




Dissection of the domestication-shaped genetic architecture of lettuce primary metabolism

Weiwei Wen^{1,*} , Xiang Zhu⁴, Qinghua Zhang⁵, Alisdair R. Fernie^{2,3,*} , Hanhui Kuang^{1,*} and Weiwei Wen^{1,*} 

¹Key Laboratory of Horticultural Plant Biology (MOE), College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan 430070, China,

²Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, Potsdam-Golm 14476, Germany,

³Center of Plant Systems Biology and Biotechnology, Plovdiv 4000, Bulgaria,

⁴Thermo Fisher Scientific, Shanghai 201206, China, and

⁵National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

Received 1 November 2019; revised 14 July 2020; accepted 21 July 2020.

*For correspondence (e-mail wwen@mail.hzau.edu.cn; fernie@mpimp-golm.mpg.de; kuangfile@mail.hzau.edu.cn).

[†]These authors contributed equally to this work.

SUMMARY

Lettuce (*Lactuca sativa* L.) is an important vegetable crop species worldwide. The primary metabolism of this species is essential for its growth, development and reproduction as well as providing a considerable direct source of energy and nutrition for humans. Here, through investigating 77 primary metabolites in 189 accessions including all major horticultural types and wild lettuce *L. serriola* we showed that the metabolites in *L. serriola* were different from those in cultivated lettuce. The findings were consistent with the demographic model of lettuce and supported a single domestication event for this species. Selection signals among these metabolic traits were detected. Specifically, galactinol, malate, quinate and threonate were significantly affected by the domestication process and cultivar differentiation of lettuce. Galactinol and raffinose might have been selected during stem lettuce cultivation as an adaptation to the local environments in China. Furthermore, we identified 154 loci significantly associated with the level of 51 primary metabolites. Three genes (*LG8749721*, *LG8763094* and *LG5482522*) responsible for the levels of galactinol, raffinose, quinate and chlorogenic acid were further dissected, which may have been the target of domestication and/or affected by local adaptation. Additionally, our findings strongly suggest that human selection resulted in reduced quinate and chlorogenic acid levels in cultivated lettuce. Our study thus provides beneficial genetic resources for lettuce quality improvement and sheds light on the domestication and evolution of this important leafy green.

Keywords: domestication, genome-wide association studies, *Lactuca sativa*, primary metabolism.

INTRODUCTION

The daily intake of a variety of vegetables is highly recommended by dietary guidelines in many countries owing to their health promoting properties (U.S. Department of Health and Human Services and U.S. Department of Agriculture, 2015; Chinese Nutrition Society, 2016). Vegetables are important sources of energy, dietary fibers, minerals, and other beneficial phytochemicals such as antioxidants (Slavin and Lloyd, 2012). Primary metabolites are direct sources of energy and nutrition for humans and are also essential for plant growth, development and reproduction (Rojas *et al.*, 2014; Sulpice and McKeown, 2015). Understanding the natural variation and genetic bases of plant

primary metabolism will thus ultimately contribute to biofortification of plants in a manner that is beneficial to humans in terms of both food security and quality (Fernie and Tohge, 2017).

In order to decipher the genetic basis of primary metabolism, both linkage mapping and genome-wide analysis study have been performed in a variety of plant species (Strauch *et al.*, 2015; Wen *et al.*, 2015, 2018). Variation of primary metabolites tends to be affected by multiple loci with small effect (Rowe *et al.*, 2008; Chan *et al.*, 2010; Wen *et al.*, 2018). Most identified candidate genes for primary metabolites were structural genes in biosynthetic pathways whilst genes which are not essential to the metabolic

biosynthesis may also relate with the contents of primary metabolites (Wen *et al.*, 2015, 2018). For instance, *ACD6* (*ACCELERATED CELL DEATH6*) balances primary metabolism with biotic stress defense in *Arabidopsis* (Fusari *et al.*, 2017). Metabolomics has been used to investigate the genetic architecture of metabolism of plant species with high quality reference genome and well sequenced genetic populations, such as rice, maize and tomato (Fang and Luo, 2019). On the other hand, the diversity and genetic basis of metabolism of many important but less studied species remain to be unveiled.

Lettuce (*Lactuca sativa*) is an important vegetable worldwide and belongs to the Compositae family, which is one of the largest, most widespread and successful families of flowering plants on earth (Funk *et al.*, 2009). As a popular vegetable worldwide, lettuce has diversified genotypes and is rich in nutrition for daily consumption such as fibers, vitamins and amino acids (Yang *et al.*, 2018). In addition, recent research revealed that lettuce could be a potential natural manufacturer of pharmaceuticals for the treatment of hepatitis B virus by producing small artificial RNA, which indicates a bright prospect of lettuce industry (Lei, 2019). The initial domestication of lettuce is believed to have begun in ancient Egypt as early as 4500 years ago with the evidence of wall paintings in tombs, and *Lactuca serriola* is commonly believed to be the direct ancestor of cultivated lettuce (Kesseli *et al.*, 1991; deVries, 1997). According to demographic modeling, lettuce had a single domestication event. The ancient cultivated lettuce probably originated in the Fertile Crescent ~10 800 years B.P and several lettuce types such as butterhead, and romaine were developed in the following few thousands of years in Europe (Zhang *et al.*, 2017). Primitive cultivars were introduced into China and subsequent selection led to the formation of stem lettuce. In the sixteenth century, cultivated lettuce was brought to America and as a product of selection a modern type of cultivar, crisphead, was subsequently generated (Zhang *et al.*, 2017). Domestication results in numerous biochemical and physiological changes in lettuce including variation in metabolite composition and abundance (Zhang *et al.*, 2017; Yang *et al.*, 2018). Increasing studies have demonstrated that many agriculturally and economically important traits especially flavors were tightly associated with alterations in metabolite composition or abundance (Shang *et al.*, 2014; Ye *et al.*, 2017; Sanchez-Perez *et al.*, 2019). For instance, the cyanogenic diglucoside amygdalin is a bitter toxic compound in almond and a non-synonymous point mutation in a *bHLH2* gene leads to reduced bitterness in the domesticated almond (Sanchez-Perez *et al.*, 2019). On the other hand, alteration of metabolite profiles may also be the consequence of selection on other classical traits during domestication (Beleggia *et al.*, 2016; Zhu *et al.*, 2018). Furthermore, Kleessen *et al.* (2012) demonstrated that

metabolic phenotypes could be tightly linked to the geographic distribution in *Arabidopsis*, which indicates that metabolite profiling is also a powerful tool to explore plant evolutionary and domestication processes. Recently, Yang *et al.* (2018) reported a non-targeted metabolomic analysis of 30 lettuce cultivars and another research group compared the levels of sesquiterpene lactones, phenolic acids and flavonoids in mature and blotting stage of 22 lettuce cultivars (Assefa *et al.*, 2019). However, the limited sample size in these studies did not fully cover all horticultural types of lettuce, especially the evolutionarily and economically important type - stem lettuce, and the wild ancestral species *L. serriola* were not included in other studies either. Furthermore, the lack of genetic data in these studies hindered the identification of causal genes and mechanistic investigation. To better understand the underlying genetic basis of and domestication related effects on natural metabolic variation of *Lactuca* necessitates a detailed systematic investigation using a diverse population coupled with high-throughput genomic information.

Generation of a comprehensive lettuce reference genome was challenging due to its large size and high proportion of repetitive regions. The genome of *L. sativa* has been sequenced and assembled recently, which provides high-quality, comprehensive reference genome for analysis of the Compositae family (Reyes-Chin-Wo *et al.*, 2017). Soon after the release of the lettuce reference genome, Zhang *et al.* (2017) reported the RNA sequencing of 240 lettuce accessions sampled from the major horticultural types and wild relatives. This RNA sequencing generated 1.1 million single-nucleotide polymorphisms (SNPs) and expression data for 22 039 genes across the lettuce genome. In addition, genome-wide association studies (GWAS) identified 5311 expression quantitative trait loci (eQTL) affecting the expression of 4105 genes, including nine eQTL associated with flavonoid biosynthesis. Moreover, GWAS for leaf color detected six candidate loci responsible for the variation of anthocyanin in lettuce leaves (Zhang *et al.*, 2017). This study thus provides a rich resource for lettuce genetic studies and will facilitate the breeding of cultivars for improved traits such as nutritional value.

Here we used a collection of 189 wild and cultivated lettuce accessions to identify metabolites that are involved in domestication and cultivar differentiation. We also unravel the genetic basis and molecular mechanisms underlying the naturally occurring variation in primary metabolism in this large lettuce association panel by integrating analyses of metabolomics, genome-wide association mapping, eQTL and transcriptional network. Our metabolome-based experiment profiles the largest yet panel of lettuce accessions used for this purpose allowing the characterization of the genetic architecture of the primary metabolism of lettuce and providing evolutionary insights into this important vegetable.

RESULTS

Metabolic composition of all horticultural types and wild lettuce

A total of 189 lettuce accessions from the above-mentioned GWAS panel were used to study the metabolism of lettuce. This sub-panel consisted of 130 cultivars, 33 accessions from a RIL population, 6 intermediate accessions (i.e., accessions with characteristics of both wild and cultivated lettuce, likely derived from crosses between wild and cultivated lettuce) and 20 wild accessions, including 16 *L. serriola*, 2 *L. saligna* and 2 *L. virosa* accessions. We identified and quantified 77 metabolites from leaves of 3-month-old lettuce. Of them, 69 metabolites were chemically identified and were classified into six groups: amino acids, organic acids, sugars, polyols, polyamines and purines. Detailed information of these metabolites is provided in Table S1. The heritability (H^2) of all 77 metabolites was greater than 0.5 while 32 metabolites (41.6%) displayed an H^2 of greater than 0.7 (Table 1). In order to explore the natural variations of these metabolites among different types of lettuce, we first performed unsupervised PCA using cultivated lettuce and wild lettuce *L. serriola* (Figure 1a,b). The PCA results based on metabolites is quite different from those based on the genetic data, in which each type of lettuce formed a distinct group (Zhang *et al.*, 2017). Nevertheless, *L. serriola* were slightly separated from other types based on metabolites. We next conducted PLS-DA, a supervised method and distinct from PCA, because this method might improve the separation among different groups. According to the first three components revealed by PLS-DA, *L. serriola* were obviously separated from other horticultural types (Figure 1c). Regarding supervised methods, it is critical to avoid overfitting problems especially for the datasets with a large number of features (Broadhurst and Kell, 2006; Rubingh *et al.*, 2006). Therefore, we performed a permutation test in order to validate our PLS-DA model. The results revealed that neither R^2 nor Q^2 of the permutation tests (100 times) reached the observed data (Figure 1d), verifying the significance of our discrimination model. In summary, both PCA and PLS-DA results suggested that the primary metabolite contents of *L. serriola* were different from cultivars.

We next calculated the intra- and inter-population Euclidean distances (EDs) in order to estimate the metabolic distance within and between populations (Figure 1e). We found that looseleaf type, atypical types and the RIL population exhibited relatively low inter-population EDs, which was probably due to an effect of the crosses for the RIL population or gene flow from other horticultural types in case of the looseleaf lettuce. *L. virosa* and *L. saligna* displayed considerably higher inter-population EDs than cultivated lettuce and *L. serriola*, which is consistent with the phylogenetic relationships among *Lactuca* species that

cultivated lettuce is closer to *L. serriola* than to either *L. virosa* or *L. saligna*. Furthermore, wild *Lactuca* species also exhibited relatively higher intra-population EDs than cultivar lettuce, indicating high level of metabolic variations within wild lettuce.

Screening selection signals of metabolic traits

During the domestication and cultivar differentiation processes, metabolite levels may be affected by adaptation to different environments and/or as a consequence of human preferences. Nested analysis of variance (NANOVA) was used to explore the potential selection signals on metabolites in lettuce. As a result, the levels of 61 metabolites were found to be significantly different ($P < 0.05$) among the six lettuce types (Table 1). We performed a Tukey's test to identify their major differences of metabolites. *L. serriola* harbored the largest number (11) of metabolites which show differences ($P < 0.05$) when compared to other types, followed by butterhead (5), stem lettuce (3) and crisphead (2). Detailed information on these metabolites is provided in Table 1. For some of the metabolites harboring large variations in the lettuce population, differentiation could also have arisen by genetic drift. We, therefore, further calculated quantitative trait differentiation of metabolites (Q_{st}). On an additive genetic basis, neutral Q_{st} is expected to be equal to the neutral F_{st} (Leinonen *et al.*, 2013). Based on this hypothesis, a trait is considered to be under directional selection if the Q_{st} of the trait is significantly greater than the neutral F_{st} . By contrast, stabilizing selection across the populations is assumed if the Q_{st} of a trait is significantly less than the neutral F_{st} . Using this method, we detected 23 metabolites may be under selection across the lettuce populations at a threshold of $P < 0.05$. Twenty-two of these 23 metabolites exhibited signals of directional selection, with the exception of 6-phospho-gluconate, which showed a Q_{st} - F_{st} significantly ($P < 0.05$) less than 0 suggesting stabilizing selection. Among these 22 metabolites, 9 were significantly ($P < 0.05$; Tukey's test) different between wild (*L. serriola*) and cultivated lettuce. Moreover, 7 of these 9 metabolites displayed higher levels in wild lettuce than in cultivars, while the remaining 2 (fucose and myo-inositol) displayed reduced levels in wild lettuce. Furthermore, glutamate and threonate not only showed the highest levels in *L. serriola* but also displayed a decrease in crisphead in comparison with other horticultural types. On the other hand, 3 and 4 metabolites exhibited obviously selection patterns in stem and romaine lettuce, respectively (Table 1). Some of those metabolites harboring significant Q_{st} ($P < 0.05$) belong to the same or related pathways. For instance, both quinate and chlorogenic acid displayed higher levels in *L. serriola* than in cultivated lettuce, while galactinol and raffinose were highly accumulated in stem lettuce. Three metabolites (malate, fumarate and succinate) which are involved in the TCA cycle showed

Table 1 List of mean levels of metabolite contents of different types of lettuce, with Tukey's test, NANOVA, *Qst-Fst* and heritability

No.	Metabolites	Relative mean contents of metabolites ^a						<i>Qst-Fst</i>	NANOVA <i>P</i> -value	<i>Qst</i> <i>P</i> -value	Heritability (<i>H</i> ²)
		Butterhead	Crisphead	Looseleaf	Romaine	Stem	<i>L. serriola</i>				
1	Adenine	1.14,a	0.61,b	1.02,ab	1.10,a	0.98,ab	1.16,a	0.16	**	0.238	0.63
2	β-alanine	1.08,a	0.51,c	1.03,ab	0.90,ab	0.77,bc	0.99,ab	0.35	***	0.030*	0.66
3	Arginine	0.41,b	0.85,b	0.99,b	0.76,b	0.72,b	3.55,a	0.21	**	0.169	0.55
4	Asparagine	0.83,ab	1.48,a	1.24,ab	0.97,ab	0.36,b	1.36,a	0.13	*	0.294	0.58
5	Aspartate	0.92,a	1.03,a	1.01,a	1.17,a	0.57,b	1.11,a	0.32	***	0.054	0.67
6	Alanine	0.78,bc	0.54,c	0.99,ac	1.23,ab	1.31,a	1.11,ac	0.21	***	0.173	0.66
7	Citrate	1.47,a	0.28,b	1.17,ab	0.75,ab	0.55,b	1.51,a	0.25	***	0.123	0.64
8	Cysteine-s-methyl	1.34,a	2.15,a	1.59,a	1.55,a	1.08,a	1.18,a	-0.07		0.371	0.61
9	Cysteine	1.82,a	0.71,b	1.02,ab	0.53,b	0.62,b	1.02,b	0.27	***	0.093	0.64
10	Dehydroascorbate dimer	1.86,ab	3.14,a	1.57,b	1.05,b	1.31,b	1.78,ab	0.21	**	0.166	0.76
11	Erythritol	1.21,a	1.19,a	0.90,a	1.01,a	0.85,a	0.96,a	-0.09		0.335	0.72
12	Fructose	1.06,a	1.05,ab	0.87,ab	0.92,ab	0.89,ab	0.87,b	0.17	**	0.217	0.66
13	Fructose-6-phosphate	0.84,b	1.22,ab	1.30,ab	1.40,a	1.32,a	1.26,ab	0.15	**	0.254	0.65
14	Fucose	1.26,a	0.99,ab	0.97,ab	1.10,ab	0.89,b	0.44,c	0.40	***	0.017*	0.82
15	Fumarate	1.36,a	1.31,a	1.10,a	1.15,a	0.78,b	0.66,b	0.41	***	0.011*	0.74
16	GABA	1.18,a	0.62,a	1.11,a	1.09,a	0.76,a	0.62,a	-0.05		0.399	0.53
17	Galactinol	0.56,cd	0.16,d	0.71,bc	0.53,cd	2.20,a	1.24,b	0.45	***	0.001**	0.93
18	Galactonate	0.88,ab	0.83,ab	0.92,ab	0.74,b	1.11,a	0.81,ab	-0.03		0.459	0.66
19	Glucose-1-phosphate	0.68,a	1.16,a	0.93,a	0.74,a	1.21,a	0.94,a	0.03		0.428	0.73
20	Glucose-6-phosphate	1.06,b	1.33,ab	1.33,ab	1.17,ab	1.44,a	1.31,ab	0.05	*	0.420	0.71
21	Glutamate	0.87,b	0.57,c	0.96,b	1.03,b	1.03,b	1.41,a	0.41	***	0.011*	0.73
22	Glutamine	0.86,bc	2.19,a	1.46,ab	0.84,bc	0.18,c	0.84,bc	0.28	***	0.091	0.65
23	Glutarate	1.91,b	6.44,a	2.15,ab	2.19,ab	4.33,ab	2.18,ab	0.07	*	0.360	0.59
24	Glycerate	1.40,a	1.16,ab	0.80,bc	0.93,bc	0.83,bc	0.63,c	0.33	***	0.058	0.76
25	Glycerol	0.94,a	0.58,a	0.92,a	0.92,a	0.79,a	0.69,a	-0.01		0.494	0.63
26	Glycerol-2-phosphate	0.94,b	1.44,a	1.16,ab	0.97,b	0.99,b	0.85,b	0.19	**	0.190	0.65
27	Glycerol-3-phosphate	1.14,a	0.67,c	1.10,ab	0.95,ac	0.98,ab	0.83,bc	0.23	***	0.160	0.66
28	Glycine	0.70,c	0.87,bc	1.10,bc	1.10,b	0.70,bc	1.61,a	0.36	***	0.030*	0.74
29	Glycolate	0.94,a	1.23,a	1.02,a	1.25,a	0.96,a	1.00,a	0.01		0.480	0.56
30	Histidine	1.26,b	0.93,b	3.30,ab	1.87,b	0.70,b	7.79,a	0.21	**	0.180	0.71
31	Homoserine	0.92,bc	1.16,ab	0.95,bc	1.04,ab	0.51,c	1.48,a	0.32	***	0.068	0.76
32	Inositol-1-phosphate	0.73,a	0.55,a	0.81,a	0.79,a	0.76,a	0.79,a	-0.01		0.473	0.6
33	Isocitrate	1.23,b	1.79,b	3.17,b	1.38,b	2.68,b	6.43,a	0.34	***	0.049*	0.6
34	Isoleucine	0.87,ab	1.47,a	0.98,ab	0.99,ab	0.52,b	0.96,ab	0.12	*	0.299	0.55
35	Leucine	0.97,a	0.90,a	0.76,a	1.14,a	0.66,a	0.87,a	-0.28		0.096	0.61
36	Lysine	1.24,a	0.38,b	0.86,ab	0.97,ab	0.54,ab	0.14,b	0.20	**	0.161	0.58
37	Maleate	1.65,a	1.23,b	1.01,bd	1.16,bc	0.86,cd	0.75,d	0.39	***	0.027*	0.81
38	Malate	1.22,a	0.97,b	0.98,b	0.94,bc	0.80,c	0.79,c	0.42	***	0.009**	0.79
39	2-methyle-malate	0.91,bc	0.61,c	0.82,bc	0.87,bc	1.10,ab	1.44,a	0.34	***	0.040*	0.79
40	Maltose	2.61,b	1.24,b	3.52,ab	5.01,a	1.62,b	2.10,b	0.23	***	0.144	0.71
41	Manitol/manose	4.88,a	1.34,b	1.79,b	1.58,b	0.86,b	2.89,ab	0.25	***	0.139	0.74
42	Methionine	1.14,a	0.80,a	0.65,a	0.89,a	0.71,a	0.99,a	-0.21		0.175	0.67
43	Myo-insitol	1.04,a	0.98,ab	0.97,ab	0.92,b	0.92,b	0.70,c	0.41	***	0.012*	0.66
44	Nicotinate	0.87,a	0.67,ab	0.83,ab	0.80,ab	0.75,ab	0.53,b	0.12	*	0.283	0.64
45	Ornithine	0.73,b	0.20,b	1.69,b	1.41,b	0.56,b	4.00,a	0.33	***	0.069	0.64
46	Phenylalanine	1.12,a	1.10,a	0.78,a	0.89,a	0.83,a	0.88,a	0.03		0.447	0.7
47	Pipecolate	1.47,ab	4.24,a	1.99,ab	2.29,ab	0.54,b	1.88,ab	0.11	*	0.291	0.65
48	Proline	0.42,d	0.39,d	0.98,bc	1.08,b	0.52,cd	1.56,a	0.40	***	0.021*	0.7
49	Putrescine	0.97,bc	1.78,a	1.25,ac	1.45,ab	0.58,c	0.87,bc	0.23	***	0.168	0.73
50	Pyroglutamate	0.90,a	1.16,a	0.99,a	0.91,a	0.42,b	0.98,a	0.38	***	0.025*	0.71
51	Pyruvate	1.80,b	9.74,a	3.28,ab	2.49,b	4.42,ab	1.98,b	0.13	*	0.270	0.67
52	Chlorogenic acid (cis)	0.96,bc	0.66,c	1.57,ab	0.49,c	1.10,bc	2.07,a	0.37	***	0.033*	0.81
53	Chlorogenic acid (trans)	0.99,ab	0.68,bc	1.27,ab	0.41,c	0.80,ac	1.37,a	0.25	***	0.124	0.67

(continued)

Table 1. (continued)

No.	Metabolites	Relative mean contents of metabolites ^a							NANOVA P-value	Qst P-value	Heritability (H ²)
		Butterhead	Crisphead	Looseleaf	Romaine	Stem	<i>L. serriola</i>	<i>Qst-Fst</i>			
54	Quinate	1.74,b	0.77,c	1.28,bc	0.94,c	1.32,bc	4.20,a	0.47	***	0.001**	0.84
55	Raffinose	0.90,cd	0.55,e	1.03,bc	0.73,de	1.37,a	1.23,ab	0.42	***	0.018*	0.84
56	Ribitol/or similar	1.39,a	1.19,ab	0.75,ab	0.72,b	0.92,ab	1.02,ab	0.05	*	0.408	0.61
57	Serine	1.21,a	0.84,cd	1.10,ab	0.89,bd	0.71,d	1.06,abc	0.38	***	0.026*	0.77
58	Succinate	1.41,a	1.03,b	0.92,b	0.76,b	0.90,b	0.83,b	0.36	***	0.037*	0.67
59	Sucrose	0.85,bc	0.67,c	1.06,a	0.99,ab	1.07,a	0.99,ab	0.34	***	0.051	0.65
60	Trehalose	0.61,a	0.56,a	0.52,a	0.60,a	0.83,a	0.49,a	0.02		0.492	0.63
61	Threonate	1.10,b	0.60,d	1.13,b	0.86,c	1.13,b	1.91,a	0.47	***	0.003**	0.81
62	Threonine	1.07,b	1.68,a	1.26,ab	0.99,b	0.48,c	0.90,bc	0.37	***	0.024*	0.67
63	Tryptophan	1.02,a	1.06,a	0.85,a	1.07,a	0.40,a	0.83,a	-0.04		0.422	0.67
64	Tyrosine	1.11,a	0.82,ac	0.77,bc	0.98,ab	0.80,bc	0.56,c	0.28	***	0.096	0.61
65	Unknown 8	0.65,c	0.88,bc	0.81,bc	0.67,c	1.38,a	1.22,ab	0.32	***	0.067	0.68
66	Unknown 1	0.72,c	0.83,bc	1.13,ab	1.15,a	0.82,c	0.91,ac	0.26	***	0.129	0.66
67	Unknown 2	0.54,c	0.91,bc	1.54,ab	1.98,a	1.19,ac	0.43,c	0.27	***	0.104	0.7
68	Unknown 3	1.65,a	2.94,a	1.60,a	1.13,a	1.63,a	2.59,a	0.04		0.427	0.52
69	Unknown 5	1.50,a	0.90,bc	1.01,bc	1.15,b	0.72,c	0.80,c	0.37	***	0.033*	0.82
70	Unknown 6	0.46,c	0.78,bc	0.67,bc	1.05,bc	1.15,b	1.98,a	0.35	***	0.038*	0.87
71	Unknown 7	1.17,a	0.79,ab	0.97,a	0.95,a	1.17,a	0.44,b	0.31	***	0.066	0.72
72	Unknown 9	1.61,a	0.91,b	0.82,b	0.97,b	0.96,b	0.42,b	0.30	***	0.069	0.71
73	Valine	0.93,a	0.91,a	1.06,a	0.91,a	0.63,a	1.16,a	-0.02		0.485	0.69
74	3-Hydroxypropanoate	0.70,b	0.69,b	0.87,ab	0.84,ab	0.92,a	0.79,ab	0.13	*	0.282	0.67
75	4-hydroxy-proline	0.72,ab	0.44,ab	0.72,ab	0.83,a	0.32,b	0.56,ab	0.08	*	0.359	0.58
76	6-phospho-gluconate	1.27,a	1.18,a	1.24,a	1.20,a	1.28,a	1.32,a	-0.38		0.027*	0.65
77	Xylose	1.28,ab	1.49,a	0.81,bc	1.04,ab	0.92,ab	0.22,c	0.32	***	0.069	0.78

^aDifferent letters indicate the significant difference (Tukey's test; P -value < 0.05).

* P -value < 0.05; ** P -value < 0.01; *** P -value < 0.001.

an apparent enrichment in butterhead, while exhibiting a decrease in stem lettuce and *L. serriola*. However, the levels of the other 2 metabolites (isocitrate and 2-methyl-malate) related to the TCA cycle were higher in *L. serriola* than in other types. This result suggested that metabolites related to TCA cycle were frequently selected during lettuce domestication. When we set a stricter threshold of $P < 0.01$, only four metabolites reached the significant level. Galactinol was specifically higher in stem lettuce than in other types. Quinate, threonate and malate had remarkable differences between *L. serriola* and cultivar types. Threonate and malate exhibited differences in crisphead and butterhead in comparison with other types, respectively (Figure 2; Table 1).

Network analysis of metabolites

In order to investigate the metabolic variations in a more systematic manner, we first constructed correlation metabolic networks in each of the five cultivated lettuce varieties and *L. serriola* (Figure 3). Nodes of these networks represent the metabolites, while edges indicate significant ($P < 0.05$) correlations between two metabolites. The largest number of significant correlations was observed in romaine (266) followed by butterhead (251), stem lettuce

(123), crisphead (122), *L. serriola* (79) and looseleaf (68). In other words, there are more than four times as many correlations in romaine as in looseleaf lettuce. Among these correlations, positive correlations were dramatically greater than negative correlations in all types of lettuce (Figure 3; Table S2). The correlation network exhibited a great diversity among different types of lettuce (Figure 3). To further investigate the relatedness of these networks, we calculated the dispersion indices to evaluate the difference between each pair of lettuce types. Dispersion indices were used to quantify the difference between two co-expression networks with higher dispersion indices indicating more variations between two networks and it could be affected by a small set of metabolites showing large correlation difference between two types of lettuce or many metabolites showing moderate correlation difference, or both. *L. serriola* showed relatively high dispersion indices in comparison to cultivated lettuce, while the lowest dispersion indices were observed between butterhead and romaine lettuce (Figure 4a). In addition, we performed permutation tests (1000 times) between each pair of lettuce types in order to assess the statistical significance of the dispersion indices. We discovered that the observed dispersion indices were strikingly greater than the

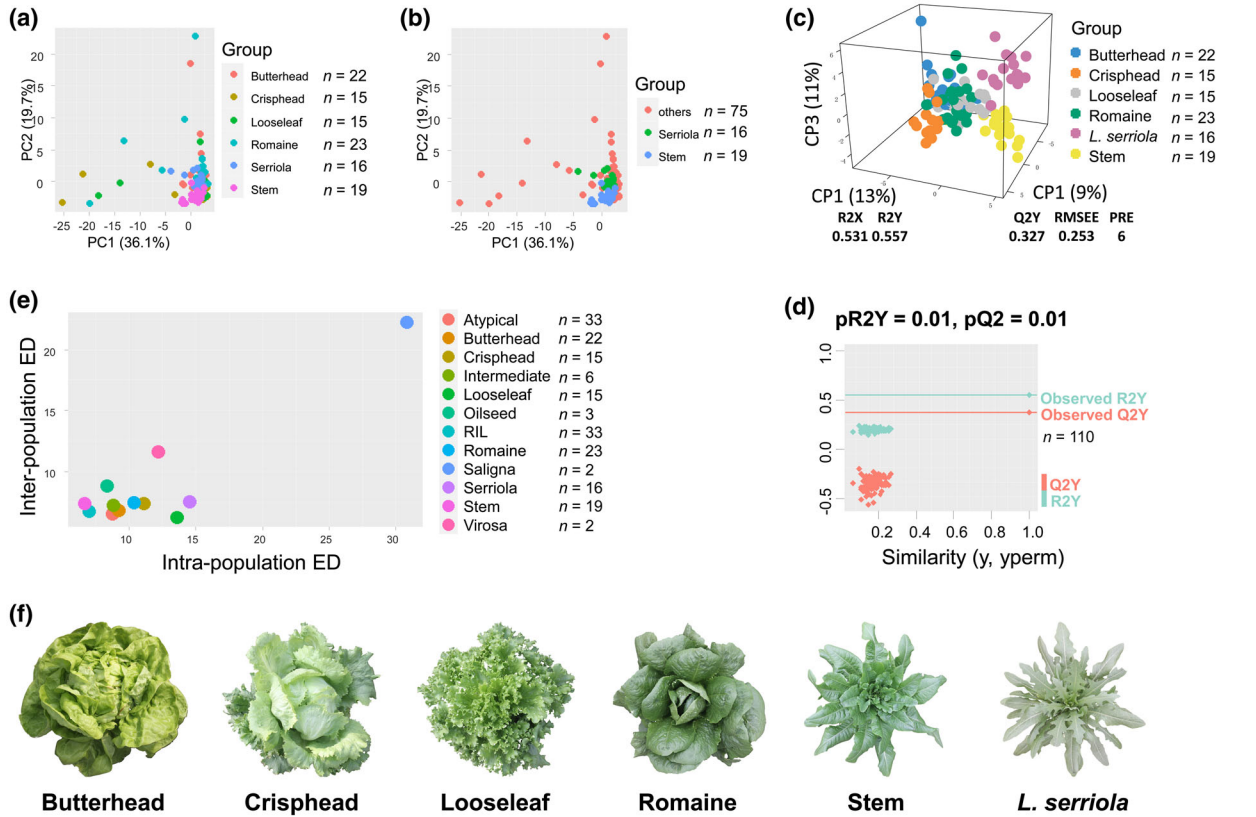


Figure 1. Metabolic profiles of different types of lettuce. (a, b) PCA results of primary metabolic profile of lettuce. Points indicate independent accessions of *Lactuca* and colors represent different types of *Lactuca*. (c) 3D scatter plot of PLS-DA results of primary metabolic profile of lettuce. The points in the plot indicate the accessions of *Lactuca* and different colors display the different types of *Lactuca*. (d) Permutation test of PLS-DA. The Blue and coral points represent the R2 and Q2 of permutation results, respectively. The blue and coral lines indicate the R2 and Q2 of observed data, respectively. (e) Scatter plot of intra and inter-group Euclidean distances which are calculated using primary metabolic profile of lettuce. (f) Picture of all horticultural types and wild lettuce.

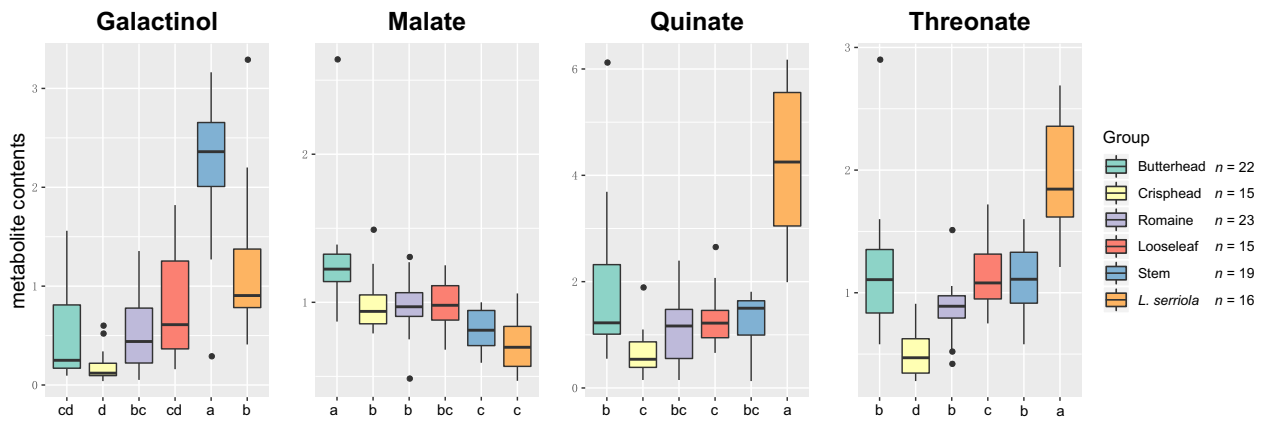


Figure 2. Boxplot of four metabolites harbored significant Qst (P -value < 0.01) across multiple types of lettuce. The letters below indicate the statistical significance (Tukey's test; P -value < 0.05).

permutation results in the comparison between cultivated varieties and *L. serriola*, and similar patterns were observed in romaine, crisphead and looseleaf when

compared with stem lettuce, indicating that the pattern of metabolite network was remarkably distinct between different types of lettuce (Figure 4a).

In order to investigate which pairs of metabolites displayed the greatest variances in correlation patterns among different types of lettuce, we used the R package DiffCoEx (Tesson *et al.*, 2010) to identify differential co-expression modules of metabolites in lettuce. Instead of discovering differential co-expression modules between two types of lettuce, we extended the DiffCoEx (Tesson *et al.*, 2010) method to all types of lettuce, which allowed us to identify the most variable modules across the lettuce population. In total, we detected three differential co-expression modules (Figure S1) and subsequent application of permutation tests to assess statistical significance all three modules showed significantly different co-expression patterns across the populations (P -value of module 1 = 0.002, P -value of module 2 = 0.008 and P -value of module 3 = 0.001). In order to display these results in a more intuitive manner, we used a heatmap to present the correlation changes in each module (Figure 4b–d). Six metabolites were involved in module 1 including adenine, glycerol, glycerol-2-phosphate, leucine, lysine and phenylalanine; module 2 contained five metabolites, citrate, erythritol, malate, 2-methyl-malate and threonine; module 3 harbored six metabolites, Cysteine-s-methyl, glutamine,

histidine, pyroglutamate, valine and 6-phospho-gluconate. In module 1, each type exhibits a unique co-expression pattern (Figure 4b). However, in module 2, butterhead, crisphead, looseleaf and romaine shared similar co-expression patterns. The first four metabolites exhibited strong positive correlations with each other and negatively correlated with threonine. By contrast, the correlation patterns were absent in stem lettuce and *L. serriola* (Figure 4c). In module 3, the first four metabolites were negatively correlated with 6-phospho-gluconate in *L. serriola*, while romaine showed positive correlations between the first four and 6-phospho-gluconate. Moreover, the connectivity of the first five was weaker in the stem lettuce when compared with the other lettuce types (Figure 4d).

Identification of QTL and candidate genes for primary metabolic variation

Due to the strong population differentiation between cultivated and wild lettuce, we conducted genome-wide association studies only using cultivated lettuce varieties to identify metabolic quantitative trait loci (mQTL). We identified 154 significant mQTL for 51 metabolites ($P \leq 6.02 \times 10^{-5}$; Figure S2; Table S3) and 88 of these 154

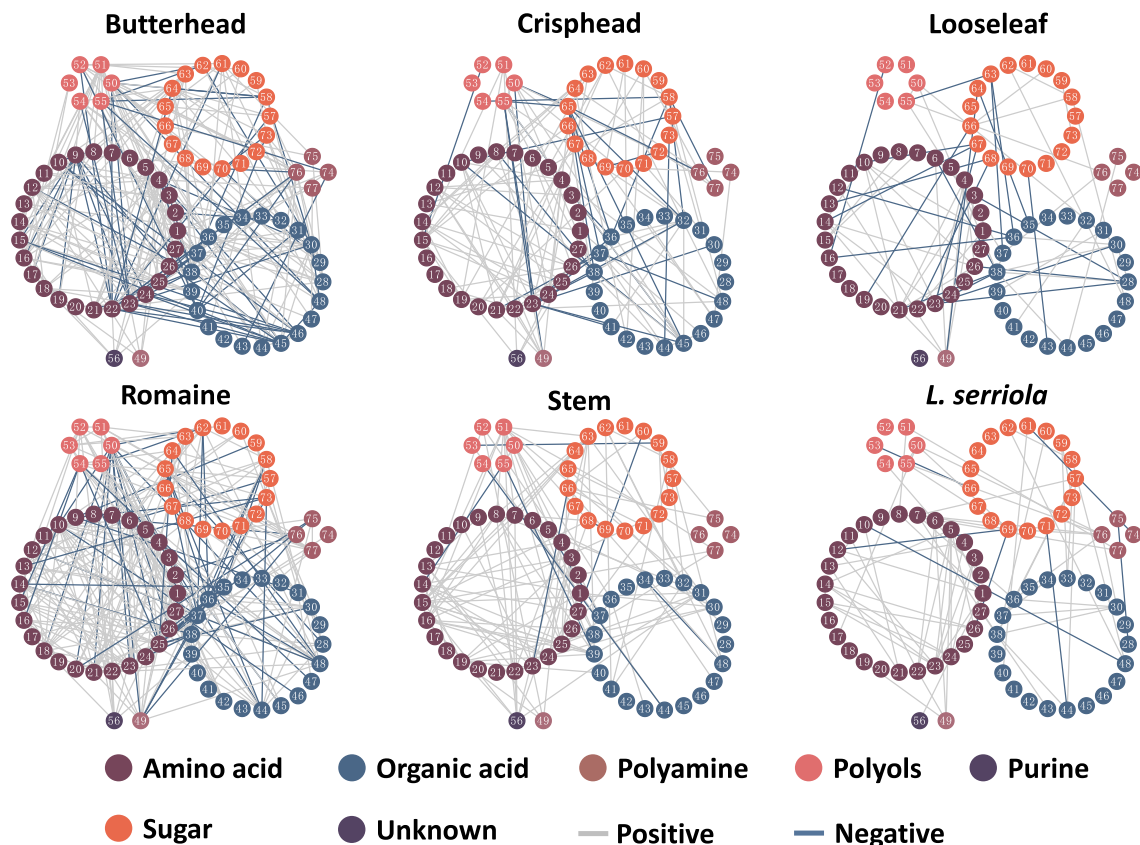


Figure 3. Metabolite correlation networks for butterhead, crisphead, looseleaf, romaine, stem lettuce and *L. serriola*.

Colors in each network display the classification of metabolites. Numbered nodes indicate independent metabolites and the corresponding names can be found in Table 1.

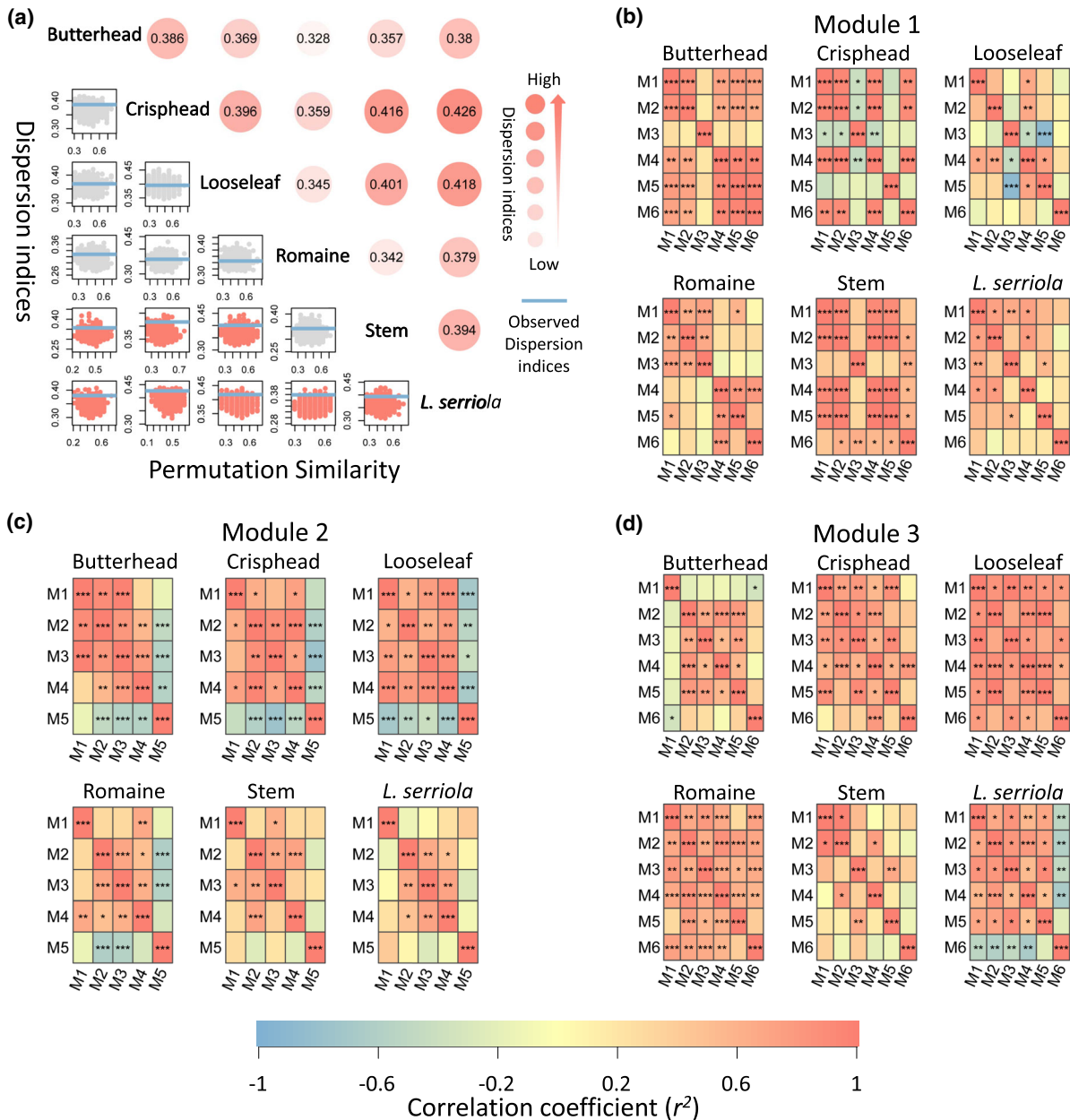


Figure 4. Similarity analysis of metabolite correlation networks and the detection of differential co-expression modules among different types of lettuce. (a) The similarity of networks of different types of lettuce. The upper right corner showed the dispersion indices between each two types of lettuce. The colors, circle sizes and the numbers represent the degree of dispersion indices. Scatter plot in the lower-left corner showed the permutation test of dispersion indices between each two types of lettuce. The blue lines in each plot indicate the observed dispersion indices. X axis displays the dispersion indices of each permutation results and Y axis shows the similarity between the rearranged and real data sets. Coral colors show that the observed dispersion index is greater than most permutation results (>95%; P -value < 0.05). (b–d) Heatmaps of metabolites in differential co-expression modules. M1 to M6 in (b) represent adenine, glycerol, glycerol-2-phosphat, leucine, lysine and phenylalanine, respectively. M1 to M5 in (c) represent citric acid, erythritol, malate, malate-2-methyl and threonine, respectively. M1 to M6 in (d) represent cysteine-s-methyl, glutamine, histidine, pyroglutamate, valine and 6-phospho-gluconate, respectively. Colors from blue to coral indicate the level of PCC from -1 to 1 and the asterisk denotes the significant level of correlations calculated from 1000 times permutation test (* P -value < 0.05; ** P -value < 0.01; *** P -value < 0.001).

mQTL overlapped with selective sweeps identified by our previous RNA-seq study (Zhang *et al.*, 2017). Among these 51 metabolites, 76.5% (34 metabolites) had at least 2 associated loci and 23.5% (12) possessed more than 5 mQTL. We determined the candidate genes for each locus by

integrating the gene functional annotations, eQTL and the correlation between metabolite and gene expression levels (Table S3).

Galactinol and raffinose family oligosaccharides (RFOs) are important carbohydrates in higher plants. It has been

reported that galactinol and raffinose are involved in the response to abiotic and biotic stresses in many species (Taji *et al.*, 2002; Zhuo *et al.*, 2013). In this study, galactinol and raffinose harbored high heritability in the lettuce population (0.93 and 0.84, respectively) and also exhibited specific accumulation in stem lettuce (Figure 2; Table 1), indicating they may be involved in the differentiation between stem lettuce and other lettuce types. Furthermore, GWAS results showed galactinol and raffinose had nine and three strikingly associated loci, respectively. Moreover, all three loci for raffinose overlapped with the loci for galactinol (m43 and m46 in Figure S2; Table S3). In order to ensure the reliability of our candidate gene discovery, we focused on the overlapping loci of galactinol and raffinose. One overlapping locus was localized on chromosome 8 from 28 380 334 to 28 514 448 bp, and we detected a strong LD block covered the entire candidate region. This candidate region contained seven genes (Figure 5a). Among them, *LG8749721* encoded a galactinol synthase which is the ortholog of Arabidopsis *GalS2*. In plants, galactinol synthases catalyze the formation of galactinol from UDP-D-galactose and myo-inositol (Taji *et al.*, 2002; Nishizawa *et al.*, 2008) (Figure 5c). We first classified haplotypes in the candidate region, and a total of 5 distinct haplotypes were detected with haplotypes H2 and H4 as the two major ones. Galactinol and raffinose contents of the H2 haplotype were significantly higher than those of the H4 haplotype ($P < 0.001$; *t*-test; Figure 5d). To further validate the function of *LG8749721*, we overexpressed this gene in tobacco, and significant increase of galactinol and raffinose were detected in the transgenic lines ($P < 0.05$; Figure 5b).

Detection of metabolite related genes affected by domestication

We developed a pipeline to discover genes under selection or affected by domestication (see method). In our previous RNA-seq study, a total of 889 candidate selective sweeps ranging from 10 to 160 Kb (with an average of 41 Kb in length) were detected (Zhang *et al.*, 2017). We combined these selective sweeps with gene annotation, gene expression levels and metabolite content and then assessed the correlations between them. Taking into account the well-studied primary metabolism in model plants, we first used the PMN (Plant Metabolic Network) database (Schlapfer *et al.*, 2018) to search for genes in Arabidopsis related to the 69 annotated metabolites in this study. We detected a total of 1170 genes in Arabidopsis, and found their homologs in lettuce using OrthoMCL (Li *et al.*, 2003). A total of 1252 genes were discovered in lettuce, of which 201 were located in the regions with selective sweeps. A gene is considered as a candidate gene if: (i) it is related to a metabolite with significant *Qst* (P -value < 0.05) and (ii) its expression level has a significant *Qst* (P -value < 0.05) and

(iii) its expression level significantly correlates with the metabolite contents (P -value < 0.05 ; 1000 times permutation test). Four candidate genes met the above-mentioned criteria, and notably three of them were related with quinate, which displayed high accumulation in *L. serriola* but low in cultivated lettuce (Table 1).

Dissection of genes involved in the quinate-chlorogenate pathway

Of the four metabolite related genes potentially associated with domestication, *LG8763094* and *LG5482522* were chosen for further analysis due to the high correlation levels between their expression amount and metabolite contents (*LG8763094*: PCC = 0.475; P -value = 2.11×10^{-9} ; *LG5482522*: PCC = 0.398; P -value = 2.67×10^{-6} ; Table S4).

LG8763094 was annotated as a hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase (HQT) like protein, which is known to play a key role in chlorogenic acid (CGA) biosynthesis (Niggeweg *et al.*, 2004). Further phylogenetic analysis revealed that *LG8763094* was similar to an artichoke (*Cynara cardunculus* subsp. *scolymus*) *HQT* gene, which catalyzed the synthesis of the quinate esters of *p*-coumaroyl and caffeoyl from *p*-coumaroyl-CoA and caffeoyl-CoA, respectively, in artichoke (Comino *et al.*, 2009; Sonnante *et al.*, 2010; Moglia *et al.*, 2016) (Figure 6a; Figure S3). Based on the RNA-seq data, a total of seven SNPs were identified in 173 accessions and 5 of them caused amino acid changes in lettuce population (Figure 6b). However, all these seven SNPs were with minor allele frequency (MAF) $< 5\%$ in our association panel. In order to validate whether *LG8763094* is located in the selective sweep we further obtained its sequences from 29 cultivars and 26 *L. serriola* accessions, and constructed a phylogenetic tree. All the cultivars as well as seven wild accessions belong to the same clade, while the majority of wild accessions formed distinct clades (Figure S4a). To test if the functional variants also exist in the promoter region of this gene we next re-sequenced the 2 Kb upstream of *LG8763094* and detected 73 variants in 39 accessions including eight accessions from *L. serriola* and 31 accessions from different types of cultivars. A neighbor-joining tree obtained from the re-sequencing data revealed that all 31 cultivars harbored exactly the same sequence in the re-sequencing data. All *L. serriola* accessions except W25 formed two distinct well-supported clades. We used the 73 variants in 39 accessions to fit a linear regression model to identify variants that were associated with the contents of quinate and chlorogenic acid and the expression of *LG8763094*. Several variants showed association with quinate and chlorogenic acid contents and gene expression levels, including a 12 bp InDel at -381 (from initiation codon) of *LG8763094* and a strong LD block was detected in this region (Figure 6e). In addition, we developed a PCR-based marker based on the 12-bp InDel to genotype a large

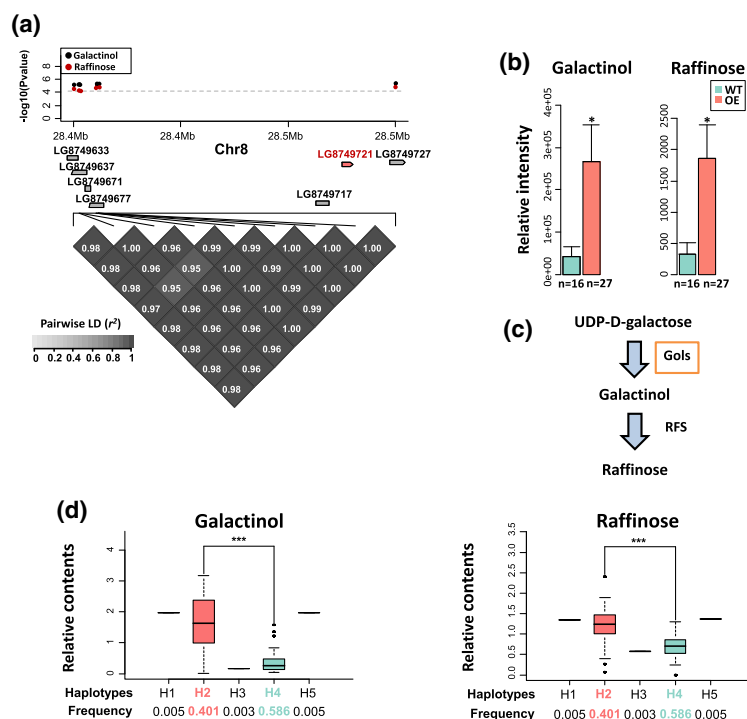


Figure 5. Validation of *LG8449721* (*LsGols2*) as a candidate gene responsible for the contents of galactinol and raffinose.

(a) Manhattan plot (upper) shows an mQTL region for galactinol and raffinose contents on chromosome 8. Gene distributed in this region are displayed, and the candidate gene (*LG8449721*) is shown in red. The lower panel is LD heatmap showing the pairwise r^2 among all polymorphic sites identified by RNA-seq in this region.

(b) Relative intensity of galactinol, raffinose and methionine in wild type (WT) and over-expression (OE) individuals of five lines. Value represent mean \pm SE (* P -value < 0.05; t -test).

(c) Biosynthetic pathway of galactinol and raffinose. UGE, Gols and RFS indicate UDP-glucose epimerase, galactinol synthase and raffinose synthase, respectively.

(d) Boxplot shows the difference of galactinol and raffinose contents among different haplotypes, respectively (H1: AATGCATCT; H2: ATCGCATCT; H3: TATACGCGA; H4: TATATGCGA; H5: TATGCATCT). Asterisks display the significant level of t -test (***) P -value < 0.001).

lettuce population. This 12-bp InDel significantly affected the expression levels of *LG8763094* and the contents of quinate and chlorogenic acid ($n = 85$; P -value = 0.0004975; $n = 73$; P -value = 1.394e-06; $n = 73$; P -value = 4.673e-05, respectively; t -test; Figure 6f).

Another candidate gene *LG5482522* is an ortholog of *Ara*-*bidopsis REF8* (reduced epidermal fluorescence 8) encoding a coumarate 3-hydroxylase (C³H), a P450-dependent monooxygenase. The C³H uses the products of HCT/HQT as substrates to catalyze the synthesis of shikimate and quinate esters of caffeoyl (Franke *et al.*, 2002). Using the RNA-seq data, we detected 17 SNPs in the coding region, and 15 of them were with MAF < 5% and the remaining two did not alter the amino acid sequence (Figure S5a). To investigate whether *LG5482522* was under selection during lettuce domestication we obtained its sequences from 31 cultivars and 26 wild accessions. The *LG5482522* gene from 30 of the 31 cultivars have identical sequences and formed a unique clade with the only exception from accession C21 (*Ls*-*tiva*.8), which grouped with *L. serriola* accessions (Figure S4b). *LG5482522* showed much higher diversity in wild

accessions as compared to cultivars (Figure S4b). The expression level of *LG5482522* was significantly higher in *L. serriola* than in lettuce cultivars (Figure S5c; P -value < 0.05; Tukey's test; Table S4). We sequenced 2 Kb upstream of this gene from 43 accessions and a total of 69 variants were identified (Figure S5a). A Neighbor-joining tree was constructed using all the polymorphic information. All the 27 sequenced cultivars formed a unique and well-supported clade while *L. serriola* formed several distinct clades (Figure S5b). A selection signal was identified upstream of *LG5482522*. Polymorphisms discovered by re-sequencing were used to fit the linear regression model to investigate the functional variants in this region. Four variants marked by red arrows in Figure S5d exhibited the most significant association with quinate, chlorogenic acid contents and *LG5482522* expression levels (Figure S5d).

Taken together, we conclude that the variations in the promoter regions of *LG8763094* and *LG5482522* may result in their differential expression, respectively, and consequently lead to differential accumulation of quinate and chlorogenic acid between wild and cultivated lettuce.

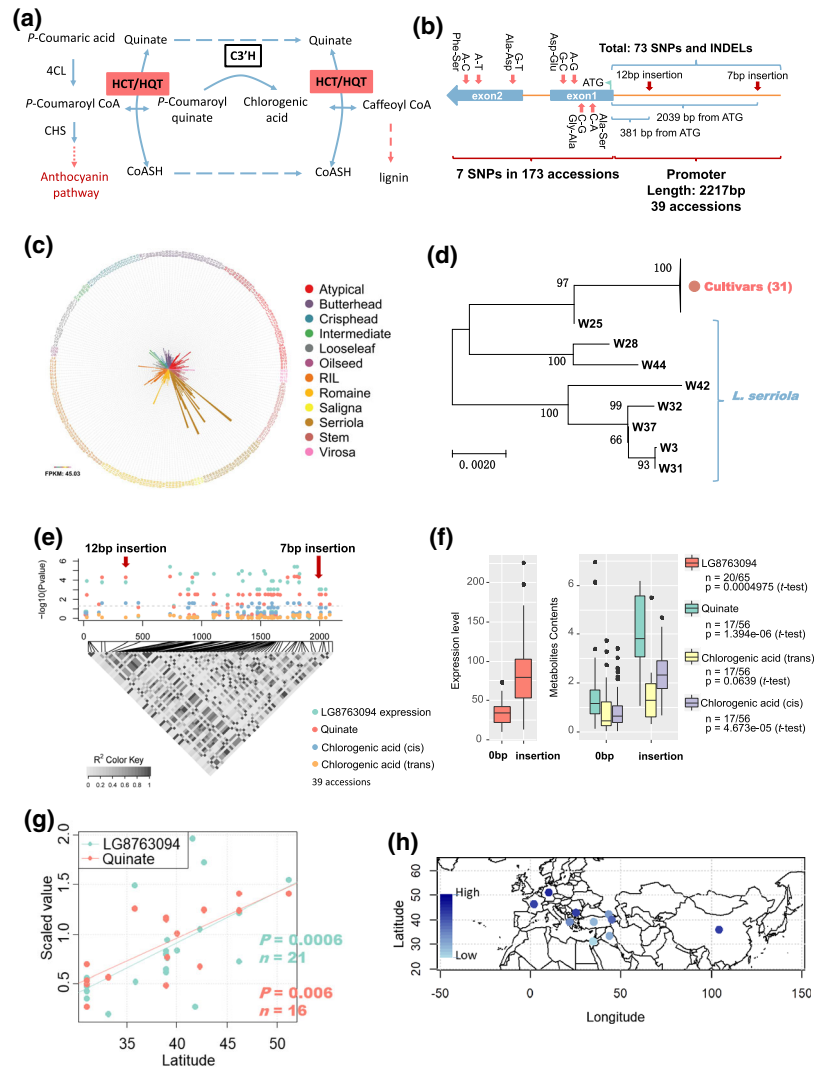


Figure 6. Identification of *LG8763094* (*LsHQT*) as a candidate affected quinate and chlorogenic acid contents.

(a) Quinate and chlorogenic acid pathway. 4CL, 4-coumaroyl-CoA ligase; HCT, hydroxycinnamoyl-CoA shikimate hydroxycinnamoyl transferase; HQT, hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase; C³H, coumarate 3-hydroxylase; Blue solid arrows indicate one-step reaction and coral dashed arrows represent reactions more than one step.

(b) Gene structure and variation of *LG8763094* in *Lactuca*. Two major InDels in the upstream of *LG8763094* are marked by red arrows. The numbers of variants and accessions are indicated below the braces.

(c) Expression profiles of *LG8763094* in 224 lettuce accessions. The radius of circle showed the expression levels of *LG8763094* in each accession and the expression levels are represented by FPKM (fragments per kilobase per million reads). Different colors indicate different types of lettuce.

(d) Neighbor-join phylogenetic tree of upstream sequences of *LG8763094* which achieved by re-sequencing of 39 accessions. The clade of cultivars is compressed and the numbers in bracket indicate the number of accessions. The bar below indicates the number of base substitutions per site.

(e) Manhattan plot for linear regression of quinate, chlorogenic acid and *LG8763094* gene expression levels against the polymorphisms identified in the upstream of *LG8763094*. The points in different colors indicate the metabolites and *LG8763094* expressions levels. The lower part displays the LD heatmap in this region.

(f) Boxplot showing the different levels of quinate, chlorogenic acid and *LG8763094* gene expression levels between the two genotypes at 12 bp Indel.

(g) Linear regression results of *LG8763094* gene expression and quinate levels against latitude. Gene expression levels and metabolites contents were scaled using scale function in R.

(h) Quinate contents and geographic distribution of the origin of *L. serriola* accessions. Each point in the map indicate a *L. serriola* accession and the color from light blue to dark blue represent the low level to high level of quinate contents in each accession.

DISCUSSION

The domestication of plants is arguably one of the most important evolutionary transitions of species (Diamond, 2002; Ross-Ibarra *et al.*, 2007). Previous research showed

that the origin of lettuce domestication is estimated to be around 10 800 years B.P. in the Middle East and the Fertile Crescent consistent with the early domestications of many plants and animals which took place during the late

Pleistocene to early Holocene transition (12 000–8200 B.P.) (Fuller, 2007; Zhang *et al.*, 2017). This thus indicates that lettuce was likely a common food of humans at a very early stage. The availability of a high quality reference genome of lettuce along with the RNA-seq study of 240 lettuce accessions published in 2017 provide us a great opportunity to explore the mystery of the lettuce domestication process and to dissect the genetic bases of domesticated traits (Reyes-Chin-Wo *et al.*, 2017; Zhang *et al.*, 2017). A large number of morphological traits in lettuce changed dramatically during domestication and subsequent cultivar differentiation, including flowering time, leaf shapes, less spine and non-shattering involucre (Devries and Vanraamsdonk, 1994; Hartman *et al.*, 2013). In addition to these traits, domestication may also reshape the compositions of small molecules which contribute to added values of vegetables such as colors, fitness, nutrition value and flavors as indicated by the present study. Through investigating 77 primary metabolites in 189 accessions including all major horticultural types and wild lettuce we here showed how domestication influenced primary metabolism in lettuce. We showed that the metabolites in *L. serriola* were different from those in cultivated lettuce (Figure 1; Figure 4). These results are consistent with demographic inferences and also support the hypothesis that lettuce has undergone a single domestication from *L. serriola* (Zhang *et al.*, 2017). However, we could not distinguish butterhead, crisphead, romaine and looseleaf lettuce varieties based on their primary metabolite contents. It is consistent with a previous conclusion that primary metabolite contents in wheat changed dramatically during the initial domestication process but did not vary much in further improvement phase (Beleggia *et al.*, 2016). We also performed *Qst-Fst* comparisons to identify the metabolites which significantly changed during the domestication process. Compared with other methods, *Qst-Fst* comparisons can distinguish natural selection from genetic drift (Leinonen *et al.*, 2013). We, therefore, used this method and identified 23 metabolites displaying dramatic changes among different cultivated lettuce types and *L. serriola*. However, changes in these metabolites may also be the consequences of other traits associated with domestication, such as the heading leaves in crisphead.

GWAS is a powerful tool to dissect the genetic basis of complex traits including metabolic traits (Strauch *et al.*, 2015; Wen *et al.*, 2015, 2018). Our previous RNA-seq study identified several loci responsible for anthocyanin accumulation in leaves (Zhang *et al.*, 2017). In this study, we identified 154 loci associated with metabolites but most of them contribute minor effects only (mean $R^2 = 5.57\%$), which is similar to previous results on primary metabolites of other species (Rowe *et al.*, 2008; Chan *et al.*, 2010; Wen *et al.*, 2018). Since RNA-seq covers only the expressed regions of a genome, it is challenging to identify the causative

variants and some QTL may be missed. On the other hand, there were usually ten or more genes in the candidate region due to high LD level in self-pollinated species like lettuce. By integrating gene expression information, we could identify the most likely candidate genes through the correlations between metabolite and gene expression levels. For instance, we discovered 15 genes located in mQTL regions whose expression levels were significantly correlated with metabolites levels (adjusted P -value < 0.05; Table S3). Interestingly, among these 15 genes, a WRKY gene's expression (*LG1100395*) was significantly positively correlated with galactinol contents. It is reported that a WRKY transcription factor affected galactinol levels by binding to the promoter of galactinol synthase in *Boea hygrometrica* (Wang *et al.*, 2009b). We also found several W-boxes (YTGACY) in the promoter regions of *LsGols1* and *LsGols2*. Further study can test whether *LG1100395* or other WRKY genes bind to the promoter of *galactinol synthase* in lettuce and influence galactinol concentrations in lettuce.

When a gene was selected during the domestication process, the nucleotide diversities of the flanking regions of this gene will be reduced, a process known as selective sweep. Selective sweeps are considered as a useful indication to determine target genes under natural selection or domestication (Hohenlohe *et al.*, 2010). We found that the level of quinate varied significantly between *L. serriola* and cultivated lettuce in *Qst-Fst* comparisons and Tukey's test (Figure 2; Table 1; P -value < 0.05). However, selective sweeps tend to cover large regions of the genome and a large number of selective sweeps were identified in our previous study (Zhang *et al.*, 2017). This makes the identification of the target genes challenging. Fortunately, as most of the genes in the pathway of primary metabolism were elucidated in other plant species, we can combine *a priori* knowledge with gene expression information to select candidate genes within selected regions. Using such an approach, we identified *LG8763094* (encoding hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase) and *LG5482522* (encoding coumarate 3-hydroxylase) as candidate genes that affect quinate and chlorogenic acid levels in lettuce. Several potential causative variants in these genes were detected using linear regression models (Figure 6; Figure S5). However, none of the variants identified by linear regression showed significant association when we applied mixed linear models. That is probably caused by the fixation of the two genes in cultivated lettuce. It is apparent that human selection resulted in reduced quinate and chlorogenic acid levels in cultivated lettuce. Undesirable flavors such as bitterness and astringency of those compounds and reduction of their levels has been documented to be paralleled on the differentiation or domestication of apple and eggplant (Clifford, 1999; Meyer *et al.*, 2015). The contents of chlorogenic acid of cider varieties are richer than culinary apple, and

chlorogenic acid levels are reduced in cultivar eggplants (Clifford, 1999; Meyer *et al.*, 2015). Interestingly, the flavor-based domestication was not consistent with the notions that breeding tends to increase functional health components in vegetables (Talavera-Bianchi *et al.*, 2010). Quinate and chlorogenic acid can benefit human in various ways given that they possess antioxidant, antiviral and anti-inflammatory activities (dos Santos *et al.*, 2006; Wang *et al.*, 2009a; Hwang *et al.*, 2014). Furthermore, chlorogenic acid also plays an important role in plant growth and development. Inhibition of chlorogenic acid synthase will lead to precocious cell death and alteration of leaf cell morphology in tobacco (Tamagnone *et al.*, 1998). In-depth understanding of the bioactivity of these compounds as well as their underlying genetic basis will aid in generating fortified cultivars through biotech-based breeding or metabolic engineering.

We also noticed that quinate and *LG8763094* expression levels exhibited great variance within *L. serriola* (Figure 2). Further analyses revealed quinate and *LG8763094* expression levels were significantly correlated with the latitude (P -value = 0.006, $n = 16$; P -value = 0.0006, $n = 21$; Figure 6g,h). Similarly, sunscreens of Arabidopsis, rice and naked barley vary with latitude and altitude (Tohge *et al.*, 2016; Peng *et al.*, 2017). Whether quinate in lettuce can influence the ability of plant to adapt local environment or it is only a byproduct of the selection of *LG8763094* require to be studied further. HQT/HCT enzymes in lettuce are encoded by a gene family which contains at least four members. The HQT enzymes are distinct from the HCT enzymes and *LG8763094* is the ortholog of artichoke *CcHQT3* according to phylogenetic analysis (Figure S3). However, none of the other three *HCT/HQT* genes showed differential expression between *L. serriola* and cultivated lettuce, which indicates that *LG8763094* may play a role in lettuce domestication process. Significantly, *CcHQT3* expression levels were notably higher in stem tissues than in leaves and bracts (Moglia *et al.*, 2016). It will be interesting to investigate whether *LG8763094* show different functions in stems.

In conclusion, this study systematically investigates the variation in primary metabolism in a diverse lettuce population including the five major horticultural types and wild relatives. We identified metabolites and their associated genes marking lettuce domestication and differentiation. These results coupled with the dissected genetic basis of these complex metabolic traits provide valuable resources for lettuce quality improvement as well as insights into how these cultivars have evolved and differentiated. Since both quinate and chlorogenic acid are largely beneficial to the plant itself and to human health, it may be important to elevate their contents in lettuce although it will be important to ensure that this is within the limit set by the fact that they should not confer pungent taste.

EXPERIMENTAL PROCEDURES

Plant materials

A total of 189 *Lactuca* accessions from a previously reported association panel were selected in this study. *Lactuca* accessions were sown in December, 2016 and grown in the plastic house on the campus of Huazhong Agricultural University, Wuhan, China. 145 of 189 accessions had two well-growing plants, and we harvested three fully expanded leaves at the same developmental stages from each individual plant of 3-month-old as two biological replicates. The remaining 44 accessions had one plant showing similar growth state as the other 145 ones for metabolite profiling.

Metabolite profiling and outlier filtration

Leaf samples were harvested and immediately frozen in liquid nitrogen and stored at -80°C until further analysis. Metabolite contents were determined according to (Roessner *et al.*, 2001; Liscic *et al.*, 2006). Briefly, the extracted residue was derivatized at 37°C for 120 min (in $40\ \mu\text{l}$ of $20\ \text{mg}\ \text{ml}^{-1}$ methoxyamine hydrochloride in pyridine), followed by a 30-min treatment at 37°C with $70\ \mu\text{l}$ of *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide. The GC-MS system used was a gas chromatograph coupled to a time-of-flight mass spectrometer (Leco Pegasus HT TOF-MS; Leco). The samples were injected with a Gerstel MultiPurpose autosampler system. Helium was used as the carrier gas at a constant flow rate of $2\ \text{ml}\ \text{sec}^{-1}$ and GC was performed on a 30-m DB-35 column. The injection temperature was 230°C and the transfer line and ion source were set to 250°C . The initial temperature of the oven (85°C) increased at a rate of $15^{\circ}\text{C}/\text{min}$ up to a final temperature of 360°C . After a solvent delay of 180 sec, mass spectra were recorded at 20 scans per s with an m/z of 70–600. Chromatograms and mass spectra were evaluated by CHROMA TOF 4.5 (Leco) and TAGFINDER 4.2⁶¹. To ensure the suitability of the method for quantifying lettuce metabolites we additionally performed a recombination experiment where we ran a mixture of lettuce and Arabidopsis alongside samples of each plant independently and assessed their quantitative similarity. Data is presented for each compound measured in Table S5.

In order to obtain the high-quality data, we define outliers based on the interquartile range (IQR) which was calculated as follows:

$$\text{IQR} = Q3 - Q1 \quad (1)$$

where $Q3$ and $Q1$ refer to the 75th and 25th percentile, respectively. Therefore, if the metabolite contents are below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$, it will be deemed as an outlier. According to the definition above, we first removed the outliers for the data without replication. For the data with replication, only those following the criteria below were retained: (i) the ratio of two replications was below 5; (ii) the ratio of two replications was larger than 5 but one of the data was an outlier. For cases like those defined in criteria 2, we removed the outlier and used another value to represent this sample. Finally, we took the mean contents of metabolites for the further analysis.

Statistical analysis of metabolic variation

Missing values of the metabolites were imputed using the R package missForest (Stekhoven and Bühlmann, 2012), following the default parameters. Principal components analysis was conducted with the *prcomp* function in R and the first two components were plotted using the R package *ggplot2* (Ito and Murphy, 2013). PLS-DA and permutation testing was performed using the R package

mixOmics (Rohart *et al.*, 2017) and ropls (Thévenot *et al.*, 2015). *Lactuca* group was calculated using the average Euclidean distance crosswise samples in that group, and the Inter-group distance of a *Lactuca* group was calculated with two steps. First, we calculated the average samples for each group and then the inter-distances were defined as the average Euclidean distances between one lettuce type and the rest (Li *et al.*, 2015).

Heritability for each metabolic trait was calculated following the equation: $H^2 = V_g/(V_g + V_e)$ which used one-way analysis of variance (ANOVA) by setting the accessions as a random effect. Where V_g and V_e are variance of genetic and environmental effects, respectively (Chen *et al.*, 2014).

Analysis of directional selection of metabolites across *Lactuca* population

Nested ANOVA and Tukey's test were conducted using the R package lme4 (Bates *et al.*, 2015) and multcomp (Hothorn *et al.*, 2008) was performed to test the differences of metabolites contents among different types of lettuce. The R code was downloaded and modified from the website <http://www.biostathandbook.com/> (McDonald, 2014).

For detecting the selection signature at metabolites level, we performed $Q_{st} - F_{st}$ analysis. Q_{st} was an analog of F_{st} for the phenotypic traits, defined as follow (Whitlock and Guillaume, 2009):

$$Q_{st} = \frac{\sigma_b^2}{(\sigma_b^2 + 2\sigma_w^2)} \quad (2)$$

where σ_b^2 was between population variation, and σ_w^2 was within population variation.

To test the significance of the observed Q_{st} , we conducted the method of Whitlock and Guillaume (2009). Instead of the comparison of observed Q_{st} with mean F_{st} , this approach generating a neutral Q_{st} distribution which can be derived using the observed F_{st} , and then comparing observed Q_{st} with this distribution to conclude whether the observed Q_{st} was greater or lower than the expected under neutrality. Specifically, we followed these steps:

1 In order to get neutral F_{st} without bias, only neutral SNPs (4DAT, four-fold synonymous transversion) without missing value were used for this analysis (Zhang *et al.*, 2017). Furthermore, we removed 4DAT SNPs in the selection region identified by previously study (Zhang *et al.*, 2017). As a result, 9009 SNPs were remained and F_{st} distribution was calculated according to Weir and Cockerham (Whitlock and Guillaume, 2009).

2 Distribution of σ_w^2 was obtained through multiplying observed σ_w^2 with a random number draw from a χ^2 distribution with a degree of freedom of 5 (number of population - 1), then divided by 5 (Whitlock and Guillaume, 2009).

3 Distribution of σ_b^2 was given by (Whitlock and Guillaume, 2009):

$$\sigma_b^2 = \frac{2F_{st}\sigma_w^2}{1 - F_{st}} \quad (3)$$

where σ_w^2 was calculated following the step 2) as shown above.

4 Expected neutral Q_{st} was calculated using σ_w^2 and σ_b^2 , following the equation (2).

For each metabolite, we repeated above 4 steps for 1000 times to generate the neutral distribution of $Q_{st} - F_{st}$. The resulting P -value was determined by the percentage of the neutral $Q_{st} - F_{st}$ distribution which exhibited more extreme values than the observed $Q_{st} - F_{st}$ value.

Metabolic network construction

First, we calculated the pair-wise Pearson's correlation coefficient (PCC) using the metabolite profile for each type of *Lactuca*. The significance of PCC was determined by 1000 permutations and Hochberg-Benjamini adjustment (Hochberg and Benjamini, 1990). Consequently, the threshold of PCC was set to -0.45 and 0.60 for butterhead, -0.63 and 0.73 for crisphead, -0.67 and 0.66 for looseleaf, -0.51 and 0.56 for romaine, -0.56 and 0.63 for Stem, -0.60 and 0.66 for *Lactuca serriola*, respectively (FDR-adjusted P -value < 0.05). Network was displayed using an in-house R script.

Dispersion indices between every two networks were computed following Choi and Kendzioriski (2009). The First step of permutation test of dispersion indices was randomly permuting the samples between the two types of lettuce. We then calculated the dispersion indices of these two rearrangement groups. By repeating these two procedures for 1000 times, we generated the distribution of dispersion indices and the P -value was determined by comparing the random distribution and observed dispersion indices.

Differential co-expression modules were discovered using R package DiffCoEx (Tesson *et al.*, 2010). In order to extend this method to 5 types of lettuce, we replaced the matrix of adjacency differences to the following equation (Tesson *et al.*, 2010):

$$D_{ij} = \left(\sqrt{\frac{1}{n-1} \sum_k \frac{\text{sign}(c_{ij}^{k|}) * (c_{ij}^{k|})^2 - (c_{ij}^{0|})^2}{2}} \right)^\beta \quad (4)$$

where $c_{ij}^{k|}$ indicates the correlation matrix of each types of lettuce, and (Tesson *et al.*, 2010)

$$c_{ij}^{0|} = \frac{1}{n} \sum_k \left(\text{sign}(c_{ij}^{k|}) * (c_{ij}^{k|})^2 \right) \quad (5)$$

We set the soft threshold β in equation (4) to (5), and we performed Dynamic tree cut by setting the minimum cluster size and cut tree height to 4 and 0.996, respectively. The heatmap in Figure 4 was plotted using an in-house R script, and the significance of correlation was obtained by the 1000 times permutation test.

Genome-wide association studies

EMMAX (Kang *et al.*, 2010) software was used for genome-wide association studies. We performed BN matrix in EMMAX to calculate the population structure and kinship matrix. To determine the threshold of P -value, we first used GEC (Li *et al.*, 2012) software to calculate the effective numbers of independent markers, and then the significant P -value was $1/\text{effective numbers of markers} = 6.02 \times 10^{-5}$. In order to identify the candidate region for mGWAS, we first detected the significant SNPs for each metabolite, and then merged significant SNPs based on their physical distance and pair-wise LD. SNPs with Physical distance ≤ 4.7 Mb or $r^2 \geq 0.1$ were grouped together. Finally, candidate region with at least two significant SNPs were remained for further analysis.

Selection of candidate genes in GWAS

Candidate genes were selected according to the function annotation and the correlation between metabolites and genes in candidate region. For gene annotation, we used nucleotide sequence of candidate genes to BLASTX (Altschul *et al.*, 1990) against Arabidopsis protein database to find genes functional related with the

target metabolites. Correlation between metabolites and genes were calculated using R *cor* function, and the significance of correlation was assigned based on the 1000 permutations, with a false discovery rate of 0.05. LD and haplotype analysis was performed through R package *genetics* (Warnes *et al.*, 2019) and *haplo.stats* (Sinnwell and Schaid, 2018), respectively. Based on the identified haplotypes, boxplot and *t*-test of metabolite contents were conducted using R package.

Vector construction and tobacco transformation

The full-length cDNA of candidate genes was amplified from S40 (a stem lettuce), and then transformed into pri101 vector using Treliel™ SoSoo Cloning Kit after Sanger sequencing to ensure the correct sequence. Confirmed clones were transferred into *Agrobacterium* GV3101 by heat shock and used to infect tobacco leaf disk. After co-culture and selection, confirmed positive T₀ individuals were moved into glasshouse to generate the T₁ generation. T₁ individuals were planted and the transgene positive (over-expression individuals; OE) and negative individuals (wild types; WT) were identified. Leaf samples were harvested from one-month old seedling and metabolic profiling was performed to compare OE and WT individuals. Primers used in these experiments are listed in Table S6.

Identification of mQTL and metabolite related genes affected by domestication and improvement

We first downloaded the metabolite related genes in Arabidopsis from the PMN database, and used orthoMCL (Li *et al.*, 2003) (−1.5 in mcl) to identify the homologs of these genes in lettuce. Genes meeting the following four conditions were selected as candidate genes. Firstly, candidate genes must be related with a metabolite with significant *Qst* – *Fst* (*P*-value < 0.05). Secondly, selected genes had to be located in a selective sweep identified in previous study. Thirdly, expression levels of candidate genes ought to harbor the significant *Qst* – *Fst* (*P*-value < 0.05). *Qst* – *Fst* of gene expression levels were calculated using the same method as for metabolites described in the previous section. Finally, expression levels of these genes should also significantly correlate with the metabolite contents (Pearson correlation, *P*-value < 0.05). For the selected candidate genes, re-sequencing was conducted in the 2 Kb region upstream of the gene and the phylogenetic tree was constructed using MEGA7 (Kumar *et al.*, 2016). Furthermore, a linear regression was done using the polymorphisms found by re-sequencing in order to identify the functional genetic variations. For the detection of functional variations, we developed the specific markers to screen the lettuce population to perform the candidate association analysis. Accessions in another panel used for validating genes under selection and primers are listed in Table S7 and S8, respectively.

ACKNOWLEDGMENTS

This work was supported by grants from the National Key Research and Development Program of China (2018YFD1000800) and the Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation to Dr. Weiwei Wen. ARF and SA were supported by the PlantaSYST project by the European Union's Horizon 2020 Research And Innovation Programme (SGA-CSA nos. 664621 and 739582 under FPA no. 664620).

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

WW, HK and ARF designed and managed this project. WYZ, QHZ, HK and WW performed field experiments and sample preparation, SA and XZ performed metabolite profiling; WYZ, QHZ and WW performed molecular experiments and data analysis. WYZ, SA and WW wrote the manuscript; HK and ARF edited the manuscript.

DATA AVAILABILITY STATEMENT

All other data mentioned in this study are available either in supporting information or from the corresponding author upon request.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Detection of metabolite differential co-expression modules in different types of lettuce. Three differential co-expression modules of metabolites were identified by DiffCoex. Modules show in different colors.

Figure S2. Chromosomal distribution of mQTL identified in this study. X axis shows the physical distance of nine chromosomes in lettuce. Y axis indicates independent metabolites, one on each line, and the classification of metabolites is showed by distinct colors. QTL regions are displayed in the dark blue box and are identified by the criterion described in the method. Detailed information of QTLs is shown in Table S4. Heatmap below indicates the density of mQTL distribution across the chromosome obtained through a sliding window algorithm (window size = 4.6 Mb; window step = 2.3 Mb). m1, alanine; m2, asparagine; m3, aspartate; m4, cysteine; m5, GABA, m6, glutamine, m7 histidine, m8, homoserine; m9, lysine; m10, methionine; m11, ornithine; m12, phenylalanine; m13, proline; m14, pyroglutamate; m15, serine; m16, threonine; m17, tryptophan; m18, tyrosine; m19, valine; m20, 4-hydroxy-proline; m21, citrate; m22, dehydroascorbate dimer; m23, fumarate; m24, galactonate; m25, glutarate; m26, glycolate; m27, isocitrate; m28, maleate; m29, malate-2-methyl; m30, nicotinate; m31, piperolate; m32, pyruvate; m33, chlorogenic acid (cis); m34, quinate; m35, succinate; m36, threonate; m37, erythritol; m38, glycerol; m39, manitol/manose; m40, ribitol/or similar; m41, adenine; m42, fucose; m43, galactinol; m44, inositol-1-phosphate; m45, maltose; m46, raffinose; m47, xylose; m48, unknown 2; m49, unknown 3; m50, unknown 5; m51, unknown 9.

Figure S3. Phylogenetic analysis of HCT/HQT gene family. Protein sequence of HQT and HCT genes from several species were used to construct Neighbor-joining tree. Red circle show the *LG8763094* (*LsHQT*) gene identified in this study. The numbers next to the branch show the bootstrap test (100 replicates) of the percentage of replicate trees. The bar below indicates the number of amino acid substitutions per site.

Figure S4. Neighbor-join phylogenetic tree based on sequences of *LG8763094* and *LG5482522* in cultivar and wild accessions. Red and black IDs indicate *L. sativa* and *L. serriola* accessions, respectively. (a) Phylogenetic tree of *LG8763094*. (b) Phylogenetic tree of *LG5482522*. The number of *L. sativa* and *L. serriola* accessions in the phylogenetic tree are shown in the brackets. The bars below the trees represent base substitution numbers per site.

Figure S5. Validation of *LG5482522* as a candidate gene affected quinate and chlorogenic acid contents. (a) Gene structure and variants identified at *LG5482522* locus. (b) Phylogenetic analysis of the upstream sequences of *LG5482522*. The neighbor-join tree is constructed using the re-sequencing data of 43 accessions. The cultivar clade is compressed and the numbers of cultivars show in the bracket. The bar indicates the number of base substitutions per site. (c) Expression profiles of *LG5482522*. Radius represents FPKM of *LG5482522* and colors indicate the different types of lettuce. (d) Manhattan for QTL and LD block of polymorphisms identified by re-sequencing in the upstream of *LG5482522*. The colored points display $-\log_{10}$ of *P*-value obtained by linear regression of quinate, chlorogenic acid and gene expression levels against variants. The proposed functional variants are labeled and marked by red arrows.

Table S1. Metabolites and their relative contents measured in this study. ^aA, B and D follow the dash line of sample IDs indicate different biological replicates.

Table S2. Summary of positive and negative metabolic correlations for each type of lettuce.

Table S3. Integrated information of candidate genes in each mQTL identified by genome wide association study. Dashed lines indicate no information found and NAs in correlation and correlation *P*-value columns indicate the gene was not expressed.

Table S4. Summary of candidate genes potentially involved in lettuce domestication and/or cultivar differentiation process. ^aThe number 1 in C to G columns indicate the genes are located in the selection sweep regions for corresponding types of lettuce. ^bLetters indicate the statistical significance of Tukey's test. ^c**P*-value <0.05; ***P*-value <0.01. NA indicates the genes are not expressed in RNA-seq data.

Table S5. Summary of GC-MS reporting metabolites data and recombination experiment results for evaluating the method of lettuce metabolites profiling.

Table S6. Primers used for re-sequencing and vector construction of *LG8749721*.

Table S7. Samples for validating selective sweeps.

Table S8. Primers used for re-sequencing the promoter and full length of *LG8763094* and *LG5482522*.

OPEN RESEARCH BADGES



This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at <https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.14950&file=tpj14950.sup-0002-TableS1-S8>. This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at <https://cgngenis.wur.nl/>.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Assefa, A.D., Choi, S., Lee, J.E., Sung, J.S., Hur, O.S., Ro, N.Y., Lee, H.S., Jang, S.W. and Rhee, J.H. (2019) Identification and quantification of selected metabolites in differently pigmented leaves of lettuce (*Lactuca sativa* L.) cultivars harvested at mature and bolting stages. *BMC Chem.* **13**, 56.
- Bates, D., Machler, M., Bolker, B.M. and Walker, S.C. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.
- Beleggia, R., Rau, D., Laido, G. et al. (2016) Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. *Molecular Biology and Evolution*, **33**, 1740–1753.
- Broadhurst, D.I. and Kell, D.B. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, **2**, 171–196.
- Chan, E.K.F., Rowe, H.C., Hansen, B.G. and Kliebenstein, D.J. (2010) The complex genetic architecture of the metabolome. *PLoS Genetics*, **6**, e1001198.
- Chen, W., Gao, Y., Xie, W. et al. (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics*, **46**, 714–721.
- Chinese Nutrition Society (2016) *The Chinese Dietary Guidelines*. Beijing: People's Medical Publishing House.
- Choi, Y. and Kendziorski, C. (2009) Statistical methods for gene set co-expression analysis. *Bioinformatics*, **25**, 2780–2786.
- Clifford, M.N. (1999) Chlorogenic acids and other cinnamates – nature, occurrence and dietary burden. *Journal of the Science of Food and Agriculture*, **79**, 362–372.
- Comino, C., Hehn, A., Moglia, A., Menin, B., Bourgaud, F., Lanteri, S. and Portis, E. (2009) The isolation and mapping of a novel hydroxycinnamoyltransferase in the globe artichoke chlorogenic acid pathway. *BMC Plant Biol.* **9**, 30.
- deVries, I.M. (1997) Origin and domestication of *Lactuca sativa* L. *Genet. Resour. Crop. Ev.* **44**, 165–174.
- Devries, I.M. and Vanraamsdonk, L.W.D. (1994) Numerical morphological analysis of lettuce cultivars and species (*Lactuca* Sect *Lactuca*, *Asteraceae*). *Plant Systematics and Evolution*, **193**, 125–141.
- Diamond, J. (2002) Evolution, consequences and future of plant and animal domestication. *Nature*, **418**, 700–707.
- Fang, C. and Luo, J. (2019) Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *The Plant Journal*, **97**, 91–100.
- Fernie, A.R. and Tohge, T. (2017) The genetics of plant metabolism. *Annual Review of Genetics*, **51**, 287–310.
- Franke, R., Humphreys, J.M., Hemm, M.R., Denault, J.W., Ruegger, M.O., Cushman, J.C. and Chapple, C. (2002) The Arabidopsis REF8 gene encodes the 3-hydroxylase of phenylpropanoid metabolism. *Plant J.* **30**, 33–45.
- Fuller, D.Q. (2007) Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the old world. *Annals of Botany*, **100**, 903–924.
- Funk, V., Susanna, A., Stuessy, T. and Bayer, R.J. (2009) *Systematics, Evolution, and Biogeography of Compositae*. Vienna: International Association for Plant Taxonomy.
- Fusari, C.M., Kooke, R., Lauxmann, M.A. et al. (2017) Genome-wide association mapping reveals that specific and pleiotropic regulatory mechanisms fine-tune central metabolism and growth in Arabidopsis. *The Plant Cell*, **29**, 2349–2373.
- Hartman, Y., Hooftman, D.A.P., Schranz, M.E. and van Tienderen, P.H. (2013) QTL analysis reveals the genetic architecture of domestication traits in Crisphead lettuce. *Genet. Resour. Crop. Ev.* **60**, 1487–1500.
- Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.
- Hohenlohe, P.A., Phillips, P.C. and Cresko, W.A. (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences*, **171**, 1059–1071.
- Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometrical J.* **50**, 346–363.
- Hwang, S.J., Kim, Y.W., Park, Y., Lee, H.J. and Kim, K.W. (2014) Anti-inflammatory effects of chlorogenic acid in lipopolysaccharide-stimulated RAW 264.7 cells. *Inflammation Research*, **63**, 81–90.
- Ito, K. and Murphy, D. (2013) Application of ggplot2 to pharmacometric graphics. *CPT: Pharmacometrics Syst. Pharmacol.* **2**, 7e79.
- Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354.
- Kesseli, R., Ochoa, O. and Michelmore, R. (1991) Variation at Rflp LOCI in *Lactuca* Spp and origin of cultivated Lettuce (*L-Sativa*). *Genome*, **34**, 430–436.

- Kleessen, S., Antonio, C., Sulpice, R., Laitinen, R., Fernie, A.R., Stitt, M. and Nikoloski, Z. (2012) Structured patterns in geographic variability of metabolic phenotypes in *Arabidopsis thaliana*. *Nat. Commun.* **3**, 1319.
- Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**, 1870–1874.
- Lei, L. (2019) Lettuce-manufactured pharmaceuticals. *Nat. Plants*, **5**, 646.
- Leinonen, T., McCairns, R.J., O'Hara, R.B. and Merila, J. (2013) Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics*, **14**, 179–190.
- Li, L., Stoekert, C.J. and Roos, D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189.
- Li, M.X., Yeung, J.M., Cherny, S.S. and Sham, P.C. (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*, **131**, 747–756.
- Li, D., Baldwin, I.T. and Gaquerel, E. (2015) Navigating natural variation in herbivory-induced secondary metabolism in coyote tobacco populations using MS/MS structural analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E4147–E4155.
- Lisee, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A.R. (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols*, **1**, 387–396.
- McDonald, J.H. (2014) *Handbook of Biological Statistics*, 3rd edn. Baltimore: Sparky House Publishing.
- Meyer, R.S., Whitaker, B.D., Little, D.P., Wu, S.B., Kennelly, E.J., Long, C.L. and Litt, A. (2015) Parallel reductions in phenolic constituents resulting from the domestication of eggplant. *Phytochemistry*, **115**, 194–206.
- Moglia, A., Acquadro, A., Eljounaidi, K., Milani, A.M., Cagliero, C., Rubiolo, P., Genre, A., Cankar, K., Beekwilder, J. and Comino, C. (2016) Genome-wide identification of BAHD acyltransferases and in vivo characterization of HQT-like enzymes involved in caffeoylquinic acid synthesis in globe artichoke. *Front. Plant Sci.* **7**, 1424.
- Niggeweg, R., Michael, A.J. and Martin, C. (2004) Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature Biotechnology*, **22**, 746–754.
- Nishizawa, A., Yabuta, Y. and Shigeoka, S. (2008) Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiology*, **147**, 1251–1263.
- Peng, M., Shahzad, R., Gul, A. *et al.* (2017) Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. *Nature Communications*, **8**, 1975.
- Reyes-Chin-Wo, S., Wang, Z.W., Yang, X.H. *et al.* (2017) Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, **8**, 14953.
- Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A.R. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11–29.
- Rohart, F., Gautier, B., Singh, A. and Le Cao, K.A. (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, **13**, e1005752.
- Rojas, C.M., Senthil-Kumar, M., Tzin, V. and Mysore, K.S. (2014) Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. *Front. Plant Sci.* **5**, 17.
- Ross-Ibarra, J., Morrell, P.L. and Gaut, B.S. (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(Suppl 1), 8641–8648.
- Rowe, H.C., Hansen, B.G., Halkier, B.A. and Kliebenstein, D.J. (2008) Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *The Plant Cell*, **20**, 1199–1216.
- Rubingh, C.M., Bijlsma, S., Derks, E.P.P.A., Bobeldijk, I., Verheij, E.R., Kochhar, S. and Smilde, A.K. (2006) Assessing the performance of statistical validation tools for megavariable metabolomics data. *Metabolomics*, **2**, 53–61.
- Sanchez-Perez, R., Pavan, S., Mazzeo, R. *et al.* (2019) Mutation of a bHLH transcription factor allowed almond domestication. *Science*, **364**, 1095–1098.
- dos Santos, M.D., Almeida, M.C., Lopes, N.P. and de Souza, G.E.P. (2006) Evaluation of the anti-inflammatory, analgesic and antipyretic activities of the natural polyphenol chlorogenic acid. *Biological and Pharmaceutical Bulletin*, **29**, 2236–2240.
- Schlapfer, P., Zhang, P., Wang, C. *et al.* (2018) Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants (vol 173, pg 2041, 2017). *Plant Physiology*, **176**, 2583.
- Shang, Y., Ma, Y., Zhou, Y. *et al.* (2014) Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science*, **346**, 1084–1088.
- Sinnwell, J.P. and Schaid, D.J. (2018) haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. R Package.
- Slavin, J.L. and Lloyd, B. (2012) Health benefits of fruits and vegetables. *Adv. Nutr.* **3**, 506–516.
- Sonnante, G., D'Amore, R., Bianco, E., Pierri, C.L., De Palma, M., Luo, J., Tucci, M. and Martin, C. (2010) Novel hydroxycinnamoyl-coenzyme a quinate transferase genes from artichoke are involved in the synthesis of chlorogenic acid. *Plant Physiol.* **153**, 1224–1238.
- Stekhoven, D.J. and Buhlmann, P. (2012) MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.
- Strauch, R.C., Svedin, E., Dilkes, B., Chapple, C. and Li, X. (2015) Discovery of a novel amino acid racemase through exploration of natural variation in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 11726–11731.
- Sulpice, R. and McKeown, P.C. (2015) Moving toward a comprehensive map of central plant metabolism. *Annual Review of Plant Biology*, **66**, 187–210.
- Taji, T., Ohsumi, C., Seki, M., Iuchi, S., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant and Cell Physiology*, **43**, S233.
- Talavera-Bianchi, M., Chambers, E. and Chambers, D.H. (2010) Lexicon to describe flavor of fresh leafy vegetables. *Journal of Sensory Studies*, **25**, 163–183.
- Tamagnone, L., Merida, A., Stacey, N., Plaskitt, K., Parr, A., Chang, C.F., Lynn, D., Dow, J.M., Roberts, K. and Martin, C. (1998) Inhibition of phenolic acid metabolism results in precocious cell death and altered cell morphology in leaves of transgenic tobacco plants. *Plant Cell*, **10**, 1801–1816.
- Tesson, B.M., Breitling, R. and Jansen, R.C. (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinform.* **11**, 497.
- Thévenot, E.A., Roux, A., Ying, X., Ezan, E. and Junot, C. (2015) Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *Journal of Proteome Research*, **14**, 3322–3335.
- Tohge, T., Wendenburg, R., Ishihara, H. *et al.* (2016) Characterization of a recently evolved flavonol-phenylacyltransferase gene provides signatures of natural light selection in Brassicaceae. *Nature Communications*, **7**, 12399.
- U.S. Department of Health and Human Services and U.S. Department of Agriculture (2015) *2015–2020 Dietary Guidelines for Americans*, 8th edn. Washington, DC: U.S. Government Publishing Office.
- Wang, G.F., Shi, L.P., Ren, Y.D. *et al.* (2009a) Anti-hepatitis B virus activity of chlorogenic acid, quinic acid and caffeic acid in vivo and in vitro. *Antiviral Research*, **83**, 186–190.
- Wang, Z., Zhu, Y., Wang, L., Liu, X., Liu, Y., Phillips, J. and Deng, X. (2009b) A WRKY transcription factor participates in dehydration tolerance in *Boea hygrometrica* by binding to the W-box elements of the galactinol synthase (BhGolS1) promoter. *Planta*, **230**, 1155–1166.
- Warnes, G., Gorjanc, G., Leisch, F. and Man, M. (2019) genetics: Population Genetics. R Package.
- Wen, W., Li, K., Alseekh, S. *et al.* (2015) Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. *The Plant Cell*, **27**, 1839–1856.
- Wen, W., Jin, M., Li, K. *et al.* (2018) An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. *The Plant Journal*, **93**, 1116–1128.

- Whitlock, M.C. and Guillaume, F.** (2009) Testing for spatially divergent selection: comparing QST to FST. *Genetics*, **183**, 1055–1063.
- Yang, X., Wei, S., Liu, B., Guo, D., Zheng, B., Feng, L., Liu, Y., Tomas-Barberan, F.A., Luo, L. and Huang, D.** (2018) A novel integrated non-targeted metabolomic analysis reveals significant metabolite variations between different lettuce (*Lactuca sativa* L) varieties. *Hort. Res.* **5**, 33.
- Ye, J., Wang, X., Hu, T. et al.** (2017) An InDel in the promoter of Al-ACTIVATED MALATE TRANSPORTER9 selected during tomato domestication determines fruit malate contents and aluminum tolerance. *The Plant Cell*, **29**, 2249–2268.
- Zhang, L., Su, W.Q., Tao, R. et al.** (2017) RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature Communications*, **8**, 2264.
- Zhu, G., Wang, S., Huang, Z. et al.** (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell*, **172**, 249–261.e12.
- Zhuo, C., Wang, T., Lu, S., Zhao, Y., Li, X. and Guo, Z.** (2013) A cold responsive galactinol synthase gene from *Medicago falcata* (MfGolS1) is induced by myo-inositol and confers multiple tolerances to abiotic stresses. *Physiologia Plantarum*, **149**, 67–78.