GENOME WIDE ASSOCIATION STUDY OF BASIC FRUIT CHEMISTRY IN THE COLD CLIMATE WINE GRAPES (*VITIS* SPP.)

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Venkateswara Rao Kadium

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Plant Sciences

March 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Genome Wide Association Study of Basic Fruit Chemistry in the Cold Climate Wine Grapes (*Vitis* spp.)

By

Venkateswara Rao Kadium

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the Degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Harlene Hatterman-Valenti
Chair

Dr. Xuehui Li

Dr. Gregory R. Cook

Approved:

April 14, 2022
Date

Dr. Richard Horsley
Department Chair

ABSTRACT

To overcome some of these challenges posed by ND climate, the utilization of native wild *Vitis*-derived varieties is the best possible option available. Despite advantageous environmental tolerances of native wild *Vitis* spp. derived crosses, their acid and sugar concentrations often deviate from expectations set for *V. vinifera*. Identifying the genetic determinants of titratable acidity (TA), pH, and total soluble solids (TSS/°Brix) in interspecific hybrid populations can help improve new hybrid cultivars. For this purpose, an incomplete diallel mapping population with substantial *riparia* and other wild *Vitis* spp. in its background was used to perform association studies. The population is genotyped with single nucleotide polymorphism (SNP) markers and phenotyped over two years. Genome wide association analysis (GWAS) identified a significant association on chromosomes 6 and 16 for all three traits in both years. Candidate gene identification under the significant region revealed multiple glucose, fructose, and amino acid metabolism genes.

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest gratitude to my advisor Dr. Harlene Hatterman-Valentin for giving me the opportunity to be a part of this research project. The invaluable and positive things I have learned during my days as her student are something that will persist with me for a very long time. The amount of freedom she gave me to grow as a researcher made my life easier in many ways. I am also deeply indebted to my committee member Dr. Xuehui Li who has been an invaluable resource and whose help and patience cannot be overestimated. I would like to extend my sincere thanks to Dr. Gregory R. Cook, who has also graciously served on my thesis committee and whose viticultural knowledge has been a great asset to this study.

I would also like to express my sincere gratitude to Dr. Andrej Svyantek for the guidance and the knowledge he shared with me throughout this research. I am thankful to the high-value crop research group members for the valuable time and assistance they offered me to complete this research work. I am also thankful to the members of the department of plant sciences in NDSU for the assistance they provided during my degree.

I would like to thank my mom for always believing in me and supporting me through different phases of my life. Finally, I'd like to express my gratitude to all of the professors who have helped me develop as a student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

LITERATURE REVIEW

Grapevine introduction

The grapevine is one of the significant global horticulture crops. Grapevine cultivation is almost as old as civilization, beginning around 8000 years ago (McGovern et al., 2003). In 2020, grapevines were cultivated roughly in 6.95 million ha worldwide and produced roughly 78 million metric tons (Mg) of grapes (FAO, 2020). In the United States, the grapevine is the leading fruit in terms of production quantity, with some 6 million Mg produced in 2020 nationally. California is the leading state contributing up to 94% of the national production, followed by Washington (NASS, 2020). Most of the fruit produced is being used for winemaking, and the remainder is being consumed as a variety of products such as table grapes, raisins, bottled fresh grape juice for consumption, processed into jellied products like ‘Concord’ jelly, or concentrated into syrups like "Pecmez" (Bouquet., 2011).

Grapevine belongs to the family *Vitis*, which consist of approximately 60 interfertile species that have evolved worldwide. *Vitis vinifera* L., indigenous to Eurasia, has been used most extensively for wine and table grape purposes (This et al., 2006; Reynolds., 2017). Grapes can be cultivated at latitudes ranging from 50°N to 40°S and up to 3,000 meters above sea level, with *Vitis vinifera* L. ssp. *sativa* varietals of Eurasian descent representing nearly 98 % of the vines worldwide (McGovern et al., 2003). *Vitis vinifera* originated from its progenitor species, *Vitis vinifera subsp. sylvestris* (Heywood and Zohary., 1995). As per historical evidence, the foundations of grape domestication are believed to have occurred in the South Caucasus between the Caspian and Black Seas (Terral et al., 2010; Myles et al., 2011). During domestication, *Vitis vinifera* underwent many morphological and physiological changes, such as a shift from the

dioecious nature of *V. vinifera* ssp. *sylvestris* to hermaphroditism in *Vitis vinifera* ssp. *sativa* by developing perfect flowers to reduce reliance on pollen donors and ensure self-pollination.

Additionally, due to human selection pressure, the size of berries and clusters increased significantly with high levels of sugar content for better fermentation (Martin et al., 2009; Zecca et al., 2010; Myles et al., 2011). Many wild *Vitis* species (*V. riparia, V. rupestris, V. lubrasca, V. aestivalis*, etc.) have been long grown under different biotic and abiotic stresses of North America. Two-thirds of the total *Vitis* species are native to the North American region (Aradhya et al., 2003). These wild relatives provide promising germplasm for breeders to develop resistance against different pests and diseases and adapt to unfavorable environmental conditions. During the mid-19th century, the accidental introduction of pests and diseases such as phylloxera (*Daktulosphaira vitifoliae* Fitch), downy mildew (*Plasmopara viticola* Berl.), and powdery mildew (*Uncinula necator* Burr) led to extensive destruction of vineyards in Europe due to the sensitive nature of *Vitis vinifera* which drastically reduced genetic diversity of the species (This et al., 2006; Myles et al., 2011). These adverse events lead to the widescale adaption of rootstock breeding programs utilizing resistance North American rootstock, including *V. riparia, V. rupestris, V. lubrasca,* and *V. aestivalis* (Cousins and Striegler., 2005; Terral et al., 2010). Additionally, many hectares of French American hybrids (*V. vinifera* x Wild American *Vitis* spp.) were planted in Europe during the same period. Many of the French-American hybrids lasted until the late twentieth century, thanks to scion breeding operations. Many of those French American hybrid cultivars were also transported to eastern North America in the late 1940s and established a critical element of the industry in the US (Reynolds., 2015). However, rootstock and scion breeding operations stagnated over the twentieth century. The varietal range in commercial vineyards declined dramatically, owing to establishing a wine trade

centered on a few internationally renowned cultivars, among other factors. Thousands of *V. vinifera* cultivars exist currently, but only a handful of cultivars dominate the global wine market (This et al., 2006; Bouquet., 2011).

Grapevine berry developmental stages

Grapes are classified as berries because their seeds are wrapped in a thick fleshy pericarp. Grape berries are organized in clusters, with a pedicel containing vascular bundles connecting each berry to the cluster. The growth of grape berries consists of two sigmoidal phases separated by a lag phase (Coombe and McCarthy., 2000; Robinson and Davies.,2000). The first growth stage lasts around 60 days from the start of the bloom. During this stage, berries and seed embryos develop from the flowers. Within the first several weeks of this stage, rapid cell division occurs, and the extent of this division influences the final size of the berries. Along with cell division, various nutrients are supplied to developing berries through the vascular system during this stage. The xylem transports water, nutrients, minerals, and growth regulators from the roots to the developing fruits early in the stage. The buildup of these solutes, particularly malic and tartaric acid, causes the berry volume to expand (Kennedy., 2002). These two acids make up to 90 % of the total acid content of the berries. The distribution and accumulation of these acids within the berries vary based on location and time. Tartaric acid tends to build on the outside of the berries towards skins, while malic acid tends to accumulate inside the flesh. Tartaric acid accumulates early in the growth stage, while most malic acid accumulates near the end of the first growth stage, shortly before veraison (Kennedy., 2002). Tannins, another essential component that gives the wine its bitter and astringent flavor, accumulates in the skin and seed during the early stage of growth.

Veraison, the second stage of berry development, is marked by the softening and coloration of the berries. The berries double in size between the start of the second growth stage and harvest. Most of the compound produced during the first growth stage remains until the end of the second phase, but their concentration dilutes due to the doubling of berry size. However, during the second phase, the malic acid content in berries declines, and the extent of this decrease is mainly determined by environmental conditions during veraison. Typically, warmer regions tend to have less malic acid when compared to cooler regions. Another significant change during veraison is the influx of sugars into the berries. Sucrose produced by photosynthesis in the leaves is actively transported into the berries through the phloem. Once transported into berries, this sucrose hydrolyzes into reduced sugars such as glucose and fructose (Coombe et al., 1987; Kennedy., 2002). The amount of sugar that accumulates during veraison is determined by various factors, including crop load, canopy management, disease pressure, and the length of time the berries are kept on the vine. Aside from sugar accumulation, another critical alteration that affects wine quality during veraison is the accumulation of secondary metabolites. These mainly include anthocyanins in red varieties and terpenoids in white varieties (Coombe et al., 2000, Kennedy., 2002).

Fruit quality parameters

Total soluble solids (TSS) measurements are commonly used to determine the sugar content of berries. Glucose and fructose make up nearly 99% of grape juice's total carbohydrates, representing a significant amount of the total soluble solids (Kliewer., 1966). Soluble solids, measured in ^{o}Brix, are an estimate for sugar concentration based on the juice's refractive index. Soluble solids describe a juice sample's relative sugar weight; for example, 1^{o}Brix denotes 1 % sugar by weight (Jackson and Lombard, 1993). Glucose and fructose are essential components of

fermentation in winemaking since yeast processes sugars into alcohol. Glucose and fructose also have an impact on wine quality because they are responsible for the sweet flavor and aid in the balance of sourness, bitterness, and astringency. As the grape berries ripen, soluble solids rise to a level that can indicate the optimum harvest ripeness (Jackson and Lombard, 1993). Wine made from grapes with a high °Brix can have a high alcohol content, which can mask other quality characteristics. As a result, a 24 °Brix upper limit is commonly utilized to signify adequate ripeness for quality wine production (Winkler., 1974).

Fruit pH is another vital component that can determine wine quality. A wine's pH level of 3.60 or above may cause some quality issues. High pH levels promote the relative activity of microbes such as bacteria that impair the color intensity of red wines, absorb more sulphur dioxide, and reduce free SO_2 concentration, and can slow down the wine's aging process (Jackson and Lombard, 1993). Increases in pH are generally correlated with increases in soluble solids (°Brix) during maturity and can be used to determine the best time to harvest.

Acidity levels fluctuate throughout berry growth stages as a result of metabolic activities. Typically, 90 % of the acids contained in grapes are tartaric and malic acids, with tartaric acid being the most common (Kliewer., 1965; Lamikanra et al., 1995). Succinic, acetic, citric, lactic, fumaric, and shikimic acids, among others, can be present in varying amounts depending on cultivar and environmental factors (García-Ruiz et al., 2008). Grape acidity is commonly expressed as titratable or total acidity (TA). The TA is a critical metric that grape growers use to assess the quality of their juice and wine. The cultivar, growing region, and environmental conditions such as sunshine, precipitation, and temperature can all affect the composition of organic acids. (Lamikanra et al., 1995). In general, *V. riparia* derived grapes have more malic acid than tartaric acid, which could explain why *V. riparia* hybrid cultivars have such high

titratable and sensory acidity. The TA value was reduced rapidly during maturation because of dilution with sugars and other components during veraison. During maturation, the drop in TA was a function of temperature and is linked to the berry's respiration rate (Winkler., 1974). The primary acid that was impacted by respiration was malic acid, and the main difference between cool and warm climates was that malate concentration decrease slowly in cool climates but quickly in warm climates.

Grapevine genome

The grapevine genome is thought to have formed due to an ancient polyploidization event involving the fusion of three genomes or a hexaploidization event. (Jaillon et al., 2007; Malacarne et al., 2012). The grapevine genome is diploid (2n = 38chromosomes) in a contemporary breeding sense and relatively small, at approximately 475-500 Mb in size (Lodhi and Reisch, 1995). There is a total of 19 haploid linkage groups. This is small compared to other commonly cultivated crops (approximately one-sixth of the corn genome), making it more appealing for genetic research. (This et al., 2006). Even though grapevine is hermaphrodite, outcrossing by means of wind or insects is the primary mode of pollination. As a result, cultivars are highly heterozygous and contain a significant number of harmful recessive mutations. Due to the high levels of heterozygosity, inbreeding depression is severe, and sterility often increases from the second generation when selfed, which would make it difficult for whole-genome sequencing of grapevine (Rick and Simmonds, 1976). PN40024, a selection derived from 'Pinot noir'. By successive selfings, this line has been bred near full homozygosity (estimated at 93 %), enabling high-quality whole-genome shotgun assembly (Jaillon et al., 2007). In the same year, another Pinot noir-derived variety 'ENTAV 115' was successfully sequenced using whole-genome shotgun sequencing (Velasco et al., 2007) and was the first perennial crop whose

genome was sequenced completely (Bouquet., 2011). Two high-quality reference genomes assisted breeding operations worldwide. The reference genome aided in the discovery of genes that underpin different cultural and quality characteristics and opened up new avenues for molecular breeding.

Genotyping-by-sequencing

Plant genetics and crop improvement programs rely heavily on genomic variation studies. DNA polymorphisms can be linked to phenotype differences or reflect the relation between individuals in the populations (Rafalski., 2002). Genotyping has aided the mapping of several candidate genes and metabolic pathways, as well as the study of evolution, diversity, and marker-assisted selection (MAS) in a variety of crop species over the last few decades (Deschamps et al., 2012). The first plant DNA markers used were restriction fragment length polymorphisms (RFLPs). Later due to their inherent challenges, RFLP's were replaced by PCR-based markers such as simple sequence repeats (SSR), random amplification of polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), and others. The PCR-based markers are relatively abundant in the genome, less expensive, polymorphic, and co-dominant in nature (Williams et al., 1990; Paran and Michelmore., 1993). Later the introduction of next-generation sequencing (NGS) technology in the early 21st century allowed for the detection of genetic variation at a single base-pair resolution, which led to the development of a new type of marker known as single nucleotide polymorphisms (SNPs) (Deschamps et al., 2012).

Genotyping by Sequencing (GBS) is a novel application of NGS technologies for identifying and genotyping SNPs in crop genomes and populations without going through the complete marker assay development stage (Elshire et al., 2011; Deschamps et al., 2012). Low cost, simplified handling, fewer PCR and purification steps, no reference sequence

restrictions, no size stratification, fast barcoding, and convenience of scaling up are all critical features of GBS technology (Davey et al., 2011). Elshire et al. (2011) were the first to describe and test GBS in maize and barley recombinant inbred lines populations. GBS is becoming more and more essential in various plant species as a cost-effective and unique method for genomics-assisted breeding (He et al., 2014). However, missing data and heterozygote under-calling affected the progress of GBS genetic maps in highly heterozygous species like grapevine

GBS has several advantages, including a highly multiplexed and shallow sequencing that simplifies library construction and lowers per-sample costs. However, this key advantage becomes a significant disadvantage when it comes to heterozygous crops like grapevine since shallow sequencing leads to genotyping errors, under-calling heterozygous sites, and missing data (Hyma et al., 2015). In heterozygous samples, imputation of missing data is more complex than in homozygous samples, resulting in a more significant number of imputation errors (Swarts et al., 2014). To overcome these difficulties of GBS in heterozygous crops, programs and functions, such as Heterozygous Mapping Strategy (HetMappS) have been developed to handle errors associated with heterozygosity and missing data while constructing genetic maps (Hyma et al. 2015). There are already a variety of grapevine populations that have been genotyped using GBS. These populations have been used to examine a variety of traits of interest, including quality attributes, pests, and disease resistance.

GBS libraries can now be sequenced on a variety of platforms. Illumina Genome Analyzer is one of them, and it is based on the notion of sequencing by synthesis (Mardis, 2008). The widescale availability of this novel NGS protocol at a cheaper cost makes the GBS an appealing strategy to map the breeding population with a high density of SNP markers. This opens the doors for more widespread use of genome-wide association studies (GWAS), genomic

selection (GS), and marker-assisted selection (MAS) to better understand a variety of traits in diverse crop species (He et al., 2014).

Association studies and linkage disequilibrium

The terms association mapping and linkage disequilibrium (LD) are often misunderstood. Linkage disequilibrium refers to the non-random association between two genes, two markers, or two QTLs within a population. In contrast, association mapping refers to a strong association of a genetic marker to a specific phenotype (Gupta et al., 2005). Association mapping is quickly becoming a popular tool for analyzing several complex traits in crop plants. It offers a benefit over linkage mapping, such as increased mapping resolution without a substantial increase in population size (Owens, 2011). In linkage mapping, the resolution of the genetic map is based on the recombination events that happened in the bi-parental population, which generally involves only one generation/round of recombinations, particularly in perennial crops. In comparison, association mapping is based on recombination events in a group of unrelated individuals in the past (historical recombinations) (Altshuler et al., 2008). This approach has numerous advantages in long-lived perennial crops like grape, where establishing and maintaining a mapping population is both time-consuming and expensive (Myles et al., 2011; Nicolas et al., 2016).

LD and genetic diversity within the association panel play a pivotal role in the association analysis. LD is influenced by various components in a particular species, including population structure, admixture, mutations, drift, and selection. Henceforth it is essential in determining domestication, evolution, and breeding patterns of plants and animals (Amaral et al., 2008; Slatkin., 2008). The number and density of markers in an association panel will be determined by the distance over which LD persists. In general, mutation contributes to the creation of LD between loci, and recombination is the critical mechanism that weakens the LD between two loci

(Flint-Garcia et al., 2003; Zhu et al., 2008). The rate of LD decay across chromosomal segments is of critical importance for improving mapping resolution (Falconer and Mackay, 1996; Mackay and Powell, 2007). Typically, LD decays occur faster in outcrossing species than in selfing species. In selfing species, where individuals are more likely to be homozygous, the chance of recombination is less than outcrossing species (Flint-Garcia et al., 2003). Grape is an outcrossing species; it is expected to have rapid LD decay, and this has been proven in several studies (Barnaud et al., 2006; Lijavetzky et al., 2007; Barnaud et al., 2010; Zhang et al., 2017).

Barnaud et al. (2006) published the first study on LD within cultivated *V. vinifera* L. subsp. *Vinifera* using 38 microsatellite loci spread across five linkage groups. They observed that LD has occurred over 16.8 cM regions within five linkage groups. In contrast, LD decay occurred much more rapidly when they studied the wild population of 85 French *Vitis vinifera* L. subsp. *silvestris* selections, demonstrating a potentially restricted genetic base following domestication, with few recombination events (Barnaud et al., 2010). Another detailed investigation by Lijavetzky et al. (2007), using 11 genotypes and 1500 SNP's, revealed rapid decay of LD in over 200 random loci, representing over 1Mb length of the sequence. Interestingly, following potential bottleneck events associated with the introduction of pests and diseases, genetic diversity was found to be reduced in wild grapevines when compared to cultivated. Using 160 SSR markers, two grouping clusters were estimated based on 81 Chinese native selections composed of 15 *Vitis* species, and LD was estimated to occur up to 14.13 cM (Zhang et al., 2017).

In humans and animals, and more recently in plants, genome-wide association studies (GWAS) have been shown to be a viable technique for mapping associations (Begum et al., 2015). The initial success of GWAS came from its application in human studies that resulted in a

better understanding of Type 2 diabetes risk factors and the discovery of over 100 schizophrenia risk loci (Visscher et al., 2017). After that, GWAS adaption in plants increased tremendously and became a standard tool in dissecting different complex traits. Examples of GWAS in crops include the discovery of significant SNPs linked to soybean resistance to bacterial, fungal, nematode, and viral diseases (Chang et al., 2016) and identification of SNP associations related to flowering time and plant height-related traits in maize (Xiao et al., 2017)

The majority of economically important traits in the grapevine are quantitative in nature, following a complex inheritance pattern involving many genes. The use of GWAS and candidate-gene association analysis have emerged as valuable approaches to study these complex traits effectively. The GWAS analysis using a subset of individuals from USDA grape core collection focused on berry color, identified 5 Mb genomic region on Chromosome-2. Later it was determined that MYB transcriptome genes in this 5 Mb region control the color of berries in grapes (Myles et al., 2011). Association study on a panel of 148 genotypes revealed that gain of function point mutation from G to T in the candidate gene *VvDXS* resulted in muscat flavor in grapes (Emanuelli et al., 2010). Other noticeable GWAS studies focused on different traits of grapes, such as determining the genetic basis of leaf shape (Chitwood et al., 2014), seedlessness (Zhang et al., 2017), establishing *VviUCC1* gene role in cluster architecture traits (Tello et al., 2020), stomatal conductance under drought stress response (Trenti et al., 2021) and many other trait associations.

GAPIT

Genome Association Prediction Integrated Tool (GAPIT) is a statistical package for Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) that is runs in R statistical software (https://zzlab.net/GAPIT/gapit_help_document.pdf). GAPIT is simple to use

and generates extensive data interpretation reports in a publishable format. GAPIT handles both numeric and HapMap genotypic formats as input data. Individuals in the phenotypic file do not have to be in the same order as those in the genotype file. GAPIT can implement a wide range of models such as general linear model (GLM), mixed linear model (MLM), compressed mixed linear model (CMLM), enhanced compressed mixed linear model (ECMLM), multiple loci mixed model (MLMM), fixed and random model circulating probability unification (FarmCPU), Bayesian-information and linkage disequilibrium iteratively nested keyway (BLINK) and genomic best linear unbiased prediction (gBLUP) to perform GWAS and GS in a user-defined way. The CMLM and MLMM models were adopted in the current project, and they will be explored in detail.

CMLM and MLMM

Although GWAS offers the ability to uncover genetic polymorphisms that underlie various traits, the false discovery rate is a key issue, which can be attributed partly to spurious associations caused by population structure and unequal relatedness among individuals in a population. To address these issues, several statistical methods have been proposed so far. The GLM was the initial model adapted to address population structure in GWAS by integrating population structure as a cofactor with marker data (Li et al., 2014).

$$Y = S_i + Q + e$$

Where 'Y' denotes phenotype, 'S_i' is the marker under test, 'Q' is population structure, and 'e' is residual error. Here, the cofactor 'Q' aids in adjusting effects that are not related to the testing markers, resulting in fewer false positives. The MLM applies the same principle by jointly adding a genetic marker-based kinship matrix (K) with the population structure (Zhang et al., 2010)

$$Y = S_i + Q + K + e$$

However, because kinship is obtained from all markers, applying kinship for the testing marker in an MLM creates conflict between the testing markers and the genetic effects specified by kinship. To get around this, Individuals are compressed into groups in CMLM to reduce kinship and testing marker confounding (Zhang et al., 2010). The user can specify the desired number of groups. Summary statistics of kinship between and within groups are applied as elements of a reduced kinship matrix after lines are divided into a given number of groups. To determine the best compression level, several mixed models will be applied. For each model, the log-likelihood function values will be determined, and the best compression level is decided as the one with the most significant log-likelihood function value.

All the mentioned approaches are based on single-locus tests to identify associations between polymorphisms and traits. These techniques, however, may not be well suited for complex traits regulated by multiple large-effect loci, especially in the presence of population structure. Using several cofactors directly in the statistical model is an easy way to increase efficiency and has indeed become the norm in modern linkage mapping, where both multiple–quantitative trait locus mapping and composite interval mapping have been proven to outperform basic interval mapping. The justification for incorporating multiple loci in GWAS is perhaps much more substantial because background loci might cause confounding effects across the genome owing to linkage disequilibrium, rather than just locally due to linkage. Using MLMM is a new technique that permits effects from multiple loci simultaneously and includes associated markers as covariates by employing forward-backward stepwise linear mixed-model regression. The MLMM investigations using human and Arabidopsis thaliana data outperformed existing approaches in terms of power and false discovery rate (Segura et al., 2012).

Current study focus

Previously an incomplete diallel mapping population of 1064 individuals was created by crossing three different parents with significant *riparia* and other *Vitis* sps. backgrounds in them. The population was genotyped with 25490 GBS-derived SNP markers to perform linkage and association studies focused on traits such as fruit quality and cold hardiness.

The present study investigated the genetic determinants of degree °Brix, pH, and total acidity (TA) in the incomplete diallel population. The population was genotyped using GBS-derived SNP markers and phenotyped for °Brix, pH, and TA over two years. A genome-wide association study was performed using genotypic and phenotypic data collected from the population.

MATERIALS AND METHODS

The incomplete diallel population development

A mapping population of 1064 F1 individuals was created by crossing three different parents in a diallel mating design. This population was called an incomplete diallel because it only includes three out of nine possible families of a 3×3 diallel mating design without including self and reciprocal crosses (Table 1). Three parents used in the development of the population were ND 213, SKND.009.41, and ND.054.27.

Table 1. Incomplete diallel population mating design with the number of individuals in each family

		♀		
		ND.213	**ND.054.27**	**SKND.009.41**
♂	**ND.213**	×	99	618
	ND.054.27	×	×	347
	SKND.009.41	×	×	×

The parent ND 213 was created by crossing female parent 'Alpenglow (ES 2-8-1)' with the pollen donor 'C14' (Figure 1). Parent ND.054.27 was created by crossing female parent 'Frontenac Gris' with the pollen donor 'Adalmiina (ES 6-16-30)'. Third parent SKND.099.41 was gifted to NDSU-Grape Germplasm Enhancement Program (NDSU-GGEP) by the University of Saskatchewan, created by crossing female parent 'Perle de Csaba' with the pollen donor 'Riparia L,' a *Vitis riparia* accession from Saskatchewan. This population consisted of three distinct families, with SKND.009.41 × ND.213 as the biggest family with 618 individuals. The cross SKND.009.41 × ND.054.27 was the second biggest family with a total of 347 individuals, and finally, the cross ND.054.27 × ND.213 was the smallest family with a total of 99 individuals.

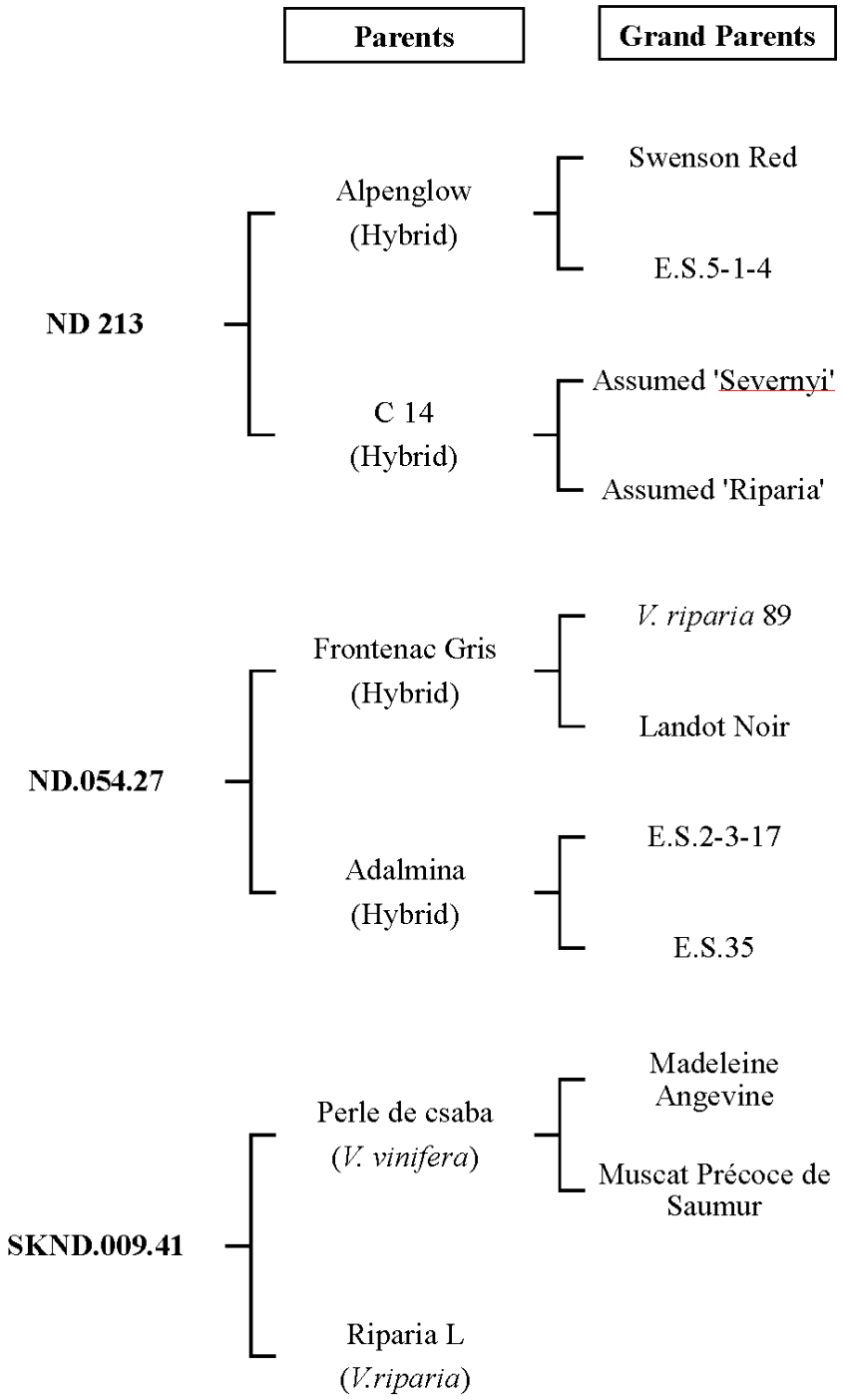

Figure 1. Pedigree history of the parents used in the creation of population

Seeds of Individuals were created by careful hand pollination. After germination, individual plants were grown in greenhouse conditions for about a month before acclimating to outdoor conditions for 14 d and then transplanting to the field environment. Vines were

transplanted in 2017, using an un-replicated design, in a research vineyard located at the NDSU Agriculture Experiment Station, Fargo, North Dakota (46°53'28.9"N 96°48'46.9"W). Vines were trained on a single-wire bilateral cordon system consisting of single trellis wire at 5ft height to support the cordons and crop. Row spacing was 1.52 m (5 feet) while vine spacing was 0.91 m (3 feet) within the row. Due to winter injury, rabbit damage during early years, and variable environmental factors, many individuals failed to establish a healthy vine. After excluding dead/damaged vines, the current population consists of around ~750 individuals in the vineyard. Vines were supported by drip irrigation during the initial two years of management to support good trunk establishment. A total of ~750 genotypes of the population with good establishment were selected to construct the genetic maps. Plant protection chemicals used include glufosinate (Rely® 280, BASF Ag Products, Research Triangle Park, NC, USA) to control weed competition during the summer season. No fungicides were used diseases such as powdery mildew or downy mildew were not present. Vines were covered with bird netting from veraison until harvest to prevent bird damage to the clusters.

Phenotypic data collection

The primary phenotypic data collected measured the basic fruit chemistry (°Brix, pH, and TA) of individual vines in the population. The population was adequately pruned and maintained throughout the growing season to establish an optimum fruit set. The number of vines that produced fruit varied from year to year and depended on many factors such as reaching reproductive maturity, winter survival, and other environmental factors during the growing season (temperature and photoperiod). In 2020, three years after transplanting, a total of 255 genotypes of the population produced fruit for the first time. That the following year (2021), most of the individuals (~ 575) of the population can set the fruit.

To measure the changes in basic fruit chemistry during veraison, berry samples for testing were collected at three different times each year during the ripening period. Harvest dates varied from year to year, depending on photoperiod variations and visible assessments of grape maturity (Stage 85-89 in the BBCH-scale for grapevine). In both years, sample collection was mainly based on the decreasing photoperiod with an average gap of ten days between one harvest and another. During the 2020 growing season, harvests one, two, and three were completed at decreasing photoperiods of 14.5, 14, and 13.5, respectively (Table 2). In 2021, genotypes matured a week earlier than the previous year due to drier and warmer conditions during growing season. In 2021 berry sampling during harvest one, two, and three were completed at decreasing photoperiods 15, 14.5, and 14, respectively (Table 3).

Table 2. Photoperiod, number of genotypes sampled, number of berries sampled, and harvest date for sample collection in 2020.

Description	**The year 2020**		
	Harvest 1	**Harvest 2**	**Harvest 3**
Photoperiod	14.5	14	13.5
Number of genotypes sampled	269	237	213
Number of berries sampled	5-10	5-10	60-80
Date	08/07/2020	08/18/2020	08/31/2020

Table 3. Photoperiod, number of genotypes sampled, number of berries sampled, and harvest date for sample collection in 2021.

Description	**The year 2021**		
	Harvest 1	**Harvest 2**	**Harvest 3**
Photoperiod	15	14.5	14
Number of genotypes sampled	574	547	528
Number of berries sampled	5-10	10-20	60-80
Date	08/02/2021	08/11/2021	08/24/2021

The number of berries sampled for testing during each harvest varied primarily due to genotype-specific fruit availability. The third harvest was the most extensive sample collection in both years and was done after the berries attained optimum harvest ripeness. During both growing years, a ten-milliliter (ml) juice sample from the third sample collection was kept at -20°C for HPLC testing, and a 5-milliliter (ml) juice sample was saved to determine titratable acidity using a pH meter. For HPLC testing, samples were shipped with dry ice to the Iowa State wine laboratory (Ames, IA, USA) and Northern Crop Institute (Fargo, ND, USA) during the years 2020 and 2021, respectively (Data not included here).

Berry samples from each fruiting genotype of the population were collected using plastic bags in the early morning hours of the respective harvest dates. To prevent desiccation, collected samples were immediately transported to a walk-in cooler (maintained at 4°C) in the nearby greenhouse. After collecting the required sample, the weight of the collected berries was measured using a Mettler AE100 Analytical Balance (Mettler-Toledo, Toledo, OH, USA). Single berry mass (SBM) from each genotype was obtained by dividing the berries measured weight by the number of berries in that respective sample. To measure the samples °Brix, pH, and TA content, the juice was extracted by hand pressing the berries in the plastic collection bags. Fruit °Brix was measured using a Pocket Refractometer for grape & wine (Atago PAL-1 | 3810, Atago, Bellevue, WA, USA). Fruit pH was measured using a Digital Pocket pH Meter for grape & wine (PAL-pH | 4311, Atago, Bellevue, WA, USA). TA was measured using a Pocket Acidity Meter for grape & wine (PAL-Easy ACID2, Atago, Bellevue, WA, USA). Dilution used for TA measurement was ten µl grape juice in 490 µl distilled water.

Leaf tissue collection and DNA extraction

Leaf tissue was collected from the selected ~750 vines that had good establishment for DNA extraction. A dime-sized piece of freshly formed tender leaf tissue without any primary vein or tendrils was collected to get an optimum quality DNA library. Collected leaf tissue from each individual was placed in a single 2.2 ml deep well of 96-well plates with silica balls to grind the tissue. After tissue collection, DNA plates were immediately stored in ice and later freeze-dried to -80℃ using a LABCONCO Freeze Dryer (LABCONCO, Kansas City, MO, USA). Lyophilized tissue was ground into a fine powder using a Retsch mixer mill (Retsch, Haan, Germany). Sample DNA was extracted from the grounded tissue using an in-house DNA extraction protocol.

GBS marker genotyping

Single nucleotide polymorphisms (SNPs) were generated from the extracted DNA of the population using the Brummer procedure. Sample GBS was performed utilizing restriction enzyme *ApeKI* for digestion and sequencing platform Illumina HiSeq 2000 (Illumina, San Diego, CA, USA) for sequencing. A 20 µl of diluted DNA was transferred to a new plate. Transferred DNA was restricted with *ApeKI* restriction enzymes (Incubated for two hours at 75 ℃ and then cooled to 4 ℃). Within a single well of a 96-well plate, each genotype was allocated a unique barcode ranging from 5-10 bp. A 30 µl of DNA ligase master mix and common adaptors together with the barcode adapters were added for ligation step (22 ℃ for 2 h, 65 ℃ for 30 min and cooled to 4 ℃ to hold). With barcodes added, DNA fragments were cleaned using beads that eliminated fragments < 300 bp in size. The success of digestion and ligation was tested to ensure the equivalent amount of DNA from each genotype using the Qiagen PCR cleanup kit following the kit instructions. The 50 ng cleaned DNA, 25 µl Kapa library amplification ready mix and

primers for the barcoded adaptors were used for PCR test (5 min at 72 °C, 30 sec at 98 °C, 10 cycles of 10 sec at 98 °C, 30 sec at 65 °C, and 30 sec at 72 °C each, 5 min at 72 °C then finally hold at 4 °C). Finally, the resulting libraries were validated using the Bioanalyzer. After validation, a total of eight DNA libraries were sent to the Univ. Texas Southwestern Medical Center (Dallas, Texas, USA) for sequencing using the sequencing platform Illumina HiSeq 2000.

The Univ. Texas Southwestern Medical Center provided the sequence data in a compressed file format. Individual vines in the population had their sequence data aligned to the 12x v2 *V.vinifera'* PN40024' for SNP calling. The TASSEL 5.2.79 software (Buckler lab) was used to filter the resulting SNP data, saved in VCF file format (Bradbury et al., 2007). Then, using in-house scripts, markers with ≥ 50% missing data were eliminated. Finally, genotypes that had more than 50% of their genotypic information missing were discarded. After removing genotypes with more missing reads, 605 individuals' genotypic information remained. The LD KNNi method in TASSEL 5.2.79 software was used to impute the SNPs with missing data (Figure A3). A total of 25490 SNPs having a minor allele frequency (MAF) of 0.05 were included in the final HapMap (Figure A4).

Genome-Wide Association Study (GWAS)

Phenotypic results of basic fruit chemistry (°Brix, pH, and TA) from three different harvests of both years and the GBS marker data set were used to perform GWAS analysis. Phenotypic data were checked for outlier and exceptional values by generating histograms, QQ plots, and box plots using R statistical software. Data was not checked for normality due to expected differences in phenotypic values due to the varying maturity of individuals in the population. Genome Association and Prediction Integrated Tools version 3 (GAPIT3) (Wang and Zhang, 2021) package in R statistical software were used to perform association analysis.

Principal component analysis (PCA) was used to estimate population structure, and it was performed using the GAPIT3 package (Figure A1). The kinship matrix or population relatedness was calculated using the GAPIT3 package from marker data. Multiple statistical methods can be implemented to perform association analysis using the GAPIT3 package. The models Compressed MLM (CMLM) (Zhang et al., 2010) and multiple loci MLM (MLMM) (Segura et al., 2012) were used in this study to perform GWAS analysis. After running both models in GAPIT3, a series of output files (.cvs and .pdf) were generated containing information such as Manhattan plots, Q-Q plots, significant SNPs with their p-values, MAF, genotypic best linear unbiased estimates (BLUEs) and best linear unbiased predictions (BLUPs), heritability, and population structure graphs (PCA and VanRaden) (Figure A2). Candidate gene scan under the significant genomic regions was done with the help of gene annotated *Vitis.vinifera* 12X reference genome file (Grimplet and Fennel., 2011).

RESULTS

Trait '°Brix'

The °Brix phenotypic distribution in 2020 ranged from 1.98 to 30.15 across various harvests, whereas it ranged from 2.07 to 32.85 in 2021 (Table 4). The °Brix phenotypic data in both cropping seasons highly correlated with different harvests and years (Tables 5, 6, and 7). The mean °Brix levels in the first, second, and third harvests of 2020 were 13.1, 17.1, and 20.3, respectively. In 2021, mean °Brix levels were 11.6, 16.3, and 21.4 during harvests one, two, and three, respectively, exhibiting similar trends as in 2020 (Table 4). The °Brix values varied from harvest to harvest, with harvest one having the lowest mean value and harvest three having the highest in both growing seasons (Figures 2 and 3). After reaching harvest maturity in 2020, 63 % of individuals had a °Brix value greater than 20, but in 2021, 70 % of individuals had a °Brix value greater than 20. Compared to the previous year, more accessions produced fruit in 2021, enabling the phenotypic screening of more genotypes of the population.

Table 4. Summary statistics of trait °Brix.

Trait	Year	Harvest	N		Mean	Maximum	Median	Minimum
			Total	IWG				
°Brix	2020	1	268	195	13.13	30.15	13.5	1.98
		2	237	173	17.17	25.25	18.05	3.2
		3	205	157	20.27	26.1	20.24	5.6
	2021	1	565	404	11.59	28.4	10.65	3.85
		2	535	386	16.32	30.2	16.65	2.07
		3	521	378	21.44	32.85	21.92	5.15

Note: N = number of individuals sampled; IWG = Individuals with GBS markers.

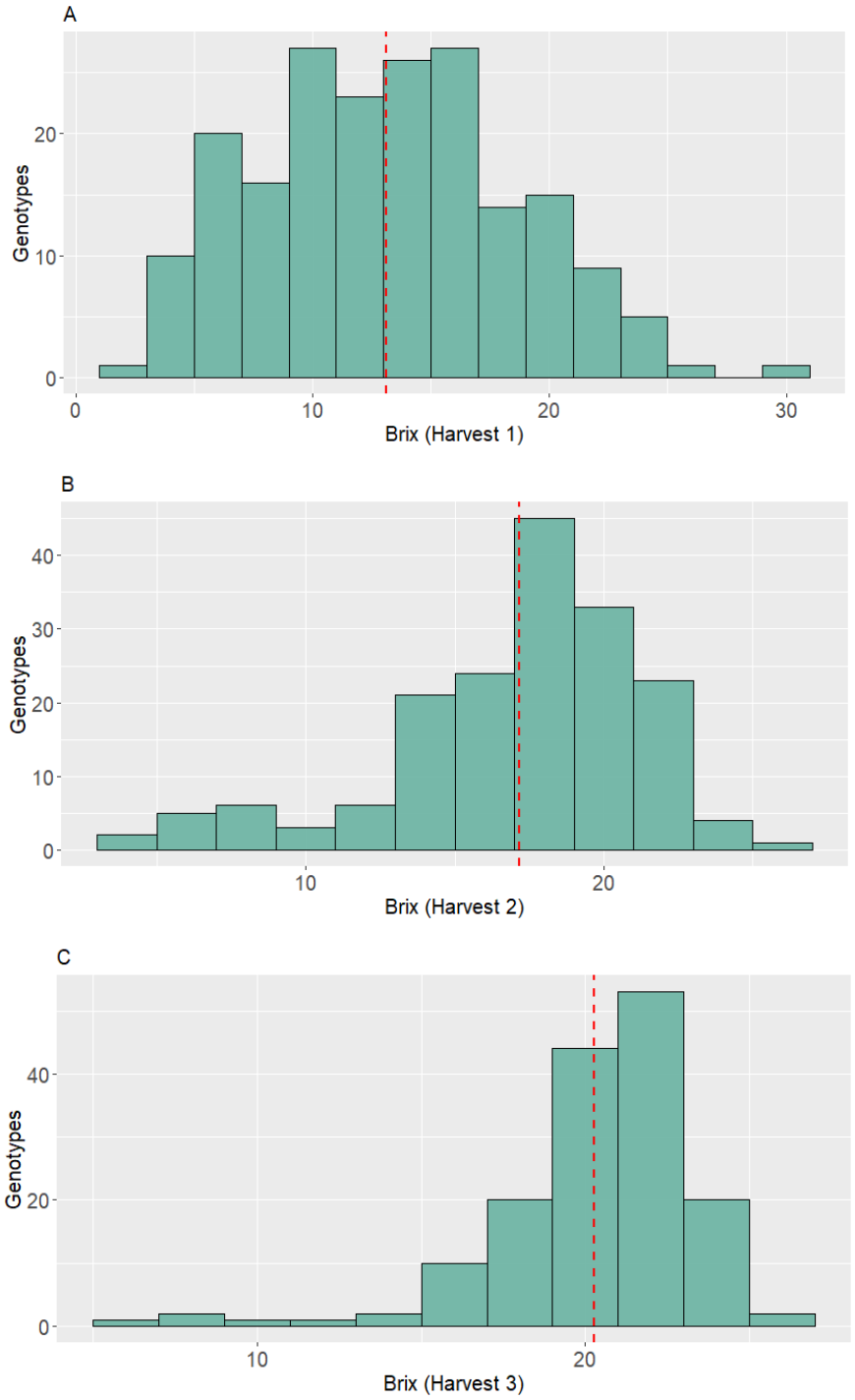

Figure 2. Histogram showing the phenotypic distribution of °Brix in year 2020. A) Histogram of °Brix in harvest one. B) Histogram of °Brix in harvest two. C) Histogram of °Brix in harvest three. Dashed vertical red line indicating the mean value of the trait.

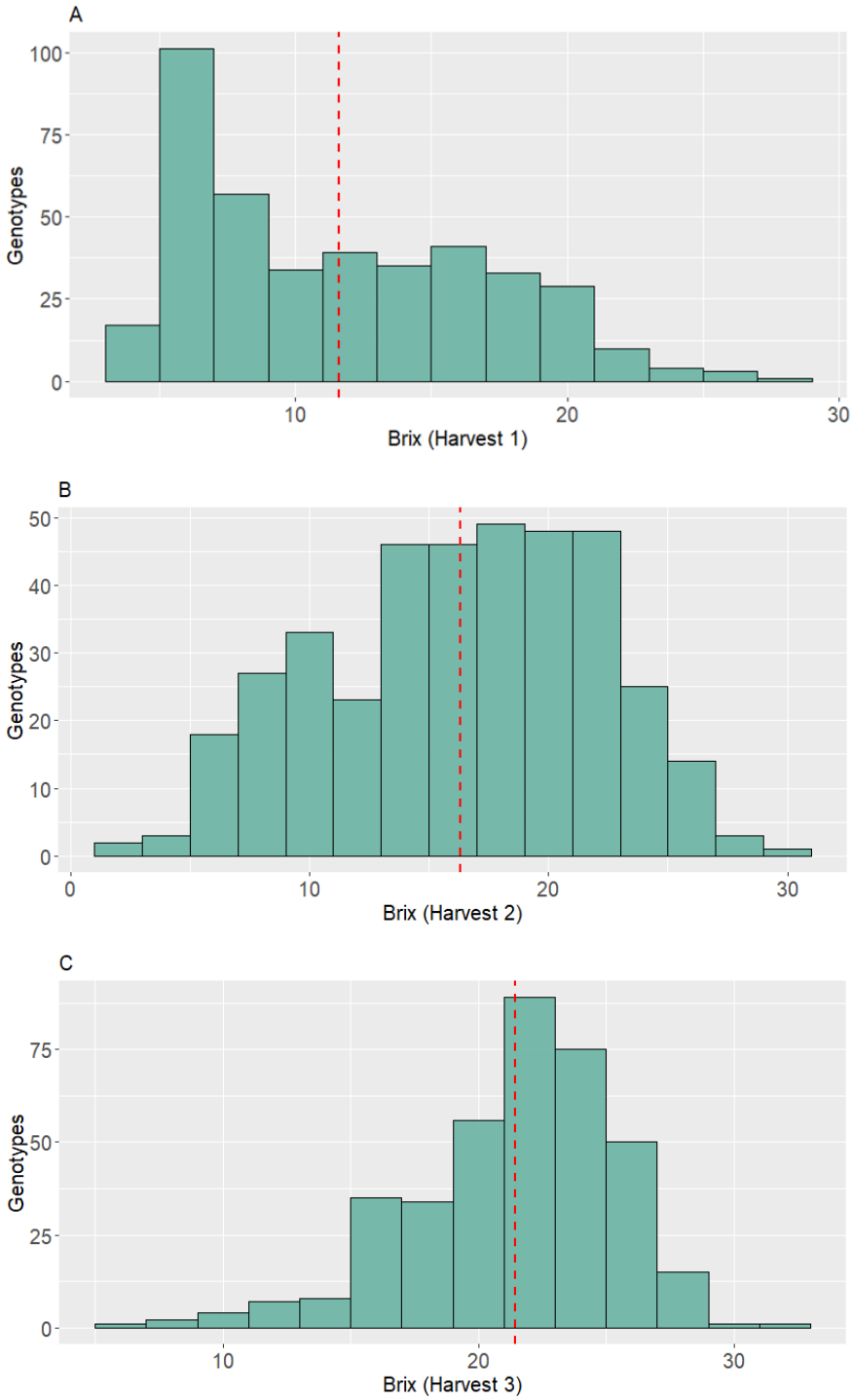

Figure 3. Histogram showing the phenotypic distribution of °Brix in year 2021. A) Histogram of °Brix in harvest one. B) Histogram of °Brix in harvest two. C) Histogram of °Brix in harvest three. Dashed vertical red line indicating the mean value of the trait.

Table 5. Pearson's correlation coefficient and significant estimates for °Brix in the year 2020.

Year		Harvest 2	Harvest 3
2020	**Harvest 1**	0.967 ***	0.895 ***
	Harvest 2		0.972 ***

Note: P-value 0.05*, 0.01**, 0.001***

Table 6. Pearson's correlation coefficient and significant estimates for °Brix in the year 2021.

Year		Harvest 2	Harvest 3
2021	**Harvest 1**	0.980 ***	0.919 ***
	Harvest 2		0.958 ***

Note: P-value 0.05*, 0.01**, 0.001***

Table 7. Pearson's correlation coefficient and significant estimates for °Brix between years.

Year		2020		
		Harvest 1	**Harvest 2**	**Harvest 3**
2021	**Harvest 1**	0.959 ***	0.919 ***	0.821 ***
	Harvest 2	0.961 ***	0.942 ***	0.851 ***
	Harvest 3	0.925 ***	0.945 ***	0.884 ***

Note: P-value 0.05*, 0.01**, 0.001***

GWAS analysis of '°Brix in the year 2020'

In 2020, phenotypic data from 195, 173, and 157 individuals with GBS markers from harvests one, two, and three were included in the GWAS study (Table 4). GWAS analysis of the trait °Brix was conducted using two different models, CMLM and MLMM. On chromosome 16, a strong relationship between traits and markers was discovered using the CMLM model. The GWAS of harvest one data exhibited the SNPs at a higher log-likelihood p-value of 6.06E-07, above the threshold $-\log_{10}$ p-value cut off, indicating that this association is most significant. In subsequent harvests, the log-likelihood p-value of the most significant SNP is less than harvest

one and a little below the significant threshold, indicating that the association had been reduced slightly. For the third harvest, the association shifted to chromosome 2, with some SNPs at a log-likelihood p-value of 4.34E-06, below the threshold cutoff (Table 8 and Figure 4).

The MLMM model produced comparable results to the CMLM model, with the same SNPs as the most significant ones during respective harvests. However, the significance level of the SNPs improved greatly. All significant SNPs were above the threshold level in all three harvests with the MLMM model, and an additional SNP above the cutoff was discovered on chromosome 17 for harvest one (Table 8 and Figure 5).

Table 8. Peak SNPs associated with °Brix in the incomplete-diallel population during growing season 2020 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 1	CMLM	S16_14593162	16	14.59	6.98E-07	0.238	4.186087
		S16_15731027	16	15.73	2.51E-06	0.238	3.8631
		S16_15991560	16	15.99	5.40E-06	0.246	3.658164
		S16_16345474*	16	16.34	6.06E-07	0.269	4.31868
	MLMM	S16_16345474*	16	16.34	1.55E-10	0.269	NA
		S17_5458267	17	54.58	1.98E-07	0.382	NA
Harvest 2	CMLM	S16_21082304*	16	21.08	2.71E-06	0.257	2.993647
		S16_21082329	16	21.08	2.71E-06	0.257	2.993647
		S16_16345474	16	16.34	4.57E-06	0.260	2.97309
	MLMM	S16_15731027*	16	15.73	3.85E-07	0.234	NA
Harvest 3	CMLM	S2_4112015*	2	41.12	4.34E-06	0.134	2.681323
	MLMM	S2_4112015*	2	41.12	4.34E-06	0.134	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

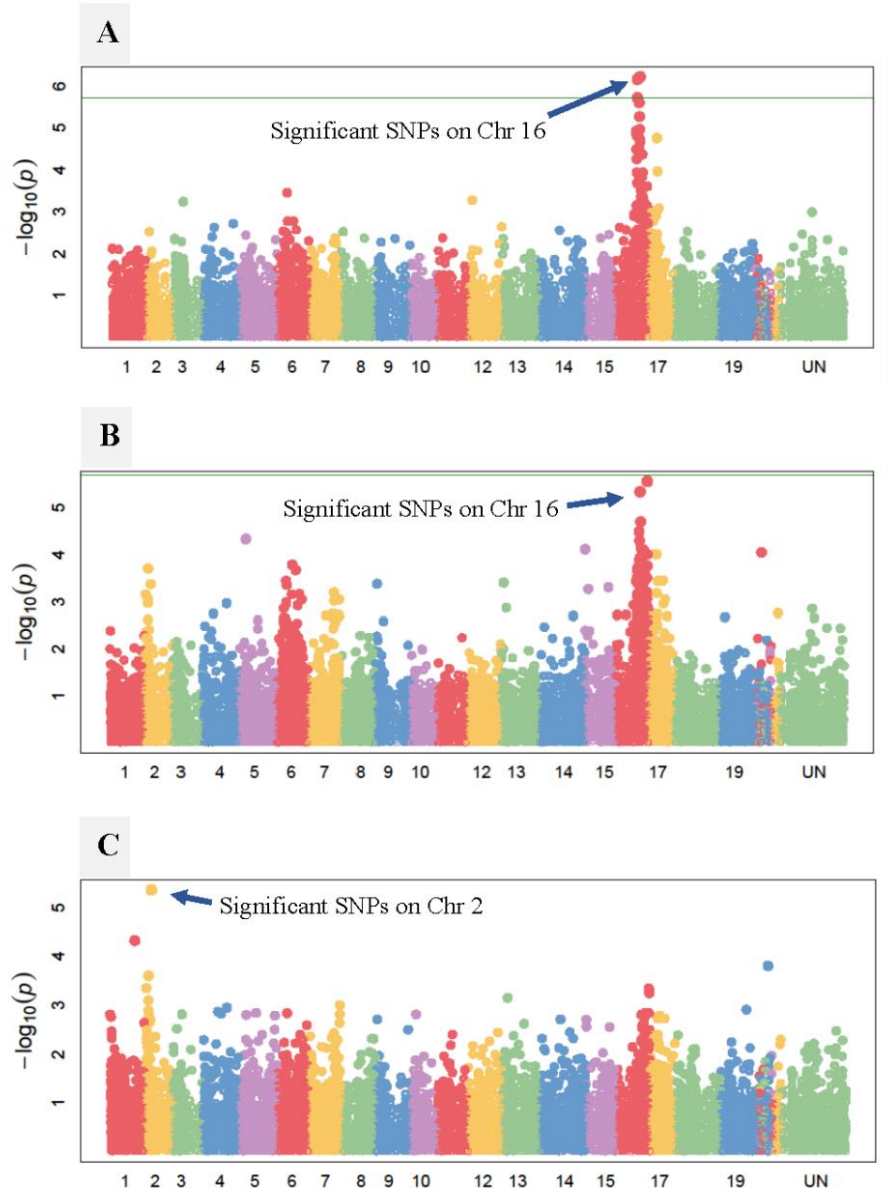

Figure 4. Manhattan plots of ‘°Brix in 2020’ using CMLM model. A) °Brix in harvest one of year 2020. B) °Brix in harvest two of year 2020. and C) °Brix in harvest three of year 2020. Green horizontal line indicating the threshold cutoff -log10 p value.

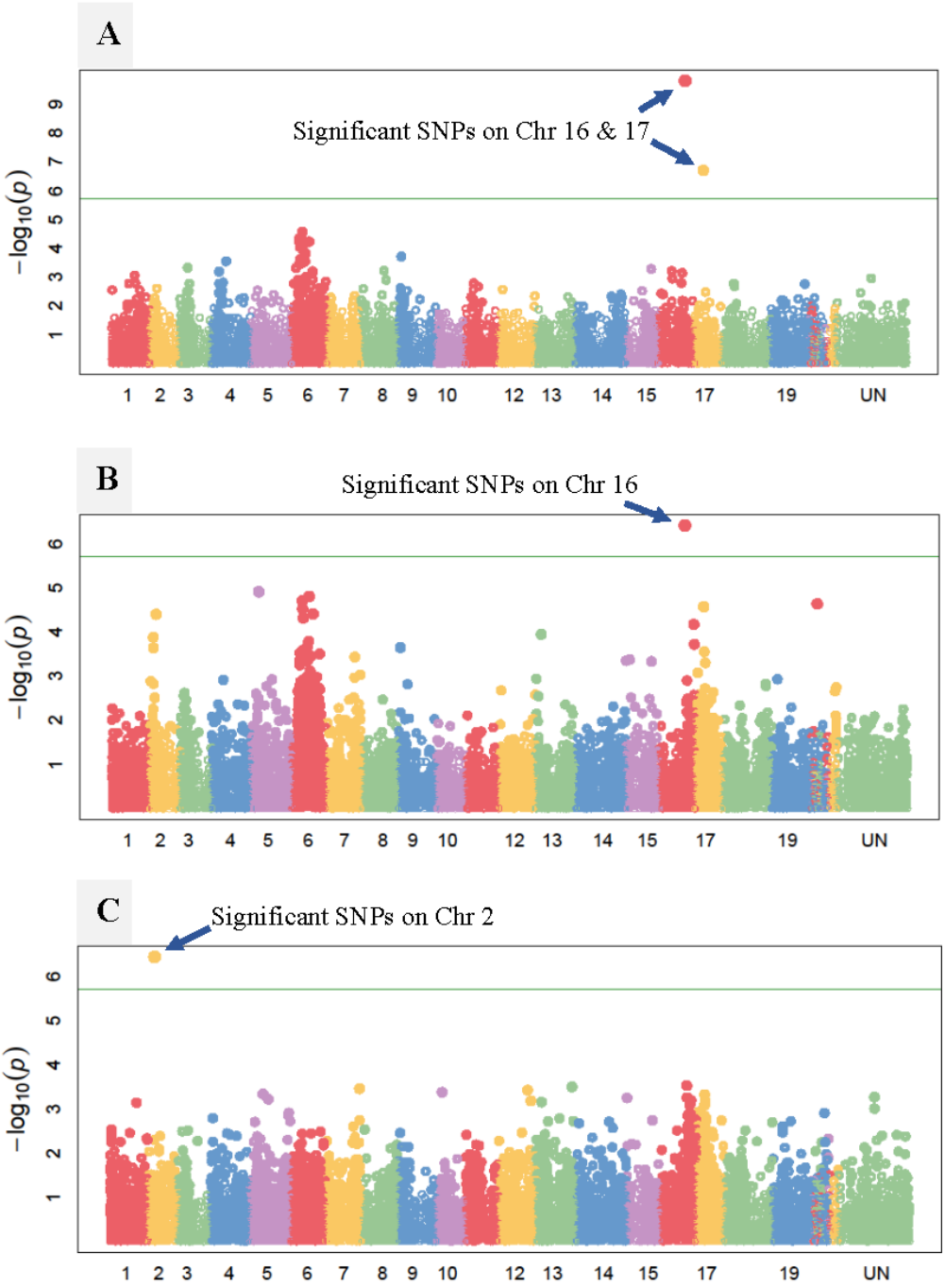

Figure 5. Manhattan plots of '°Brix in 2020' using MLMM model. A) °Brix in harvest one of year 2020. B) °Brix in harvest two of year 2020. C) °Brix in harvest three of year 2020. Green horizontal line indicating the threshold cutoff -log10 p value.

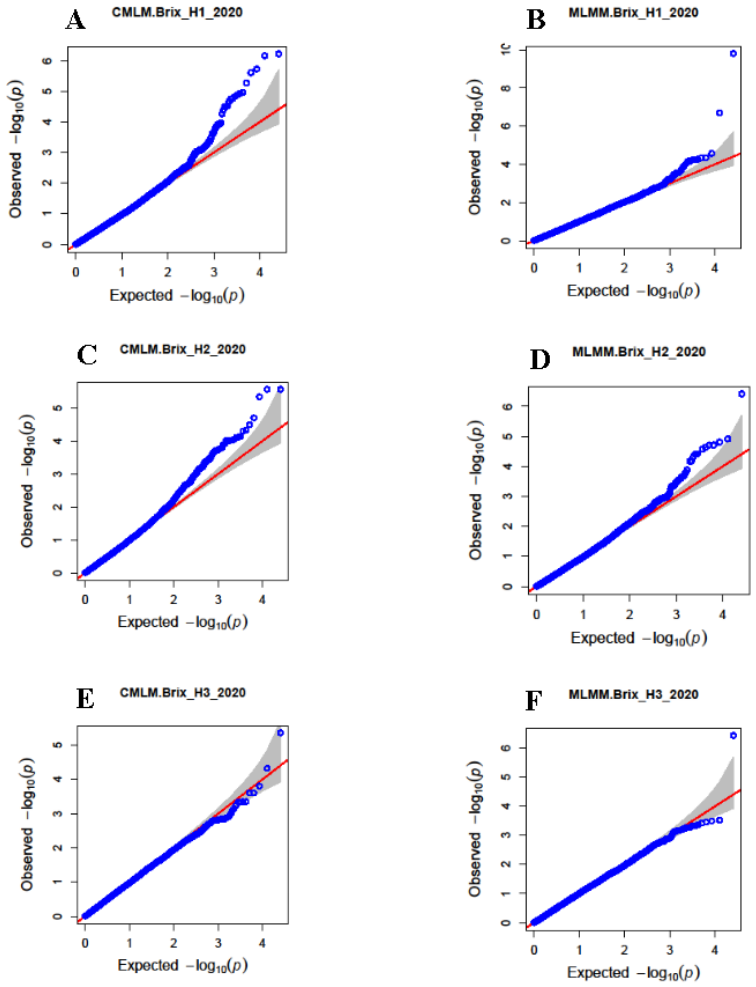

Figure 6. Q-Q plots of °Brix in 2020. A) CMLM of °Brix harvest one phenotypic distribution. B) MLMM of °Brix harvest one phenotypic distribution. C) CMLM of °Brix harvest two phenotypic distribution. D) MLMM of °Brix harvest two phenotypic distribution. E) CMLM of °Brix harvest three phenotypic distribution. F) MLMM of °Brix harvest three phenotypic distribution.

The QQ plot is a graphical representation of the deviation of the observed *P* values from the null hypothesis. Observed P values were more significant than expected for some SNPs under the null hypothesis. Those significant SNPs were moved away from the line, which means they were associated with the trait. SNPs in the MLMM model moved further away from the line than the CMLM model, resulting in a high significance level (Figure 6).

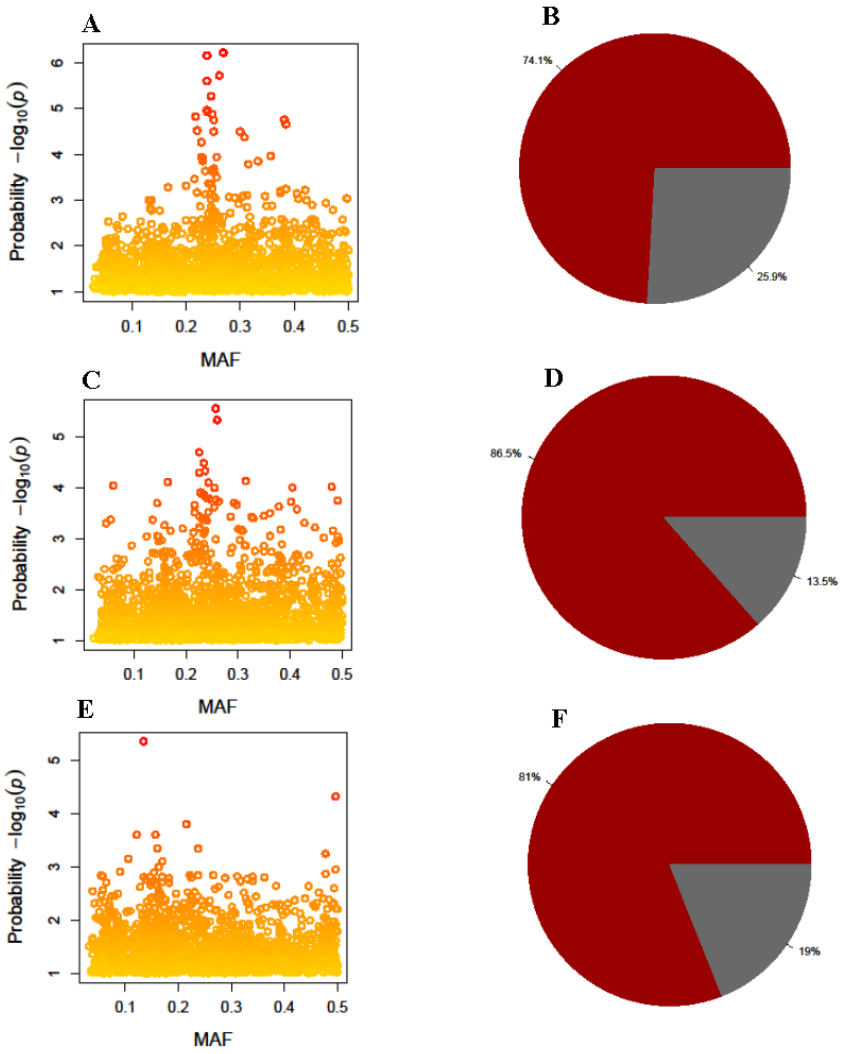

Figure 7. Minor allele frequency and heritability plots of the population for °Brix in year 2020. A) MAF of °Brix harvest one phenotypic distribution. B) Heritability of °Brix harvest one phenotypic distribution. C) MAF of °Brix harvest two phenotypic distribution. D) Heritability of °Brix harvest two phenotypic distribution. E) MAF of °Brix harvest three phenotypic distribution. F) Heritability of °Brix harvest three phenotypic distribution. Red shaded area in the pie chart indicating the heritability of the trait. *MAF = Minor allele frequency.

The °Brix phenotype data in 2020 showed substantial heritability estimates ranging from 0.74 to 0.86 across different harvests. This indicated that the trait was controlled to a high degree by genetic variance rather than environmental influences (Figure 7 and Table 9).

Table 9. Heritability estimates for °Brix in the year 2020.

Trait	Year	Harvest	Genetic variance (σ^2_G)	Residual variance (σ^2_e)	$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_e}$
°Brix	2020	Harvest 1	24.92	8.7	0.741
		Harvest 2	17.39	2.71	0.865
		Harvest 3	10.27	2.41	0.81

Note: H^2 = Broad sense heritability.

GWAS analysis of '°Brix in the year 2021'

In 2021, phenotypic data from 404, 386, and 378 individuals with GBS markers from harvests one, two, and three were included in the GWAS study (Table 4). The GWAS analysis of °Brix in 2021 was performed using the same two models, CMLM and MLMM, as earlier described. Similar to 2020 results, the CMLM model found a significant association between markers and the trait on chromosome 16. Harvest one results were more significant than the rest of the harvests, which exhibited SNPs at a higher log-likelihood p-value of 1.48E-13. Also, a significant number of SNPs in Harvest one surpassed the cutoff $-\log_{10}$ p-value threshold (Table 10). The GWAS analysis of harvest two was also identified as a significant association in the same region on chromosome 16. However, the level of the most significant SNP was slightly less than harvest one, and, a slightly smaller number of SNPs were able to surpass the $-\log_{10}$ p-value cutoff (Table 11). The association was still exhibited on the same chromosome in harvest three, unlike the 2020 results where the association was shifted to chromosome 2. Compared to the first two harvests the log-likelihood p-value of significant SNPs was reduced by half, and only two SNPs were able to surpass the cutoff $-\log_{10}$ p-value in harvest three (Table 12 and Figure 8).

Analysis using the MLMM model produced the exact same results, but the log-likelihood p-value of the most significant SNPs improved tremendously for all three harvests. In MLMM, a

new association was found on chromosome 6 for the °Brix for harvest two, along with the stable association on chromosome 16 (Table 11 and Figure 9). A clear-cut deviation in the log-likelihood p-value of the SNPs was observed from the expected p-value in both models, which was accurately depicted in the Q-Q plots (Figure 10). The °Brix heritability estimates ranged from 78.6 to 83.2% across the three harvests of 2021 (Table 13 and Figure 11).

Table 10. Peak SNPs associated with °Brix in the incomplete-diallel population during harvest one of the year 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 1	CMLM	S16_14021985	16	14.02	1.65E-06	0.256	2.64224
		S16_14154471	16	14.15	3.83E-07	0.235	2.957013
		S16_14156209	16	14.15	8.24E-06	0.245	2.563899
		S16_14170272	16	14.17	7.27E-06	0.224	2.565626
		S16_14446661	16	14.44	8.63E-09	0.248	3.269763
		S16_14578110	16	14.57	1.93E-06	0.247	2.688043
		S16_14593121	16	14.59	7.49E-06	0.222	2.566139
		S16_14593159	16	14.59	8.09E-07	0.225	2.88612
		S16_14593162	16	14.59	2.82E-08	0.242	3.351063
		S16_14757290	16	14.75	4.32E-07	0.226	3.018021
		S16_14992417	16	14.99	2.36E-08	0.433	2.472278
		S16_15731027	16	15.73	9.98E-09	0.246	3.21474
		S16_15855853	16	15.85	1.42E-05	0.368	1.993878
		S16_15855878	16	15.85	1.74E-05	0.283	1.966421
		S16_15872050	16	15.87	5.01E-11	0.232	3.86917
		S16_15872893	16	15.87	5.03E-06	0.5	1.670673
		S16_15991560*	16	15.99	1.48E-13	0.237	4.332953
		S16_15992710	16	15.99	5.41E-08	0.246	3.134357
		S16_15992931	16	15.99	1.64E-09	0.242	3.704784
		S16_16345474	16	16.34	8.97E-09	0.259	3.453306
	MLMM	S16_15991560*	16	15.99	2.15E-16	0.237	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = centimorgan, NA = Not available.

Table 11. Peak SNPs associated with °Brix in the incomplete-diallel population during harvest two of the year 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 2	CMLM	S16_14021985	16	14.02	1.16E-05	0.253	2.640424
		S16_14022036	16	14.02	1.39E-05	0.363	2.388652
		S16_14156209	16	14.15	2.42E-05	0.244	2.600947
		S16_14446661	16	14.44	2.63E-08	0.253	3.371863
		S16_14578110	16	14.57	1.49E-06	0.246	2.895624
		S16_14593121	16	14.59	7.98E-07	0.229	3.021784
		S16_14593159	16	14.59	6.04E-07	0.229	3.08065
		S16_14593162	16	14.59	6.36E-07	0.246	3.10455
		S16_14992417	16	14.99	1.16E-06	0.440	2.291631
		S16_15731027	16	15.73	5.43E-07	0.247	2.969624
		S16_15855853	16	15.85	7.62E-06	0.379	2.189805
		S16_15872050	16	15.87	7.60E-07	0.237	3.040421
		S16_15991560*	16	15.99	9.60E-09	0.240	3.533564
		S16_15992710	16	15.99	1.60E-06	0.251	2.935211
		S16_15992931	16	15.99	2.68E-07	0.246	3.288058
		S16_16345474	16	16.34	4.11E-08	0.262	3.519691
	MLMM	S16_15991560*	16	15.99	1.02E-09	0.240	NA
		S16_20841143	16	20.84	1.06E-06	0.386	NA
		S6_8828811	6	8.82	8.00E-08	0.418	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

Table 12. Peak SNPs associated with °Brix in the incomplete-diallel population during harvest three of the year 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 3	CMLM	S16_15731027*	16	15.73	3.18E-07	0.252	2.257613
		S16_15991560	16	15.99	3.71E-05	0.246	1.837448
		S16_15992931	16	15.99	9.83E-07	0.255	2.297578
	MLMM	S16_15731027*	16	15.73	1.16E-07	0.252	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

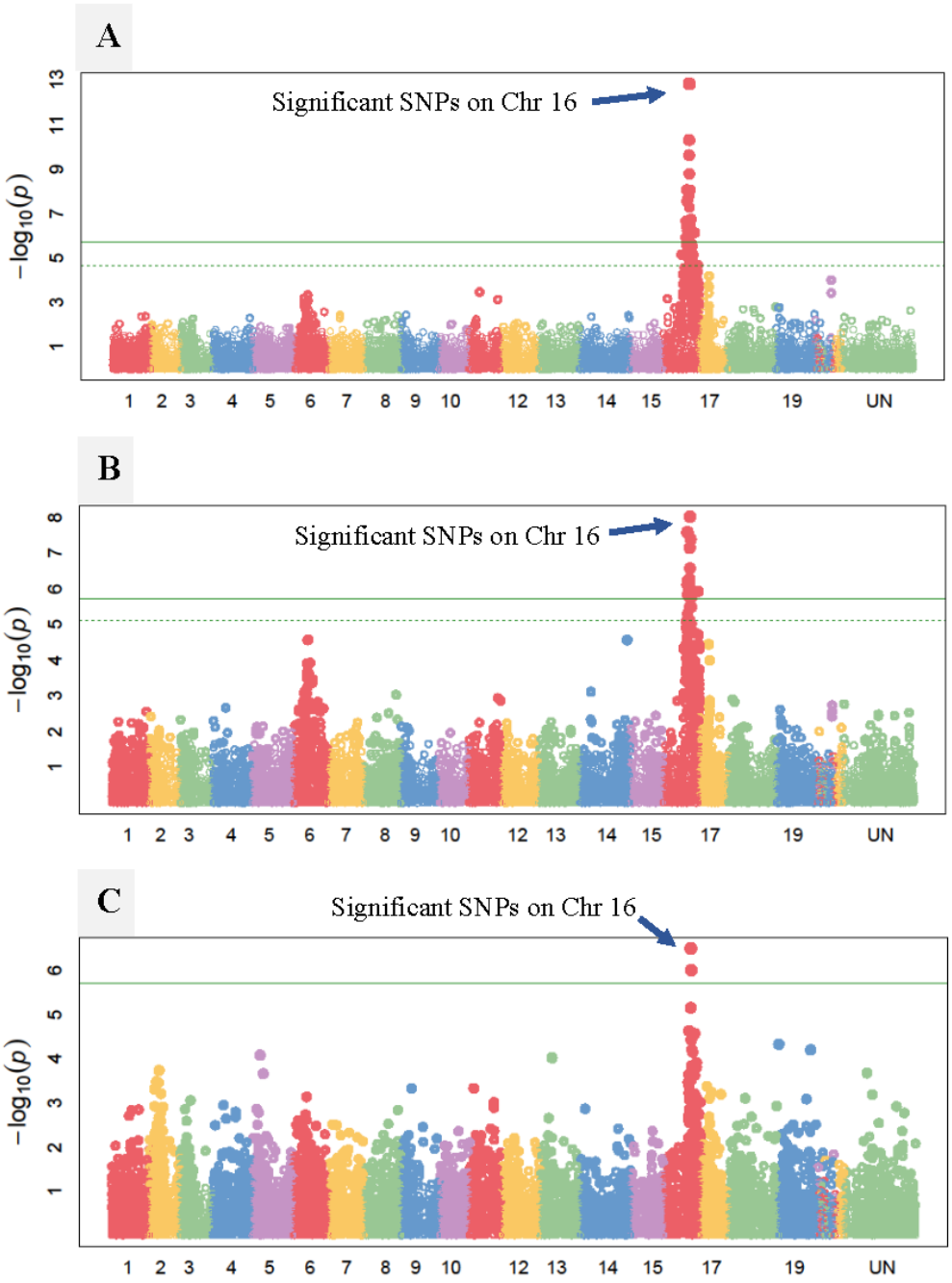

Figure 8. Manhattan plots of ‘°Brix in 2021’ using CMLM model. A) °Brix in harvest one of year 2021. B) °Brix in harvest two of year 2021. C) °Brix in harvest three of year 2021. Green horizontal line indicating the threshold cutoff -log10 p value.

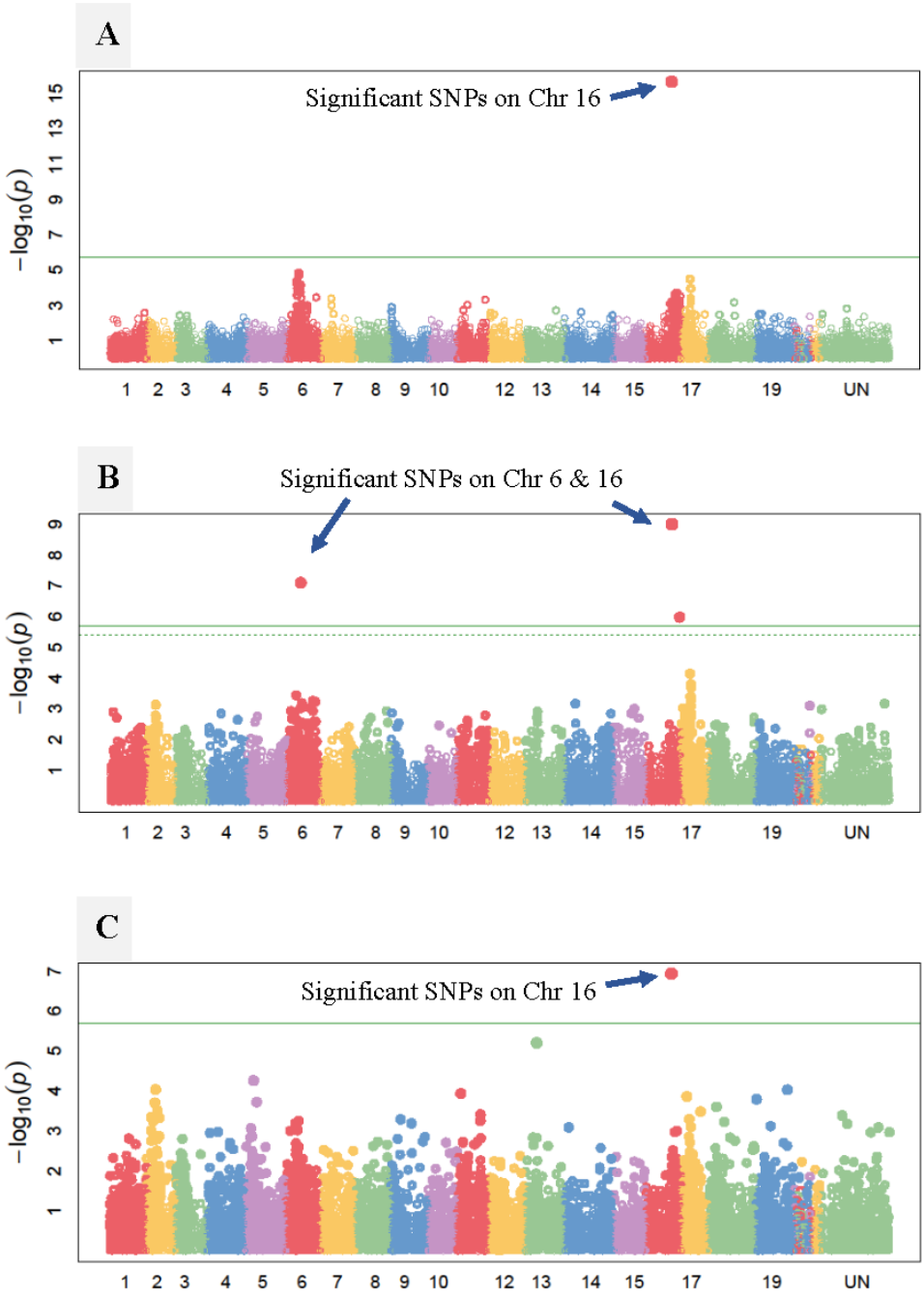

Figure 9. Manhattan plots of '°Brix in 2021' using MLMM model. A) °Brix in harvest one of year 2021. B) °Brix in harvest two of year 2021. C) °Brix in harvest three of year 2021. Green horizontal line indicating the threshold cutoff -log10 p value.

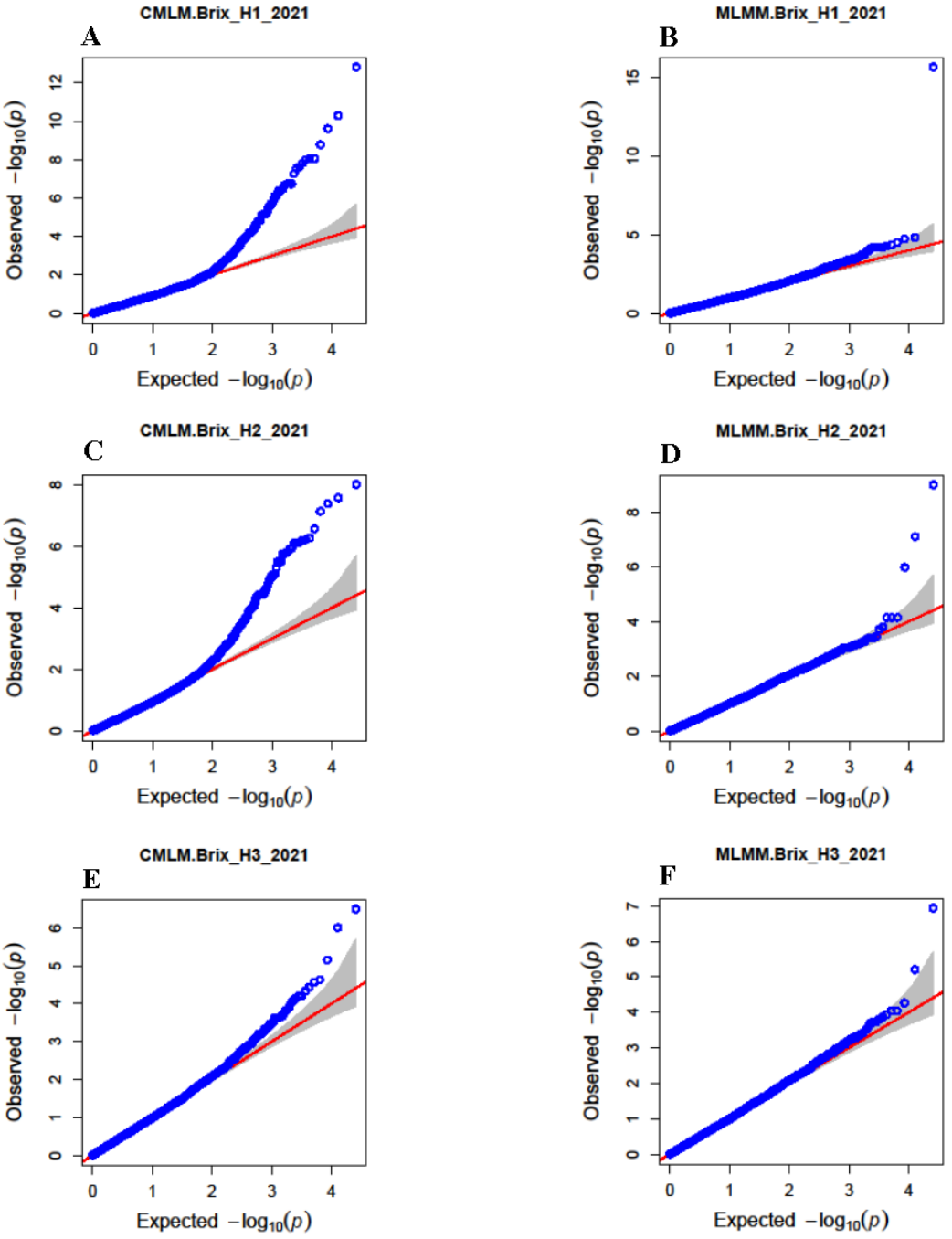

Figure 10. Q-Q plots of °Brix in 2021. A) CMLM of °Brix harvest one phenotypic distribution. B) MLMM of °Brix harvest one phenotypic distribution. C) CMLM of °Brix harvest two phenotypic distribution. D) MLMM of °Brix harvest two phenotypic distribution. E) CMLM of °Brix harvest three phenotypic distribution. F) MLMM of °Brix harvest three phenotypic distribution.

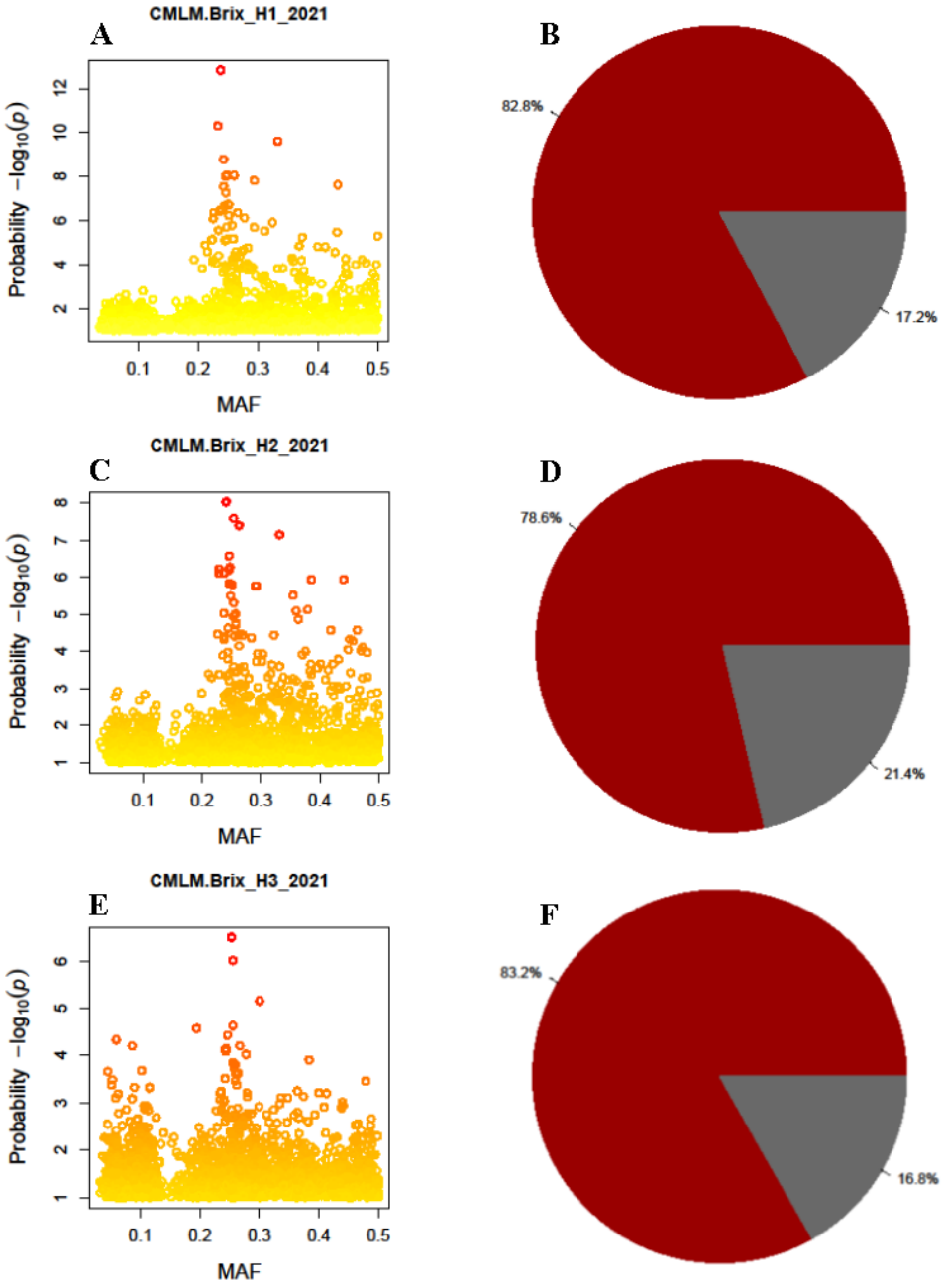

Figure 11. Minor allele frequency and heritability plots of the population for °Brix in year 2021. A) MAF of °Brix harvest one phenotypic distribution. B) Heritability of °Brix harvest one phenotypic distribution. C) MAF of °Brix harvest two phenotypic distribution. D) Heritability of °Brix harvest two phenotypic distribution. E) MAF of °Brix harvest three phenotypic distribution. F) Heritability of °Brix harvest three phenotypic distribution. Red shaded area in the pie chart indicating the heritability of the trait. *MAF = Minor allele frequency.

Table 13. Heritability estimates for °Brix in the year 2021.

Trait	Year	Harvest	Genetic variance (σ^2_G)	Residual variance (σ^2_e)	$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_e}$
°Brix	2021	Harvest 1	26.47	5.49	0.828
		Harvest 2	27.22	7.42	0.786
		Harvest 3	15.61	3.15	0.832

Note: H^2 = Broad sense heritability.

Trait 'pH'

Fruit pH phenotypic distribution in 2020 ranged from 1.7 to 4.62 across various harvests, whereas it ranged from 2.39 to 3.96 in 2021 (Table 14). Fruit pH phenotypic data in both cropping seasons was highly correlated within different harvests and years (Table 15, 16, and 17). In 2021, mean pH values were slightly higher than mean pH values for 2020 during harvests one and two, while the mean pH value for the third harvest, was higher in 2020. In both years, mean pH values increased to level desired for winemaking by the later harvests. During the first, second, and third harvests of 2020, the mean pH levels were 2.26, 2.35, and 3.34, respectively (Table 14 and Figure 12). In 2021, mean pH values of harvest one, two, and three were 2.73, 2.92, and 3.10, respectively (Table 14 and Figure 13). About 92 % and 73 % of the accessions had pH values greater than 3.0 after harvest maturity in 2020 and 2021, respectively.

Table 14. Summary statistics of trait pH.

Trait	Year	Harvest	N		Mean	Maximum	Median	Minimum
			Total	IWG				
pH	2020	1	268	195	2.26	2.83	2.24	1.8
		2	237	173	2.35	2.95	2.38	1.7
		3	204	156	3.34	4.62	3.23	2.7
	2021	1	565	403	2.73	3.96	2.69	2.46
		2	535	386	2.92	3.8	2.9	2.39
		3	521	378	3.10	3.7	3.1	2.57

Note: N = Number of individuals sampled, IWG = Individuals with GBS markers.

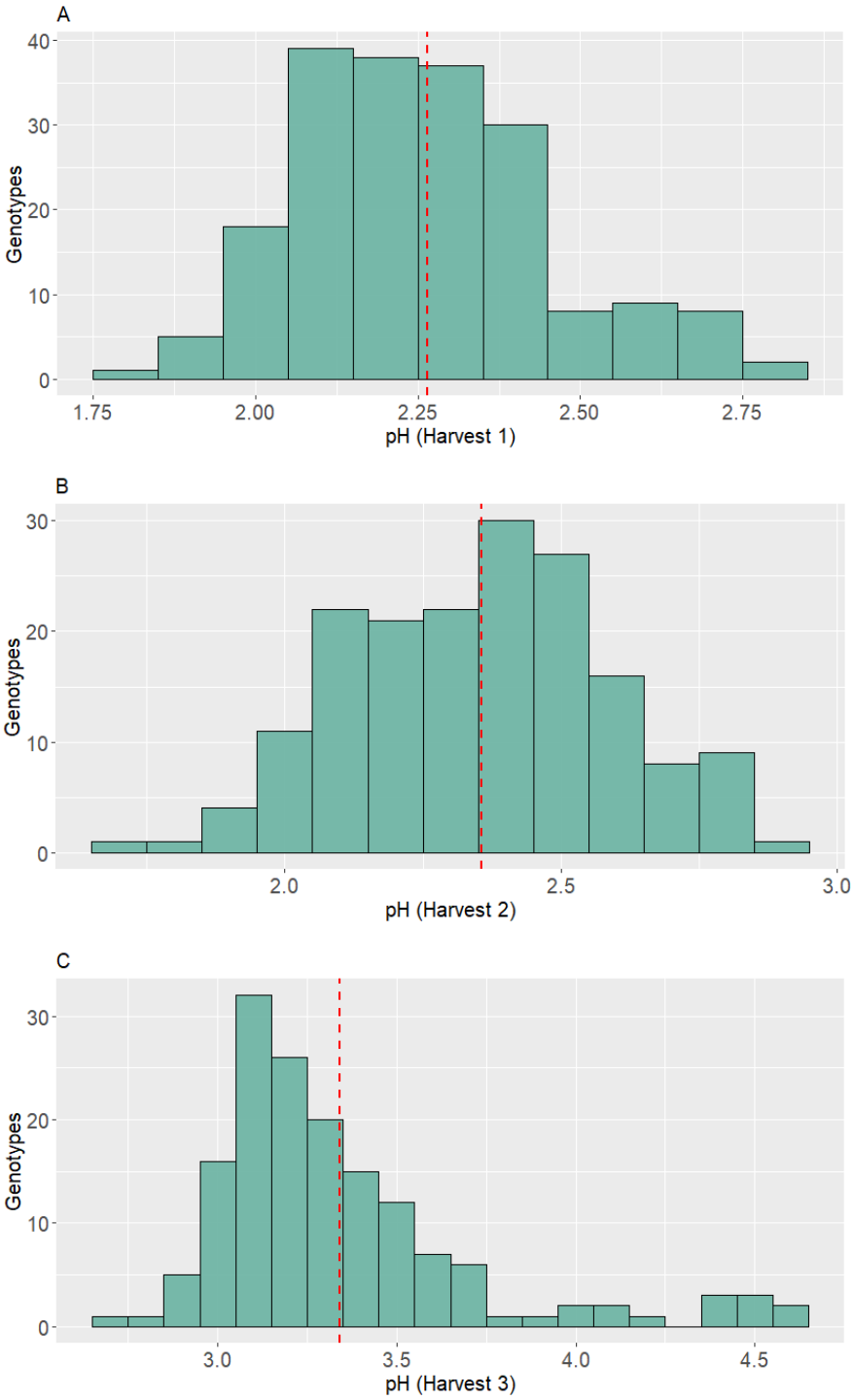

Figure 12. Histogram showing the phenotypic distribution of pH in year 2020. A) Histogram of pH in harvest one. B) Histogram of pH in harvest two. C) Histogram of pH in harvest three. Dashed vertical red line indicating the mean value of the trait.

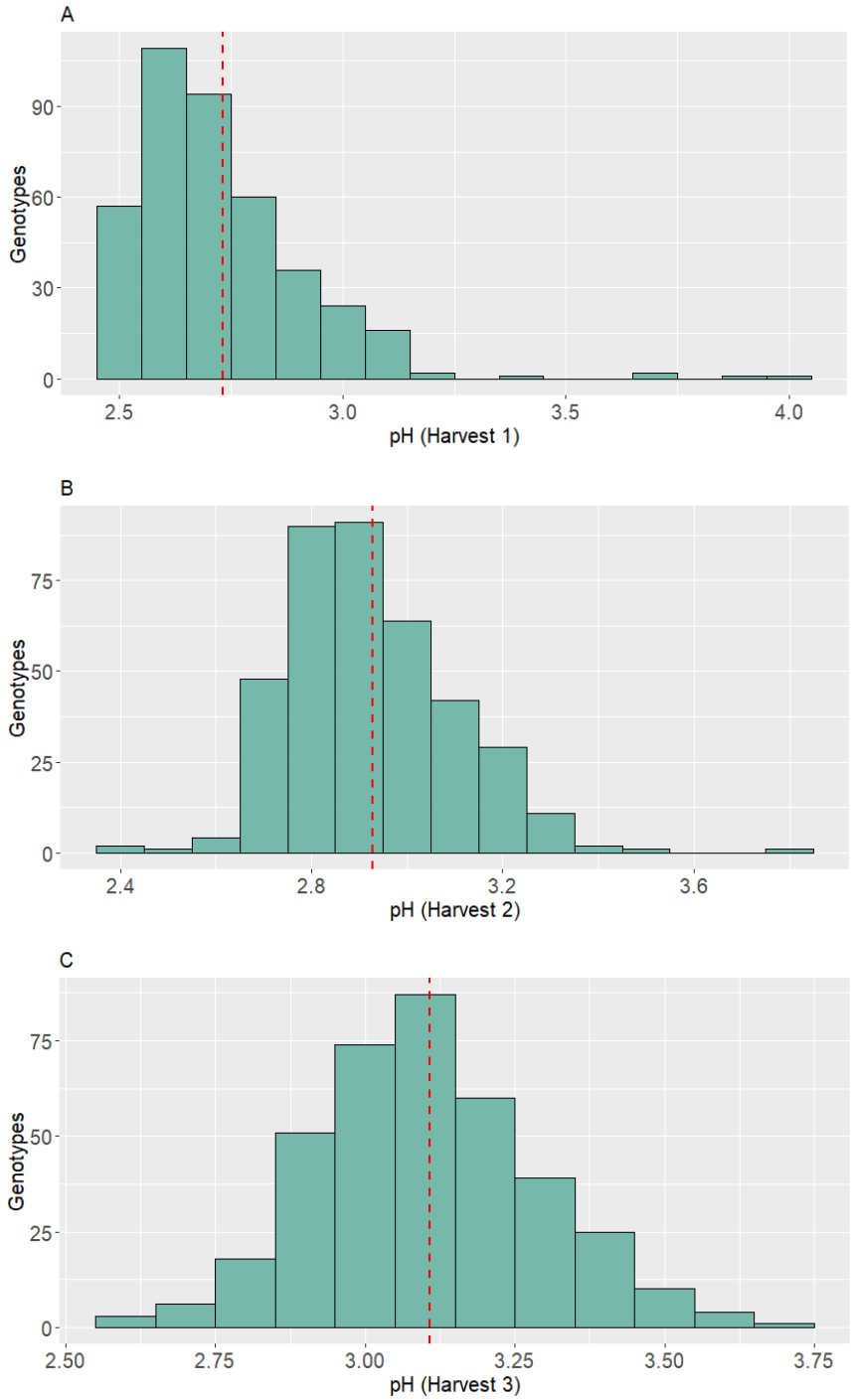

Figure 13. Histogram showing the phenotypic distribution of pH in year 2021. A) Histogram of pH in harvest one. B) Histogram of pH in harvest two. C) Histogram of pH in harvest three. Dashed vertical red line indicating the mean value of the trait.

Table 15. Pearson's correlation coefficient and significant estimates for pH in the year 2020.

Year		Harvest 2	Harvest 3
2020	Harvest 1	0.870 ***	0.684 **
	Harvest 2		0.772 ***

Note: P-value 0.05*, 0.01**, 0.001***

Table 16. Pearson's correlation coefficient and significant estimates for pH in the year 2021.

Year		Harvest 2	Harvest 3
2021	Harvest 1	0.962 ***	0.901 ***
	Harvest 2		0.957 ***

Note: P-value 0.05*, 0.01**, 0.001***

Table 17. Pearson's correlation coefficient and significant estimates for pH between years.

Year		2020		
		Harvest 1	Harvest 2	Harvest 3
2021	Harvest 1	0.953 ***	0.826 ***	0.631 **
	Harvest 2	0.970 ***	0.840 ***	0.656 **
	Harvest 3	0.957 ***	0.854 ***	0.712 ***

Note: P-value 0.05*, 0.01**, 0.001***

GWAS analysis of 'pH in the year 2020'

For GWAS analysis of pH in 2020, phenotypic data from 195, 173, and 156 individuals with GBS markers from harvest one, two, and three were used, respectively (Table 14). As previously described, GWAS was performed using two different models, CMLM and MLMM. No significant associations above the threshold cutoff were discovered using the CMLM model in any of the three harvests. But in harvest one, some SNPs were placed almost close to the cutoff line with a log-likelihood p-value of 5 on chromosome 16, which is in the same interval as identified for °Brix earlier (Table 18 and Figure 14). In the MLMM model, a significant

association on chromosome 16 above the threshold cutoff $-\log_{10}$ p-value is identified for pH in harvest one. For the remaining two harvests, no significant associations were found in either model (Table 18 and Figure 15).

Except in the harvest one, observed p values of all the SNPs were in line with the expected p values in both models indicating no significant relationship between trait and markers. In harvest one, p values of some significant SNPs were deviated slightly from the expected range, indicating a somewhat relation between marker and phenotype (Figure 16). Heritability and MAF ranges were generally good for pH in 2020 harvests (Figure 17). Still, the failure to find a significant association even with the good heritability (H^2) and MAF is mainly due to the availability of limited phenotypic data and possible error associated with precocity.

Table 18. Peak SNPs associated with pH in incomplete-diallel population during growing season 2020 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 1	CMLM	S16_16975630*	16	16.97	4.47E-06	0.241	0.136743
		S16_16975634	16	16.97	4.47E-06	0.241	0.136743
	MLMM	S16_16975630*	16	16.97	4.29E-07	0.241	NA
Harvest 2	CMLM	No significant associations					
	MLMM	No significant associations					
Harvest 3	CMLM	No significant associations					
	MLMM	No significant associations					

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

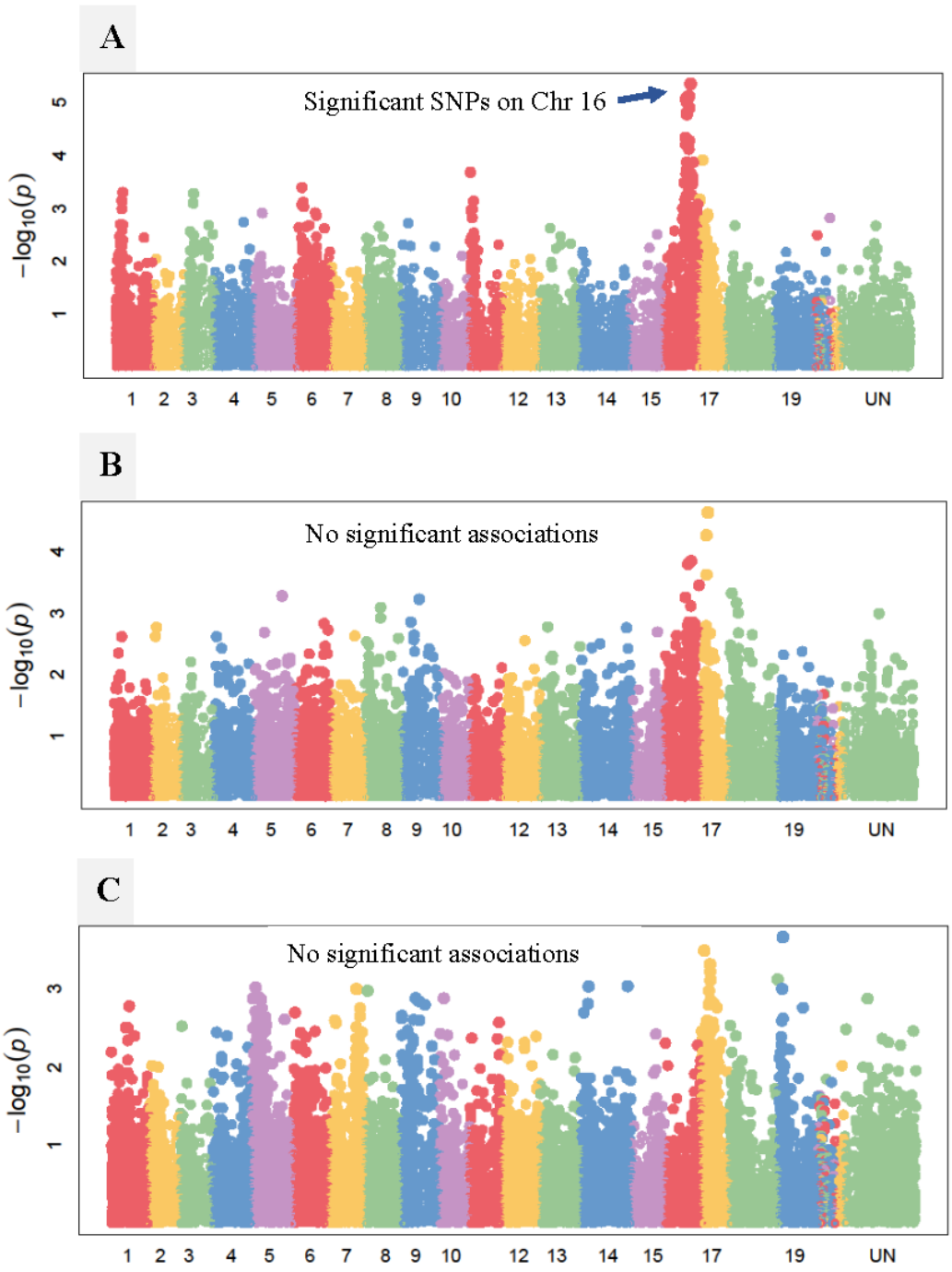

Figure 14. Manhattan plots of 'pH in 2020' using CMLM model. A) pH in harvest one of year 2020. B) pH in harvest two of year 2020. C) pH in harvest three of year 2020. Green horizontal line indicating the threshold cutoff -log10 p value.

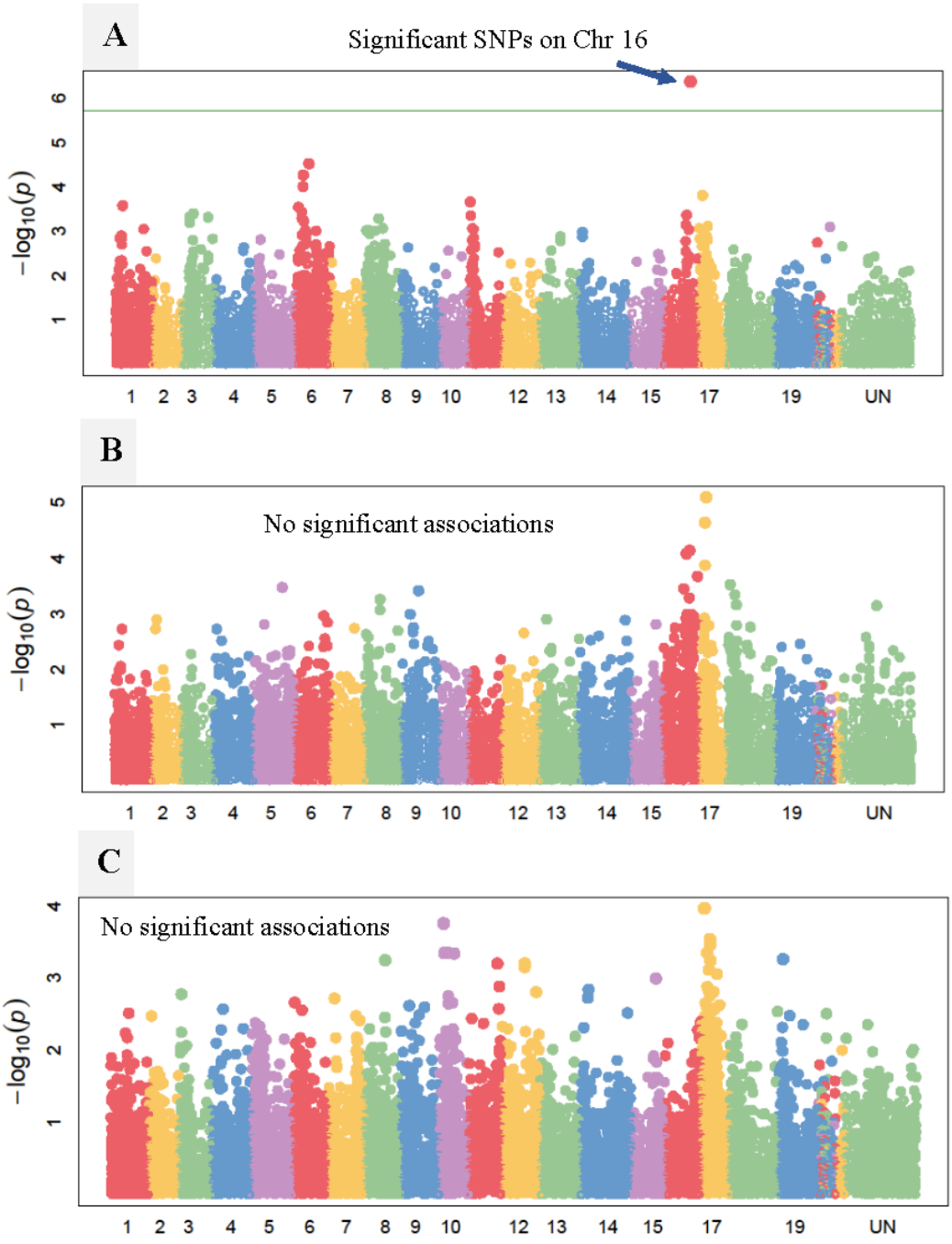

Figure 15. Manhattan plots of 'pH in 2020' using MLMM model. A) pH in harvest one of year 2020. B) pH in harvest two of year 2020. C) pH in harvest three of year 2020. Green horizontal line indicating the threshold cutoff -log10 p value.

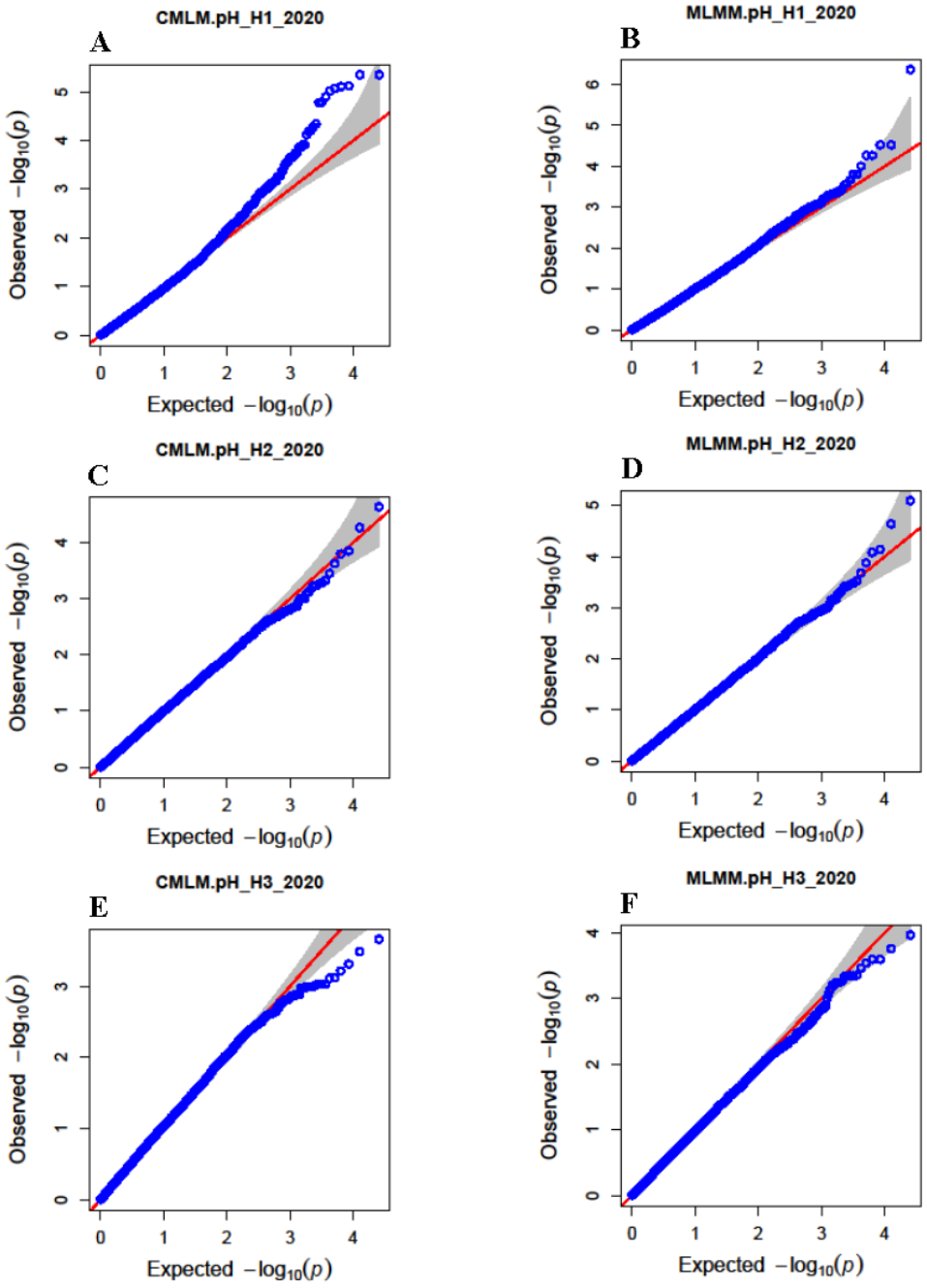

Figure 16. Q-Q plots of pH in 2020. A) CMLM of pH harvest one phenotypic distribution. B) MLMM of pH harvest one phenotypic distribution. C) CMLM of pH harvest two phenotypic distribution. D) MLMM of pH harvest two phenotypic distribution. E) CMLM of pH harvest three phenotypic distribution. F) CMLM of pH harvest three phenotypic distribution.

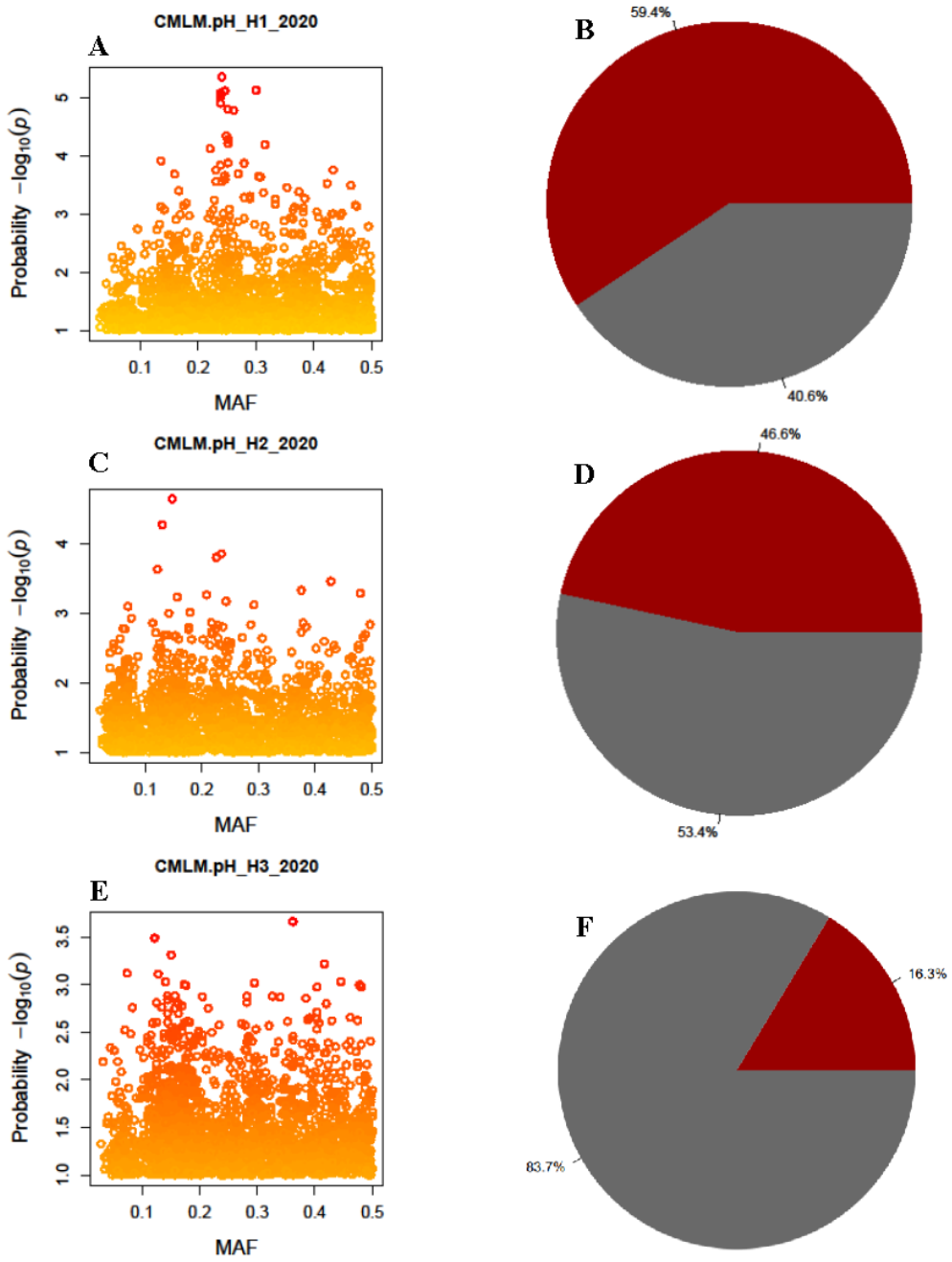

Figure 17. Minor allele frequency and heritability plots of the population for pH in year 2020. A) MAF of pH harvest one phenotypic distribution. B) Heritability of pH harvest one phenotypic distribution. C) MAF of pH harvest two phenotypic distribution. D) Heritability of pH harvest two phenotypic distribution. E) MAF of pH harvest three phenotypic distribution. F) Heritability of pH harvest three phenotypic distribution. Red shaded area in the pie chart indicating the heritability of the trait. *MAF = Minor allele frequency.

GWAS analysis of 'pH in the year 2021'

For GWAS analysis of pH in 2021, phenotypic data from 403, 386, and 378 individuals with GBS markers from harvest one, two, and three were used, respectively (Table 14). As previously described, GWAS was performed using two different models, CMLM and MLMM. Using the CMLM model, a significant association between pH and markers was found on chromosome 16 in all three harvests. This is a good improvement compared to the 2020 results, where both models failed to detect significant associations. This improvement is mainly due to the availability of phenotype from more individuals of the population. Even though log-likelihood p-values of significant SNPs were well above the cutoff $-\log_{10}$ p-value in all three harvests of 2021, SNPs in harvest two are most significant than the rest of the harvests. Log-likelihood p-values of the significant SNPs increased from harvest one to harvest two. Then decreased slightly in harvest three but still higher than harvest one (Table 19 to 21, and Figure 18).

The MLMM model GWAS analysis also produced the same results as CMLM by identifying the association on the same chromosomal region. Still, log-likelihood p-values of the significant SNPs improved tremendously in all three harvests. Along with the improved significance of SNPs on chromosome 16, an additional association was found on chromosome 6 for pH in harvest two. These results of pH are almost identical with °Brix in 2021 results (Table 20 and Figure 19).

Q-Q plots explained the apparent deviation in observed p values of SNPs from expected p values, affirming the association between significant SNPs with the trait pH (Figure 20). Heritability estimates of pH in 2021 were relatively high, ranging from 64% to 76.5%, showing a strong genetic influence on the trait (Figure 21).

Table 19. Peak SNPs associated with pH in incomplete-diallel population during harvest one of the year 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 1	CMLM	S16_14446661	16	14.44	1.68E-06	0.249	0.102227
		S16_15855853	16	15.85	1.62E-05	0.369	0.075189
		S16_15872050*	16	15.87	3.41E-08	0.233	0.121129
		S16_15991560	16	15.99	1.38E-06	0.238	0.10496
		S16_15992710	16	15.99	1.49E-05	0.246	0.094212
	MLMM	S16_15872050*	16	15.87	2.88E-09	0.233	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

Table 20. Peak SNPs associated with pH in incomplete-diallel population during harvest two of the year 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 2	CMLM	S16_14156209	16	14.15	5.27E-06	0.244	0.082463
		S16_14446661	16	14.44	1.97E-07	0.253	0.092566
		S16_14578110	16	14.57	2.27E-06	0.246	0.085025
		S16_14593159	16	14.59	3.00E-06	0.229	0.085779
		S16_14593162	16	14.59	2.99E-07	0.246	0.090815
		S16_14757290	16	14.75	3.74E-06	0.226	0.085669
		S16_14992417	16	14.99	3.85E-06	0.440	0.064296
		S16_15731027	16	15.73	3.54E-06	0.247	0.080624
		S16_15872050	16	15.87	4.45E-09	0.237	0.107436
		S16_15872893	16	15.87	5.30E-06	0.497	0.054057
		S16_15991560*	16	15.99	1.98E-10	0.240	0.115686
		S16_15992710	16	15.99	1.12E-07	0.251	0.096329
		S16_15992931	16	15.99	4.20E-08	0.246	0.104238
		S16_16345474	16	16.34	1.53E-08	0.262	0.107877
		S16_16918882	16	16.91	7.70E-07	0.243	0.087501
		S16_16918954	16	16.91	7.70E-07	0.243	0.087501
	MLMM	S16_15991560*	16	15.99	3.46E-15	0.240	NA
		S6_6641054	6	6.64	9.21E-07	0.265	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

Table 21. Peak SNPs associated with pH in incomplete-diallel population during harvest three of the year 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 3	CMLM	S16_14021985	16	14.02	2.00E-06	0.259	0.098265
		S16_14446661	16	14.44	9.99E-06	0.259	0.089473
		S16_15731027	16	15.73	1.24E-07	0.252	0.108309
		S16_15872050	16	15.87	3.88E-07	0.242	0.105602
		S16_15991560	16	15.99	2.69E-08	0.246	0.115662
		S16_15992710	16	15.99	3.19E-07	0.255	0.108039
		S16_15992931	16	15.99	3.04E-08	0.255	0.12056
		S16_16345474*	16	16.34	6.11E-09	0.267	0.128089
	MLMM	S16_16345474*	16	16.34	4.04E-10	0.267	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

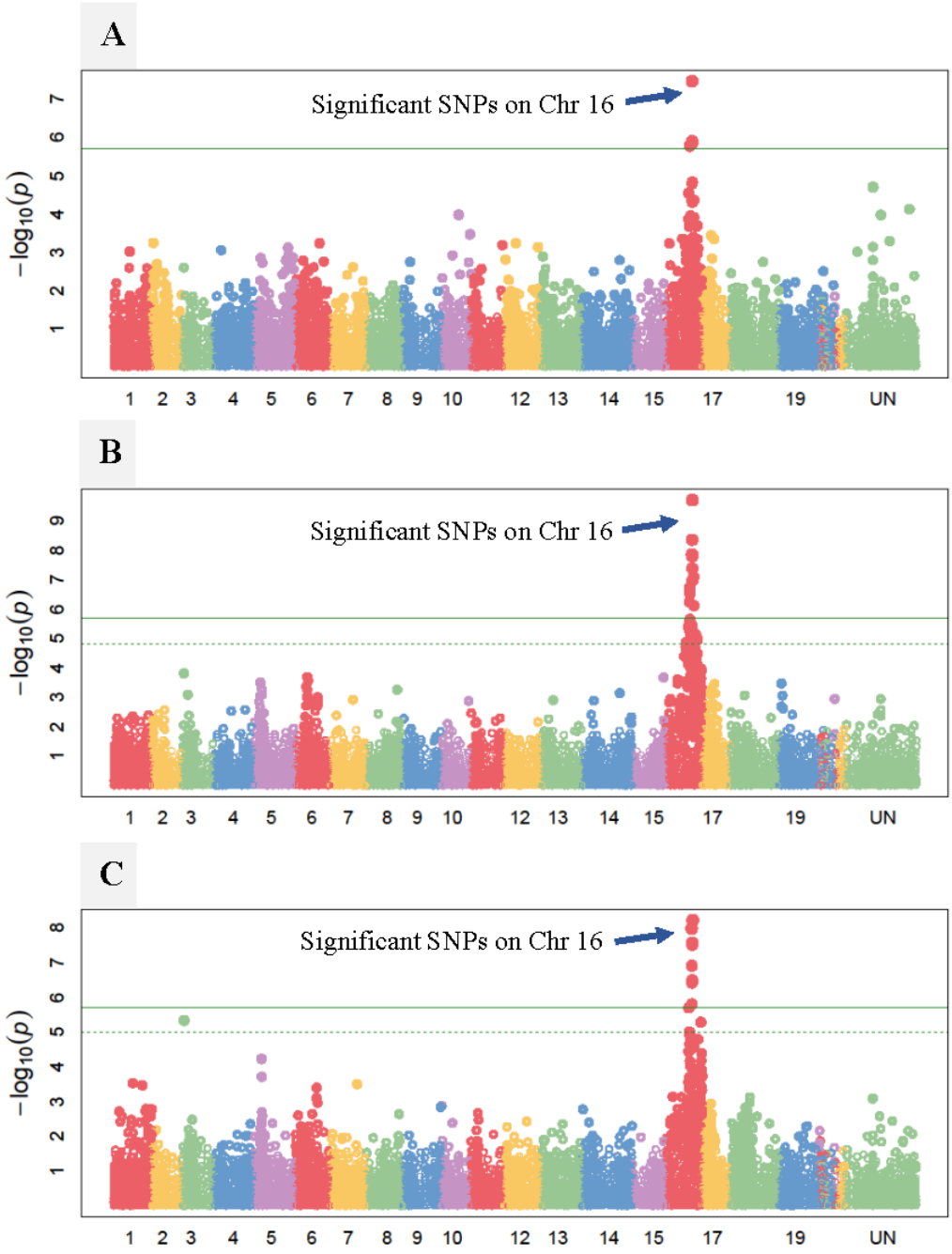

Figure 18. Manhattan plots of 'pH in 2021' using CMLM model. A) pH in harvest one of year 2021. B) pH in harvest two of year 2021. C) pH in harvest three of year 2021. Green horizontal line indicating the threshold cutoff -log10 p value.

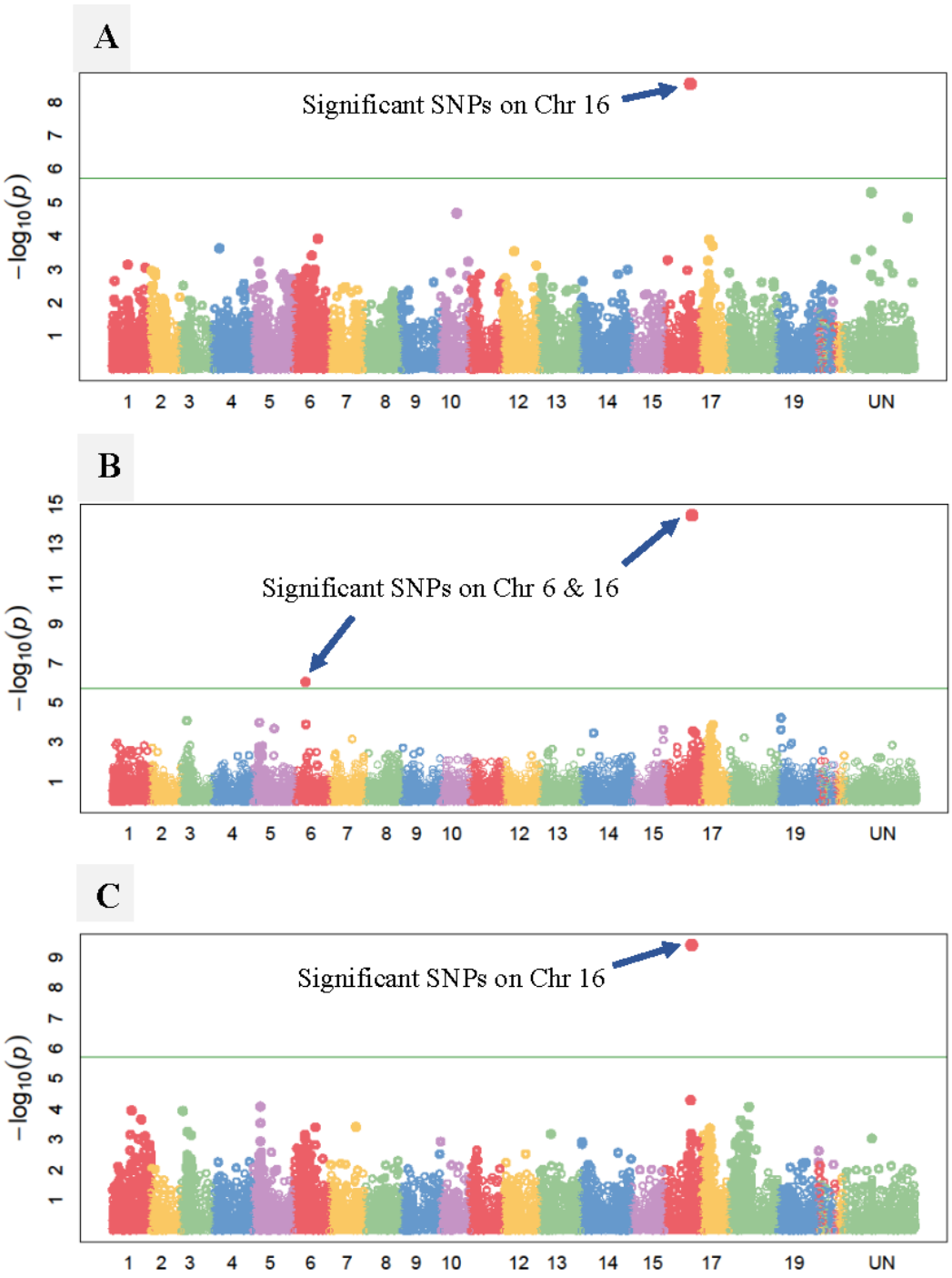

Figure 19. Manhattan plots of 'pH in 2021' using MLMM model. A) pH in harvest one of year 2021. B) pH in harvest two of year 2021. C) pH in harvest three of year 2021. Green horizontal line indicating the threshold cutoff -log10 p value.

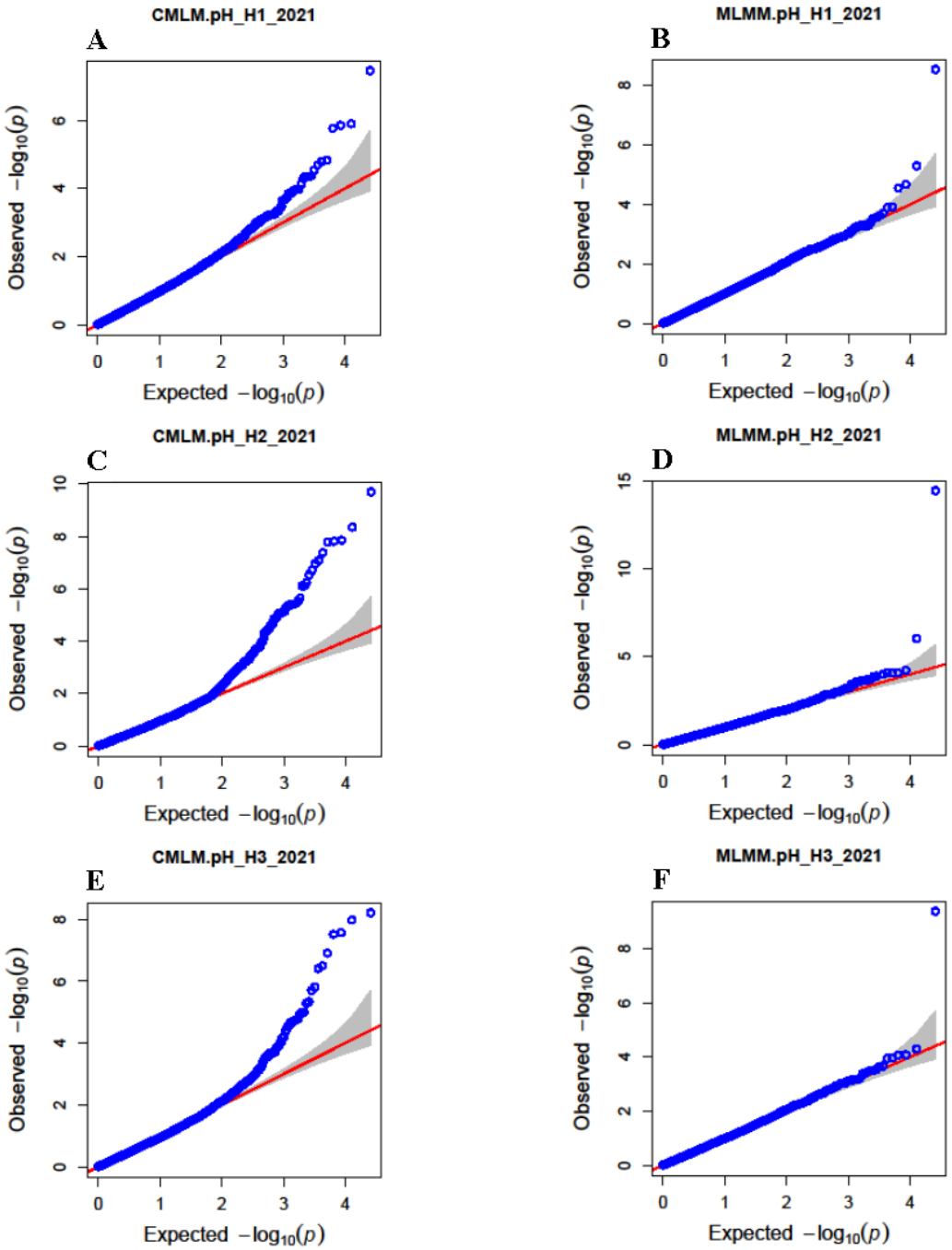

Figure 20. Q-Q plots of pH in 2021. A) CMLM of pH harvest one phenotypic distribution. B) MLMM of pH harvest one phenotypic distribution. C) CMLM of pH harvest two phenotypic distribution. D) MLMM of pH harvest two phenotypic distribution. E) CMLM of pH harvest three phenotypic distribution. F) CMLM of pH harvest three phenotypic distribution.

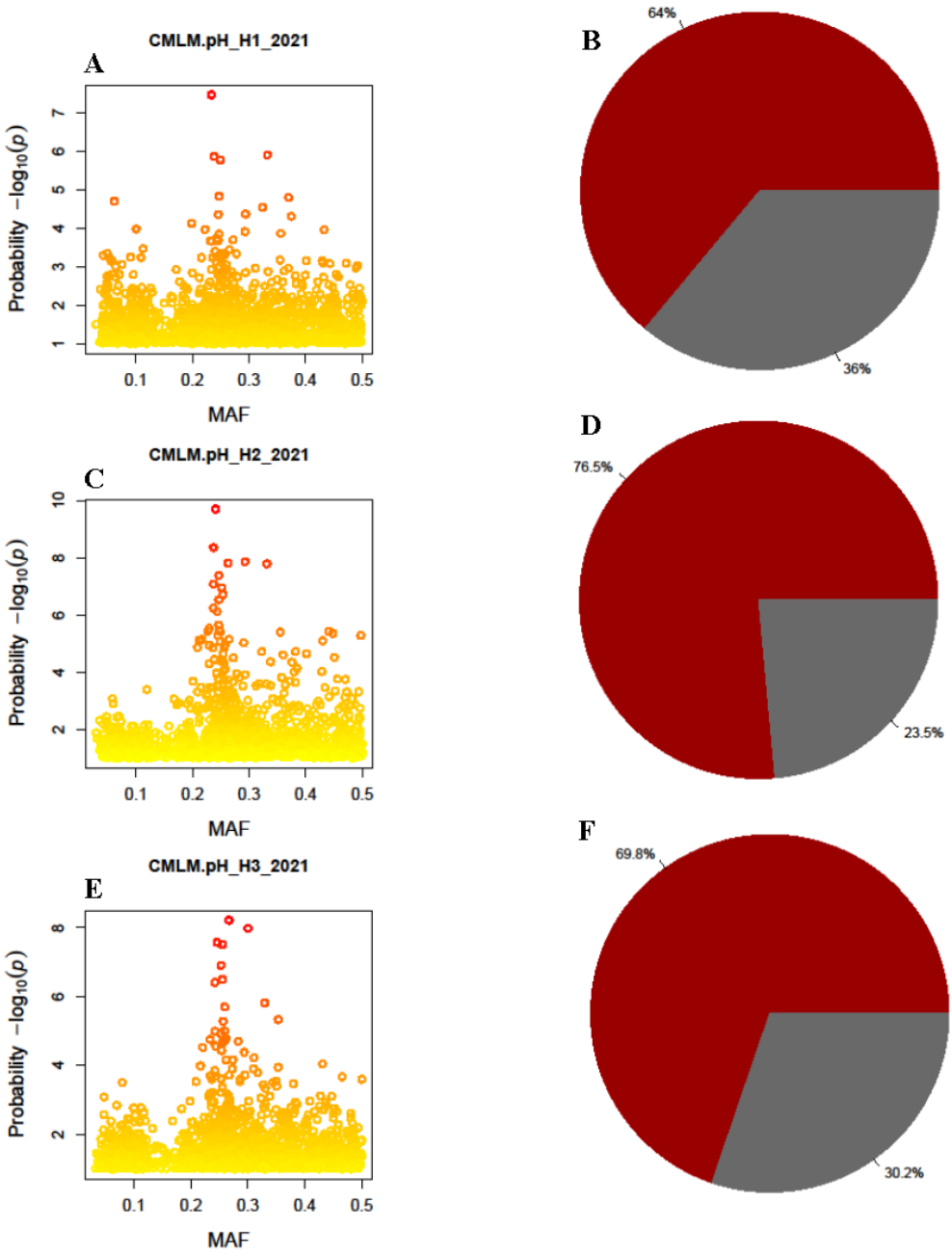

Figure 21. Minor allele frequency and heritability plots the population for pH in 2021 growing season. A) MAF of pH harvest one phenotypic distribution. B) Heritability of pH harvest one phenotypic distribution. C) MAF of pH harvest two phenotypic distribution. D) Heritability of pH harvest two phenotypic distribution. E) MAF of pH harvest three phenotypic distribution. F) Heritability of pH harvest three phenotypic distribution. Red shaded area in the pie chart indicating the heritability value of the trait. *MAF = Minor allele frequency.

Trait 'TA'

Phenotypic distribution of TA in the year 2020 ranging from 0.97 to 5.39 across different harvests (Table 22). In 2021, TA values ranged from 0.57 to 6.82. TA phenotypic data in both growing seasons were highly correlated within different harvests and years (Table 23 to 25). In 2021, mean TA values were slightly higher than 2020 except in harvest three. The mean TA values decreased with time and later harvests in both years. During the first, second, and third harvests of 2020, the mean TA levels were 2.88, 2.57, and 1.96, respectively (Figure 22). In 2021 mean TA values of harvest one, two, and three were 3.24, 2.62, and 1.83, respectively (Table 22 and Figure 23). In 2020 only three individuals had a TA value less than one, whereas, in 2021, approximately 30 individual vines had a TA value less than one.

Table 22. Summary statistics of trait TA.

Trait	Year	Harvest	N		Mean	Maximum	Median	Minimum
			Total	IWG				
TA	2020	1	268	195	2.88	5.26	2.85	1.4
		2	237	173	2.57	5.39	2.54	0.98
		3	204	156	1.96	3.76	1.86	0.97
	2021	1	565	402	3.24	6.82	3.18	1.19
		2	535	382	2.62	5.86	2.52	0.68
		3	521	378	1.83	4.66	1.75	0.57

Table 23. Pearson's correlation coefficient and significant estimates for TA in the year 2020.

Year		Harvest 2	Harvest 3
2020	Harvest 1	0.910 ***	0.757 ***
	Harvest 2		0.769 ***

Note: P-value 0.05*, 0.01**, 0.001***

Table 24. Pearson's correlation coefficient and significant estimates for TA in the year 2021.

Year		Harvest 2	Harvest 3
2021	**Harvest 1**	0.866 ***	0.808 ***
	Harvest 2		0.806 ***

Note: P-value 0.05*, 0.01**, 0.001***

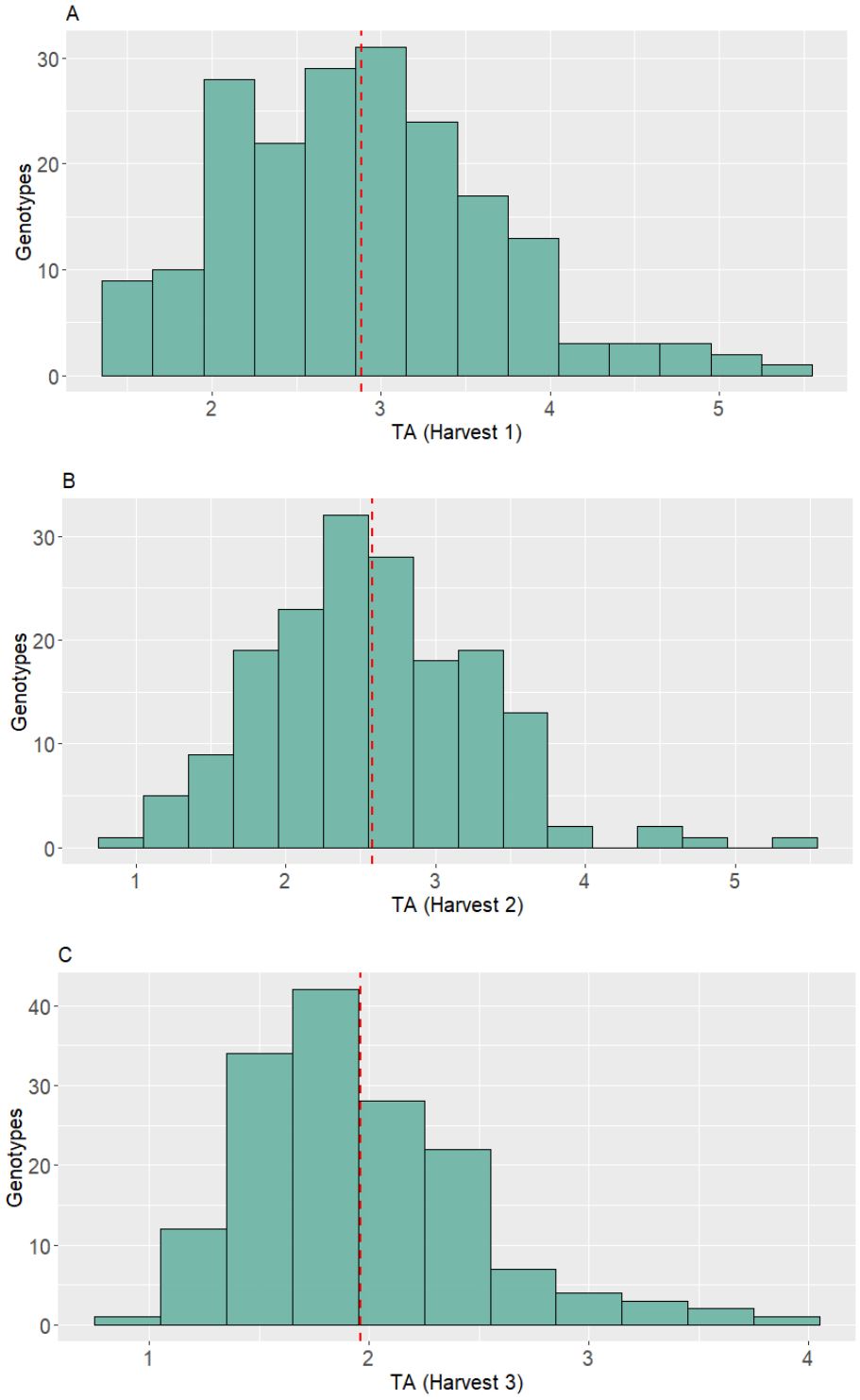

Figure 22. Histogram showing the phenotypic distribution of TA in year 2020. A) Histogram of TA in harvest one B) Histogram of TA in harvest two C) Histogram of TA in harvest three. Dashed vertical red line indicating the mean value of the trait.

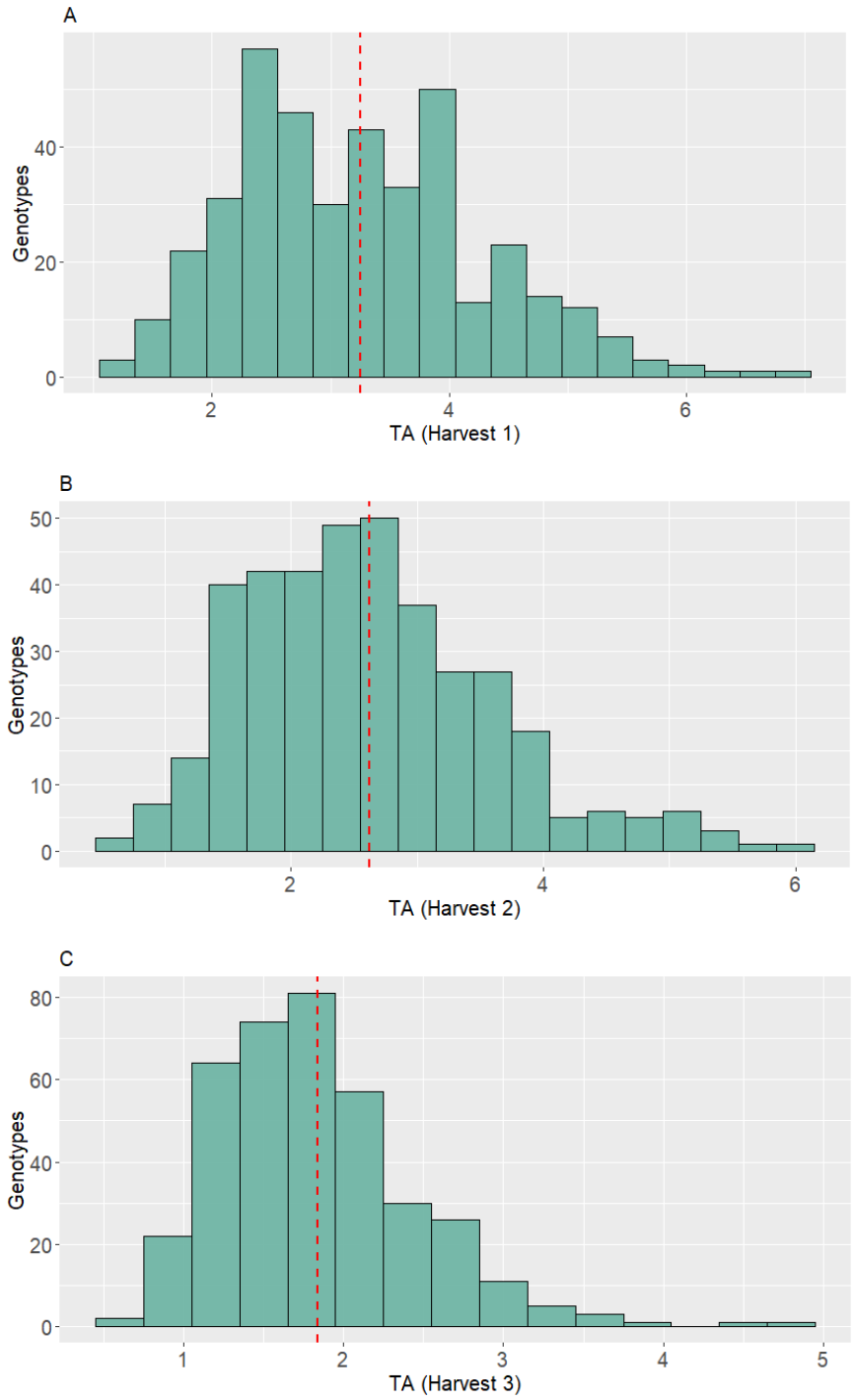

Figure 23. Histogram showing the phenotypic distribution of TA in year 2021. A) Histogram of TA in harvest one B) Histogram of TA in harvest two C) Histogram of TA in harvest three. Dashed vertical red line indicating the mean value of the trait.

Table 25. Pearson's correlation coefficient and significant estimates for TA between years.

Year		**2020**		
		Harvest 1	**Harvest 2**	**Harvest 3**
2021	**Harvest 1**	0.912 ***	0.865 ***	0.745 ***
	Harvest 2	0.824 ***	0.821 ***	0.822 ***
	Harvest 3	0.862 ***	0.872 ***	0.824 ***

Note: P-value 0.05*, 0.01**, 0.001***

GWAS analysis of 'TA in the year 2020'

For GWAS analysis of TA in 2020, phenotypic data from 195, 173, and 156 individuals with GBS markers from harvest one, two, and three were used, respectively (Table 22). As previously described, GWAS analysis was performed using two different models, CMLM and MLMM. No significant associations were found in any harvests using either CMLM or MLMM. But in harvest three, some SNPs on chromosome 6 were placed at a higher log-likelihood p-value level than the other two harvests in 2020 (Table 26, Figures 24 and 25). The Q-Q plots of both models displayed the same trend as the observed p values and were aligned with expected p values (Figure 26).

Table 26. Peak SNPs associated with TA in incomplete-diallel population during growing season 2020 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 1	CMLM	No significant associations					
	MLMM	No significant associations					
Harvest 2	CMLM	No significant associations					
	MLMM	No significant associations					
Harvest 3	CMLM	S6_2785463*	6	2.78	6.67E-05	0.230	0.286241
	MLMM	S6_5522070*	6	5.52	2.27E-05	0.278	0.230798

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = centimorgan, NA = Not available.

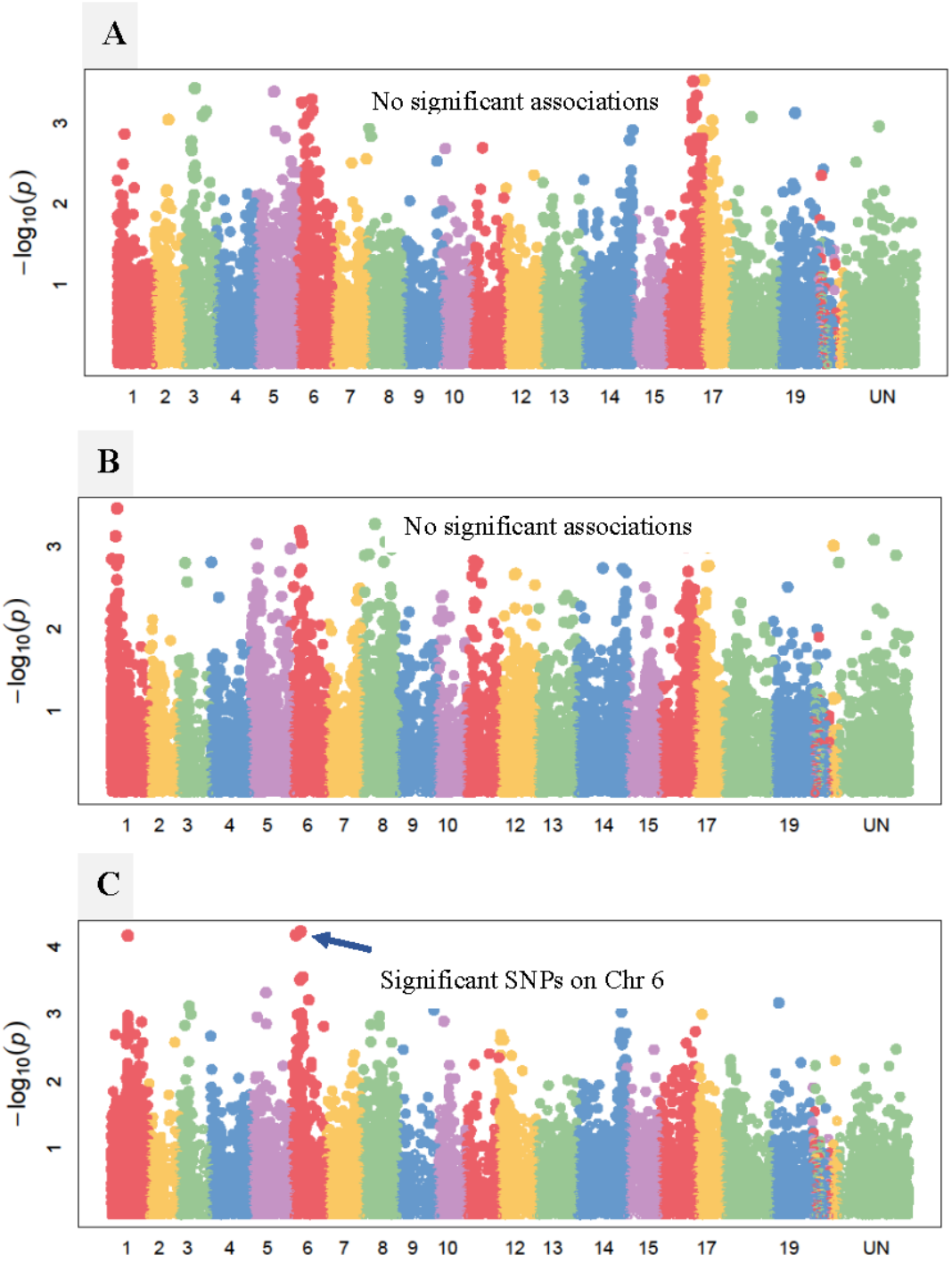

Figure 24. Manhattan plots of 'TA in 2020' using CMLM model. A) TA in harvest one of year 2020. B) TA in harvest two of year 2020. C) TA in harvest three of year 2020. Green horizontal line indicating the threshold cutoff -log10 p value.

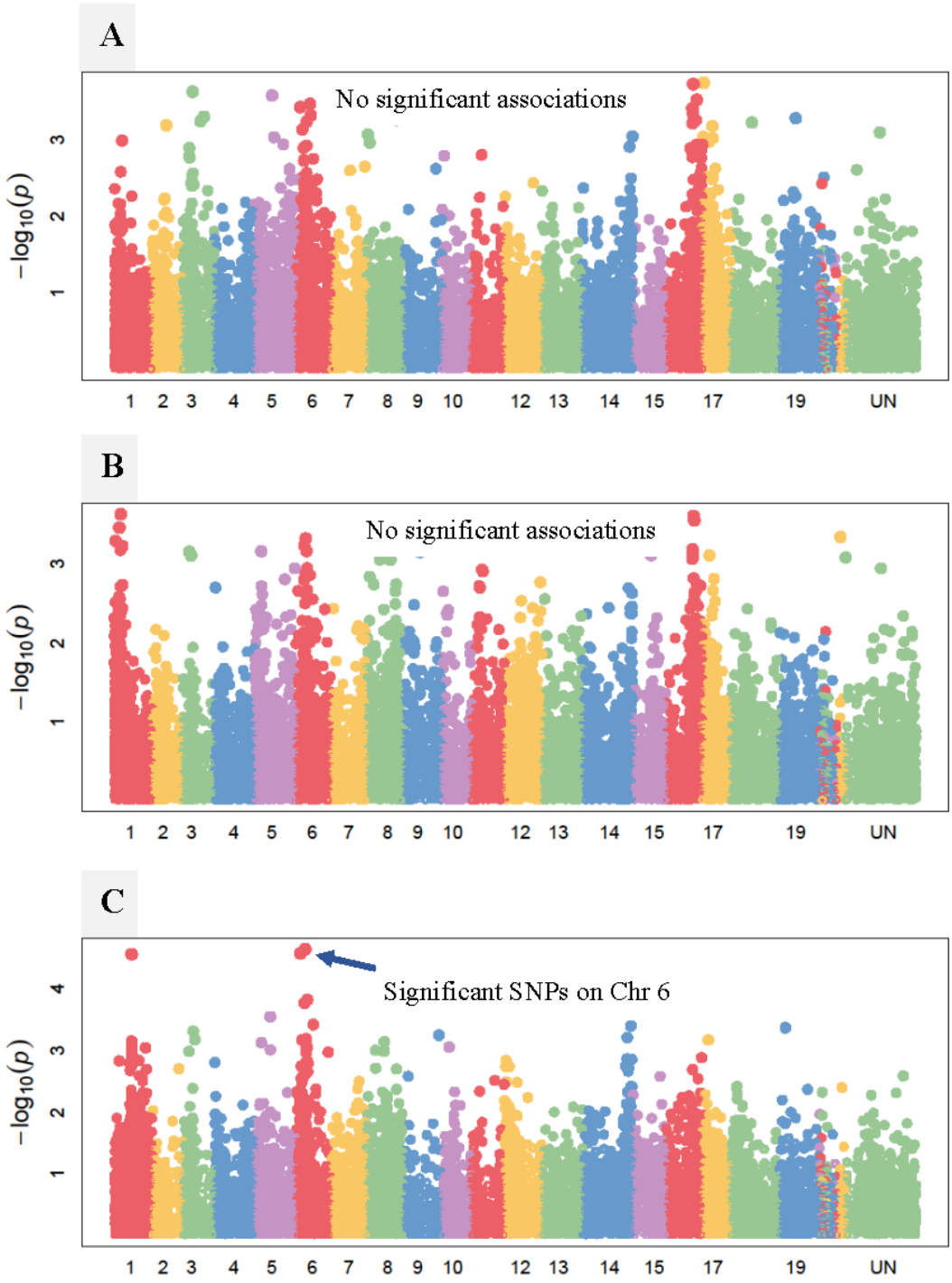

Figure 25. Manhattan plots of 'TA in 2020' using MLMM model. A) TA in harvest one of year 2020. B) TA in harvest two of year 2020. C) TA in harvest three of year 2020. Green horizontal line indicating the threshold cutoff -log10 p value.

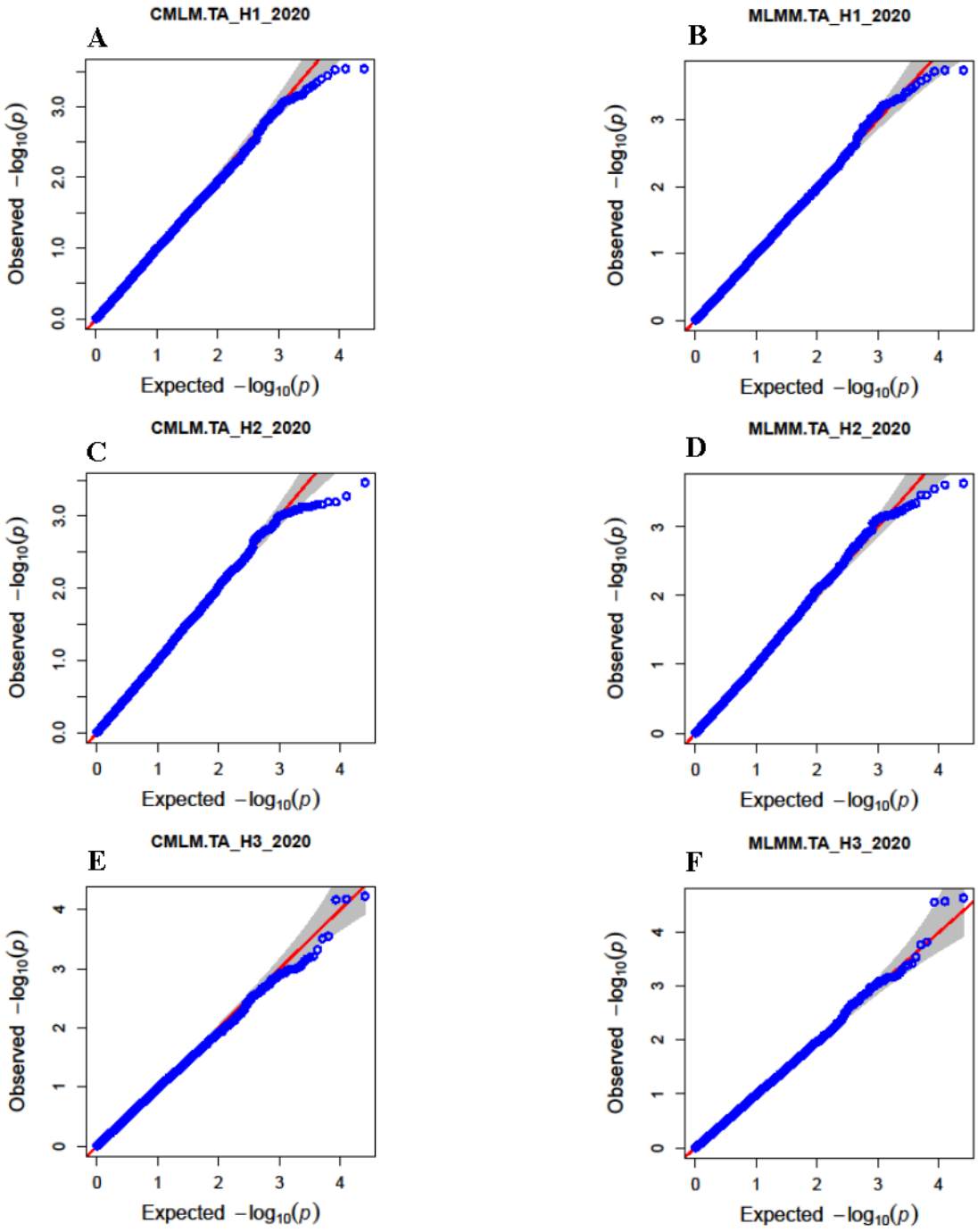

Figure 26. Q-Q plots of TA in year 2020. A) CMLM of TA harvest one phenotypic distribution. B) MLMM of TA harvest one phenotypic distribution. C) CMLM of TA harvest two phenotypic distribution. D) MLMM of TA harvest two phenotypic distribution. E) CMLM of TA harvest three phenotypic distribution. F) MLMM of TA harvest three phenotypic distribution.

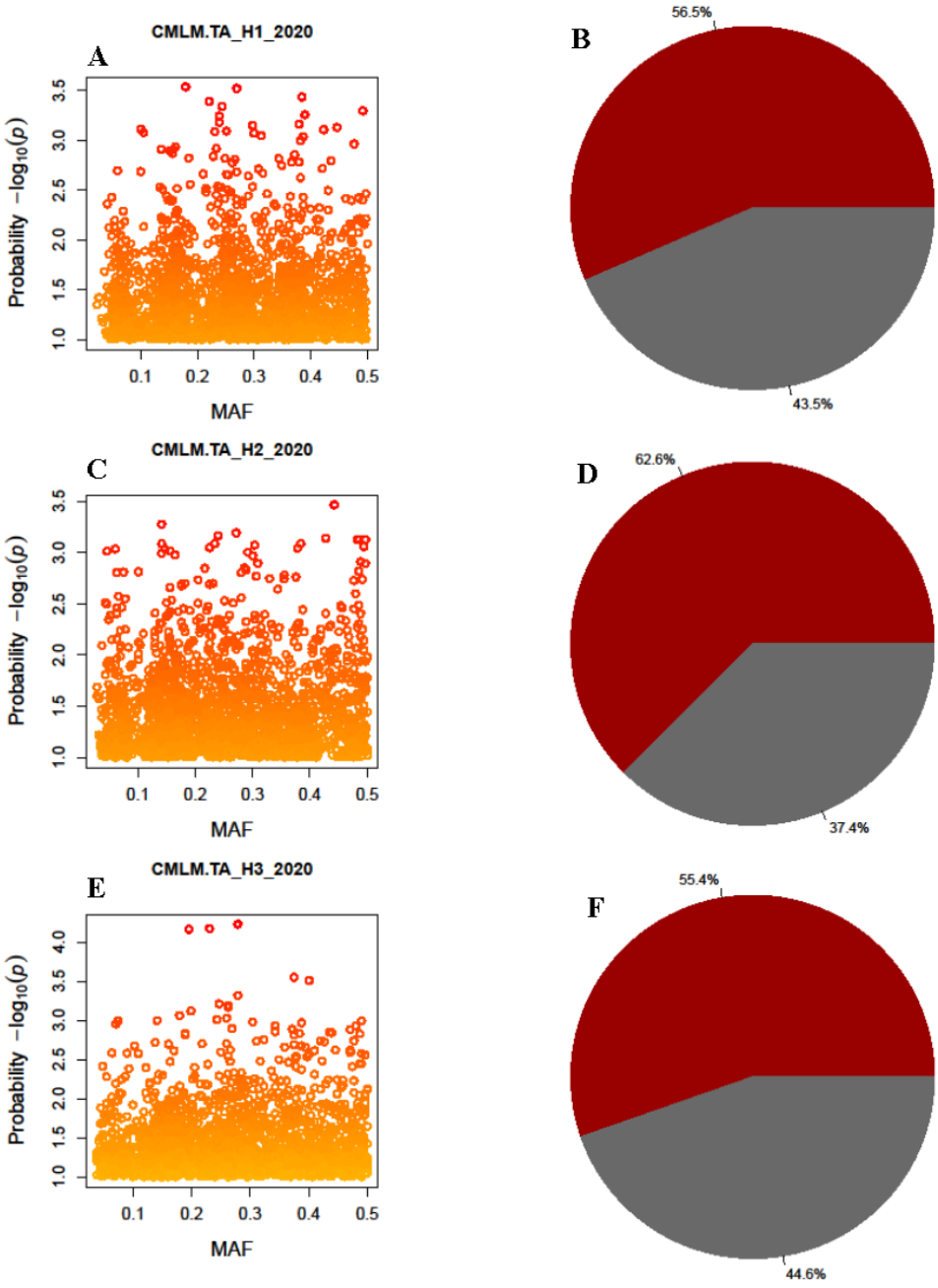

Figure 27. Minor allele frequency and heritability plots the population for TA in 2020 growing season. A) MAF of TA harvest one phenotypic distribution. B) Heritability of TA harvest one phenotypic distribution. C) MAF of TA harvest two phenotypic distribution. D) Heritability of TA harvest two phenotypic distribution. E) MAF of TA harvest three phenotypic distribution. F) Heritability of TA harvest three phenotypic distribution. Red shaded area in the pie chart indicating the heritability value of the trait. *MAF = Minor allele frequency.

GWAS analysis of 'TA in the year 2021'

For GWAS analysis of TA in 2021, phenotypic data from 402, 382, and 378 individuals with GBS markers from harvest one, two, and three were used, respectively (Table 22). As previously described, GWAS was performed using two different models such as CMLM and MLMM. In the CMLM model, unlike 2020 results, significant SNPs above the cutoff $-\log_{10}$ p-value were identified on chromosome 16 during the first two harvests. Significant SNPs in harvest one were placed at a higher log-likelihood p-value than those in harvest two. In the third harvest, no association was observed above the cutoff threshold using CMLM, but similar to 2020 results, some SNPs are placed almost close to the cutoff line on chromosome 6 (Table 27 and Figure 28).

Analysis using the MLMM model was able to find the significant association in all three harvests of 2021. Similar to CMLM results, MLMM identified a relation between SNPs and the trait on chromosome 16 during the first two harvests. Still, log-likelihood p-values of the significant SNPs were slightly higher. Unlike CMLM, in harvest three, the association found on chromosome 6 is above the cutoff $-\log_{10}$ p-value in MLMM (Table 27 and Figure 29).

Except in the third harvest of the CMLM model, observed p values were deviated from the expected p values in all other harvests using both models. This indicated a significant association between the deviated SNPs and trait TA (Figure 30). Heritability values of trait TA range from 50% to 60%, indicating the trait was majorly influenced by genetic variance (Figure 31).

Table 27. Peak SNPs associated with TA in incomplete-diallel population during growing season 2021 using two different models (CMLM and MLMM).

Harvest number	Model	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
Harvest 1	CMLM	S16_10023969	16	10.02	1.11E-06	0.317	0.497087
		S16_13846780	16	13.84	9.50E-07	0.440	0.441889
		S16_13846784	16	13.84	9.50E-07	0.440	0.441889
		S16_15872050*	16	15.87	1.83E-07	0.232	-0.61615
		S16_15872778	16	15.87	8.43E-06	0.333	0.462167
		S16_15992697	16	15.99	1.46E-06	0.373	0.464762
	MLMM	S16_15872050*	16	15.87	6.46E-08	0.232	NA
Harvest 2	CMLM	S16_21085301*	16	21.08	1.35E-06	0.269	0.491869
	MLMM	S16_21085301*	16	21.08	4.13E-07	0.269	NA
Harvest 3	CMLM	No significant associations					
	MLMM	S6_5521338*	6	5.52	4.31E-07	0.267	NA

Note: * Most significant SNP in the respective model, MAF = Minor allele frequency, cM = Centimorgan, NA = Not available.

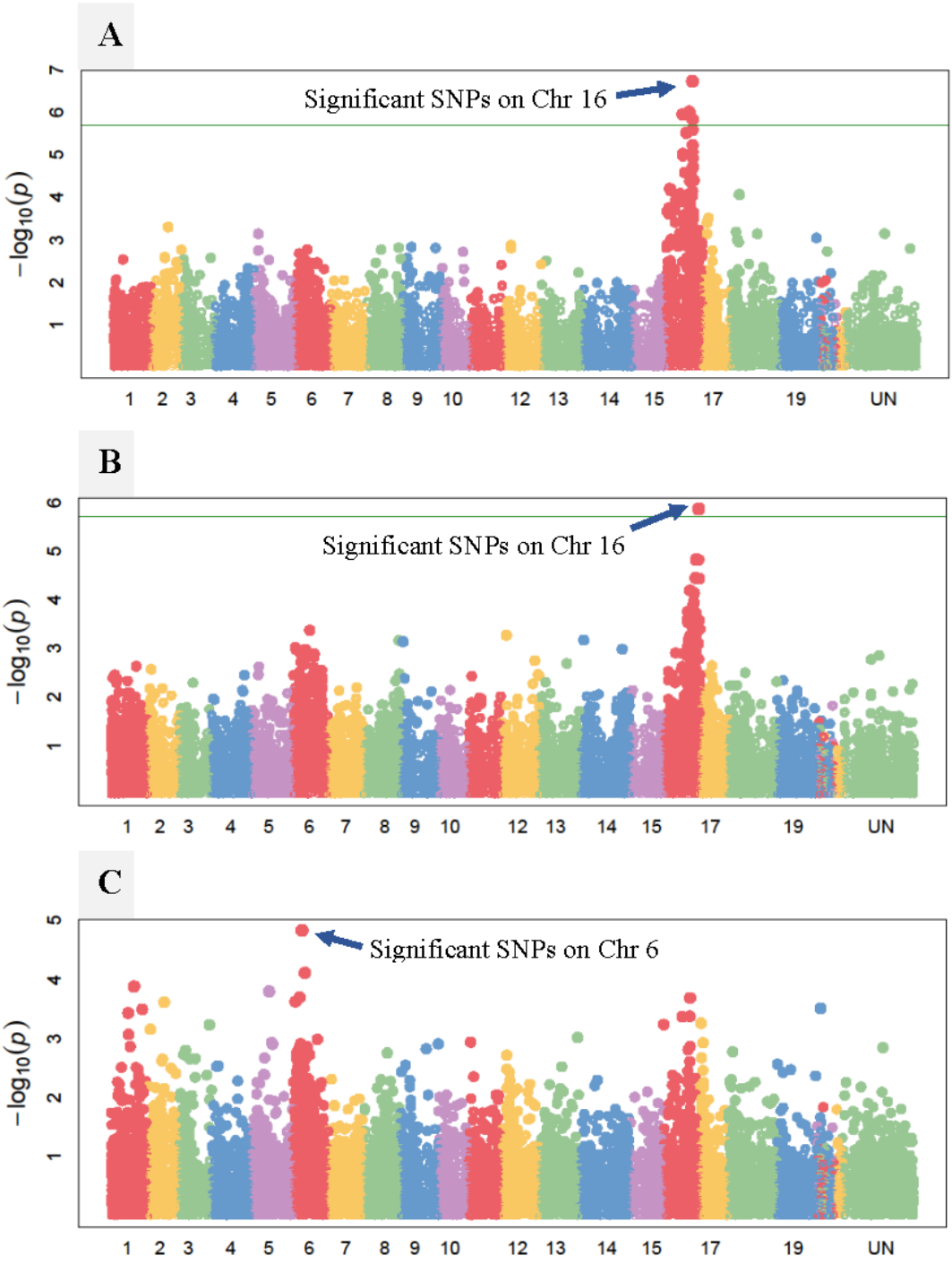

Figure 28. Manhattan plots of 'TA in 2021' using CMLM model. A) TA in harvest one of year 2021. B) TA in harvest two of year 2021. C) TA in harvest three of year 2021. Green horizontal line indicating the threshold cutoff -log10 p value.

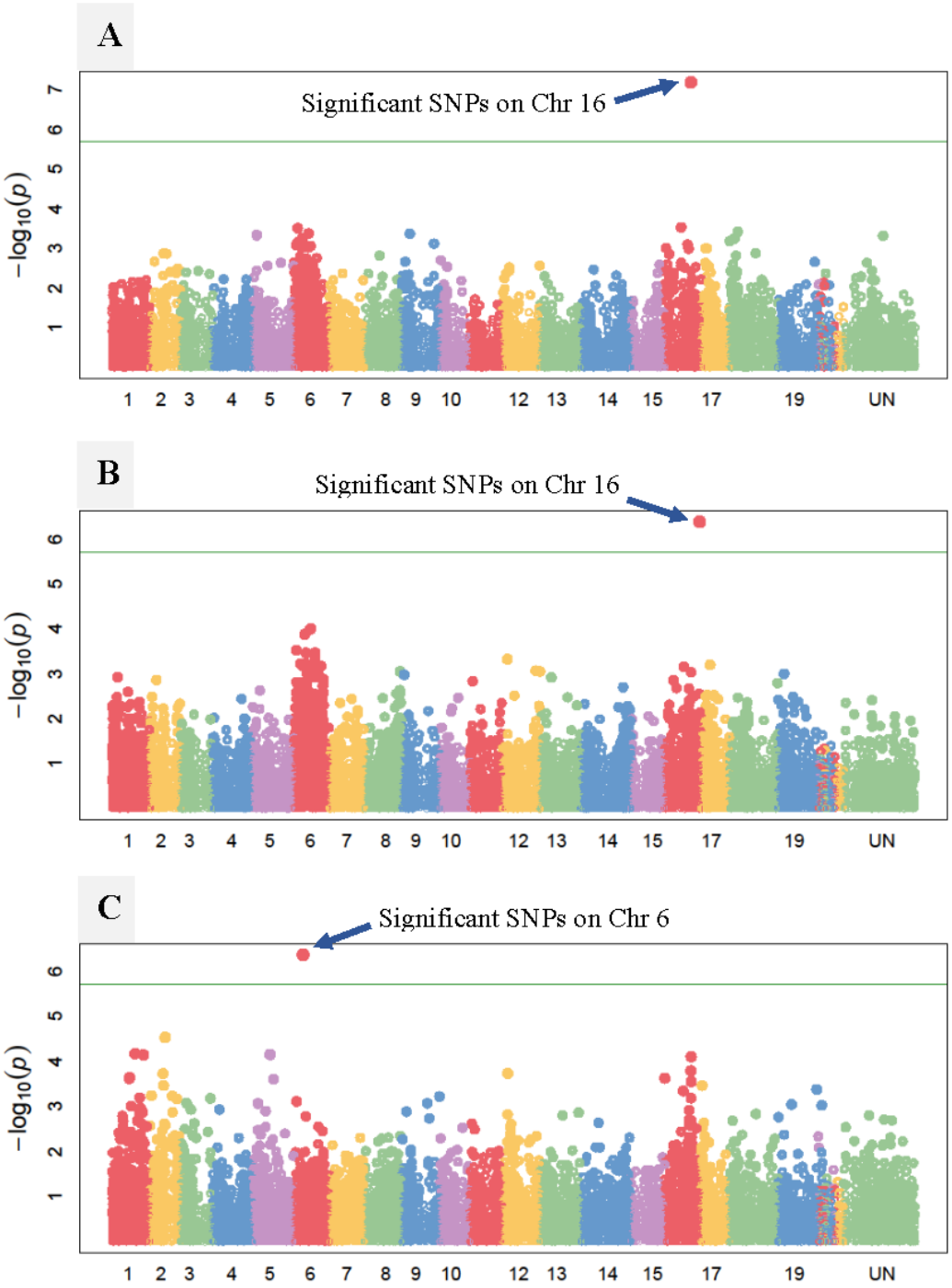

Figure 29. Manhattan plots of 'TA in 2021' using MLMM model. A) TA in harvest one of year 2021. B) TA in harvest two of year 2021. C) TA in harvest three of year 2021. Green horizontal line indicating the threshold cutoff -log10 p value.

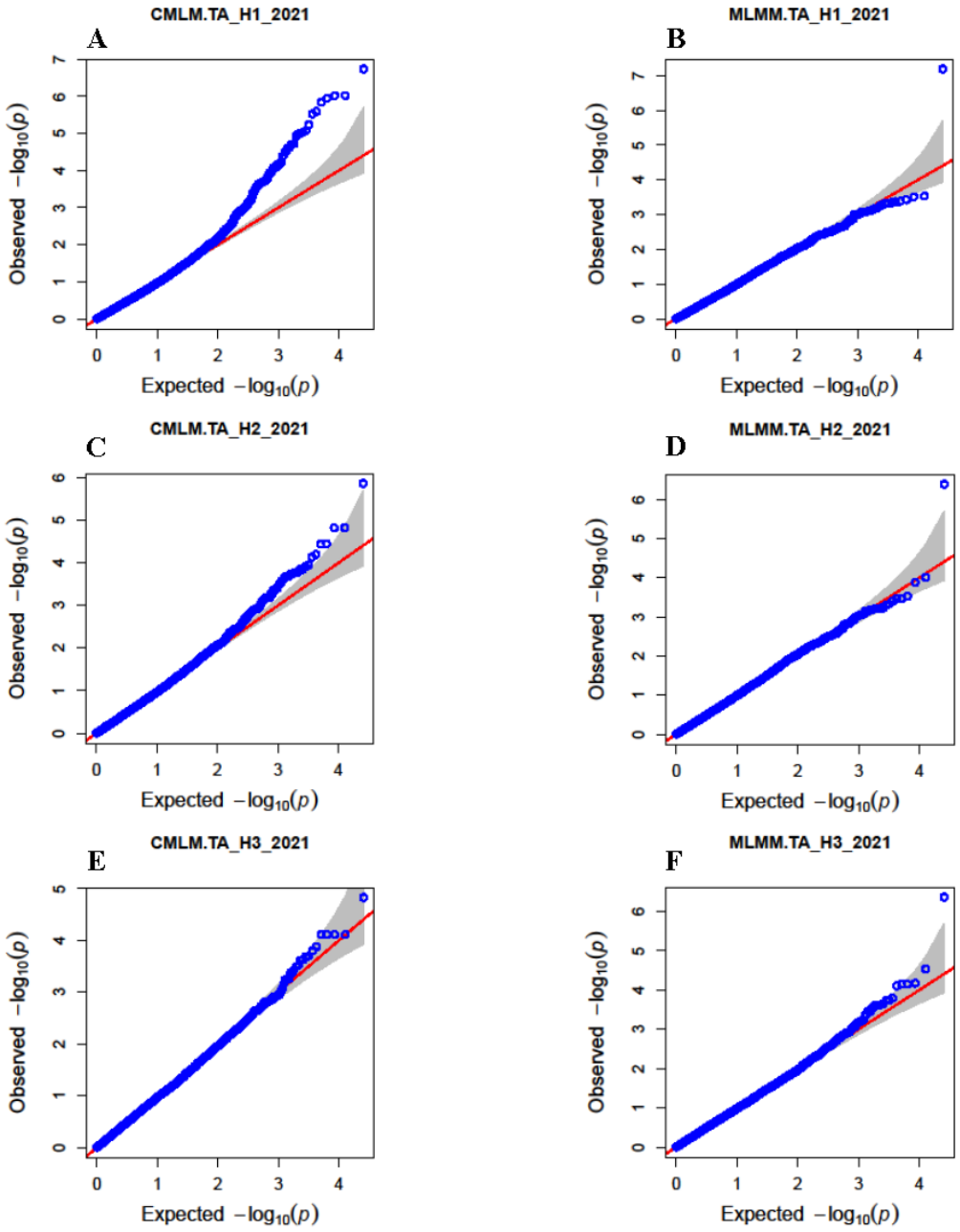

Figure 30. Q-Q plots of TA in 2021. A) CMLM of TA harvest one phenotypic distribution. B) MLMM of TA harvest one phenotypic distribution. C) CMLM of TA harvest two phenotypic distribution. D) MLMM of TA harvest two phenotypic distribution. E) CMLM of TA harvest three phenotypic distribution. F) MLMM of TA harvest three phenotypic distribution.

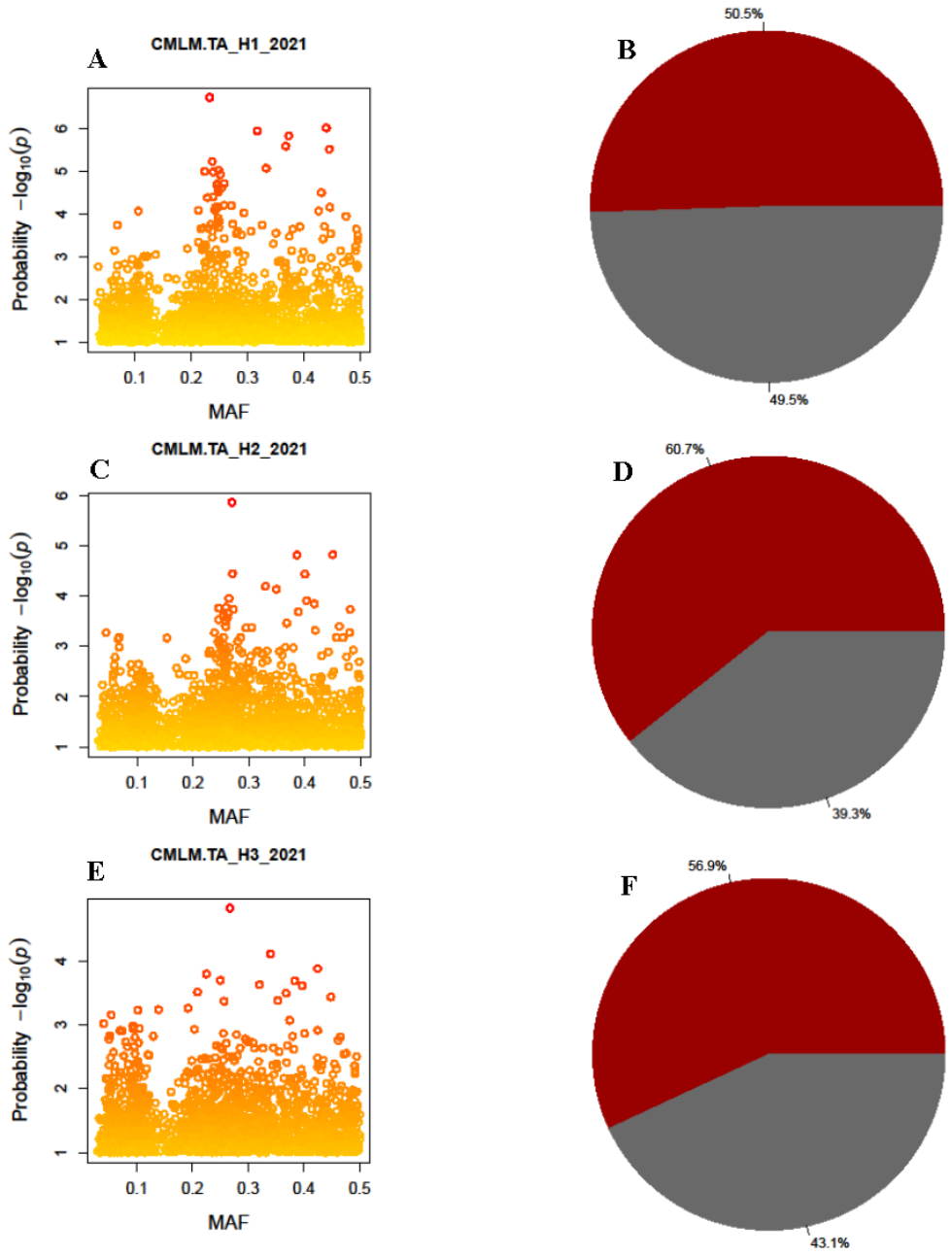

Figure 31. Minor allele frequency and heritability plots the population for TA in 2021 growing season. A) MAF of TA harvest one phenotypic distribution. B) Heritability of TA harvest one phenotypic distribution. C) MAF of TA harvest two phenotypic distribution. D) Heritability of TA harvest two phenotypic distribution. E) MAF of TA harvest three phenotypic distribution. F) Heritability of TA harvest three phenotypic distribution. Red shaded area in the pie chart indicating the heritability value of the trait. *MAF = Minor allele frequency.

DISCUSSION

In the continental climates of North Dakota, grape growers face many production issues such as winter injury, poor fruit quality, short growing season, and frost damage. Due to these obstacles, cold-hardy hybrid grape cultivars are the only suitable grapevines for production in this region. Most of the cultivars grown in this area are hybrids of different native and Eurasian *Vitis* spp. Diverse genetic backgrounds of these cultivars resulted in atypical fruit ripening profiles with high titratable acidity that tend to make wines less appealing to the consumer. To overcome some of these difficulties, in this current study, I focused on understanding the genetic determinants of the basic fruit chemistry (˚Brix, pH, and TA) through an association study in a cold-hardy hybrid population suitable to this region.

Efficient phenotypic data collection was the first step to conducting an association study. The collection of phenotypic data at three different times during each year greatly helped to document both gradual changes in trait value and changes in genetic factors responsible for the trait. In both years, ˚Brix and pH values were improved substantially from initial sample collection to the later harvest. Total acidity levels were decreased greatly from the earliest harvest to the latter as expected. The population at maturity still had a higher TA mean value than the generally acceptable range, and only a handful of individual vines had TA values under one. Phenotypic data collected in different years was consistent with each other and no significant variations in traits were found between years (Table A1). Phenotypic data of all three traits are highly correlated between years and different harvests of each year. Mean phenotypic distribution of population in comparison with parental phenotypic distribution was clearly expressed in table A2.

GWAS analysis on an incomplete diallel population identified marker-trait association on chromosome 16 for all three traits in both years. Along with this, an additional association on chromosome 6 was also identified for all three traits in at least one harvest of 2021. All the associations related to sugar/°Brix were identified on chromosomes 2, 6, 16, and 17 in the present study. Previous linkage mapping studies that focused on interspecific hybrid populations, showed QTLs related to sugar were identified on chromosomes 4, 11, 14, and 17 by Chen et al., 2015 and on chromosome 6 by Yang et al., 2016 and on chromosome 2 by Bayo-Canha et al., 2019. The association identified on chromosome 16 for sugar in the present study is novel and hasn't been previously reported.

In the present study, associations related to trait TA were only identified on chromosomes 6 and 16. Previous linkage mapping studies on interspecific hybrid populations reported QTLs related to acid on chromosome 6 by Chen et al., 2015; Duchêne et al., 2020 and Negus et al., 2021. The association identified on chromosome 6 for trait TA in the current study is a repetition of previous findings and signifies the stable nature of this association across different populations. The QTLs related to acid were also previously identified on other linkage groups 5, 8, 15, and 18 (Chen et al., 2015, Bayo-Canha et al., 2019, Duchêne et al., 2020 and Negus et al., 2021). However, the association found on chromosome 16 for trait TA in the current study is novel, not previously been reported anywhere.

For trait pH, associations were only identified on chromosomes 6 and 16 in the current study. Very recently, a stable QTL for pH on chromosome 6 in two different years was reported in a *Vitis. aestivalis* derived 'Norton' population (Negus et al., 2021). Repetition of this association for pH on chromosome 6 in our mapping population shows the importance of this QTL in various interspecific hybrid populations. Again, the association on chromosome 16 for

trait pH is a novel finding that has not been identified previously. Almost 90% of the significant SNPs identified in the current GWAS study of these three traits were in a narrow ~2.5 cM region on chromosome 16 between 14446661 to 16918954 base pairs. Also, the most significant SNPs of these traits on chromosome 6 were in a short 3 cM region between 5522070 to 8828811 base pairs. Candidate gene scanning in these two genomic regions using a gene annotated *Vitis.vinifera* 12X reference genome file (Grimplet and Fennel., 2011) revealed multiple genes related to carbohydrate and amino acid metabolism (Tables 28 to 31).

The SNPs found in these significant regions of the genome have a differential effect on the three traits. Some SNPs showed a strong positive effect on ˚Brix and pH while having a negative effect on trait TA and vice versa (Table 32 and 33). SNP 'S16_15991560' was the most significant in multiple GWAS analyses. It showed a higher log-likelihood p-value above the cutoff in almost every harvest for ˚Brix and pH. This SNP has a big effect on the traits, and it alone contributes more than four units of ˚Brix and 0.11 units of pH. Along with this, other SNPs such as S16_15731027, S16_15872050, S16_16345474, S16_16975630 were found significant in multiple harvests. Gene scanning under these most significant SNPs revealed important genes in their proximity, such as fructokinase-2, glutamine synthetase B1 GLB1, UDP-glycosyltransferase 88A4, and sorbitol dehydrogenase, etc. which has a function in glucose, fructose, and amino acid metabolism.

Besides identifying new novel regions, this research also validates some QTLs identified previously. Based on its significance and metabolism (carbohydrate & amino acid) related genes located in the targeted genomic region, markers S6_5521338, S16_15991560, S16_15731027, S16_15872050, S16_16345474, and S16_16975630 are promising to be helpful in breeding purposes to improve these traits further (Table A3).

Table 28. Genes facilitating carbohydrate metabolism in the significant region on chromosome 6.

Chr.	Position (cM)	Gene annotation	Function
6	5.386 - 5.387	Lactoylglutathione lyase	Monosaccharide and pyruvate metabolism
6	5.683 - 5.684	Beta-1,3 glucanase	Polysaccharide metabolism
6	5.717 - 5.725	Triosephosphate isomerase, cytosolic	Fructose and mannose metabolism
6	5.873 - 5.875	Exostosin (Xyloglucan galactosyltransferase KATAMARI 1)	Oligosaccharide metabolism
6	6.106 - 6.109	Ribulose bisphosphate carboxylase/oxygenase activase, chloroplast	Photosynthesis and Calvin cycle
6	6.364 - 6.368	Serine/threonine-protein phosphatase PP1	Starch and sucrose metabolism
6	6.472 - 6.474	Glycosyltransferase family 1 protein	Carbohydrate metabolism
6	6.648 - 6.656	Phosphopyruvate hydratase.	Glycolysis and Gluconeogenesis
6	6.901 - 6.906	Exo-1,3-beta-glucanase	Starch and sucrose metabolism
6	7.027 - 7.045	SETH2; transferase, Glycosyl transference	transferring glycosyl groups
6	7.366 - 7.369	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase	Monosaccharide metabolism
6	7.370 - 7.375	Phosphoinositide-specific phospholipase C	Monosaccharide metabolism
6	7.629 - 7.641	fructose-6-phosphate-2-kinase	Fructose and mannose metabolism
6	8.023 - 8.024	UDP-glucuronosyl/UDP-glucosyltransferase	Pentose glucuronate interconversion and sucrose metabolism
6	8.395 – 8.397	Ribokinase	Monosaccharide metabolism
6	8.784 – 8.785	Beta-fructofuranosidase	Starch, sucrose and galactose metabolism

Table 29. Genes facilitating carbohydrate metabolism in the significant region on the chromosome 16.

Chr.	Position (cM)	Gene annotation	Function
16	14.269 - 14.271	Anthocyanidin 3-O-glucosyltransferase	Flavonoid biosynthesis
16	14.530 - 14.532	Exostosin family protein	Oligosaccharide and carbohydrate metabolism
16	14.619 - 14.632	Alpha-L-fructosidase	Glycan degradation and oligosaccharide metabolism
16	14.959 - 14.964	fructokinase-2	Sucrose, fructose, mannose, and starch metabolism
16	15.621 - 15.623	Glucose-methanol-choline (GMC) oxidoreductase family protein	Amino acid metabolism
16	15.651 - 15.669	L-idonate dehydrogenase	Fructose and mannose metabolism
16	15.675 - 15.679	Sorbitol dehydrogenase	Fructose and mannose metabolism
16	16.080 - 16.081	Glutamine synthetase B1 GLB1	Carbohydrate and amino acid metabolism
16	16.717 - 16.718	6-phospho 3-hexuloisomerase	Monosaccharide metabolism
16	17.046 - 17.047	UDP-glycosyltransferase 88A4	Carbohydrate metabolism
16	18.582 - 18.583	UDP-glucose: isoflavone 7-O-glucosyltransferase	Isoflavonoid biosynthesis
16	18.585 - 18.586	UDP-glucose: anthocyanidin 5,3-O-glucosyltransferase	Anthocyanin-glycoside biosynthesis
16	19.150 - 19.157	Pyruvate kinase isozyme G, chloroplast precursor	Glycolysis and carbon fixation
16	19.180 - 19.188	Acidic endochitinase (CHIB1)	Amino sugar metabolism
16	19.193 - 19.194	Chitinase [Vitis vinifera]	Amino sugar metabolism
16	19.585 - 19.587	Trehalose-6-phosphate phosphatase	Starch and sucrose metabolism
16	19.670 - 19.676	Pyruvate kinase	Glycolysis and carbohydrate metabolism

Table 30. Genes facilitating amino acid metabolism in the significant region on the chromosome 6.

Chr.	Position (cM)	Gene annotation	Function
6	5.967 - 5.968	Protein-S-isoprenylcysteine O-methyltransferase	Amino acid (Methionine) metabolism
6	6.168 - 6.171	Tropinone reductase	Amino acid (Arginine and proline) metabolism
6	6.883 - 6.887	Diphenol oxidase	Aromatic amino acid (Tyrosine) metabolism
6	7.107 - 7.113	Anthranilate synthase component I-1, chloroplast precursor	Tyrosine and tryptophan biosynthesis
6	7.765 - 7.771	Ethylene overproducer 1 (ETO1)	Ethylene signaling
6	8.386 - 8.387	Aspartyl-tRNA synthetase	Amino acid (Alanine and aspartate) metabolism
6	8.408 - 8.409	Acetohydroxy acid reductoisomerase	Valine leucine and isoleucine biosynthesis
6	8.450 - 8.454	Alanine--glyoxylate aminotransferase 2 2, mitochondrial	Glycine, serine, alanine, and threonine metabolism
6	8.523 - 8.524	S-adenosyl-L-methionine-dependent methyltransferase mraW	Amino acid derivative metabolism
6	8.542 - 8.545	Cationic peroxidase 1 precursor	Aromatic amino acid (Phenylalanine) metabolism
6	8.545 - 8.546	TPA: class III peroxidase 40	Aromatic amino acid (Phenylalanine) metabolism
6	8.561 - 8.564	Peroxidase	Aromatic amino acid (Phenylalanine) metabolism
6	8.882 - 8.885	Trans-cinnamate 4-monooxygenase	Aromatic amino acid (Phenylalanine) metabolism

Table 31. Genes facilitating amino acid metabolism in the significant region on the chromosome 16.

Chr.	Position (cM)	Gene annotation	Function
16	14.302 – 14.307	1-aminocyclopropane-1-carboxylate synthase	Organic acid metabolism and ethylene signaling
16	15.385 – 15.386	Cationic peroxidase	Aromatic amino acid metabolism
16	15.623 – 15.628	Mandelonitrile lyase-like protein	Amino acid derivative metabolism
16	15.763 – 15.813	Valyl-tRNA synthetase	Amino acid (Valine, leucine, and isoleucine) metabolism
16	15.957 – 15.959	Dihydrodipicolinate reductase	Lysine biosynthesis
16	16.028 – 16.030	S-N-methylcoclaurine 3'-hydroxylase	Primary amino acids derivate metabolism and alkaloid biosynthesis
16	17.588 – 17.591	Prolyl 4-hydroxylase	Amino acid and flavonoid metabolism
16	19.678 – 19.685	Spermine synthase	Amino acid (Arginine and proline) metabolism
16	20.666 – 20.679	GLT1 (NADH-dependent glutamate synthase one gene)	Amino acid (Glutamate) biosynthesis
16	21.152 – 21.156	Peroxidase 3	Aromatic amino acid (Phenylalanine) metabolism

Table 32. Significant SNPs and their effect on °Brix, pH, TA of different harvests in the year 2020.

SNP	Effect on °Brix			Effect on pH			Effect on TA		
	Harvest 1	Harvest 2	Harvest 3	Harvest 1	Harvest 2	Harvest 3	Harvest 1	Harvest 2	Harvest 3
S16_14446661	2.99	2.10	0.88	0.12	0.09	0.06	-0.26	-0.31	-0.01
S16_15731027	3.86	2.56	1.31	0.10	0.11	0.06	-0.41	-0.27	-0.18
S16_15872050	3.50	1.54	0.99	0.11	0.11	0.02	-0.26	-0.25	-0.10
S16_15991560	3.65	2.44	1.11	0.12	0.09	0.04	-0.28	-0.25	-0.14
S16_15992931	3.68	2.43	1.40	0.11	0.15	0.06	-0.32	-0.39	-0.14
S16_16345474	4.31	2.97	1.50	0.10	0.09	0.00	-0.45	-0.33	-0.22
S16_16975630	2.99	2.18	1.53	0.13	0.07	0.10	-0.27	-0.36	-0.14
S16_21082304	2.74	2.99	1.94	0.07	0.06	0.12	-0.35	-0.20	-0.12
S16_21085301	-2.18	-1.47	-0.79	-0.06	-0.05	-0.08	0.21	0.09	0.03
S6_5521338	1.09	1.55	0.92	0.04	0.02	0.13	-0.24	-0.31	-0.19
S6_5522070	2.13	1.08	0.38	0.06	0.07	0.02	-0.26	-0.24	-0.35
S6_8828811	-1.64	-1.21	-1.01	-0.01	-0.03	-0.03	0.22	0.07	0.09

Table 33. Significant SNPs and their effect on °Brix, pH, TA of different harvests in the year 2021.

SNP	Effect on °Brix			Effect on pH			Effect on TA		
	Harvest 1	Harvest 2	Harvest 3	Harvest 1	Harvest 2	Harvest 3	Harvest 1	Harvest 2	Harvest 3
S16_14446661	3.26	3.37	1.60	0.10	0.09	0.08	-0.48	-0.40	-0.14
S16_15731027	3.21	2.96	2.25	0.08	0.08	0.10	-0.48	-0.29	-0.09
S16_15872050	3.86	3.04	1.61	0.12	0.10	0.10	-0.61	-0.38	-0.17
S16_15991560	4.33	3.53	1.83	0.10	0.11	0.11	-0.52	-0.34	-0.13
S16_15992931	3.70	3.28	2.29	0.08	0.10	0.12	-0.49	-0.39	-0.19
S16_16345474	3.45	3.51	1.89	0.07	0.10	0.12	-0.51	-0.43	-0.17
S16_16975630	2.09	1.35	0.96	0.04	0.05	0.06	-0.33	-0.24	-0.13
S16_21082304	1.83	2.28	1.13	0.02	0.04	0.04	-0.17	-0.22	-0.02
S16_21085301	-1.99	-2.27	-0.84	-0.02	-0.05	-0.05	0.17	0.49	0.08
S6_2785463	-0.33	-0.56	-0.31	-0.00	-0.00	-0.00	0.04	0.10	0.03
S6_5521338	0.81	0.54	0.51	0.03	0.00	0.03	-0.09	-0.05	-0.30
S6_5522070	1.08	0.94	0.92	0.06	0.01	0.05	-0.21	-0.19	-0.17
S6_8828811	-1.33	-1.92	-0.82	-0.03	-0.04	-0.02	0.15	0.24	0.11

REFERENCES

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. Science (80-) 322:881–888.

Amaral AJ, Megens H-J, Crooijmans RPMA, Heuven HCM, Groenen MAM. 2008. Linkage Disequilibrium Decay and Haplotype Block Structure in the Pig. Genetics 179:569–579.

Aradhya MK, Dangl GS, Prins BH, Boursiquot JM, Walker MA, Meredith CP, Simon CJ. 2003. Genetic structure and differentiation in cultivated grape, Vitis vinifera L. Genet Res 81:179–192.

Barnaud A, Laucou V, This P, Lacombe T, Doligez A. 2010. Linkage disequilibrium in wild French grapevine, Vitis vinifera L. subsp. silvestris. Heredity (Edinb) 104:431–437.

Barnaud A, Lacombe T, Doligez A. 2006. Linkage disequilibrium in cultivated grapevine, Vitis vinifera L. Theor Appl Genet 112:708–716.

Bayo-Canha A, Costantini L, Fernández-Fernández JI, Martínez-Cutillas A, Ruiz-García L. 2019. QTLs Related to Berry Acidity Identified in a Wine Grapevine Population Grown in Warm Weather. Plant Mol Biol Report 37:157–169.

Begum H, Spindel JE, Lalusin A, Borromeo T, Gregorio G, Hernandez J, Virk P, Collard B, McCouch SR. 2015. Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS One 10:1–19.

Bouquet A. 2011. Grapevines and Viticulture. Genet Genomics, Breed Grapes:1–29.

Bradbury, P.J. et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633-2635.

Chang HX, Lipka AE, Domier LL, Hartman GL. 2016. Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Phytopathology 106:1139–1151.

Chen J, Wang N, Fang LC, Liang ZC, Li SH, Wu BH. 2015. Construction of a high-density genetic map and QTLs mapping for sugars and acids in grape berries. BMC Plant Biol 15:1–14.

Chitwood DH, Ranjan A, Martinez CC, Headland LR, Thiem T, Kumar R, Covington MF, Hatcher T, Naylor DT, Zimmerman S, et al. 2014. A modern ampelography: A genetic basis for leaf shape and venation patterning in grape. Plant Physiol 164:259–272.

Coombe B. 1987. Distribution of solutes within the developing grape berry in relation to its morphology. Am J Enol Vitic 38:120–127.

Coombe BG, McCarthy MG. 2000. Dynamics of grape berry growth and physiology of ripening. Aust J Grape Wine Res 6:131–135.

Cousins P, Striegler RK. 2005. Grapevine Rootstocks : Current Use , Research , and Application. 2005 Rootstock Symp:116.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510.

Deschamps S, Llaca V, May GD. 2012. Genotyping-by-sequencing in plants. Biology (Basel) 1:460–483.

Duchêne É, Dumas V, Butterlin G, Jaegli N, Rustenholz C, Chauveau A, Bérard A, Le Paslier MC, Gaillard I, Merdinoglu D. 2020. Genetic variations of acidity in grape berries are controlled by the interplay between organic acids and potassium. Theor Appl Genet 133:993–1008.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6.

Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot JM, This P, Grando MS. 2010. A candidate gene association study on muscat flavor in grapevine (Vitis vinifera L.). BMC Plant Biol 10:1–17.

Falconer DS, Mackay FC. 1996. Introdutction to Quantitative Genetics. *In* Introdutction to Quantitative Genetics. p. 464.

Flint-Garcia SA, Thornsberry JM, Edward IV SB. 2003. Structure of Linkage Disequilibrium in Plants. Annu Rev Plant Biol 54:357–374.

García-Ruiz A, Bartolomé B, Martínez-Rodríguez AJ, Pueyo E, Martín-Álvarez PJ, Moreno-Arribas M V. 2008. Potential of phenolic compounds for controlling lactic acid bacteria growth in wine. Food Control 19:835–841.

Grimplet, J., & Anne Fennell, D. (n.d.). 2011. VitisNet 12X: Reference file - vitis vinifera genome manual annotation file: Genes annotation. Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. Retrieved March 22, 2022, from https://openprairie.sdstate.edu/vitisnet-12x_files/84

Gupta PK, Rustgi S, Kulwal PL. 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Mol Biol 57:461–485.

He J, Zhao X, Laroche A, Lu ZX, Liu HK, Li Z. 2014. Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:1–8.

Heywood VH, Zohary D. 1995. A Catalogue of the Wild Relatives of Cultivated Plants Native to Europe. Flora Mediterr 5:375–415.

Hyma KE, Barba P, Wang M, Londo JP, Acharya CB, Mitchell SE, Sun Q, Reisch B, Cadle-Davidson L. 2015. Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: A case study in grapevine.

Jackson DI, Lombard PB. 1993. Environmental and Management Practices Affecting Grape Composition and Wine Quality - A Review. Am J Enol Vitic 44:409 LP – 430.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467.

Kennedy J. 2002. Understanding grape berry development. Practical winery & vineyard, 4, 1-5.

Kliewer WM. 1965. Changes in the Concentration of Malates, Tartrates, and total free Acids in Flowers and Berries of Vitis Vinifera. Am J Enol Vitic 16:92–100.

Kliewer WM. 1966. Sugars and Organic Acids of Vitis vinifera . Plant Physiol 41:923–931.

Lamikanra O, Inyang ID, Leong S. 1995. Distribution and Effect of Grape Maturity on Organic Acid Content of Red Muscadine Grapes. J Agric Food Chem 43:3026–3028.

Li M, Liu X, Bradbury P, Yu J, Zhang YM, Todhunter RJ, Buckler ES, Zhang Z. 2014. Enrichment of statistical power for genome-wide association studies. BMC Biol 12:1–10.

Lijavetzky D, Cabezas J, Ibáñez A, Rodríguez V, Martínez-Zapater JM. 2007. High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics 8:1–11.

Lodhi MA, Reisch BI. 1995. Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. Theor Appl Genet 90:11–16.

Mackay I, Powell W. 2007. Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63.

Malacarne G, Perazzolli M, Cestaro A, Sterck L, Fontana P, van de Peer Y, Viola R, Velasco R, Salamini F. 2012. Deconstruction of the (paleo)polyploid grapevine genome based on the analysis of transposition events involving NBS resistance genes. PLoS One 7.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141.

Martin DM, Toub O, Chiang A, Lo BC, Ohse S, Lund ST, Bohlmann J. 2009. The bouquet of grapevine (Vitis vinifera L. cv. Cabernet Sauvignon) flowers arises from the biosynthesis of sesquiterpene volatiles in pollen grains. Proc Natl Acad Sci U S A 106:7245–7250.

McGovern, PE., Stuart J. Fleming, and Solomon H. Katz. 2003. eds. The origins and ancient history of wine: food and nutrition in history and antropology. Routledge.

McGovern PE. 2013. Ancient Wine: The Search for the Origins of Viniculture. Princeton University Press.

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, et al. 2011. Genetic structure and domestication history of the grape. Proc Natl Acad Sci U S A 108:3530–3535.

Negus KL, Chen L-L, Fresnedo-Ramírez J, Scott HA, Sacks GL, Cadle-Davidson L, Hwang C-F. 2021. Identification of QTLs for berry acid and tannin in a *Vitis aestivalis*-derived 'Norton'-based population. Fruit Res 1:1–11.

Nicolas SD, Péros JP, Lacombe T, Launay A, Le Paslier MC, Bérard A, Mangin B, Valière S, Martins F, Le Cunff L, et al. 2016. Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies. BMC Plant Biol 16:1–19.

Owens C. 2011. Linkage Disequilibrium and Prospects for Association Mapping in Vitis. Genet Genomics, Breed Grapes:93–110.

Paran I, Michelmore RW. 1993. Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theor Appl Genet 85:985–993.

Rafalski A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100.

Reynolds AG. 2017. The Grapevine, Viticulture, and Winemaking: A Brief Introduction. *In* Grapevine Viruses: Molecular Biology, Diagnostics and Management. B Meng, GP Martelli, DA Golino, and M Fuchs (eds.), pp. 3–29. Springer International Publishing, Cham.

Reynolds AG. 2015. Grapevine breeding in France – a historical perspective. Grapevine Breed Programs Wine Ind:65–76.

Reynolds AG, Reisch BI. 2015. Grapevine breeding in the Eastern United States. Grapevine Breed Programs Wine Ind:345–358.

Rick C. M, and N. W. Simmonds. 1976. Evolution of crop plants: 268-273.

Robinson SP, Davies C. 2000. Molecular biology of grape berry ripening. Aust J Grape Wine Res 6:175–188.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830.

Slatkin M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485.

Swarts K, Li H, Romero Navarro JA, An D, Romay C, Hearne S, Acharya C, Glaubitz JC, Mitchell SE, Elshire RJ. 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crops plants.

Tello J, Torres-Pérez R, Flutre T, Grimplet J, Ibáñez J. 2020. Vviucc1 nucleotide diversity, linkage disequilibrium and association with rachis architecture traits in grapevine. Genes (Basel) 11:1–19.

Terral JF, Tabard E, Bouby L, Ivorra S, Pastor T, Figueiral I, Picq S, Chevance JB, Jung C, Fabre L, et al. 2010. Evolution and history of grapevine (Vitis vinifera) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. Ann Bot 105:443–455.

This P, Lacombe T, Thomas MR. 2006. Historical origins and genetic diversity of wine grapes. Trends Genet 22:511–519.

This P, Zapater J, Péros J-P, Lacombe T. 2011. Natural Variation in Vitis. Genet Genomics, Breed Grapes:30–67.

Tolar Korenčič T, Jakše J, Korošec-Koruza Z. 2008. The oldest macroremains of Vitis from Slovenia. Veg Hist Archaeobot 17.

Trenti M, Lorenzi S, Bianchedi PL, Grossi D, Failla O, Grando MS, Emanuelli F. 2021. Candidate genes and SNPs associated with stomatal conductance under drought stress in Vitis. BMC Plant Biol 21:1–21.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101:5–22.

Wang J, Zhang Z. 2021. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics Bioinformatics.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey S V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535.

Winkler AJ. 1974. Development and composition of grapes. General viticulture, 138-196.

Xiao Y, Liu H, Wu L, Warburton M, Yan J. 2017. Genome-wide Association Studies in Maize: Praise and Stargaze. Mol Plant 10:359–374.

Yang S, Fresnedo-Ramírez J, Sun Q, Manns DC, Sacks GL, Mansfield AK, Luby JJ, Londo JP, Reisch BI, Cadle-Davidson LE, et al. 2016. Next generation mapping of enological traits in an F2 interspecific grapevine hybrid family. PLoS One 11:1–19.

Zecca G, De Mattia F, Lovicu G, Labra M, Sala F, Grassi F. 2010. Wild grapevine: silvestris, hybrids or cultivars that escaped from vineyards? Molecular evidence in Sardinia. Plant Biol 12:558–562.

Zhang Y, Feng L, Fan X, Jiang J, Zheng X bo, Sun H, Chonghuai L. 2017. Genome-wide assessment of population structure, linkage disequilibrium and resistant QTLs in Chinese wild grapevine. Sci Hortic (Amsterdam) 215:59–64.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360.

Zhu C, Gore M, Buckler ES, Yu J. 2008. Status and Prospects of Association Mapping in Plants. Plant Genome 1:5–20.

APPENDIX

Table A1. Phenotype summary statistics.

Trait	Year	Harvest	N		Mean	Maximum	Median	Minimum
			Total	IWG				
°Brix	2020	1	268	195	13.13	30.15	13.5	1.98
		2	237	173	17.17	25.25	18.05	3.2
		3	205	157	20.27	26.1	20.24	5.6
	2021	1	565	404	11.59	28.4	10.65	3.85
		2	535	386	16.32	30.2	16.65	2.07
		3	521	378	21.44	32.85	21.92	5.15
pH	2020	1	268	195	2.26	2.83	2.24	1.8
		2	237	173	2.35	2.95	2.38	1.7
		3	204	156	3.34	4.62	3.23	2.7
	2021	1	565	403	2.73	3.96	2.69	2.46
		2	535	386	2.92	3.8	2.9	2.39
		3	521	378	3.10	3.7	3.1	2.57
TA	2020	1	268	195	2.88	5.26	2.85	1.4
		2	237	173	2.57	5.39	2.54	0.98
		3	204	156	1.96	3.76	1.86	0.97
	2021	1	565	402	3.24	6.82	3.18	1.19
		2	535	382	2.62	5.86	2.52	0.68
		3	521	378	1.83	4.66	1.75	0.57

*N = Number of individuals sampled; IWG = Individuals with GBS markers.

Table A2. Parent phenotypic distribution in comparison with mean phenotype values of the population in year 2021.

Harvest	Trait	ND 213	ND.054.21	SKND.009.41	Population mean
1	Brix	7.5	7.6	12	11.59
	pH	2.68	2.65	2.61	2.73
	TA	3.34	3.25	3.97	3.24
2	Brix	12.3	12.2	20.1	16.32
	Ph	2.91	2.75	2.91	2.92
	TA	2.30	2.61	3.30	2.62
3	Brix	16.45	19.4	25.3	21.44
	pH	3.17	3.04	3.08	3.10
	TA	1.18	1.69	1.76	1.83

Table A3. Most significant SNPs associated with °Brix, pH, TA in the incomplete-diallel populations using compression mixed model.

Trait	Year	Harvest	Significant SNPs	Chr.	Position (cM)	P values	MAF	Effect
°Brix	2020	1	S16_16345474*	16	16.34	6.06E-07	0.269	4.31868
		2	S16_21082304*	16	21.08	2.71E-06	0.257	2.99365
		3	S2_4112015*	2	41.12	4.34E-06	0.134	2.68132
°Brix	2021	1	S16_15991560*	16	15.99	1.48E-13	0.237	4.33295
		2	S16_15991560*	16	15.99	9.60E-09	0.24	3.53356
		3	S16_15731027*	16	15.73	3.18E-07	0.252	2.25761
pH	2020	1	S16_16975630*	16	16.97	4.47E-06	0.241	0.13674
		2	No significant associations					
		3	No significant associations					
pH	2021	1	S16_15872050*	16	15.87	3.41E-08	0.233	0.12113
		2	S16_15991560*	16	15.99	1.98E-10	0.24	0.11569
		3	S16_16345474*	16	16.34	6.11E-09	0.267	0.12809
TA	2020	1	No significant associations					
		2	No significant associations					
		3	S6_2785463*	6	2.78	6.67E-05	0.23	0.28624
TA	2021	1	S16_15872050*	16	15.87	1.83E-07	0.232	-0.6162
		2	S16_21085301*	16	21.08	1.35E-06	0.269	0.49187
		3	S6_5521338*	6	5.52	4.31E-07	0.267	NA

*MAF: Minor allele frequency; NA: Not available.

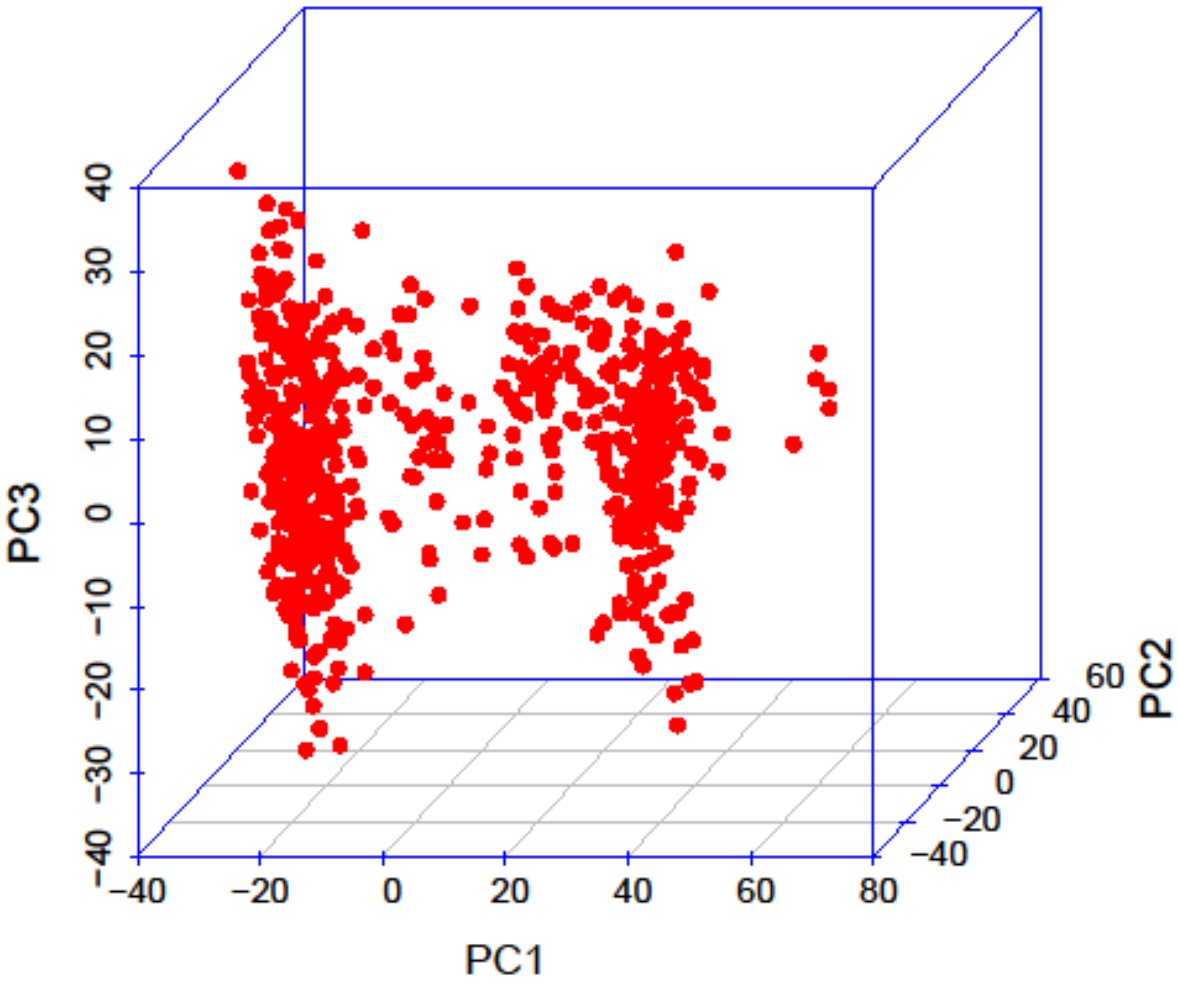

Figure A1. Principal component analysis 3D plot of incomplete-diallel population showing population structure using marker data.

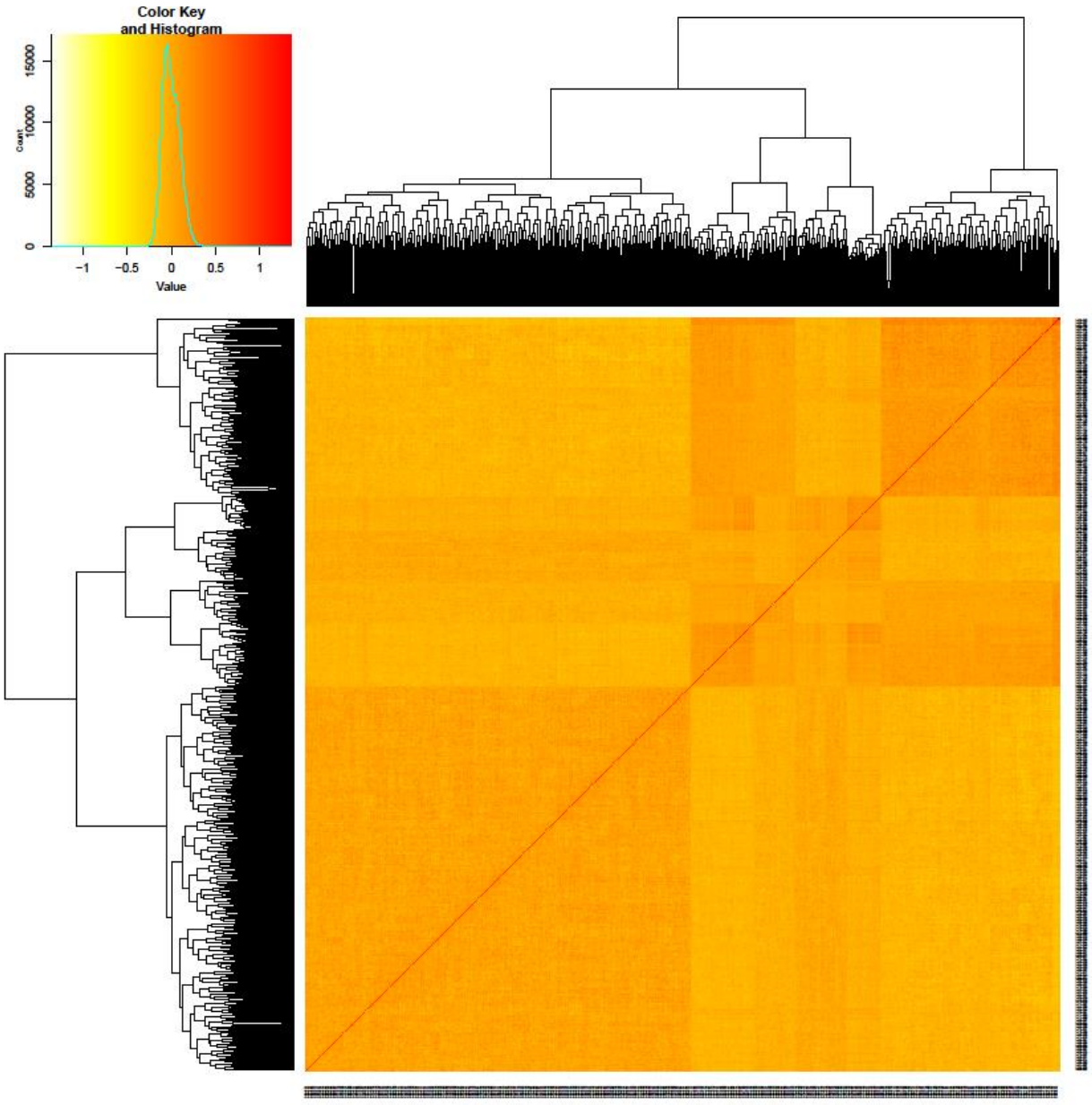

Figure A2. VanRaden plot of incomplete-diallel population showing population structure using marker data.

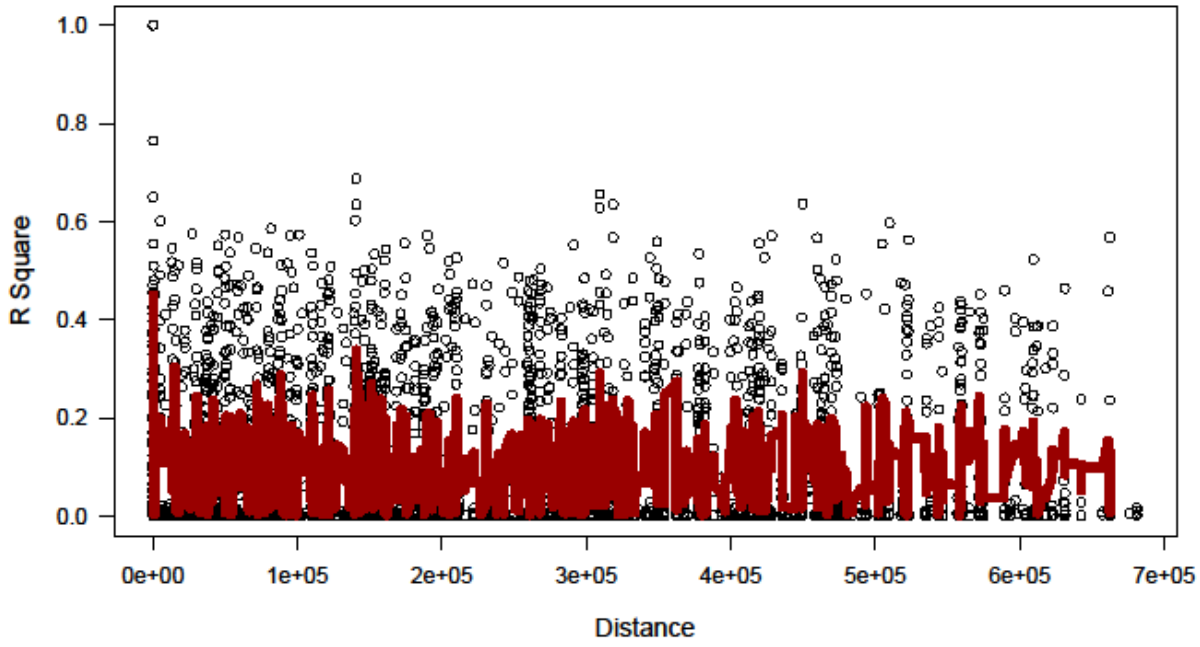

Figure A3. Linkage disequilibrium decay of incomplete-diallel population.

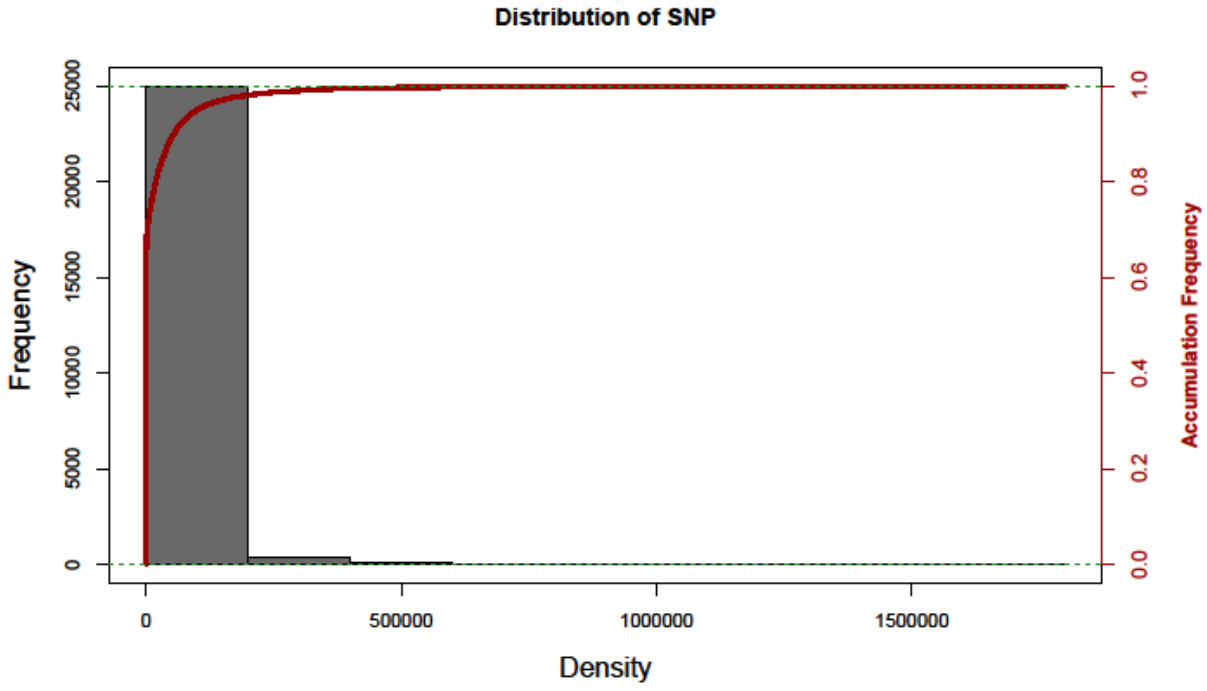

Figure A4. Marker density of incomplete-diallel population.