Van Arkel G et al (1998) Aligning male and female linkage  
1080 maps of apple (*Malus pumila* Mill.) using multi-allelic markers. *Theor Appl Genet*  
1081 97:60–73

1082 Mansfeld BN, Yocca A, Ou S, Harkess A, Burchard E, Gutierrez B, van Nocker S,  
1083 Gottschalk C (2023) A haplotype resolved chromosome-scale assembly of North  
1084 American wild apple *Malus fusca* and comparative genomics of the fire blight Mfu10  
1085 locus. *Plant J* 116:989–1002

1086 Marić, S., Lukić, M., Cerović, R., Mitrović, M., and Bošković, R.I. 2010. Application  
1087 of molecular markers in apple breeding. *Genetika* 42:359–375

1088 Migicovsky Z, Gardner KM, Richards C et al (2021) Genomic consequences of apple  
1089 improvement. *Hort Res* 8:9

1090 Norelli JL, Wisniewski M, Fazio G, Burchard E, Gutierrez B et al (2017) Genotyping-  
1091 by-sequencing markers facilitate the identification of quantitative trait loci controlling  
1092 resistance to *Penicillium expansum* in *Malus sieversii*. *PLoS One* 12:e0172949

1093 Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution  
1094 in R language. *Bioinformatics* 20:289–290

1095 Peace C, Norelli J (2009) Genomics approaches to crop improvement in the Rosaceae.  
1096 In: Foltá KM, Gardiner SE (eds) *Genetics and Genomics of Rosaceae*. Springer, New  
1097 York, pp 19–53

1098 Peace CP, Bianco L, Troggio M et al (2019) Apple whole genome sequences: recent  
1099 advances and new prospects. *Hort Res* 6:59

1100 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using  
1101 multilocus genotype data. *Genetics* 155:945–959

1102 Qin S, Xu G, He J, Li L, Ma H, Lyu D (2023) A chromosome-scale genome assembly  
1103 of *Malus domestica*, a multi-stress resistant apple variety. *Genomics* 115:110627

1104 Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing  
1105 genomic features. *Bioinformatics* 26:841–842

1106 Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z (2017) Crop breeding  
1107 chips and genotyping platforms: Progress, challenges, and perspectives. *Mol Plant*  
1108 10:1047–1064

1109 Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z (2017) Crop breeding  
1110 chips and genotyping platforms: Progress, challenges, and perspectives. *Mol Plant*  
1111 10:1047–1064

- 1112 Rosenberg NA (2004) distruct: a program for the graphical display of population  
1113 structure. *Mol Ecol Notes* 4:137–138
- 1114 Rymenants M, van de Weg E, Auwerkerken A et al (2020) Detection of QTL for  
1115 apple fruit acidity and sweetness using sensorial evaluation in multiple pedigreed full-  
1116 sib families. *Tree Genet Genomes* 16:71
- 1117 Sabety J, Švara A, Tegtmeier R, Feulner H, Cho P, Sakina A, Hickok D, Khan A  
1118 (2024) Unlocking diversity from wild relatives of perennial fruit crops in the pan-  
1119 genomics era. *Curr Opin Plant Biol* 82:102652
- 1120 Sansaloni C, Petroli C, Jaccoud D et al (2011) Diversity Arrays Technology (DArT)  
1121 and next-generation sequencing combined: genome-wide, high throughput, highly  
1122 informative genotyping for molecular breeding of Eucalyptus. *BMC Proc* 5(Suppl  
1123 7):P54
- 1124 Schouten HJ, van de Weg WE, Carling J, Khan SA, McKay SJ, van Kaauwen MPW,  
1125 Wittenberg AHJ, Durel CE (2012) Diversity arrays technology (DArT) markers in  
1126 apple for genetic linkage maps. *Mol Breeding* 29:645–660
- 1127 Sun X, Jiao C, Schwaninger H, Chao CT, Ma Y, Duan N, Khan A, Ban S, Xu K,  
1128 Cheng L, Zhong GY, Fei Z (2020) Phased diploid genome assemblies and pan-  
1129 genomes provide insights into the genetic history of apple domestication. *Nat Genet*  
1130 52:1423–1432
- 1131 Švara A, Sun H, Fei Z, Khan A (2023) Chromosome-level phased genome assembly  
1132 of ‘Antonovka’ identified candidate apple scab resistance genes highly homologous to  
1133 *HcrVf2* and *HcrVf1* on linkage group 1. *G3 Genes|Genomes|Genetics* 14:jkad253
- 1134 Švara A, Sun H, Fei Z, Khan A (2024a) Advancing apple genetics research: *Malus*  
1135 *coronaria* and *Malus ioensis* genomes and a gene family-based pangenome of native  
1136 North American apples. *DNA Res* 31:dsae026
- 1137 Švara A, Sun H, Fei Z, Khan A (2024b) Chromosome-level phased genome assembly  
1138 of “Antonovka” identified candidate apple scab-resistance genes highly homologous  
1139 to *HcrVf2* and *HcrVf1* on linkage group 1. *G3 Genes Genomes Genet* 14:jkad253
- 1140 Wang J, Zhang Z (2021) GAPIT Version 3: Boosting power and accuracy for  
1141 genomic association and prediction. *Genom Proteom Bioinf.* 19:629–640
- 1142 Wang L, Li Z, Wang D, Tian W, Sun S, Wang G, Lu X, Liu Z, Wang K, Gao Y (2024)  
1143 Genetic diversity and geographic distribution patterns of *Malus* landraces from five  
1144 cultivated species under climate change in China. *Sci Hortic* 336:113389
- 1145 Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer-Verlag, New  
1146 York
- 1147 Wickham H, François R, Henry L, Müller K, Vaughan D (2023) dplyr: A grammar of  
1148 data manipulation. R package version 1.1.4

1149 Wickham H, Vaughan D, Girlich M (2024) tidy: Tidy messy data. R package version  
1150 1.3.1

1151 Yang S, Fresnedo-Ramírez J, Wang M, Cote L, Schweitzer P, Barba P, Takacs EM,  
1152 Clark M, Luby J, Manns DC, Sacks G, Mansfield AK, Londo J, Fennell A, Gadoury  
1153 D, Reisch B, Cadle-Davidson L, Sun Q (2016) A next-generation marker genotyping  
1154 platform (AmpSeq) in heterozygous crops: A case study for marker-assisted selection  
1155 in grapevine. *Hortic Res* 3:16002

1156 Zhang Q, Li J, Zhao Y et al (2012) Evaluation of genetic diversity in Chinese wild  
1157 apple species along with apple cultivars using SSR markers. *Plant Mol Biol Rep*  
1158 30:539–546

1159 Zhang Q, Ma B, Li H et al (2012) Identification, characterization, and utilization of  
1160 genome-wide simple sequence repeats to identify a QTL for acidity in apple. *BMC*  
1161 *Genomics* 13:537

1162 Zhao D, Mejia-Guerra KM, Mollinari M, Samac DA, Irish BM, Heller-Uszynska K,  
1163 Beil CT, Sheehan MJ (2023) A public mid-density genotyping platform for alfalfa  
1164 (*Medicago sativa* L.). *Genet Resour* 4:55–63

1165 Zhao D, Sandercock AM, Mejia-Guerra KM, Mollinari M, Heller-Uszynska K, Wadl  
1166 PA, Webster SA, Beil CT, Sheehan MJ (2024) A public mid-density genotyping  
1167 platform for hexaploid sweetpotato (*Ipomoea batatas* [L.] Lam). *Genes* 15:1047

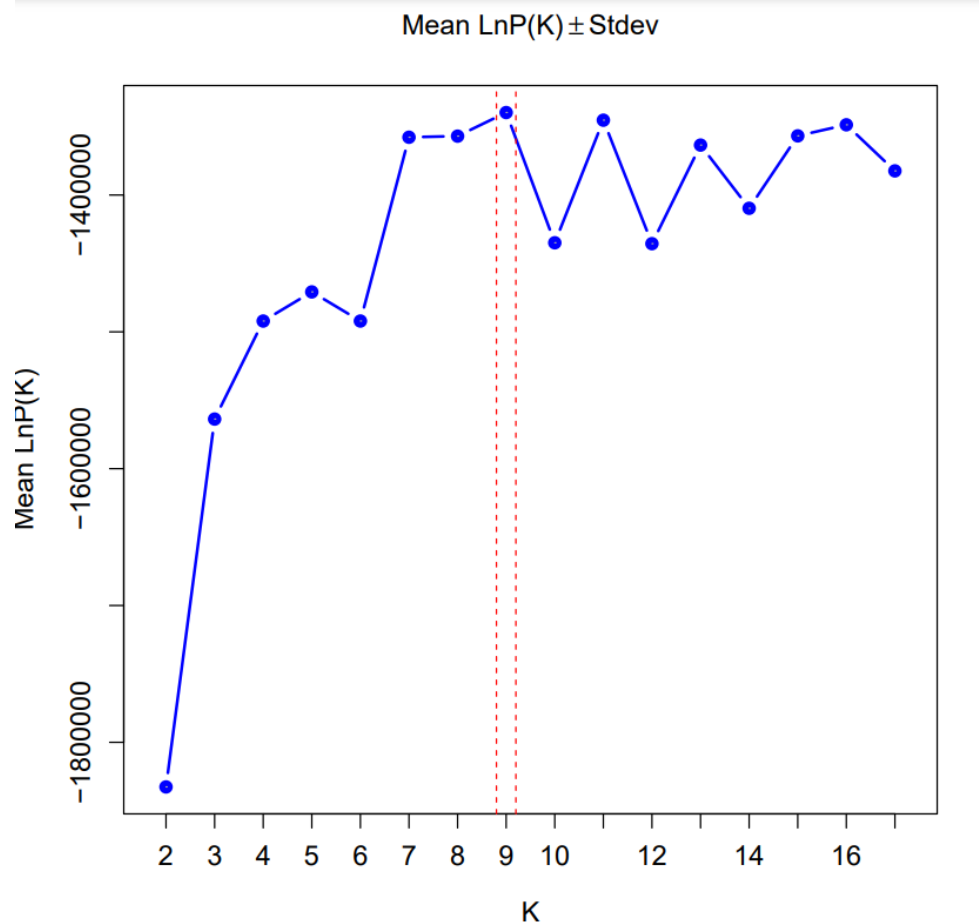
1168 Zhao D, Sapkota M, Glaubitz J, Bassil N, Mengist M et al (2024) A public mid-  
1169 density genotyping platform for cultivated blueberry (*Vaccinium* spp.). *Genet Resour*  
1170 5:36–44

1171 Zou C, Karn A, Reisch B, Nguyen AH, Sun Y, Bao Y, Owens CL, Brown PJ, Londo  
1172 JP, Xu Y (2020) Haplotyping the *Vitis* collinear core genome with rhAmpSeq  
1173 improves marker transferability in a diverse genus. *Nat Commun* 11:413

1174

1175

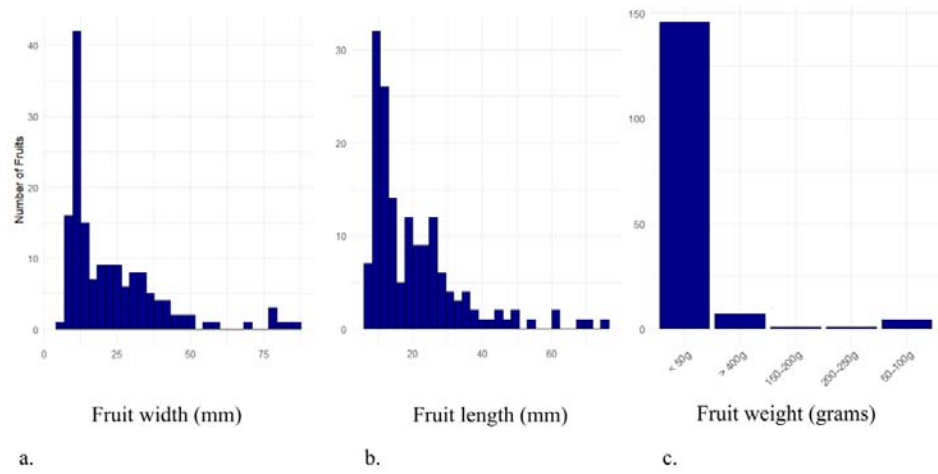
1176 **Supplementary Figure 1:** Evanno plot showing  $\Delta K$  values for determining the  
1177 optimal number of genetic clusters. Mean log probability of the data  $[\text{LnP}(K)] \pm$   
1178 standard deviation is plotted against the number of clusters ( $K$ ). Vertical red lines  
1179 indicate the optimal  $K$  as determined by the Evanno  $\Delta K$  method, supporting the  
1180 presence of nine genetically distinct populations in the dataset.



1181

1182 **Supplementary Figure 2: Distribution of fruit morphological traits across *Malus***  
1183 **accessions.**

1184 Histograms show the frequency of fruits measured for a. fruit width, b. fruit length,  
1185 and c. fruit weight. Width and length are shown in centimeters, with most fruits  
1186 clustering below 40 cm in both dimensions. Fruit weight is categorized into discrete  
1187 bins:  $<50$  g, 50–100 g, 150–200 g, 200–250 g, and  $>400$  g, with the majority  
1188 of fruits weighing less than 50 g. The y-axis represents the number of fruits observed  
1189 within each bin.



1190