Received: 8 July 2024 | Accepted: 31 August 2024

DOI: 10.1002/tpg2.20518

The Plant Genome

ORIGINAL ARTICLE

Special Section: Enhancing Food Security through Innovative Agricultural Management

Studies of genetic diversity and genome-wide association for vitamin C content in lettuce (*Lactuca sativa* L.) using high-throughput SNP arrays

Inés Medina-Lozano[1,2] | **Juan Ramón Bertolín**[2,3] | **Jörg Plieske**[4] | **Martin Ganal**[4] | **Heike Gnad**[4] | **Aurora Díaz**[1,2]

[1]Department of Plant Sciences, Agrifood Research and Technology Centre of Aragon (CITA), Zaragoza, Spain

[2]AgriFood Institute of Aragon–IA2 (CITA-University of Zaragoza), Zaragoza, Spain

[3]Department of Animal Sciences, Agrifood Research and Technology Centre of Aragon (CITA), Zaragoza, Spain

[4]SGS Institut Fresenius GmbH TraitGenetics Section, Seeland, Germany

Correspondence
Aurora Díaz, Department of Plant Sciences, Agrifood Research and Technology Centre of Aragon (CITA), Avd. Montañana 930, 50059, Zaragoza, Spain.
Email: adiazb@cita-aragon.es; adiazb@unizar.es

Assigned to Associate Editor Lubin Tan.

Funding information
Spanish Ministry of Science and Innovation (MCIN); European Union (EU) NextGenerationEU: PRTR-C17.I1, Grant/Award Number: AGROALNEXT; MCIN; State Research Agency (AEI), Grant/Award Number: PID2022-138484OR-I00; EU; Government of Aragón: "Grupo de investigación en Fruticultura - Caracterización, Adaptación y Mejora Genética", Grant/Award Number: A12_23R; Spanish Ministry of Science, Innovation and Universities (MCIU); AEI, Grant/Award Number: I.M-L. predoctoral contract

Abstract

Lettuce (*Lactuca sativa* L.) is a source of beneficial compounds though they are generally present in low quantities. We used 40K Axiom and 9K Infinium SNP (single nucleotide polymorphism) arrays to (i) explore the genetic variability in 21 varieties and (ii) carry out genome-wide association studies (GWAS) of vitamin C content in21 varieties and a population of 205 plants from the richest variety in vitamin C ('Lechuga del Pirineo'). Structure and phylogenetic analyses showed that the group formed mainly by traditional varieties was the most diverse, whereas the red commercial varieties clustered together and very distinguishably apart from the rest. GWAS consistently detected, in a region of chromosome 2, several SNPs related to dehydroascorbic acid (a form of vitamin C) content using three different methods to assess population structure, subpopulation membership coefficients, multidimensional scaling, and principal component analysis. The latter detected the highest number of SNPs (17) and the most significantly associated, 12 of them showing a high linkage disequilibrium with the lead SNP. Among the 84 genes in the region, some have been reported to be related to vitamin C content or antioxidant status in other crops either directly, like those encoding long non-coding RNA, several F-box proteins, and a pectinesterase/pectinesterase inhibitor, or indirectly, like extensin-1-like

Abbreviations: AA, ascorbic acid; AMOVA, analysis of molecular variance; CR, call rate; CWR, crop wild relatives; DHAA, dehydroascorbic acid; FarmCPU, fixed and random model circulating probability unification; FLD, Fisher's linear discriminant; H_O, observed heterozygosity; HomFLD, homozygous Fisher's linear discriminant; GBS, genotyping-by-sequencing; tGBS, tunable genotyping-by-sequencing; GD, gene diversity; GDSL, glycine-asparagine-serine-leucine; GWAS, genome-wide association study; IBS, identity by state; LD, linkage disequilibrium; MAF, minor allele frequency; MDS, multidimensional scaling; PCA, principal component analysis; Q, subpopulation membership coefficient; QC, quality control; QQ, quantile-quantile; SA, salicylic acid; SNP, single nucleotide polymorphism; TAA, total ascorbic acid; UPLC, ultra performance liquid chromatography; UV, ultaviolet; WPP, tryptophan-proline-proline.

protein and endoglucanase 2 genes. The involvement of other genes identified within the region in vitamin C levels needs to be further studied. Understanding the genetic control of such an important quality trait in lettuce becomes very relevant from a breeding perspective.

Plain Language Summary

Domestication and breeding have impoverished many crops, like lettuce, in terms of nutritional value. We have explored the genetic and nutritional diversity in cultivated lettuces, finding that traditional varieties are richer in both aspects than those that are commercial. The variety with the highest content in vitamin C, a traditional one, was self-pollinated to create descendants with high level of vitamin C and genetic homogeneity. Using this breeding population and the set of diverse varieties mentioned above, we have found a region of the lettuce genome associated to the amount of one of the forms of vitamin C. Among the genes in the region, there are some that have become strong putative candidates to be involved in vitamin C accumulation as they play similar roles in other crops, like wheat, tomato, and potato, among others. Knowing those genes and understanding how they work is challenging, though it will boost the breeding toward lettuce varieties biofortified in vitamin C and hence more nutritious.

1 | INTRODUCTION

Global production of lettuce (*Lactuca sativa* L.), together with chicory, exceeded 27 million tonnes, with a harvested area of 1.24 million ha in 2022 (FAOSTAT, 2022). These figures reflect the high demand among consumers. This makes lettuce an ideal candidate crop to start a breeding program aimed at improving its nutritional value and health properties as it will certainly have a strong and positive effect on a large part of the global population. Although lettuce is a source of a big range of compounds, which confer benefits to human health, such as vitamins, fiber, and phenolic compounds, they are generally only present in low quantities (USDA, 2022). In fact, when compared to other leafy vegetables, especially spinach, chard, cabbage, or watercress, lettuce is the poorest in vitamin C (USDA, 2022), which is a key indicator of the quality of fruit and vegetables. The process of wild *Lactuca* spp. domestication, which led to the present cultivated lettuce, entailed some collateral effects, like a higher susceptibility to some pathogens and pests, a loss in phytonutrients and beneficial compounds, and a decrease in genetic diversity, among others, as has also happened in other crops. The last one is particularly true in lettuce as it is a predominantly autogamous species. Breeding programs have made use of crop wild relatives (CWR) to introduce disease- and pest-resistance genes in the crop, but they are underused resources with regard to the improvement of the nutritional quality (Dempewolf et al., 2017). This could result surprising in lettuce, especially in the light of studies that have compared the content of vitamin C in different *Lactuca* spp. and found that the wild lettuce relatives are generally richer than the cultivated varieties (Medina-Lozano et al., 2021, 2024; van Treuren et al., 2018). However, linkage drag of undesirable traits linked to the genes of interest (especially those related to organoleptic or sensory attributes) could have discouraged their inclusion in modern breeding programs, in which the flavor has been prioritized over the nutritional quality in the selection. In this sense, landraces and traditional varieties have been proposed as shortcuts when compared to CWR to be used in breeding programs (Medina-Lozano & Díaz, 2021) as they do not carry detrimental and/or maladaptive variants and the reproductive barriers sometimes hindering interspecific crossings between cultivated and wild forms are avoided. Even if they do not harbor as much genetic variability as the wild germplasm, they are still more diverse than the modern commercial varieties (Flint-Garcia et al., 2023). Furthermore, in the case of vitamin C content in lettuce, they have revealed themselves to be also richer than the commercial varieties tested (Medina-Lozano et al., 2020, 2021).

Vitamin C or total ascorbic acid (TAA) is composed of ascorbic acid (AA) and its oxidation product, dehydroascorbic acid (DHAA). AA is a powerful antioxidant, which contributes to a healthy state by preventing common diseases (Granger & Eck, 2018) and must be supplemented in the diet

mainly through fruits and vegetables as humans are unable to synthesize it. DHAA is also a bioactive compound, which can be converted into AA in the human body. That is why DHAA has been suggested to serve as vitamin C reservoir under some adverse conditions, such as those causing oxidative stress, in which AA get transformed into DHAA as a consequence of its antioxidant activity (Medina-Lozano et al., 2021, 2024).

The use of genomic technologies to explore biodiversity and to associate the traits of interest to the genomic regions responsible for them (genome-wide association studies [GWASs]) are among the most used tools nowadays in early stages of crop biofortification. The predominant markers to assess genomic diversity and genetic structure are SNPs (single nucleotide polymorphisms), thanks to their abundance in all species, and the availability of high-throughput SNP genotyping platforms in many food crops (Medina-Lozano & Díaz, 2022). At the moment, more than 70K SNP data are available at the Lettuce Genome Database (LettuceGDB, https://www.lettucegdb.com). Thousands of SNPs have been obtained and used to explore *Lactuca* diversity, mainly among lettuce cultivars but also including CWR, with different technologies such as (i) microarrays like Illumina GoldenGate (S. J. Kwon et al., 2012) or Affymetrix GeneChip (Stoffel et al., 2012) assays and (ii) next-generation sequencing like single primer enrichment technology (Tripodi et al., 2023), genotyping-by-sequencing (GBS) (J. S. Park et al., 2022), tunable GBS (tGBS) (S. Park et al., 2021; Simko, 2023), and even whole-genome resequencing (Wei et al., 2021), which has rendered 179 million SNPs, as well as other types of variants. Some of these studies have also inquired into lettuce domestication history using the diversity panel of 445 *Lactuca* accessions firstly characterized by Wei et al. (2021).

There are a substantial number of works on GWAS in lettuce regarding different types of traits, like those related to tolerance to biotic or abiotic stress, development, nutrient efficiency use, postharvest behavior, and morphology, among others. However, GWAS of characters related to lettuce health-promoting properties or nutritional value are not common. Interestingly, one of the few types of metabolites targeted by GWAS in lettuce, anthocyanins, are bioactive compounds with antioxidant activity (as it is the case of vitamin C), though they were originally addressed as morphological traits, either as a qualitative character, leaf color (L. Zhang et al., 2017), or measuring their content in leaves (Tripodi et al., 2023; Wei et al., 2021). Other GWAS have been carried out to dissect the genetic basis of primary metabolite content, partly responsible (together with others) for the nutritional value of lettuce (W. Zhang et al., 2020). Information on both types of compounds, those with positive effects on human health and nutrients, is actually very helpful to pursue the biofortification of the crop.

To date, different approaches have been adopted to enhance vitamin C content in lettuce, either conventional, by supplying UV (ultra-violet)-B radiation (H. Zhou et al., 2023) or applying minerals (Dyląg et al., 2023) to the plants, or based on genetic engineering techniques (Guo et al., 2013). Both present disadvantages such as temporary effects, in the case of conventional strategies, and legislative issues in some countries, in the case of modern genetic engineering methods. However, classic genetic breeding has also limitations, for instance, only the variability present in plants from sexually compatible groups can be used.

Core Ideas

- Lettuce is ideal to be biofortified as it is nutritionally poor (i.e., vitamin C) but highly demanded by consumers.
- Traditional varieties harbor a great genetic diversity, essential for the breeding of this autogamous species.
- These are the first genetic associations with dehydroascorbic acid content found in lettuce: a 5.1Mb region in chromosome 2.
- High linkage disequilibrium values were only found between the lead single nucleotide polymorphism (SNP) and other significantly associated SNPs.
- Some of the candidate genes in the region of interest are involved in vitamin C metabolism in other crops.

In this work, we explore and exploit the diversity, both genetic and nutritional, within the cultivated lettuce, including traditional varieties. With the richest accession in vitamin C, which happened to be a traditional variety, breeding populations were built. Those, together with a diversity panel of commercial and traditional varieties, were used to find a genomic region associated to vitamin C content and identify putative candidate genes to boost the breeding toward lettuce varieties biofortified in vitamin C and thereupon healthier and more nutritious.

2 | MATERIALS AND METHODS

2.1 | Plant material and trait evaluation

The present study was divided into two different parts: (i) a genetic diversity analysis and (ii) a GWAS of vitamin C content.

First, a total of 21 lettuce accessions were used in the genetic diversity analysis (Table 1). These included 10 commercial varieties (4 green and 6 red) and 11 Spanish

TABLE 1 Description of the plant material and the assays carried out in the present study.

Variety or population name	Assay (No. of SNPs)	Description	Type[a]	Origin	Source[b]	Accession number
'Begoña'	Genetic diversity (13,026) and GWAS (9,242)	Green commercial	Batavia	Spain	Ramiro Arnedo Semillas S.A.	–
'Dolomiti G12'	Genetic diversity (13,026) and GWAS (9,242)	Green commercial	Gem	Spain	Ramiro Arnedo Semillas S.A.	–
'Lechuga de Beceite'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Cos	Spain	BGHZ	BGHZ2006
'Lechuga de Bureta'	Genetic diversity (13,026) and GWAS (9,242)	Semi-red traditional	Cos	Spain	BGHZ	BGHZ4927
'Lechuga de Ensalada'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Cos	Spain	BGHZ	BGHZ2031
'Lechuga de Híjar'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Cos	Spain	BGHZ	BGHZ0529
'Lechuga de Subías'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Cos	Spain	BGHZ	BGHZ1852
'Lechuga del Pirineo'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Cos	Spain	BGHZ	BGHZ2229
'Lechuga del Valle de Tena'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Butterhead	Spain	BGHZ	BGHZ1850
'Lengua de Buey'	Genetic diversity (13,026) and GWAS (9,242)	Green traditional	Cos	Spain	BGHZ	BGHZ2004
'Likarix'	Genetic diversity (13,026) and GWAS (9,242)	Red commercial	Frisée d'Amérique	Netherlands	CGN	CGN24522
'Lollo Rosso'	Genetic diversity (13,026) and GWAS (9,242)	Red commercial	Lollo	Italy	CGN	CGN09385
'Morada de Belchite'	Genetic diversity (13,026) and GWAS (9,242)	Semi-red traditional	Cos	Spain	BGHZ	BGHZ0527
'Morada de Bernués'	Genetic diversity (13,026) and GWAS (9,242)	Semi-red traditional	Batavia	Spain	BGHZ	BGHZ2097
'Morada de Sorripas'	Genetic diversity (13,026) and GWAS (9,242)	Semi-red traditional	Cos	Spain	BGHZ	BGHZ2026
'Nestorix'	Genetic diversity (13,026) and GWAS (9,242)	Red commercial	Lollo	Netherlands	CGN	CGN24712
'Red Sails'	Genetic diversity (13,026) and GWAS (9,242)	Red commercial	Lollo	Germany	CGN	CGN19014
'Revolution'	Genetic diversity (13,026) and GWAS (9,242)	Red commercial	Lollo	Netherlands	CGN	CGN20714
'Romana Inverna'	Genetic diversity (13,026) and GWAS (9,242)	Green commercial	Cos	Spain	BGHZ	BGHZ3604
'Romired'	Genetic diversity (13,026) and GWAS (9,242)	Red commercial	Lollo	Netherlands	CGN	CGN24713
'Winter Crop'	Genetic diversity (13,026) and GWAS (9,242)	Green commercial	Butterhead	Hungary	CGN	CGN05853
'Lechuga del Pirineo' S0	GWAS (9,242)	205 plants (green traditional)	Cos	Spain	–	–
'Lechuga del Pirineo' S1	–[c] (9,242)	239 full-sibs (green traditional)	Cos	Spain	–	–
'Lechuga del Pirineo' S2	–[c] (9,242)	179 full-sibs (green traditional)	Cos	Spain	–	–

Abbreviations: GWAS, genome-wide association study; SNP, single nucleotide polymorphism.
[a]According to the International Union for the Protection of New Varieties of Plants (UPOV, 2021).
[b]BGHZ: Vegetable Germplasm Bank of Zaragoza (Spain); CGN: Centre for Genetic Resources (Wageningen, Netherlands).
[c]Populations originally built but not included in the GWAS because resulted monomorphic for all the markers.

traditional varieties (7 green and 4 semi-red), representing six types of lettuce attending to their morphology: Batavia, Butterhead, Cos, Frisée d'Amérique, Gem, and Lollo (UPOV, 2021).

Second, for the GWAS, the plant material originally included target populations coming from a breeding program aimed at enhancing the vitamin C content in lettuce. The plant with the highest vitamin C content according to a previous study with *Lactuca* germplasm, these 21 lettuce varieties (among others) and some wild relatives, belonged to the traditional variety 'Lechuga del Pirineo' (Medina-Lozano et al., 2021). From this one, three populations were analyzed, the original variety population (S0) made up of 205 plants (presumably coming from seeds of different plants supplied by a germplasm bank, BGHZ, Table 1), and two self-pollination generations, S1 and S2, composed of 239 and 179 plants, respectively, obtained by selfing the plant with the highest vitamin C content in each generation to increase both the metabolite concentration and the genetic homogeneity (Table 1). As lettuce is a predominantly autogamous species, the genetic variability was already negligible in S1 and, obviously, in S2 (Table S1), so they were not used in association analysis. To increase GWAS's power, the needed heterozygosity was incorporated by using the 21 varieties mentioned above from the diversity panel (Table 1), as recommended by Hamazaki et al. (2020).

For both analyses, genetic diversity (21 varieties) and GWAS (21 varieties and 205 individuals of 'Lechuga del Pirineo' S0), plants were grown in pots (30 × 25 cm and 11.7 L volume) with a mix of black and blonde peat (1:1) in a greenhouse at Agrifood Research and Technology Centre of Aragon (CITA, Zaragoza, Spain). Plants for the diversity study were cultivated in winter 2018/2019, and S0, S1, and S2 'Lechuga del Pirineo' populations were cultivated in winters 2020/2021, 2021/2022, and 2022/2023, respectively. After a period ranging from 2.5 to 4 months, depending on the accession and the population, leaves were harvested as described in the next section and immediately frozen with liquid nitrogen and kept at −80°C.

For the GWAS, the two forms of vitamin C, AA and DHAA, as well as the total content, TAA, were quantified in samples consisting of both inner and outer leaves from 644 plants (205 S0, 239 S1, 179 S2, and 21 varieties) by UPLC (ultra performance liquid chromatography) according to the method described by Medina-Lozano et al. (2020). Briefly, the extraction was conducted using 50 mg of finely powdered lyophilized samples with 5 mL of a solution of 8% acetic acid (v/v), 1% meta-phosphoric acid (w/v), and 1 mM EDTA (ethylenediaminetetraacetic acid). The mixture was vortexed for 5 s, shaken for 10 min at 2000 rpm, and then sonicated for 20 min at room temperature, and centrifuged at 4000 × *g* for 10 min at 4°C. The supernatant was filtered through a 0.22-µm regenerated cellulose filter (Agilent). The filtrate (Extract 1, E1) contained both AA and DHAA. Two 200-µL aliquots of E1 were used to determine (i) AA directly and (ii) TAA by reducing DHAA to AA adding 200 µL of a reduction solution (40 mM DTT [dithiothreitol] with 0.5 M Tris pH 9.0) and stopping the reaction after 30 min with 200 µL of 0.4 M sulfuric acid to obtain Extract 2 (E2). This last step is needed because DHAA absorptivity in the UV range of the spectrum is too low to be measured directly. A volume of 5 µL of E1 and E2 diluted with ultrapure water (1:4 v:v) was injected in a liquid chromatographer UPLC H-Class with an HSS T3 column (150 mm × 2.1 mm × 1.8 µm). The total running time was 3 min and the temperature of the samples and the column was programed at 5°C and 30°C, respectively. The wavelength of the Acquity UPLC Photodiode Array eλ detector was set at 245 nm. The mobile phase consisted of 2% methanol and 98% ultrapure water pH 2.0 acidified with formic acid at a flow rate of 0.3 mL min^{-1} in isocratic mode. For quantification of AA and TAA contents, a calibration curve from 0.5 to 25.0 µg mL^{-1} of the commercial L-ascorbic acid (≥99.9% purity, Sigma-Aldrich) was built. DHAA content was calculated by subtracting AA from TAA.

Effects of the generation (S0, S1, and S2) on vitamin C content (AA, DHAA, and TAA) were tested with an analysis of the variance by Kruskal–Wallis test and *post hoc* Dunn's test for mean comparison using a Bonferroni corrected $\alpha = 0.017$.

2.2 | DNA extraction, SNP genotyping, and quality control

DNA was extracted from young leaves of 644 plants (205 S0, 239 S1, 179 S2, and 21 varieties) as described in Doyle and Doyle (1990) with the following modifications (Díaz et al., 2017): 0.2% β-mercaptoethanol was added together with the 2% CTAB (hexadecyltrimethylammonium bromide) buffer, the washing buffer consisted of 76% ethanol and 10 mM ammonium acetate, and 0.2 µL of 10 mg × mL^{-1} RNAse A (Invitrogen) was added to 30 µL of milliQ water to dissolve the pellet. DNA quality control (QC) was carried out via electrophoresis in 1% agarose gels. DNA concentration was measured fluorometrically using a Quantiflour-ST (Promega GmbH).

DNA sample genotyping was performed using the lettuce 40K Axiom and 9K Infinium arrays developed by the SGS Institut Fresenius GmbH TraitGenetics Section, containing 41,975 and 9,381 SNPs, respectively. Axiom array was scanned with a GeneTitan Scan Instrument (Thermofisher Scientific) followed by data analysis with Axiom Analysis Suite software v5.1.1.1 (Thermofisher Scientific) using the default settings. SNPs were filtered by values of call rate (CR) >90%, Fisher's linear discriminant (FLD) >5,

and homozygous FLD (HomFLD) >10. Infinium array was scanned with an iScan system (Illumina) followed by data analysis with GenomeStudio 2.0 (Illumina).

Several QC steps were undertaken with the whole set of markers to be used in both studies, diversity and GWAS. After compiling all data with GenomeStudio software, sample clustering was manually optimized, and SNP markers were filtered by values CR >90%. To improve the data quality, a comparison between common markers from both arrays that met the QC criteria just described was carried out. The markers rendering incongruous genotypes were discarded. Another quality filter was applied by discarding the markers with spurious genotypes according to the pedigree in the 'Lechuga del Pirineo' populations. Then, only polymorphic SNPs in the corresponding set of samples for each analysis were selected: 13,026 markers in the genetic diversity study and 9,242 in the GWAS (Tables S1 and S2).

2.3 | Genetic and genomic analyses

2.3.1 | Population structure and genetic diversity and relationships

The population structure of the varieties in the diversity panel was analyzed carrying out a simulation using the Bayesian algorithm and the admixture model in STRUCTURE v2.3.4 (Pritchard et al., 2000). A burn-in period of 100,000 cycles followed by 100,000 Markov chain Monte Carlo iterations was tested with a number of subpopulations (K) set from one to six. Ten independent runs per K value were performed. The optimal number of K was inferred applying the ΔK method (Evanno et al., 2005). The analysis was repeated with the same parameters for K = 2 and K = 3. Data were plotted using the web application Structure Plot v2.0 (Ramasamy et al., 2014).

Alternatively, both a principal component analysis (PCA) and a multidimensional scaling (MDS), also known as principal coordinate analysis, were performed to visualize patterns of diversity using TASSEL v5.2 (Bradbury et al., 2007) and the software JMP v17.2 for Windows (SAS Institute Inc.) was used to plot the results.

To explore the genetic relationships between the 21 varieties included in the diversity study, a matrix of genetic distances was created from the genotypic data using the IBS (identity by state) method in TASSEL v5.2 software. A phylogenetic network was built based on the genetic distance matrix with the NeighborNet method using the SplitsTree App v6.3.12 software (Huson & Bryant, 2006).

The genetic differentiation was assessed based on the number of distinct populations obtained in the genetic structure analysis (K = 3). Wright's fixation index F_{ST} among populations was calculated. An analysis of molecular variance (AMOVA) was conducted to detect genetic variation among and within populations, as well as within individuals. Summary statistics including gene diversity (GD) and observed heterozygosity (H_O) were calculated for each subpopulation. All the above statistical analyses were performed using PowerMarker 3.25 software (K. Liu & Muse, 2005).

2.3.2 | Linkage disequilibrium, GWAS, and identification of candidate genes

In the GWAS subset of markers, SNPs were filtered for minor allele frequency (MAF) >0.01. This value was not arbitrarily chosen but justified by the nature of the samples. Since most plants belonged to the same variety, 'Lechuga del Pirineo' (205 S0 plants out of 226, Table 1), minor alleles present in that accession were overrepresented and vice versa, the frequency of non-rare alleles in the 21 varieties diversity panel was diluted in the whole set of samples. With the conventional threshold MAF >0.05, a variant should have to be present in more than 11 samples for not being filtered out, which means to be in more than half of the samples of the diversity panel, which obviously is not a minor allele. MAF >0.01 ensures that only variants appearing less than twice were eliminated. Chromosomal and physical SNP positions were determined based on lettuce reference genome Lsat_Salinas_v11 (GCF_002870075.4).

Linkage disequilibrium (LD) between SNPs on each chromosome was calculated through pairwise correlation coefficients (r^2) with an LD window size of 50 sites around each marker. LD decay was determined by plotting r^2 values against the physical distance of the SNPs and then plotted in Rstudio (Rstudio Team, 2020).

Multi-locus GWAS was conducted using GAPIT3 (Wang & Zhang, 2021) in Rstudio. Fixed and random model Circulating Probability Unification (FarmCPU) was used considering both kinship and structure of the samples (X. Liu et al., 2016). Kinship was assessed with GAPIT3 package. Three different methods of measuring the structure were used: a PCA performed with TASSEL v5.2 software, an MDS analysis also conducted in TASSEL v5.2 from the genetic distance matrix calculated using the IBS method, and subpopulation membership coefficients (Q) obtained with STRUCTURE software using the linkage model at K = 3 and the rest of the settings as described earlier. The threshold p-value and $-\log_{10}(p\text{-value})$ were determined using the Bonferroni correction with genome-wide significance level of $\alpha = 0.05$ and taking into account the total number of SNPs (9,242) as follows: $p = 0.05/9{,}242 = 5.41 \times 10^{-6}$ and $-\log_{10}(5.41 \times 10^{-6}) = 5.27$. Manhattan and quantile–quantile (QQ) plots for GWAS results were obtained using CMplot package (Yin et al., 2021) in Rstudio.

A local LD block analysis was conducted in the region of associated SNPs using the LDBlockShow software (Dong

et al., 2021) in Linux environment. The most significantly associated SNP in the GWAS analysis was set as the lead SNP.

Finally, putative candidate genes for vitamin C content (AA, DHAA, and TAA) in the region were identified in the annotated version of *L. sativa* genome (Lsat_Salinas_v11, GCF_002870075.4) available at the NCBI database.

3 | RESULTS AND DISCUSSION

3.1 | Assessment of population structure and genetic diversity and relationships

A total of 13,026 polymorphic SNPs were used to analyze the population structure of a panel of 21 lettuce varieties as well as the genetic diversity and relationships.

The best K estimated by calculating ΔK using the method described by Evanno et al. (2005) indicated that the optimal number of subpopulations for the 21 lettuce varieties was three, corresponding to the highest peak in the ΔK plot (Figure 1A; Figure S1). The first population (Pop. 1) grouped five varieties, consisting of red commercial lettuces exclusively, that came from the Netherlands and Italy. Most red commercial lettuces were entirely assigned to Pop. 1, except for 'Revolution' that still showed a Q1 of 0.69. The second population (Pop. 2) was the largest and the most diverse group, including 12 lettuce varieties. It comprised both commercial and traditional varieties, which were mainly green and semi-red (there was only one red commercial variety), most of them coming from Spain. Finally, the third population (Pop. 3) consisted of four green traditional varieties native from a small geographic region (Figure S2). Among them, three out of four were 100% assigned to Pop. 3. The other one ('Lengua de Buey') showed a high level of admixture with a high Q2 value (0.41). Except for two commercial varieties, one red and one green, both from European countries other than Spain, Pop. 2 members were closer to Pop. 3 than to Pop. 1, especially some of the traditional varieties. This, together with the fact that Pop. 3 was composed exclusively of traditional varieties, suggests that Pop. 3 could really be a subgroup of Pop. 2. Indeed, when two populations were assumed (K = 2), all the varieties of Pop. 3 became part of Pop. 2 (close genetic relatedness between these two subgroups), whereas Pop. 1 remained unchanged (Figure 1A).

To evaluate genetic relationships, a phylogenetic network was obtained from genetic distances calculated using the IBS method (Figure 1B). The most clearly differentiated cluster comprised the five red commercial varieties from Pop. 1 in the structure analysis (Figure 1A,B). The four green traditional varieties of Pop. 3 also clustered together in the phylogenetic network, again with three varieties closely related and 'Lengua de Buey' a bit apart from them (Figure 1B). This cluster was not as distant from Pop. 2 as the one formed by the varieties of Pop. 1, as previously observed in the genetic structure studies. Within varieties of Pop. 2, some clustered together in a similar way as they did in the structure analysis like 'Red Sails' and 'Begoña' or 'Winter Crop', 'Lechuga del Valle de Tena' and 'Dolomiti G12' (Figure 1A,B). Overall, the phylogenetic network was in agreement with the genetic population structure. The results from STRUCTURE were also validated with a PCA and an MDS analysis (Figure 1C,D). The first two components were represented in the PCA, explaining 20.07% and 12.36% of the variation, respectively. Likewise, the first two dimensions were represented in the MDS analysis, capturing 40.74% and 23.31% of variation, respectively. In both analyses, the 21 varieties were divided in three groups, as observed before. The red commercial varieties of the Pop. 1 were clearly grouped separately from the rest by the PC1 and Dim1 from the PCA and MDS, respectively (Figure 1C,D). The green traditional varieties of Pop. 3 were completely separated in the PCA (Figure 1C), whereas in the MDS, they clustered together but overlapped completely with the 95% confidence ellipsis of Pop. 2 (Figure 1D), reflecting the proximity between these two populations.

The results from the four approaches to assess the population structure were highly consistent. Taking all of them together, we could conclude that the 21 lettuce varieties consisted of three main groups. On the one hand, Pop. 1 was the most distinguishable population in all cases, as expected, since all the varieties belonging to similar types have the same leaf color and come from a common geographical region in Europe (Table 1). On the other hand, Pop. 2 was the most diverse group and results suggested that Pop. 3 could be a subgroup of Pop. 2, as mentioned above. This reinforces the idea of traditional varieties harboring a great diversity (Medina-Lozano & Díaz, 2021), being separated in different subpopulations, even when they all come from a small area, as it is the case of all the traditional varieties in Pop. 2 and the whole Pop. 3 (Figure S2). In previous studies carried out to assess the genetic variability of *Lactuca* spp., diversity panels were mainly composed of commercial varieties, advanced breeding lines and breeding populations, like recombinant inbred lines, and lettuce wild relatives (J. S. Park et al., 2022; Peng et al., 2022). Our results suggest that traditional varieties might be an important source of genetic variability in future breeding programs as selection has had a detrimental effect on the crop diversity (J. S. Park et al., 2022). Other studies on genetic structure and phylogenetic relationships conducted in different lettuce accessions found that varieties could group together according to their morphological type (S. Kwon et al., 2013; S. Park et al., 2021; Tripodi et al., 2023). We observed something similar for Pop. 1, composed mainly of Lollo lettuces, as well as for Pop. 3, consisting exclusively of Cos varieties (Table 1; Figure 1). Pop. 2 also included

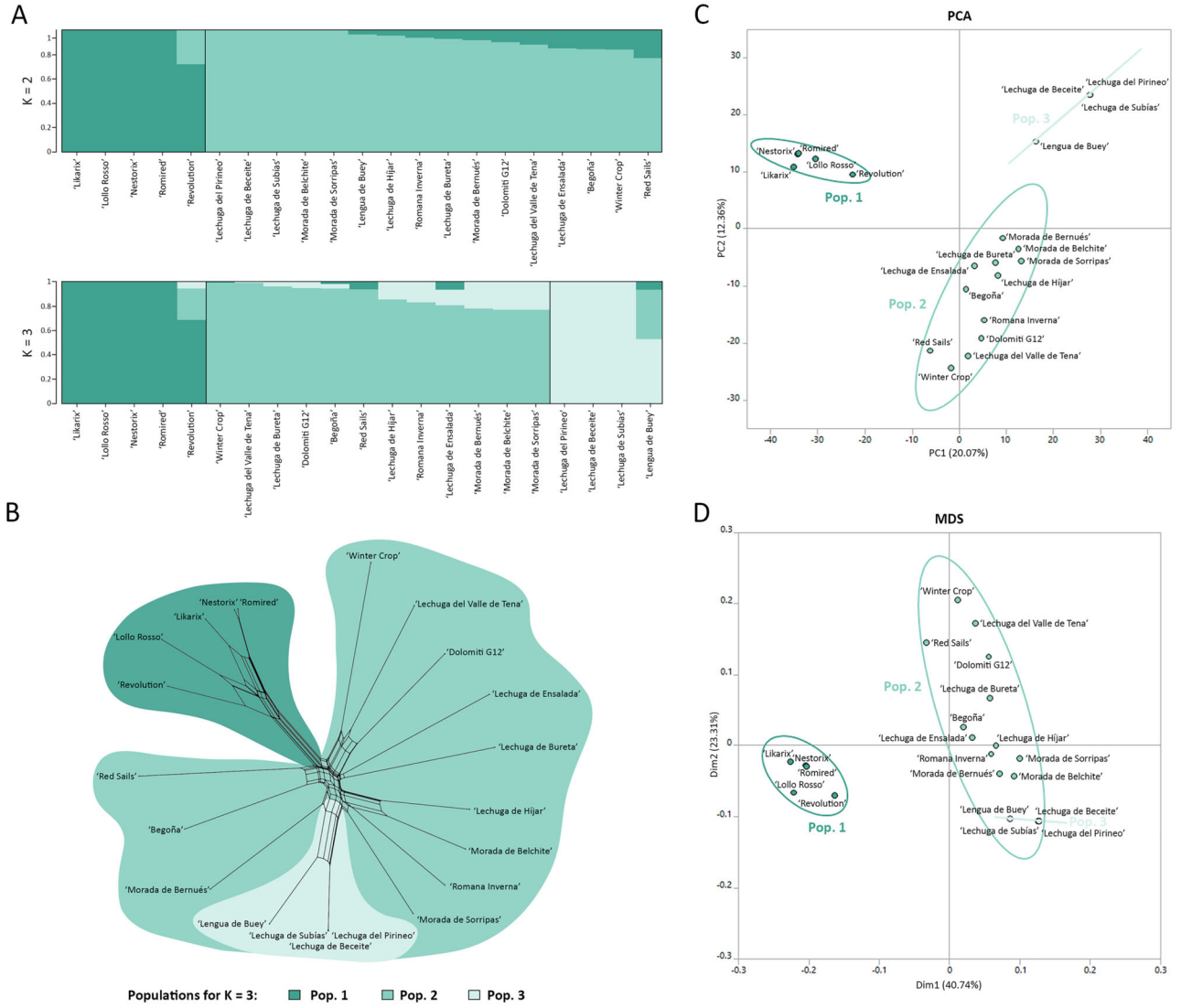

FIGURE 1 Genetic structure and relationships of 21 lettuce varieties genotyped with 13,026 single nucleotide polymorphisms (SNPs). (A) Sample membership to each of the clusters obtained with the STRUCTURE software (K = 2, 3). (B) Network built using the NeighborNet method. (C) Representation of the first two axes from a principal component analysis (PCA) and (D) a multidimensional scaling (MDS) analysis. Confidence ellipses at 95% are drawn.

Cos lettuces, what again supports the idea of Pop. 3 being a subgroup of Pop. 2. Nevertheless, more factors seemed to be influencing the population structure in our study, such as the geographical origin, especially in the case of the traditional varieties (Figure S2).

To assess the genetic variation among the populations obtained from the structure analysis, pairwise F_{ST} values were calculated, with higher F_{ST} values indicating stronger differentiation (Weir & Hill, 2002). The lowest F_{ST} values were observed between Pop. 1 and Pop. 2, followed by Pop. 2 and Pop. 3 (0.36 and 0.43, respectively) (Table 2). This could be explained by the fact that the sort of accessions in common in Pop. 1 and Pop. 2 are less diverse (commercial varieties), whereas the shared type of accessions between Pop. 2 and Pop. 3 are those with greater diversity (traditional varieties).

TABLE 2 Pairwise F_{ST} values for population differentiation.

Population	F_{ST} (Pop. 1)	F_{ST} (Pop. 2)	F_{ST} (Pop. 3)
F_{ST} (Pop. 1)	0.00		
F_{ST} (Pop. 2)	0.36	0.00	
F_{ST} (Pop. 3)	0.63	0.43	0.00

So, Pop. 2 occupied an intermediate position which agrees with being the group with more admixture (Figure 1), including varieties of all colors, types, and origins. The highest differences were obtained between Pop. 1 (exclusively formed by red commercial varieties) and Pop. 3 (only composed by green traditional varieties) ($F_{ST} = 0.63$) (Table 2), as expected since both groups were quite homogeneous and did not share

TABLE 3 Genetic variation at population (Pop.) and individual levels.

Source of variation		Sum of squares	Variation (%)
Among populations		50,660.71	28.97
Within populations	Pop. 1	17,142.40	9.80
	Pop. 2	93,025.33	53.20
	Pop. 3	6,469.75	3.70
Within individuals	Pop. 1	2,176.00	1.24
	Pop. 2	5,168.00	2.96
	Pop. 3	233.00	0.13
Total		174,875.19	100.00

Note: Analysis of molecular variance (AMOVA) using the genotyping data coming from 21 lettuce varieties.

TABLE 4 Genetic variability of 21 lettuce varieties grouped in the populations (Pops.) obtained by the structure analysis and measured as mean genetic diversity (GD) and observed heterozygosity (H_O).

Population	GD	H_O
Pop. 1	0.148a	0.033a
Pop. 2	0.314ab	0.033b
Pop. 3	0.064c	0.004c

any common characteristics between them, such as leaf color, type, or geographical origin.

The genetic differentiation at both population and individual levels was also analyzed conducting an AMOVA. The results indicated that 28.97% of the genetic variation was observed among populations, 66.70% within populations, and 4.33% within individuals (Table 3). Of the three populations, the highest percentage of variation was found within Pop. 2 (53.20%), followed far behind by Pop. 1 (9.80%), and then by Pop. 3 (3.70%) (Table 3). Within individuals, the same pattern as within populations was obtained, the highest variation was observed within individuals of Pop. 2, followed by those of Pop. 1 and Pop. 3 with values of 2.96%, 1.24%, and 0.13%, respectively (Table 3). These results confirmed that Pop. 2 was the most diverse group and that varieties within Pop. 3 exhibited a strong similarity among them, consistently with the structure and phylogenetic analyses. In the same way, GD was higher in Pop. 2 (0.314), followed by Pop. 1 (0.148), and then by Pop. 3 (0.064) (Table 4). Finally, H_O was 0.033 in both Pop. 1 and Pop. 2, and 0.004 in Pop. 3 (Table 4). These low H_O values shown by the three populations could be explained by the fact that lettuce is a predominantly autogamous crop, so homozygous genotypes are very common. Tripodi et al. (2023) also obtained average H_O values below 4% for the groups of lettuce varieties included in their study. Despite those GD and H_O values, we observed substantial genetic variability within the set of 21 lettuce varieties, especially in the subset of traditional varieties, which is of great importance in breeding programs aimed at developing new varieties with desired traits.

3.2 | LD and association analyses, and identification of putative candidate genes

LD, defined as the nonrandom association of alleles at different loci in a given population, plays an important role in association studies. Therefore, LD was estimated between pairs of SNPs along all chromosomes. The LD decay to half r^2 ranged from 8.84 to 21.79 Mb, with a global average value of 14.92 Mb (Figure S3). Although differences among the chromosomes were observed, LD decay was overall slow, as expected in a predominantly autogamous species like *L. sativa*. Similarly, a long average value of LD decay (9.6 Mb) has been previously reported in lettuce (Simko et al., 2022). In general, autogamous species have lower recombinant rates than those that are allogamous and show slower LD decay. Genetic breeding could be favored from this LD extension, as trait-marker associations could be more probably identified than in the case of species with faster LD decay, in which the regions in LD are shorter and hence could contain a lower number of markers (Flint-Garcia et al., 2003).

In the current study, a total of 9,242 markers were used for carrying out a GWAS of vitamin C content in lettuce. 'Lechuga del Pirineo' was selected to obtain breeding populations since a plant belonging to this traditional variety was the richest in vitamin C in a previous study carried out within our group (Medina-Lozano et al., 2021). Vitamin C content (AA and DHAA) was measured in three 'Lechuga del Pirineo' populations, the original S0, and two generations coming from self-fecundation of the plant with the highest content in vitamin C, S1 and S2 (Table S3; Figure S4). The differences were highly significant ($p < 0.001$) when the average contents of AA (H: 154.38), DHAA (H: 364.97), and TAA (H: 120.73) were compared among generations. Interestingly, even though TAA was lower in S2, the content of the most active form of vitamin C (AA) increased with the two rounds of

selfing (Figure S4). Genetic homogeneity was already reached at S1 for our set of markers, so polymorphisms were only detected in S0 (Table S1). Due to the limited genetic variability of the target population (S0), a diversity panel consisting of 21 lettuce varieties was included in the GWAS, as recommended by Hamazaki et al. (2020). In this way, the whole set of samples consisted of subpopulations with different genetic backgrounds (Figure 1A), which is desirable to detect new variants, and provides the necessary genetic variability and a better genome coverage with polymorphic markers. The variability in terms of vitamin C content was also a bit higher in the diversity panel (152–424 mg × 100 g^{-1}) when compared to 'Lechuga del Pirineo' S0 population (184–424 mg × 100 g^{-1}) (Table S3).

To search for marker associations with the vitamin C content in lettuce, the GWAS was performed on DHAA, AA, and TAA contents using the FarmCPU model. FarmCPU is a multi-locus method able to control false positives incorporating both population structure and kinship, preventing overfitting by the estimation of the associations through fixed and random effect models (X. Liu et al., 2016). Controlling population structure effects is essential in GWAS (Tibbs Cortes et al., 2021), as such, different approaches have been undertaken in the present study. Three different methods were used to assess population structure: obtaining Q from the analysis with STRUCTURE software, MDS, and PCA. Several significantly associated SNPs to DHAA content were found in the same region of chromosome 2 using the three methods to elucidate the aforementioned population structure (Figure 2A), though no significant associations were obtained with either AA or TAA (Figures S5 and S6). Based on the conservative Bonferroni correction, genome-wide threshold was set at 5.27 [$-\log_{10}(5.41 \times 10^{-6})$], as explained earlier. The highest number of significantly associated SNPs was obtained using the PCA method, with a total of 17 SNPs (Figure 2A; Table 5). Using Q values and MDS, five and four associated SNPs were identified, respectively. Not only the number of associated SNPs but also the significance levels were higher using PCA, that ranged from 4.30×10^{-6} to 7.48×10^{-12}, in comparison with the Q values and MDS (2.42×10^{-6} to 1.81×10^{-7} and 4.33×10^{-6} to 7.92×10^{-7}, respectively). Despite the differences, results from the three approaches were consistent given that the significant SNPs found when using Q values and MDS were among the most significant SNPs obtained using PCA, including the lead SNP, which is the most significant one (Table 5). In a previous study carried out on baby leaf lettuce for postharvest and developmental traits, a higher number of significant associations were also reported when the PCA structure was used compared to Q coefficients (Sthapit Kandel et al., 2022).

All the significant SNPs were not only located on chromosome 2, but in a particular region of the chromosome, as mentioned earlier. Therefore, LD was analyzed in detail in

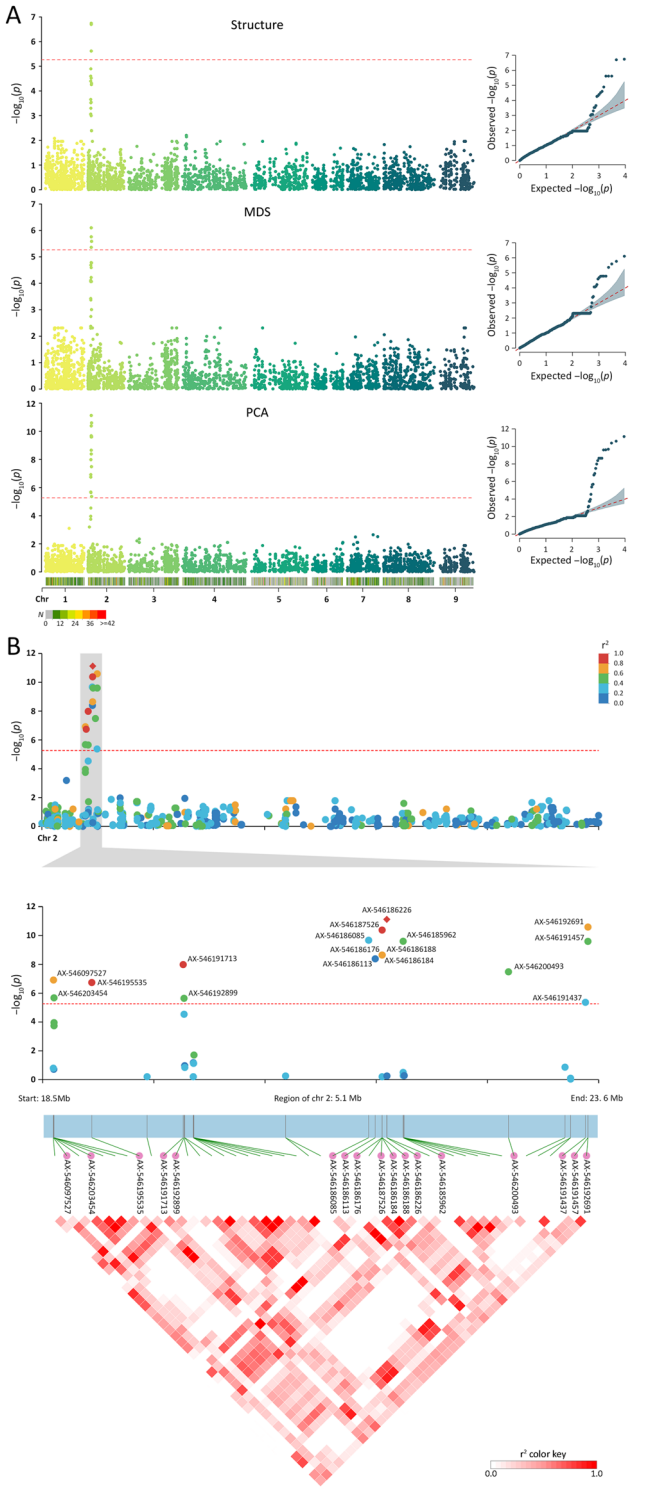

FIGURE 2 Genome-wide association study (GWAS) and linkage disequilibrium (LD) for dehydroascorbic acid (DHAA) content in lettuce samples (21 varieties and 205 plants of 'Lechuga del Pirineo' S0) genotyped with 9,242 single nucleotide polymorphisms (SNPs). (A) Manhattan and quantile–quantile (QQ) plots using the Fixed and random model Circulating Probability Unification (FarmCPU) considering both kinship and the structure of the samples measured with three methods: subpopulation membership coefficients (Q) obtained with STRUCTURE software using the linkage model and

(Continues)

TABLE 5 SNP (single nucleotide polymorphism) markers significantly associated with DHAA (dehydroascorbic acid) content in 226 lettuce plants using three approaches to assess the population structure: PCA (principal component analysis), MDS (multidimensional scaling) and STRUCTURE software.

			p-value		
SNP	**Chromosome**	**Position (bp)**	**PCA**	**MDS**	**Structure**
AX-546097527	2	18,549,215	1.21×10^{-7}	–	–
AX-546203454	2	18,553,852	2.09×10^{-6}	–	–
AX-546195535	2	18,908,075	1.81×10^{-7}	–	–
AX-546191713	2	19,764,398	1.03×10^{-8}	–	–
AX-546192899	2	19,772,368	2.26×10^{-6}	–	–
AX-546186085	2	21,502,354	2.10×10^{-10}	–	–
AX-546186113	2	21,564,067	4.08×10^{-9}	–	–
AX-546186176	2	21,628,199	2.23×10^{-9}	–	2.42×10^{-6}
AX-546187526	2	21,628,541	4.12×10^{-11}	1.73×10^{-6}	1.81×10^{-7}
AX-546186184	2	21,629,055	2.23×10^{-9}	–	2.42×10^{-6}
AX-546186188	2	21,629,299	2.23×10^{-9}	–	2.42×10^{-6}
AX-546186226	2	21,672,408	7.48×10^{-12}	7.92×10^{-7}	2.00×10^{-7}
AX-546185962	2	21,825,828	2.48×10^{-10}	4.33×10^{-6}	–
AX-546200493	2	22,812,807	3.27×10^{-8}	–	–
AX-546191437	2	23,533,970	4.30×10^{-6}	–	–
AX-546191457	2	23,557,223	2.52×10^{-10}	–	–
AX-546192691	2	23,557,727	2.55×10^{-11}	2.58×10^{-6}	–

FIGURE 2 (Continued)
K = 3, MDS (multidimensional scaling) analysis, and PCA (principal component analysis). Statistical significance threshold is shown with the horizontal line ($-\log_{10}(0.05/9242) = 5.27$). Chr: chromosome; *N*: marker density. (B) Zoom of the region harboring the significantly associated SNPs to DHAA and the squared correlation coefficients (r^2) of each marker with the lead SNP. LD patterns for the 17 SNPs significantly associated to DHAA are shown. Triangle plot depicts the LD structure of the associated region.

the region harboring the associated SNPs that covered from 18.5 to 23.6 Mb. Among them, 12 markers showed a high LD with the lead SNP ($r^2 > 0.5$) (Figure 2B). Interestingly, high LD values of the lead SNP with the rest of the SNPs present in that region were exclusively found among the significant ones. Nevertheless, a higher significance in the set of associated SNPs did not necessarily imply a higher LD with the lead SNP. To illustrate this, the second most significant SNP, and the ones following it, did not show a sequentially decreasing LD with the lead SNP (Figure 2B). These differences in the LD values observed for the most significant SNPs among the associated ones might mean that there is more than one polymorphism responsible for (or linked to) the mutation that influences DHAA content. This makes sense because vitamin C content is a complex trait controlled by multiple genes. Similarly, the associated SNPs that were physically closest to the lead SNP were not necessarily the ones with the highest LD values (Figure 2B). One reason to explain this might be that the breeding for a phenotype of a particular trait, which is controlled by more than one locus, may have resulted in the selection of variants in those loci, which will then be in high LD although they can be physically distant (Flint-Garcia et al., 2003). In addition, the significantly associated SNPs might not be within the gene responsible for the phenotypic variation observed but be in high LD with it. For this reason, genes that are in this region must be explored to find candidates related to changes in lettuce DHAA content, as not all genes were covered with SNPs.

A total of 84 genes were found in the region comprised between 18.5 and 23.6 Mb of chromosome 2 (Figure 3; Table 6). The 17 significant SNPs were within the sequence of 12 of those genes. In particular, the lead SNP was located in the uncharacterized gene of a long non-coding RNA (lncRNA) (LOC111920743), which is a class of RNA molecules of over 200 nucleotides length with none or limited coding capacity. They have been intensively studied in recent years and are known to affect gene expression in many biological processes in plants, as reviewed by J. Liu et al. (2015). Therefore, the lncRNA containing the lead SNP in the current study could be regulating the expression of genes that participate in DHAA accumulation, as it is the case of the lncRNAs targeting different genes related to vitamin C content found in kiwifruit (Deng et al., 2022). Among the other genes that harbor the associated

TABLE 6 Putative candidate genes in the 5.1 Mb region of chromosome 2 containing the 17 single nucleotide polymorphisms (SNPs) associated to DHAA (dehydroascorbic acid) content.

Gene ID	SNP	Start	End	Strand	Gene name (GFF annotation)	Abbreviated name	Function (according to UniProt)
LOC111894682	AX-546203383, **AX-546097527**, AX-546203428, **AX-546203454**, AX-546202294, AX-546097531, AX-546202300	**18,547,391**	**18,556,627**	**+**	**GDSL (glycine-asparagine-serine-leucine) esterase/lipase At3g14820 (*Arabidopsis thaliana* GDSL-like lipase/acylhydrolase superfamily)-like**	***GELP***	**Catalysis of acyltransfer or hydrolase reactions with lipid and non-lipid substrates**
LOC111894684		18,581,963	18,586,115	–	GDSL esterase/lipase EXL3 (extracellular lipase 3)	*GELP EXL3*	Catalysis of acyltransfer or hydrolase reactions with lipid and non-lipid substrates
LOC111894685		18,604,078	18,604,788	+	LOC111894685	LOC111894685	Uncharacterized protein-coding
LOC111892004	**AX-546195535**	**18,907,246**	**18,910,005**	**+**	**GDSL esterase/lipase At5g42170 (*A. thaliana* SGNH (serine-glycine-asparagine-histidine) hydrolase-type esterase superfamily)**	***GELP***	**Catalysis of acyltransfer or hydrolase reactions with lipid and non-lipid substrates**
LOC111892005		18,998,678	19,001,618	+	GDSL esterase/lipase At5g42170 (*A. thaliana* SGNH hydrolase-type esterase superfamily)	*GELP*	Catalysis of acyltransfer or hydrolase reactions with lipid and non-lipid substrates
LOC111892006		19,138,697	19,140,501	+	LOC111892006	LOC111892006	Uncharacterized long non-coding RNA (lncRNA)
LOC111917458		19,341,283	19,342,294	–	LOC111917458	LOC111917458	Uncharacterized protein-coding
LOC111917457	AX-546196703	19,416,122	19,426,404	–	GDSL esterase/lipase EXL3 (extracellular lipase) 3	*GELP EXL3*	Catalysis of acyltransfer or hydrolase reactions with lipid and non-lipid substrates
LOC111917459		19,590,644	19,590,761	+	5S ribosomal RNA (ribonucleic acid)	*5S rRNA*	Enhancement of protein synthesis by stabilizing ribosome structure
LOC111917455		19,610,262	19,616,504	–	GDSL esterase/lipase EXL3	*GELP EXL3*	Catalysis of acyltransfer or hydrolase reactions with lipid and non-lipid substrates
LOC111917456		19,666,560	19,667,356	–	LOC111917456	LOC111917456	Uncharacterized protein-coding
LOC111917454	**AX-546191713**	**19,764,319**	**19,768,646**	**–**	**FREE1 (FYVE (Fab1 (1-phosphatidylinositol 3-phosphate 5-kinase), YOTB, Vac1 (vacuolar transport protein), and EEA1 (early endosome antigen 1)-domain protein required for endosomal sorting) 1**	***FREE1***	**Endosomal sorting complex required for transport (ESCRT) component regulating multivesicular body (MVB) protein sorting, intra-luminal vesicles formation and ubiquitin-dependent protein degradation**

(Continues)

TABLE 6 (Continued)

Gene ID	SNP	Start	End	Strand	Gene name (GFF annotation)	Abbreviated name	Function (according to UniProt)
LOC111917451	**AX-546192899**, AX-546192891	**19,772,200**	**19,776,020**	**–**	**FREE1**	***FREE1***	**ESCRT component regulating MVB protein sorting, intra-luminal vesicles formation and ubiquitin-dependent protein degradation**
LOC111917453	AX-546192855, AX-546191621	19,776,750	19,779,968	–	LOC111917453	LOC111917453	Uncharacterized protein-coding
LOC111917450	AX-546192833, AX-546192818, AX-546191589, AX-546191573, AX-546080448, AX-546191538, AX-546192770	19,858,104	19,867,974	–	ABC (ATP (adenosine triphosphate)-binding cassette) transporter B family member 13	*ABCB13*	Auxin efflux transmembrane transporter that is a member of the multidrug resistance P-glycoprotein (MDR/PGP) subfamily of ABC transporters
LOC111916092		20,146,243	20,147,217	+	LOC111916092	LOC111916092	Uncharacterized protein-coding
LOC111916094		20,158,748	20,158,880	+	Small nucleolar RNA104	*snoRNA104*	rRNA processing and maturation
LOC111916081		20,339,463	20,340,294	+	Auxin-responsive protein SAUR (small auxin up-regulated RNA) 50	*SAUR50*	Type 2C phosphatases activity inhibition and cell expansion promotion
LOC111916091		20,608,411	20,611,033	+	LOC111916091	LOC111916091	Uncharacterized protein-coding
LOC111916090		20,617,214	20,620,060	+	LOC111916090	LOC111916090	Uncharacterized protein-coding
LOC111916089		20,620,463	20,621,497	+	LOC111916089	LOC111916089	Uncharacterized protein-coding
LOC111916088		20,639,474	20,647,569	–	LOC111916088	LOC111916088	Uncharacterized lncRNA
LOC111916086		20,658,193	20,660,252	+	LOC111916086	LOC111916086	Uncharacterized protein-coding
LOC128132278		20,669,279	20,681,175	+	LOC128132278	LOC128132278	Uncharacterized protein-coding
LOC128132009		20,682,450	20,683,167	+	LOC128132009	LOC128132009	Uncharacterized lncRNA
LOC111916083	AX-546206772, AX-546207933	20,719,893	20,724,733	–	ATP sulfurylase 2	*APS2*	Catalysis of the first step of the sulfate assimilation pathway
LOC111916082		20,730,886	20,739,630	+	5′−3′ exoribonuclease 3	*XRN3*	Suppression of post-transcriptional gene silencing and processing of pre-rRNAs
LOC111916085		20,792,423	20,823,413	+	UTP (uridine triphosphate):RNA uridylyltransferase 1	*URT1*	Uridylation of mRNAs to reduce accumulation of oligo(A)-tailed mRNAs, repair desadenylated mRNA ends and prevent the biogenesis of siRNAs (small interfering RNAs)

(Continues)

TABLE 6 (Continued)

Gene ID	SNP	Start	End	Strand	Gene name (GFF annotation)	Abbreviated name	Function (according to UniProt)
LOC111916084		20,850,280	20,850,687	–	LOC111916084	LOC111916084	Uncharacterized protein-coding
LOC111916093		20,855,711	20,855,817	–	Small nucleolar RNA R71	*snoRNA R71*	rRNA processing and maturation
LOC111920731		21,335,071	21,340,839	+	Cytokinin riboside 5′-monophosphate phosphoribohydrolase LOG7	*LOG7*	Phosphoribohydrolase that converts inactive cytokinin nucleotides to the biologically active free-base forms
LOC111920722		21,351,445	21,359,082	–	Endoglucanase 2	*EGL2*	Cell wall remodeling
LOC111920723		21,459,462	21,465,181	–	LOC111920723	LOC111920723	Uncharacterized protein-coding
LOC111920732		21,492,738	21,494,083	–	Flavonol 3-sulfotransferase	*FST*	O-sulfation of position 3 of flavonols
LOC111920733		21,497,994	21,499,517	–	Flavonol 3-sulfotransferase-like	*FST*	O-sulfation of position 3 of flavonols
LOC111920734	**AX-546186085**	**21,500,026**	**21,502,406**	**–**	**Diphthine methyl ester synthase**	***DPH***	**Catalysis of four methylation reactions of the modified target histidine residue in translation elongation factor 2 (EF2), to form an intermediate called diphthine methyl ester**
LOC111920735	**AX-546186113**	**21,560,289**	**21,564,252**	**+**	**HVA22 (*Hordeum vulgare* abscisic acid-induced protein) 22-like protein G**	***HVA22G***	**ABA (abscisic acid)-induced negative regulation of vesicle transport**
LOC111920737		21,592,197	21,593,540	–	Flavonol 3-sulfotransferase	*FST*	O-sulfation of position 3 of flavonols
LOC111920738		21,597,402	21,599,095	–	Flavonol 3-sulfotransferase-like	*FST*	O-sulfation of position 3 of flavonols
LOC111920739		21,599,488	21,601,854	–	Diphthine methyl ester synthase	*DPH*	Catalysis of four methylation reactions of the modified target histidine residue in translation EF2, to form an intermediate called diphthine methyl ester
LOC111920740		21,605,115	21,611,299	–	Transcription factor E2FC (Elongation Factor 2FC)	*E2FC*	Transcriptional repression of E2F-regulated genes in mature differentiated cells
LOC111920741		21,619,154	21,623,776	–	LOC111920741	LOC111920741	Uncharacterized protein-coding
LOC111920742	**AX-546186176**, AX-546186179, **AX-546187526**, **AX-546186184**, **AX-546186188**	**21,625,830**	**21,630,483**	**–**	**WPP (tryptophan-proline-proline) domain-associated protein**	***WPP***	**Regulation of the mitotic activity in roots. It plays a role with HSP (Heat Shock Protein) 70–1 in facilitating WIT (WPP domain-interacting tail-anchored protein) 1 nuclear envelope targeting**
LOC111920743	**AX-546186226**, AX-546187557	**21,671,729**	**21,675,836**	**+**	**LOC111920743**	**LOC111920743**	**Uncharacterized lncRNA**
LOC111920744		21,680,488	21,681,316	+	LOC111920744	LOC111920744	Uncharacterized lncRNA
LOC128132153		21,738,201	21,739,575	+	LOC128132153	LOC128132153	Uncharacterized lncRNA
LOC111920727		21,740,681	21,742,059	+	LOC111920727	LOC111920727	Uncharacterized protein-coding

(Continues)

TABLE 6 (Continued)

Gene ID	SNP	Start	End	Strand	Gene name (GFF annotation)	Abbreviated name	Function (according to UniProt)
LOC111920745		21,743,475	21,746,258	+	MDIS (male discoverer) 1-interacting receptor like kinase 2	*MIK2*	It acts as a receptor of SCOOPs (serine-rich endogenous peptides) regulating multiple processes including plant growth, development and stress responses
LOC128132154		21,746,263	21,749,060	–	LOC128132154	LOC128132154	Uncharacterized lncRNA
LOC111920746		21,762,945	21,767,039	–	Non-specific lipid transfer protein GPI (glycosyl-phosphatidylinositol)-anchored 20	*LTPG20*	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
LOC111920728		21,767,368	21,768,192	+	LOC111920728	LOC111920728	Uncharacterized protein-coding
LOC111920747	**AX-546185962**, AX-546058147, AX-546184536, AX-546185932, AX-546185928	**21,825,566**	**21,829,126**	**+**	**Anion transporter 4;2C chloroplastic**	***ANTR4;2C***	**Inorganic phosphate and probable anion transporter**
LOC111920748	AX-546184483	21,835,623	21,836,464	+	LOC111920748	LOC111920748	Uncharacterized lncRNA
LOC111920729		21,992,636	21,993,391	+	LOC111920729	LOC111920729	Uncharacterized protein-coding
LOC111920749		22,292,212	22,294,828	+	LOC111920749	LOC111920749	Uncharacterized protein-coding
LOC128132280		22,490,627	22,491,776	–	Pectinesterase/pectinesterase inhibitor 47	*PPE47*	Modification of cell walls through demethylesterification or inhibition of demethylesterification in cell wall pectin
LOC111920750		22,530,412	22,534,962	+	LOC111920750	LOC111920750	Uncharacterized protein-coding
LOC111920752		22,546,568	22,549,510	+	LOC111920752	LOC111920752	Uncharacterized lncRNA
LOC122196667		22,628,569	22,629,290	+	LOC122196667	LOC122196667	Uncharacterized protein-coding
LOC111917719		22,701,425	22,731,739	+	NADPH (nicotinamide adenine dinucleotide phosphate)-cytochrome P450 reductase 1-like	*NADPHP450R1*	Electron transfer from NADP to cytochrome P450 in microsomes and to heme oxygenase and cytochrome B5
LOC111917708		22,790,982	22,792,590	–	F-box protein (FBP) At2g21930 (*A. thaliana* F-box associated ubiquitination effector family protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917691		22,805,611	22,807,192	–	FBP CPR (constitutive expressor of PR (pathogenesis-related) genes) 1	*FBP CPR1*	Negative regulation of both salicylic acid (SA)-dependent and SA-independent defense signaling
LOC111917627	**AX-546200493**	**22,812,244**	**22,813,737**	**–**	**FBP At4g19940 (*A. thaliana* F-box and associated interaction domains-containing protein)**	***FBP***	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome

(Continues)

TABLE 6 (Continued)

Gene ID	SNP	Start	End	Strand	Gene name (GFF annotation)	Abbreviated name	Function (according to UniProt)
LOC111917682		22,830,699	22,831,826	–	FBP At3g52320 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC122196669		22,876,211	22,877,228	–	LOC122196669	LOC122196669	Uncharacterized lncRNA
LOC111917600		22,904,314	22,908,704	+	LOC111917600	LOC111917600	Uncharacterized protein-coding
LOC111887812		22,910,075	22,918,074	+	LOC111887812	LOC111887812	Uncharacterized protein-coding
LOC128132155		22,910,927	22,911,716	–	LOC128132155	LOC128132155	Uncharacterized lncRNA
LOC111917652		22,921,889	22,923,031	–	FBP At1g30790 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917642		22,970,133	22,971,275	–	FBP At1g11270 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917634		22,984,801	22,985,943	–	FBP At1g11270 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917737		23,205,208	23,206,723	–	FBP At4g21240 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917589		23,284,926	23,286,438	–	FBP At1g47790 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917578	AX-546192898	23,342,399	23,343,866	–	FBP At1g11270 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC111917572		23,352,039	23,353,564	+	LOC111917572	LOC111917572	Uncharacterized lncRNA
LOC111917538		23,359,493	23,364,932	+	LOC111917538	LOC111917538	Uncharacterized lncRNA
LOC111917526	AX-546192844, AX-546194047	23,394,448	23,395,853	+	FBP At1g32420 (*A. thaliana* F-box and associated interaction domains-containing protein)	*FBP*	Regulation of protein degradation tagging them for ubiquitination and subsequent degradation by 26S proteasome
LOC122196440		23,395,967	23,397,609	–	LOC122196440	LOC122196440	Uncharacterized lncRNA
LOC111917726		23,410,176	23,412,200	–	LOC111917726	LOC111917726	Uncharacterized lncRNA
LOC111905617		23,454,527	23,458,447	+	LOC111905617	LOC111905617	Uncharacterized protein-coding

(Continues)

TABLE 6 (Continued)

Gene ID	SNP	Start	End	Strand	Gene name (GFF annotation)	Abbreviated name	Function (according to UniProt)
LOC128132010		23,464,349	23,466,565	+	LOC128132010	LOC128132010	Uncharacterized lncRNA
LOC128132281		23,487,588	23,488,052	+	Extensin-1-like	*EXT1*	Structural component which strengthens the primary cell wall
LOC111903668	**AX-546191437**	**23,533,897**	**23,535,387**	**–**	**F-box/kelch-repeat protein At3g23880 (*A. thaliana* F-box and associated interaction domains-containing protein)**	***KFB***	**FBP of Kelch subfamily that regulates protein degradation by targeting specific substrates**
LOC111903669	**AX-546191457, AX-546192691**	**23,554,314**	**23,563,321**	**+**	**LOC111903669**	**LOC111903669**	**Uncharacterized protein-coding**

Note: Significantly associated SNPs with DHAA content and the genes harboring them are shown in bold.

SNPs with DHAA content, different gene products were identified, like GDSL (glycine-asparagine-serine-leucine) esterase/lipases (LOC111894682, LOC111892004), FREE1 proteins (LOC111917454, LOC111917451), a diphthine methyl ester synthase (LOC111920734), the HVA22-like protein G (LOC111920735), a WPP (tryptophan-proline-proline) domain-associated protein (LOC111920742), an anion transporter (LOC111920747), and two F-box proteins (LOC111917627, LOC111903668) (Table 6). Ten more genes that encode F-box proteins were found in the studied region (Table 6). These results are interesting because a gene encoding an F-box protein has been recently proposed as a candidate gene to regulate the ascorbate peroxidase (APX) activity in a GWAS carried out in barley (Thabet et al., 2022). The APX catalyzes the conversion of AA to DHAA (through ascorbate) as part of the reduction of hydrogen peroxide to water (Apel & Hirt, 2004). F-box proteins have been related to antioxidant status in plants in other previous studies. For example, the overexpression of the *TaFBA1* gene that encodes an F-box protein, enhanced the oxidative stress response through a higher APX activity in wheat (S. M. Zhou et al., 2015). Therefore, the genes in the region of interest in the current study encoding F-box proteins are potential candidates for the control of DHAA content in lettuce. Another interesting candidate gene is the one that encodes the pectinesterase/pectinesterase inhibitor (*PPE*) 47 (LOC128132280) (Table 6). Vitamin C biosynthesis has been widely studied in plants, being the D-manose/L-galactose or Smirnoff-Wheeler pathway (Wheeler et al., 1998) the main route for AA synthesis, though there are at least three alternative pathways. *PPEs* have been previously proposed as candidate genes involved in the increase of vitamin C content through the alternative D-galacturonic acid pathway that starts with the degradation of cell wall pectins (Di Matteo et al., 2010; Ruggieri et al., 2015). In our analysis, we found markers associated with DHAA instead of AA, the last being the compound assessed in both studies just mentioned. However, AA and DHAA contents are directly related as they are interconvertible. Plants need to maintain a balance of antioxidant compounds and DHAA is the product of the AA oxidation mediated by APX (among others), and it is recycled to AA by the action of the dehydroascorbate reductase (DHAR) in the ascorbate-glutathione cycle (Apel & Hirt, 2004). Two more candidate genes, encoding an endoglucanase 2 (EGL2) (LOC111920722) and an extensin-1-like (EXT1) protein (LOC128132281), also play a role in the cell wall remodeling, which could ultimately alter vitamin C content (Table 6). There are also 39 out of the 84 candidate genes that have not been characterized yet in lettuce, including 16 lncRNA (Table 6). They may participate in DHAA accumulation, but more efforts in gene annotation are needed to shed light on their potential functions and contribution to DHAA content in lettuce.

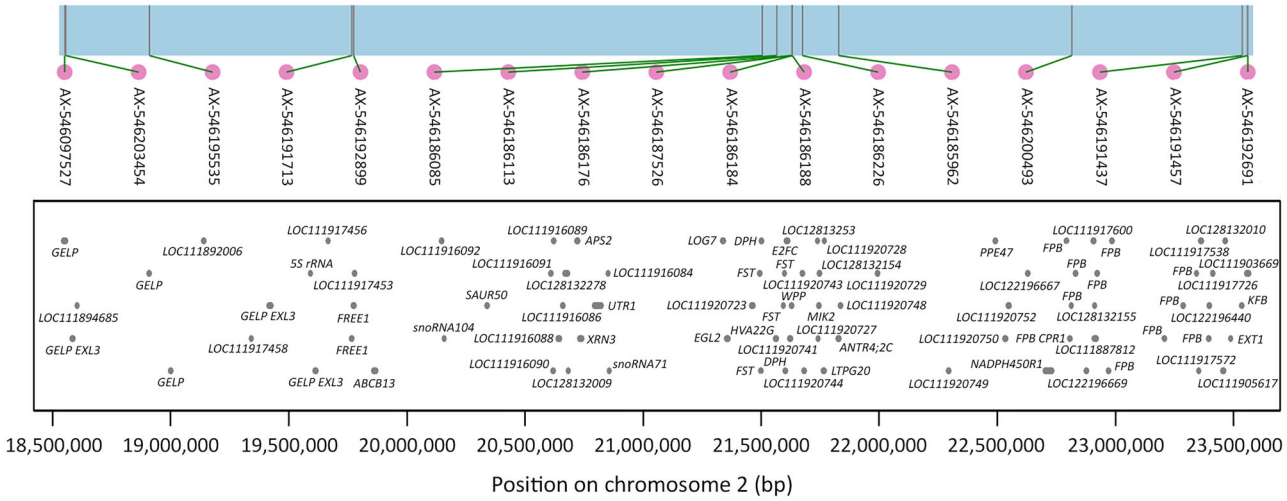

FIGURE 3 Physical map of the 5.1 Mb region containing the 17 single nucleotide polymorphisms (SNPs) significantly associated to dehydroascorbic acid (DHAA) content. Abbreviations of the gene names are described in Table 6.

As far as we know, this is the first GWAS that has found genetic associations with the DHAA content in lettuce. GWAS has been used as a powerful approach in searching for genes or genetic markers associated with different traits in lettuce. However, most studies have assessed resistance to biotic stresses (Lu et al., 2014; Simko et al., 2022, 2023) or agronomic traits like leaf color and morphology, bolting and flowering times, or shelf life and developmental rate (S. Kwon et al., 2013; S. Park et al., 2021; Sthapit Kandel et al., 2020; Wei et al., 2021). In other crops, the regulation of vitamin C content has actually been analyzed by GWAS. Candidate genes related to the common biosynthesis pathways have been reported, such as genes encoding a monodehydroascorbate reductase, a bHLH transcription factor that regulates genes of the D-mannose/L-galactose pathway, and also a PPE like the one identified in our study, what reinforces it as a candidate gene (Berdugo-Cely et al., 2023; Sauvage et al., 2014; Ye et al., 2019). New candidate genes have also been proposed to have an effect on vitamin C content, like the Fas-associated factor 1-like (*FAF1*) and ethylene responsive factor 1 (*ERF1*) genes in tomato (Ruggieri et al., 2014) or genes that encode methionine sulfoxide reductases (*MSR*) in potato (Berdugo-Cely et al., 2023). Except for Sauvage et al. (2014), these studies have evaluated associations exclusively with AA, the most biologically active form of vitamin C, and have not paid attention to DHAA. However, DHAA is essential to maintain the reduced AA pool through its recycling by the action of the DHAR, for instance. In addition, DHAA also presents some biological activity, as mentioned before. Consequently, identifying candidate genes that modulate DHAA content, and not only AA, is essential to understand vitamin C regulation in crops and, in particular, in lettuce. Furthermore, identification of molecular markers associated to DHAA content might allow the use of marker-assisted selection in breeding programs aimed at obtaining vitamin C biofortified lettuce.

AUTHOR CONTRIBUTIONS

Inés Medina-Lozano: Data curation; formal analysis; investigation; methodology; visualization; writing—original draft; writing—review and editing. **Juan Ramón Bertolín**: Data curation; formal analysis; methodology; writing—review and editing. **Jörg Plieske**: Data curation; formal analysis; methodology; writing—review and editing. **Martin Ganal**: Supervision; writing—review and editing. **Heike Gnad**: Resources; writing—review and editing. **Aurora Díaz**: Conceptualization; funding acquisition; investigation; methodology; project administration; resources; supervision; writing—original draft; writing—review and editing.

ACKNOWLEDGMENTS

We thank J. A. Aranjuelo and "laboratorio de valoración nutritiva" from CITA for technical support and the staff of SGS Institut Fresenius GmbH TraitGenetics Section for performing the Axiom and Illumina array genotyping. We are also grateful to D. L. Goodchild for reviewing the English language and to the Vegetable Germplasm Bank of Zaragoza (BGHZ-CITA, Spain), the Centre for Genetic Resources (CGN, Wageningen, Netherlands), and Ramiro Arnedo Semillas S.A. for supplying part of the seeds used here. This work was funded by the projects PID2022-138484OR-I00 from the Spanish Ministry of Science and Innovation (MCIN) and State Research Agency (AEI) and AGROALNEXT from the MCIN with funding from European Union (EU) NextGenerationEU: PRTR-C17.I1, and by the European Social Fund from the EU through the Government of Aragon (A12_23R: "Grupo de investigación en fruticultura: caracterización, adaptación y mejora genética").

Inés Medina-Lozano was supported by a predoctoral contract for training doctors from the Spanish Ministry of Science, Innovation and Universities (MCIU) and the AEI.

CONFLICT OF INTEREST STATEMENT

Jörg Plieske, Heike Gnad, and Martin Ganal are members of the company SGS Institut Fresenius GmbH TraitGenetics Section, which offers lettuce SNP genotyping. The other authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings shown in this study and needed to reproduce the analyses, including the data of the SNP markers, are included in the main manuscript and the supplemental files or are publicly available.

ORCID

Inés Medina-Lozano https://orcid.org/0000-0001-5533-3505
Aurora Díaz https://orcid.org/0000-0001-7297-1699

REFERENCES

Apel, K., & Hirt, H. (2004). Reactive oxygen species: Metabolism, oxidative stress, and signal transduction. *Annual Review of Plant Biology*, *55*, 373–399. https://doi.org/10.1146/annurev.arplant.55.031903.141701

Berdugo-Cely, J. A., Céron-Lasso, M. D. S., & Yockteng, R. (2023). Phenotypic and molecular analyses in diploid and tetraploid genotypes of *Solanum tuberosum* L. reveal promising genotypes and candidate genes associated with phenolic compounds, ascorbic acid contents, and antioxidant activity. *Frontiers in Plant Science*, *13*, 1007104. https://doi.org/10.3389/fpls.2022.1007104

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., & Guarino, L. (2017). Past and future use of wild relatives in crop breeding. *Crop Science*, *57*(3), 1070–1082. https://doi.org/10.2135/cropsci2016.10.0885

Deng, H., Xia, H., Guo, Y., Liu, X., Lin, L., Wang, J., Xu, K., Lv, X., Hu, R., & Liang, D. (2022). Dynamic changes in ascorbic acid content during fruit development and ripening of *Actinidia latifolia* (an ascorbate-rich fruit crop) and the associated molecular mechanisms. *International Journal of Molecular Sciences*, *23*(10), 5808. https://doi.org/10.3390/ijms23105808

Díaz, A., Martín-Hernández, A. M., Dolcet-Sanjuan, R., Garcés-Claver, A., Álvarez, J. M., Garcia-Mas, J., Picó, B., & Monforte, A. J. (2017). Quantitative trait loci analysis of melon (*Cucumis melo* L.) domestication-related traits. *Theoretical and Applied Genetics*, *130*, 1837–1856. https://doi.org/10.1007/s00122-017-2928-y

Di Matteo, A., Sacco, A., Anacleria, M., Pezzotti, M., Delledonne, M., Ferrarini, A., Frusciante, L., & Barone, A. (2010). The ascorbic acid content of tomato fruits is associated with the expression of genes involved in pectin degradation. *BMC Plant Biology*, *10*, Article 163. https://doi.org/10.1186/1471-2229-10-163

Dong, S. S., He, W. M., Ji, J. J., Zhang, C., Guo, Y., & Yang, T. L. (2021). LDBlockShow: A fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Briefings in Bioinformatics*, *22*(4), bbaa227. https://doi.org/10.1093/bib/bbaa227

Doyle, J. J., & Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus*, *12*, 13–15.

Dyląg, A., Smoleń, S., Wisła-Świder, A., Kowalska, I., Sularz, O., Krzemińska, J., Pitala, J., & Koronowicz, A. (2023). Evaluation of the chemical composition and nutritional value of lettuce (*Lactuca sativa* L.) biofortified in hydroponics with iodine in the form of iodoquinolines. *Frontiers in Plant Science*, *14*, 1288773. https://doi.org/10.3389/fpls.2023.1288773

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

FAOSTAT. (2022). *Statistics of the Food and Agriculture Organization of the United Nations*. https://www.fao.org/faostat

Flint-Garcia, S., Feldmann, M. J., Dempewolf, H., Morrell, P. L., & Ross-Ibarra, J. (2023). Diamonds in the not-so-rough: Wild relative diversity hidden in crop genomes. *PLoS Biology*, *21*, e3002235. https://doi.org/10.1371/journal.pbio.3002235

Flint-Garcia, S., Thornsberry, J. M., & Buckler, E. S., IV (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology*, *54*(7), 357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Granger, M., & Eck, P. (2018). Dietary vitamin C in human health. *Advances in Food and Nutrition Research*, *83*, 281–310. https://doi.org/10.1016/bs.afnr.2017.11.006

Guo, X., Liu, R. H., Fu, X., Sun, X., & Tang, K. (2013). Over-expression of l-galactono-γ-lactone dehydrogenase increases vitamin C, total phenolics and antioxidant activity in lettuce through bio-fortification. *Plant Cell, Tissue and Organ Culture*, *114*, 225–236. https://doi.org/10.1007/s11240-013-0318-y

Hamazaki, K., Kajiya-Kanegae, H., Yamasaki, M., Ebana, K., Yabe, S., Nakagawa, H., & Iwata, H. (2020). Choosing the optimal population for a genome-wide association study: A simulation of whole-genome sequences from rice. *The Plant Genome*, *13*(1), e20005. https://doi.org/10.1002/tpg2.20005

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Kwon, S., Simko, I., Hellier, B., Mou, B., & Hu, J. (2013). Genome-wide association of 10 horticultural traits with expressed sequence tag-derived SNP markers in a collection of lettuce lines. *The Crop Journal*, *1*(1), 25–33. https://doi.org/10.1016/j.cj.2013.07.014

Kwon, S. J., Truco, M. J., & Hu, J. (2012). LSGermOPA, a custom OPA of 384 EST-derived SNPs for high-throughput lettuce (*Lactuca sativa* L.) germplasm fingerprinting. *Molecular Breeding*, *29*(4), 887–901. https://doi.org/10.1007/s11032-011-9623-5

Liu, J., Wang, H., & Chua, N. H. (2015). Long noncoding RNA transcriptome of plants. *Plant Biotechnology Journal*, *13*(3), 319–328. https://doi.org/10.1111/pbi.12336

Liu, K., & Muse, S. V. (2005). PowerMaker: An integrated analysis environment for genetic maker analysis. *Bioinformatics*, *21*(9), 2128–2129. https://doi.org/10.1093/bioinformatics/bti282

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient

genome-wide association studies. *PLOS Genetics*, *12*(2), e1005767. https://doi.org/10.1371/journal.pgen.1005767

Lu, H., Hu, J., & Kwon, S. J. (2014). Association analysis of bacterial leaf spot resistance and SNP markers derived from expressed sequence tags (ESTs) in lettuce (*Lactuca sativa* L.). *Molecular Breeding*, *34*, 997–1006. https://doi.org/10.1007/s11032-014-0092-5

Medina-Lozano, I., Bertolín, J. R., & Díaz, A. (2021). Nutritional value of commercial and traditional lettuce (*Lactuca sativa* L.) and wild relatives: Vitamin C and anthocyanin content. *Food Chemistry*, *359*, 129864. https://doi.org/10.1016/j.foodchem.2021.129864

Medina-Lozano, I., Bertolín, J. R., & Díaz, A. (2024). Impact of drought stress on vitamin C and anthocyanin content in cultivated lettuces (*Lactuca sativa* L.) and wild relatives (*Lactuca* spp.). *Frontiers in Plant Science*, *15*, 3389. https://doi.org/10.3389/fpls.2024.1369658

Medina-Lozano, I., Bertolín, J. R., Zufiaurre, R., & Diaz, A. (2020). Improved UPLC-UV method for the quantification of vitamin C in lettuce varieties (*Lactuca sativa* L.) and crop wild relatives (*Lactuca* spp.). *Journal of Visualized Experiments*, *160*, e61440. https://doi.org/10.3791/61440

Medina-Lozano, I., & Díaz, A. (2021). Nutritional value and phytochemical content of crop landraces and traditional varieties. In A. Elkelish (Ed.), *Landraces —Traditional variety and natural breed* (pp. 95–116). IntechOpen. https://doi.org/10.5772/intechopen.95514

Medina-Lozano, I., & Díaz, A. (2022). Applications of genomic tools in plant breeding: Crop biofortification. *International Journal of Molecular Sciences*, *23*(6), 3086. https://doi.org/10.3390/ijms23063086

Park, J. S., Kang, M. Y., Shim, E. J., Oh, J. H., Seo, K. I., Kim, K. S., Sim, S. C., Chung, S. M., Park, Y., Lee, G. P., Lee, W. S., Kim, M., & Jung, J. K. (2022). Genome-wide core sets of SNP markers and Fluidigm assays for rapid and effective genotypic identification of Korean cultivars of lettuce (*Lactuca sativa* L.). *Horticulture Research*, *9*, uhac119. https://doi.org/10.1093/hr/uhac119

Park, S., Kumar, P., Shi, A., & Mou, B. (2021). Population genetics and genome-wide association studies provide insights into the influence of selective breeding on genetic variation in lettuce. *The Plant Genome*, *14*(2), e20086. https://doi.org/10.1002/tpg2.20086

Peng, H., Zhao, R., Smith, R., & Simko, I. (2022). Phenotypic and genetic analyses of yellow spot malady in lettuce. *Scientia Horticulturae*, *305*, 111389. https://doi.org/10.1016/j.scienta.2022.111389

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Ramasamy, R. K., Ramasamy, S., Bindroo, B. B., & Naik, V. G. (2014). STRUCTURE PLOT: A program for drawing elegant STRUCTURE bar plots in user friendly interface. *SpringerPlus*, *3*(1), Article 431. https://doi.org/10.1186/2193-1801-3-431

Rstudio Team. (2020). *RStudio: Integrated development environment for R*. RStudio, PBC. http://doi.wiley.com/10.1002/jwmg.232

Ruggieri, V., Francese, G., Sacco, A., D'Alessandro, A., Rigano, M. M., Parisi, M., Milone, M., Cardi, T., Mennella, G., & Barone, A. (2014). An association mapping approach to identify favourable alleles for tomato fruit quality breeding. *BMC Plant Biology*, *14*(1), Article 337. https://doi.org/10.1186/s12870-014-0337-9

Ruggieri, V., Sacco, A., Calafiore, R., Frusciante, L., & Barone, A. (2015). Dissecting a QTL into candidate genes highlighted the key role of pectinesterases in regulating the ascorbic acid content in tomato fruit. *The Plant Genome*, *8*(2), plantgenome2014.08.0038. https://doi.org/10.3835/plantgenome2014.08.0038

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., Fernie, A. R., & Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiology*, *165*(3), 1120–1132. https://doi.org/10.1104/pp.114.241521

Simko, I. (2023). Dataset on the single nucleotide variation in diversity panel of 500 lettuce accessions genotyped with tunable genotyping-by-sequencing (tGBS) method. *Data in Brief*, *49*, 109419. https://doi.org/10.1016/j.dib.2023.109419

Simko, I., Hasegawa, D. K., Peng, H., & Zhao, R. (2023). Genetic and physiological determinants of lettuce partial resistance to Impatiens necrotic spot virus. *Frontiers in Plant Science*, *14*, 1163683. https://doi.org/10.3389/fpls.2023.1163683

Simko, I., Peng, H., Sthapit Kandel, J., & Zhao, R. (2022). Genome-wide association mapping reveals genomic regions frequently associated with lettuce field resistance to downy mildew. *Theoretical and Applied Genetics*, *135*, 2009–2024. https://doi.org/10.1007/s00122-022-04090-3

Sthapit Kandel, J., Peng, H., Hayes, R. J., Mou, B., & Simko, I. (2020). Genome-wide association mapping reveals loci for shelf life and developmental rate of lettuce. *Theoretical and Applied Genetics*, *133*(6), 1947–1966. https://doi.org/10.1007/s00122-020-03568-2

Sthapit Kandel, J., Sandoya, G. V., Zhou, W., Read, Q. D., Mou, B., & Simko, I. (2022). Identification of quantitative trait loci associated with bacterial leaf spot resistance in baby leaf lettuce. *Plant Disease*, *106*(10), 2583–2590. https://doi.org/10.1094/PDIS-09-21-2087-RE

Stoffel, K., van Leeuwen, H., Kozik, A., Caldwell, D., Ashrafi, H., Cui, X., Tan, X., Hill, T., Reyes-Chin-Wo, S., Truco, M. J., Michelmore, R. W., & Van Deynze, A. (2012). Development and application of a 6.5 million feature Affymetrix Genechip for massively parallel discovery of single position polymorphisms in lettuce (*Lactuca* spp.). *BMC Genomics*, *13*, Article 185. https://doi.org/10.1186/1471-2164-13-185

Thabet, S. G., Alomari, D. Z., Börner, A., Brinch-Pedersen, H., & Alqudah, A. M. (2022). Elucidating the genetic architecture controlling antioxidant status and ionic balance in barley under salt stress. *Plant Molecular Biology*, *110*(3), 287–300. https://doi.org/10.1007/s11103-022-01302-8

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*(1), e20077. https://doi.org/10.1002/tpg2.2007

Tripodi, P., Beretta, M., Peltier, D., Kalfas, I., Vasilikiotis, C., Laidet, A., Briand, G., Aichholz, C., Zollinger, T., van Treuren, R., Scaglione, D., & Goritschnig, S. (2023). Development and application of Single Primer Enrichment Technology (SPET) SNP assay for population genomics analysis and candidate gene discovery in lettuce. *Frontiers in Plant Science*, *14*, 1252777. https://doi.org/10.3389/fpls.2023.1252777

UPOV. (2021). *Guidelines for the conduct of tests for distinctness, homogeneity, and stability* (Document UPOV TG/13/11 Rev 2). UPOV.

USDA. (2022). *FoodData central*. https://fdc.nal.usda.gov/

van Treuren, R., van Eekelen, H. D. L. M., Wehrens, R., & de Vos, R. C. H. (2018). Metabolite variation in the lettuce gene pool: Towards healthier crop varieties and food. *Metabolomics*, *14*(11), Article 146. https://doi.org/10.1007/s11306-018-1443-8

Wang, J., & Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genomics, Pro-*

teomics and Bioinformatics, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wei, T., van Treuren, R., Liu, X., Zhang, Z., Chen, J., Liu, Y., Dong, S., Sun, P., Yang, T., Lan, T., Wang, X., Xiong, Z., Liu, Y., Wei, J., Lu, H., Han, S., Chen, J. C., Ni, X., Wang, J., … Liu, H. (2021). Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. *Nature Genetics*, *53*(5), 752–760. https://doi.org/10.1038/s41588-021-00831-0

Weir, B. S., & Hill, W. G. (2002). Estimating F-statistics. *Annual Review of Genetics*, *36*, 721–750. https://doi.org/10.1146/annurev.genet.36

Wheeler, G. L., Jones, M. A., & Smirnoff, N. (1998). The biosynthetic pathway of vitamin C in higher plants. *Nature*, *393*(6683), 365–369. https://doi.org/10.1038/30728

Ye, J., Li, W., Ai, G., Li, C., Liu, G., Chen, W., Wang, B., Wang, W., Lu, Y., Zhang, J., Li, H., Ouyang, B., Zhang, H., Fei, Z., Giovannoni, J. J., Ye, Z., & Zhang, Y. (2019). Genome-wide association analysis identifies a natural variation in basic helix-loop-helix transcription factor regulating ascorbate biosynthesis via D-mannose/L-galactose pathway in tomato. *PLoS Genetics*, *15*(5), 1008149. https://doi.org/10.1371/journal.pgen.1008149

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics & Bioinformatics*, *19*(4), 619–628. https://doi.org/10.1016/j.gpb.2020.10.007

Zhang, L., Su, W., Tao, R., Zhang, W., Chen, J., Wu, P., Yan, C., Jia, Y., Larkin, R. M., Lavelle, D., Truco, M. J., Chin-Wo, S. R., Michelmore, R. W., & Kuang, H. (2017). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature Communications*, *8*(1), 2264. https://doi.org/10.1038/s41467-017-02445-9

Zhang, W., Alseekh, S., Zhu, X., Zhang, Q., Fernie, A. R., Kuang, H., & Wen, W. (2020). Dissection of the domestication-shaped genetic architecture of lettuce primary metabolism. *The Plant Journal*, *104*(3), 613–630. https://doi.org/10.1111/tpj.14950

Zhou, H., Yu, L., Liu, S., Zhu, A., Yang, Y., Chen, C., Yang, A., Liu, L., & Yu, F. (2023). Transcriptome comparison analyses in UV-B induced AsA accumulation of *Lactuca sativa* L. *BMC Genomics*, *24*(1), Article 61. https://doi.org/10.1186/s12864-023-09133-7

Zhou, S. M., Kong, X. Z., Kang, H. H., Sun, X. D., & Wang, W. (2015). The involvement of wheat F-box protein gene TaFBA1 in the oxidative stress tolerance of plants. *PLoS ONE*, *10*(4), e0122117. https://doi.org/10.1371/journal.pone.0122117

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Medina-Lozano, I., Bertolín, J. R., Plieske, J., Ganal, M., Gnad, H., & Díaz, A. (2024). Studies of genetic diversity and genome-wide association for vitamin C content in lettuce (*Lactuca sativa* L.) using high-throughput SNP arrays. *The Plant Genome*, e20518. https://doi.org/10.1002/tpg2.20518