

A Genome-Wide Scan of Selective Sweeps and Association Mapping of Fruit Traits Using Microsatellite Markers in Watermelon

UMESH K. REDDY*, LAVANYA ABBURI*, VENKATA LAKSHMI ABBURI, THANGASAMY SAMINATHAN, ROBERT CANTRELL, VENKATA GOPINATH VAJJA, RISHI REDDY, YAN R. TOMASON, AMNON LEVI, TODD C. WEHNER, AND PADMA NIMMAKAYALA*

From the Department of Biology, Gus R. Douglass Institute, West Virginia State University, Institute, WV 25112-1000 (Reddy, Abburi, Saminathan, Cantrell, Vajja, Reddy, Tomason, and Nimmakayala); the U.S. Vegetable Laboratory, USDA, ARS, 2875 Savannah Highway, Charleston, SC 29414 (Levi); and the Department of Horticultural Science, North Carolina State University, Raleigh, NC 27695-7609 (Wehner).

*These authors contributed equally to the work.

Author correspondence to Umesh K. Reddy at the address above, or e-mail: ureddy@wvstateu.edu.

Abstract

Our genetic diversity study uses microsatellites of known map position to estimate genome level population structure and linkage disequilibrium, and to identify genomic regions that have undergone selection during watermelon domestication and improvement. Thirty regions that showed evidence of selective sweep were scanned for the presence of candidate genes using the watermelon genome browser (www.icugi.org). We localized selective sweeps in intergenic regions, close to the promoters, and within the exons and introns of various genes. This study provided an evidence of convergent evolution for the presence of diverse ecotypes with special reference to American and European ecotypes. Our search for location of linked markers in the whole-genome draft sequence revealed that BVWS00358, a GA repeat microsatellite, is the GAGA type transcription factor located in the 5' untranslated regions of a structure and insertion element that expresses a Cys₂His₂ Zinc finger motif, with presumed biological processes related to chitin response and transcriptional regulation. In addition, BVWS01708, an ATT repeat microsatellite, located in the promoter of a DTW domain-containing protein (Cla002761); and 2 other simple sequence repeats that association mapping link to fruit length and rind thickness.

Subject areas: Genomics and gene mapping; Quantitative genetics and Mendelian inheritance

Key words: domestication, heterozygosity, linkage disequilibrium, population structure, recombination rate

Watermelon [*Citrullus lanatus* (Thunb.) Matsum. and Nakai] is a specialty crop with a genome size of 425 Mb ($2n = 2 \times = 22$). The cultivated dessert watermelon flesh contains 87–90% water, a high sugar content (8–13%), and several important health-related compounds, including lycopene, citrulline, arginine, and glutathione (Ren et al. 2012; Nimmakayala et al. 2014a). *Citrullus lanatus* subsp. *mucospermus* represent the “egusi” watermelon group, and have large edible seeds with a fleshy pericarp; *C. lanatus* subsp. *vulgaris* represent the red sweet dessert type (Fursa 1972; Jeffrey 2001). Although genetic diversity within subsp. *vulgaris* is extremely narrow, the cultivars are phenotypically diverse in fruit shape, flesh

texture and color, fruit weight, soluble solids, and rind thickness (Levi 2004).

Breeding watermelon of fairly similar backgrounds has resulted in a narrow bottleneck of genetic diversity (Levi et al. 2001). The molecular diversity of watermelon cultivars is reflected in the low number of single nucleotide polymorphisms (SNPs) between the American heirloom cultivar Charleston Gray and the Chinese elite line 97103, which show an average of 1 SNP every 1430 bp (Guo et al. 2013). Our previous study confirmed the narrow genetic diversity among American watermelon accessions (Nimmakayala et al. 2014a). One explanation for this narrow diversity is the

founder effect created by a small number of accessions being brought to a continent or region as people travel (Dane and Liu 2007; Zoltán 2007).

Excavations in Egypt and Libya have indicated that northern Africa is a primary center of domestication for watermelon (Wasylikowa and van der Veen 2004). Watermelon may have entered Europe around 512 AD, when the Moors invaded the Iberian peninsula, or during the Crusades (Zoltán 2007). In India and China, watermelon was introduced around 800 and 1100 AD, respectively (Paris et al. 2013). Watermelon cultivars were introduced into the Americas after the second voyage of Columbus and during the time of the slave trade and colonization (Romão 2000; Zoltán 2007; Paris et al. 2013).

Genetic analysis of this valuable crop involves several critical steps that include accurate phenotyping for various fruit traits along with high-throughput genotyping and sequencing (Guo et al. 2013; Ren et al. 2012). Simple sequence repeats (SSRs), also known as microsatellites, are repeats of nucleotides consisting of a variable number of repeat units that produce codominant, multiallelic, reproducible bands upon amplification (Stägel et al. 2008; Parida et al. 2009; Cavagnaro et al. 2010; Sun et al. 2013; Nimmakayala et al. 2014b). SSR markers are transferrable across *Citrullus* spp. and can be used in distantly related taxa (Jarret et al. 1997). Microsatellites have been used for the construction of genetic maps in many plant species, and provide dependable landmarks throughout the genome (Córdoba et al. 2010; Cavagnaro et al. 2011; Ren et al. 2012). Hamblin et al. (2007) concluded that SSRs were more effective in clustering germplasm into populations than SNPs, by comparing analyses based on 89 SSRs to analyses based on 847 SNPs, covering 554 SNP haplotypes, in individuals from the same 259 inbred maize lines. An ancestry matrix (Q matrix) calculated using SSRs can be used along with SNP genotypes in an association mapping model to reduce spurious associations (Nimmakayala et al. 2014b).

Currently, the most genome-wide map for watermelon involved 698 SSRs, 219 insertions or deletions, and 36 structural variants, covering 800 cM of the genome with a mean marker interval of 0.8 cM (Ren et al. 2012). The map positions 234 watermelon genome sequence scaffolds, accounting for 93.5% of the assembled 353 Mb genome size (Ren et al. 2012). More recently this map was integrated with another 386 SNP markers that were previously mapped by Sandlin et al. (2012). The additional SNP markers further reduced the mean marker interval to 0.6 cM (Ren et al. 2014). Fifty-eight previously reported quantitative trait loci (QTL) for 12 traits (Sandlin et al. 2012) were integrated into the map, and 10 new QTLs for sugar content were identified (Ren et al. 2014). Notably, some of the traits for which QTLs mapped were Brix, fructose, sucrose, glucose, fruit weight, fruit length, fruit width, fruit shape index, rind thickness, 100 seed weight, seed length, seed width, seed oil percentage, and egusi loci. This map is useful for locating recombination cold spots and the distribution of segregation distortion. Linkage disequilibrium (LD) is nonrandom association of alleles across the chromosomes and therefore features recombination

cold spots. An examination of population structure, genetic diversity at the molecular level, and LD distribution across various chromosomes allows for the identification of QTLs through the model-based association mapping of various traits (Nimmakayala et al. 2014b). A unified, mixed-model approach for association mapping combined with a population structure (Q) analysis is a dependable and robust system for identifying reliable QTLs (Yu and Buckler 2006; Zhao et al. 2007).

Our current genetic diversity study uses microsatellites of known map position to look for evidence of selection and to identify genomic regions that have undergone selection during watermelon's domestication and improvement. We also estimate genome level population structure and LD across the linkage groups, and perform association mapping for fruit traits in the watermelon.

Materials and Methods

Our study used 96 watermelon accessions collected across a wide geographic sample of the world to maximize molecular diversity (Supplementary Table S1 online). The plants from individual accessions were grown in 3 randomized plot replicates during 2 seasons (2012 and 2013) with a row-to-plant spacing of 8 × 3 ft. Standard management practices were followed throughout the growing season. Fruit length and diameter, rind thickness, and soluble solids were recorded for 5 fruits in each replicate.

Marker Resources

We used 201 microsatellites, all with known positions on the watermelon genetic and physical map. Details related to the microsatellite map locations and positions obtained from Ren et al. (2012) are presented in Supplementary Tables S2 and S3 online.

DNA Isolation, PCR, and Resolution of SSR Alleles

Genomic DNA isolation was done with a plant DNA isolation kit (Qiagen cat.#69104). The resolution of SSR polymorphisms on a high-throughput DNA fragment analyzer (AdvanCE FS; Advanced Analytical Technologies, Ames, IA) was carried out as per Nimmakayala et al. (2014a).

Statistical Analysis

Data were processed using statistical analysis of variance (ANOVA) models for all fruit traits. Structure 2.2 (Pritchard et al. 2000) was used to cluster individuals into *K* groups by estimating the posterior probability of the data for a given *K*, $\Pr(X|K)$. The number of populations (*K*) was determined using an admixture model with correlated alleles, and *K* = 2–10. Five independent runs of 100 000 Markov Chain Monte Carlo generations, after 100 000 generation burn-in periods, were used to estimate each value of *K*. The optimal *K* value was determined using the ad hoc statistic, ΔK (Evanno et al. 2005). The number of *K*s in each dataset was

evaluated using ΔK values estimated with Structure Harvester (<http://taylor0.biology.ucla.edu/structureHarvester/>), a program implementing the Evanno method for visualizing Structure output (Earl and vonHoldt 2012). Genetic distance estimates were calculated as described by Crossa and Franco (2004). The neighbor-joining algorithm was used to build a dendrogram based on genetic distances with the software MEGA 5.0 (Tamura et al. 2011). Genetic variance was partitioned among the watermelon groups identified by molecular diversity, and population structure was determined by analysis of molecular variance (AMOVA), using the Arlequin 2.0 program (Schneider 2000). Molecular genetic diversity was estimated using genetic diversity and heterozygosity (h). Estimation of F_{ST} (the proportion of genetic variance in a subpopulation relative to the total population) and F_{IS} (the inbreeding coefficient) was based on Wright's F-statistics (Weir and Cockerham 1984) with the program PopGene 1.31 (Yeh 1999). The Ewens–Watterson neutrality test (Watterson 1978) was performed using Manly's algorithm (Manly 1985) as implemented in PopGene 1.31 to investigate neutrality across the linkage groups. The P matrix for first 3 principal components was calculated from 795 SSR alleles with the TASSEL 3.0 package (<http://www.maizegenetics.net>) (Bradbury et al. 2007). LD was estimated as the correlation between all pairs of SSRs with 1000 permutations, also using TASSEL 3.0. The Q matrix was adapted from the K-5 cluster for association mapping to control spurious results arising from population stratification. The generalized linear model (GLM) of TASSEL 3.0 was used for association mapping. The linked marker P values obtained from our association study underwent false discovery rate (FDR) analysis (Benjamini and Hochberg 1995).

Results

Morphological Traits

The 96 accessions in this research included 90 sweet watermelons (var. *vulgaris*) and 6 egusi types (var. *mucospermus*), cultivated for their edible seeds rather than fruit. Egusi types are generally known for their hard rinds, firm flesh, and low total soluble solids, unlike the sweet watermelon cultivars (var. *vulgaris*). ANOVA results with various traits are presented in [Supplementary Table S4 online](#).

Basic Properties of Microsatellite Alleles

Allele numbers for the 201 microsatellites varied across the accessions. There were 135, 94, and 128 usable loci, of which 100, 57, and 107 were polymorphic for accessions belonging to Africa, Asia, and the rest of the world (Europe, and North and South America), respectively. The total number of alleles amplified for the different microsatellites is shown in [Supplementary Table S5](#). BVWS00117, located on chromosome 6, amplified 10 different alleles. A total of 795 alleles were amplified across the entire dataset. Three alleles per microsatellite were amplified, on average. Shannon indices (SI) are a measure of the level of polymorphism

in microsatellites. The SI of ours ranged from 0.06 SI for BVWS00309 to 1.36 SI for BVWS00117 ([Supplementary Table S5 online](#)).

Genetic Diversity

Molecular diversity analysis showed that mean genetic diversities were 0.26 ± 0.13 , 0.24 ± 0.12 , and 0.22 ± 0.12 for the accessions collected in Africa, Asia, and the rest of the world, respectively. The mean genetic diversity among the 90 *vulgaris* types and 6 *mucospermus* types was 0.343 ± 0.12 . AMOVA results for the *vulgaris* and *mucospermus* types, combined and separately, are shown in [Supplementary Tables S6a and S6b](#). Variance was lower between populations, rather than within populations, indicating wide introgression among all groups. Variance among *vulgaris* and *mucospermus* was 9.38, as compared with the groups from Asia, Africa, and the rest of the world, which showed a variance of 3.73. The overall F_{ST} was 0.04 among the global collections studied.

The 5 clusters in the neighbor-joining tree contained mixtures of several African and Asian accessions ([Figure 1](#)). The branches to Sugar Baby (United States) and PI 379223 (a former Serbia and Montenegro accession) were the longest, compared with the rest, indicating occurrence of rapid genetic changes. The first, third, and fifth clusters were predominantly North American accessions with a mixture of some African and Asian types. The second cluster was mostly African types, with some mixture of Asian types. The fourth cluster contained accessions belonging to all types, which were evenly spread across the cluster. Population structure analysis based on model-based assumptions was used to estimate K-2 to K-10 clusters each with 3 iterations. The results were analyzed for mean $\pm$ SD $\ln P(K)$ and ΔK values as estimated with the Structure Harvester ([Supplementary Figure S2](#)). K-5 was the most appropriate cluster size for this population, with $\Delta K = 17.3$. A population structure ancestry chart using K-5 ([Figure 2](#)) shows accessions in pink and red that can be traced to Africa. Ancestry colored in green primarily occurred in Asia, whereas blue and yellow designate accessions with ancestors in the rest of the world ([Figure 2](#)). Sugar Baby, the US variety, grouped completely with yellow-colored ancestry accessions; and similarly PI 379223, the former Serbia and Montenegro accession, grouped with green-colored ancestry accessions. The distinctiveness of these varieties is corroborated in the neighbor-joining tree by the longer branches mentioned above. PI 508441 (Korea), PI 535947 (Cameroon), Sweet Princess (USA), PI 319212 (Egypt), PI 169261 (Turkey), and PI 357731 (another former Serbia and Montenegro accession) were grouped with completely red ancestry in our analysis.

Evidence of Selection Across the Genome

F_{IS} , F_{ST} , and H_e (expected heterozygosity) were plotted across the physical maps of various chromosomes ([Figure 3](#)). Parts of the genome showed a pattern of higher F_{IS} coupled with low H_e ; this trend is considered a signature of selective sweep or purifying selection. Chromosomes 1, 2, 3, 4, 5, 6, 7,

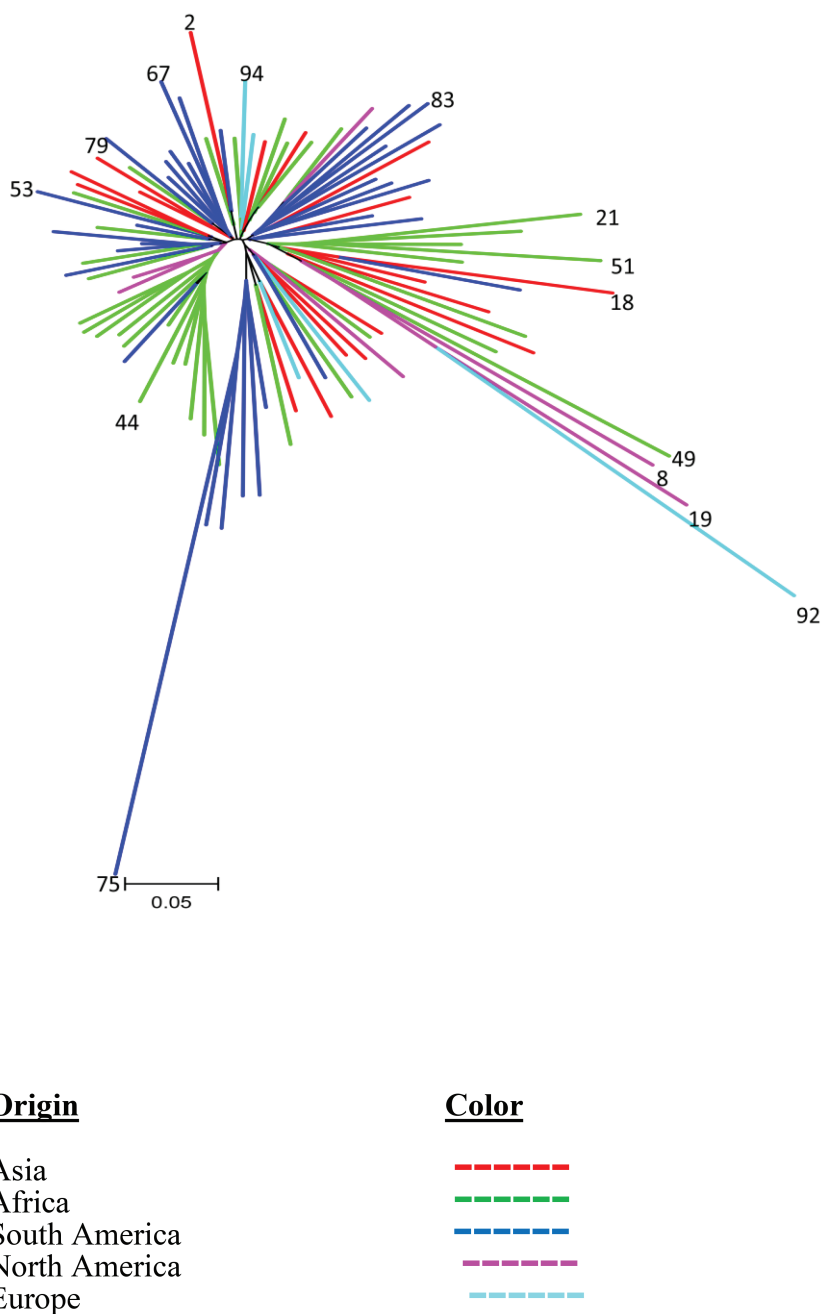


Figure 1. Neighbor-Joining tree depicting relationships estimated across 96 diverse watermelon collections. Numbers at branch termini correspond to the accessions in [Supplementary Table S1](#).

9, 10, and 11 contained 4, 6, 2, 3, 4, 2, 2, 4, 2, and 1 regions of selective sweep, respectively. Chromosome 10 showed a massive domain that had undergone neutral selection (low F_{IS} and high H_e).

The 30 regions that we determined had undergone purifying selection were scanned for the presence of candidate genes using the watermelon genome browser. Selective sweep was noted in the intergenic regions, very close to the 5' UTR or promoter, and in the exons and introns of various genes. Important proteins identified included the ripening-related

protein, basic helix-loop-helix protein, Cyclin D1, and cytochrome p450; playing varied roles in protein, nucleic acid, and ion binding, transcriptional regulation, and catalysis. Physical locations, gene IDs, and annotations of our results are presented in [Table 1](#). Combining F_{ST} test results with results from the Ewens–Watterson test enabled us to detect deviations from a neutral-equilibrium model: either as a deficit of genetic diversity relative to the scaffold size (below the curve of observed F), or as an excess of genetic diversity (above the curve) ([Supplementary Figure S3](#)).

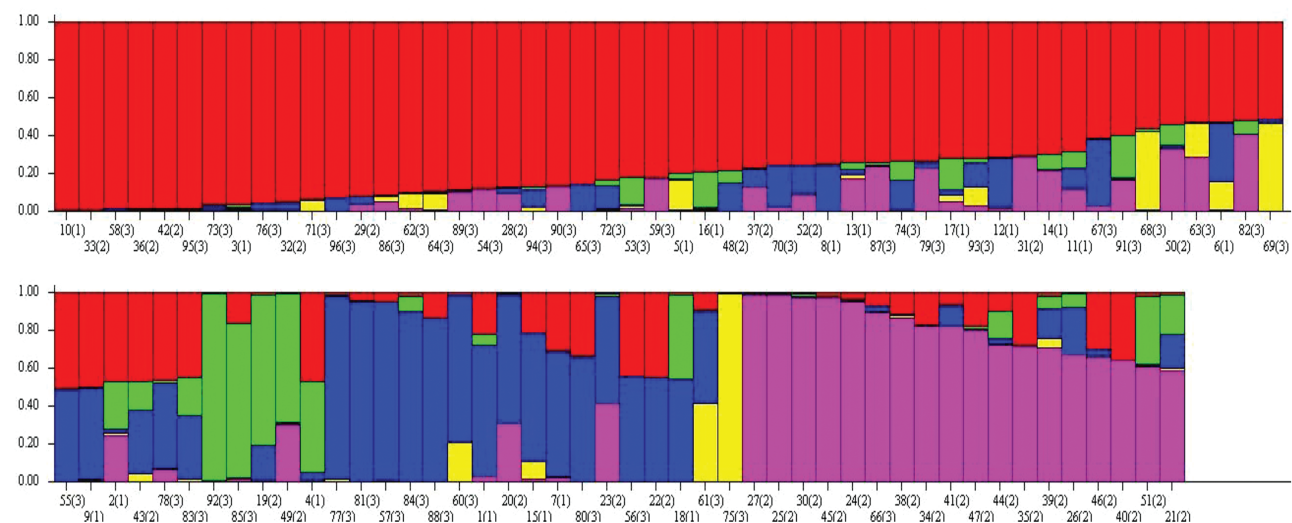


Figure 2. Shared ancestry as revealed by population structure analysis across 96 diverse watermelon collections. Numbers correspond to the accessions in [Supplementary Table S1](#). Numbers in parenthesis refers to (1) Africa; (2) Asia, and (3) Rest of the world (Europe, and North and South America).

Linkage Disequilibrium and Recombination Rate

A total 15 LD blocks of various sizes were identified across the 9 watermelon chromosomes. None of the blocks were located on chromosome 5 or 9. Genetic map distances in cM and the physical length of the various blocks are shown in [Supplementary Table S7](#). Individual correlation values pertaining to marker associations are listed in [Supplementary Table S8](#). Recombination rates estimated for the various LD blocks ranged from 0 to 7.02 (cM/Mb). In contrast, recombination rates at the ends of chromosomes ranged from 1.09 to 14.89, which indicates the presence of relative hot spots of recombination.

Association Mapping

Four linked markers were identified with the GLM approach. BVWS00358 and BVWS01708 were linked with fruit length across the seasons ([Table 2](#)). The BVWS01708 association was significant after our FDR correction during the 2013 season and still significant ($P = 0.01$) during the 2012 season. Additionally, BVWS00711 was associated with fruit width and length, but only during the 2013 season ($P = 0.02$, versus 0.05 for 2012). BVWS00681 was associated with rind thickness during the 2013 season ($P = 0.02$). Our search for the location of the linked markers in the watermelon whole-genome draft sequence showed that the BVWS00358, a GA repeat microsatellite, is a GAGA type transcription factor located in the 5' UTR of a structure and insertion element (Cla004026) that expresses a Cys₂His₂ (C2H2) Zinc finger motif, with presumed biological processes related to chitin response and transcriptional regulation. BVWS01708, an ATT repeat microsatellite, is located in the promoter of the DTW domain-containing protein (Cla002761). Another marker, BVWS00681 is located in the intergenic region between the genes Cla006870 (a nucleotide sugar transporter

protein) and Cla006868 (a pentatricopeptide repeat-containing protein).

Discussion

Our study provides comprehensive insight into the population genetic structure of sweet and unsweet cultivated forms of watermelon. Convergent evolution may have been responsible for the formation of American ecotypes such as Sugar Baby and PI 379223, 2 cultivars that do not appear to share significant ancestry with any of the African or Asian types. The distinctness of American and Asian ecotypes has been previously described in detail by [Ren et al. \(2012\)](#) and [Guo et al. \(2013\)](#). Phenotypic convergence suggests independent evolutionary lineages with parallel phenotypic evolution (the appearance of the same trait in closely related or potentially interbreeding lineages), and that this has occurred many times during crop domestication ([Gross and Olsen 2010](#)). We provide evidence for the presence of ecotype divergence at the molecular level, with special reference to American (Sugar Baby) and European ecotypes (PI 379223).

We investigated diversity in watermelon using the SSRs that have known genetic and physical locations in the genome to identify genomic regions with evidence of selection during watermelon domestication and improvement. Because these SSRs are located at regular intervals covering the whole length of various chromosomes, hitchhiking effects with the loci under selection are minimal, in contrast to what would happen were a similar analysis done with random microsatellites that do not have mapping information.

A limited number of watermelon accessions were introduced from the progenitor pools of Africa to the rest of the world. Therefore, much of the genetic diversity in the progenitor was probably left behind, which resulted in genetic bottlenecks everywhere watermelon was introduced.

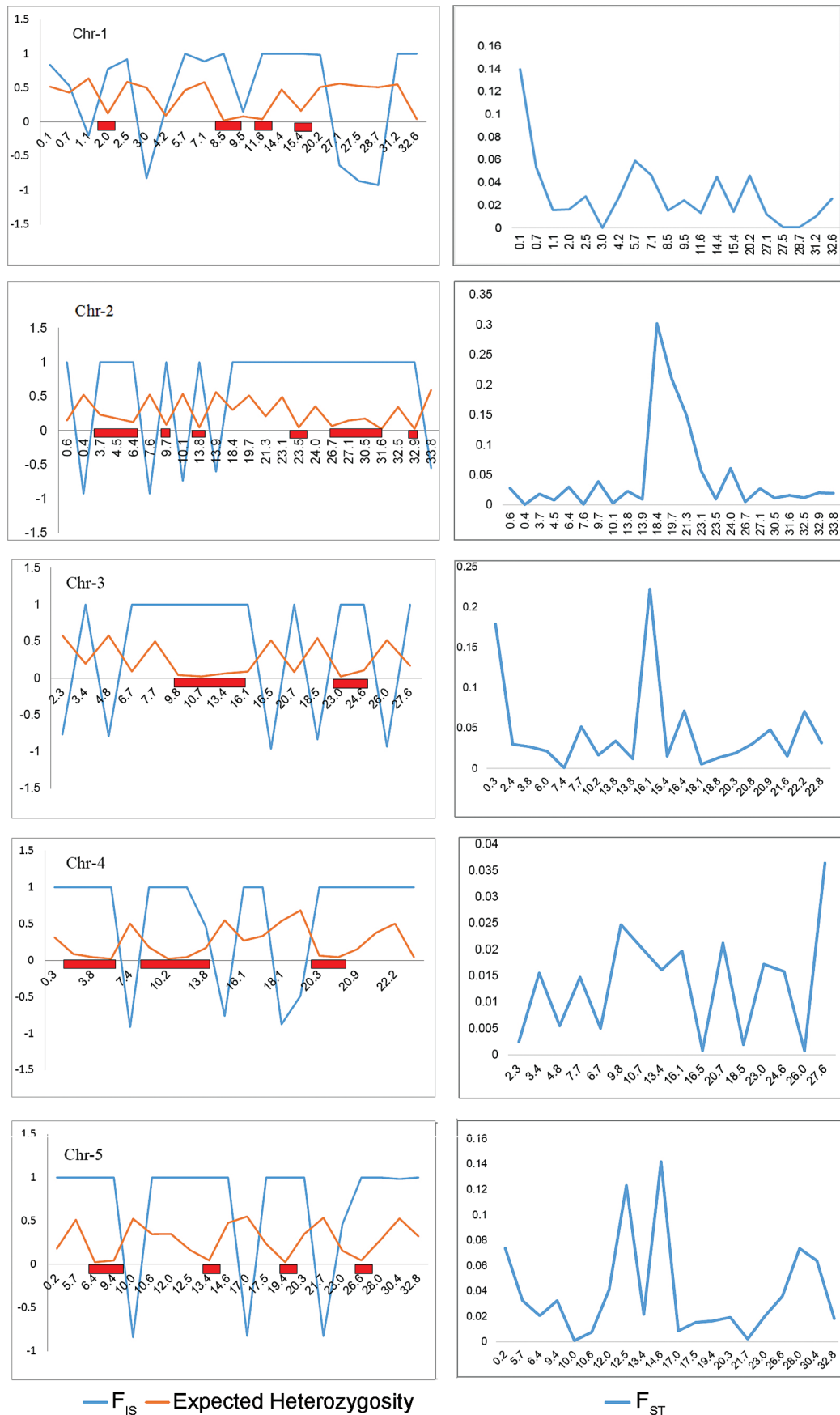


Figure 3. Mapping of F_{IS} against heterozygosity and F_{ST} along the length of the physical map for various chromosomes. Red bars indicate sites of selective sweep.



Figure 3. Continued

Table 1 Details pertaining to the genes identified at the sites of selection sweep across the genome of watermelon

Chromosome no.	Marker	Physical position of gene	Strand	Gene ID	Gene name	Insertion position	GO biological process	GO molecular function	GO cellular component
1	BVWS02029	8 480 391..8 481 329	+	Cla008509	GNS1/SUR4 membrane family protein	Intergenic	—	Long chain fatty acid elongation systems	—
2	BVWS02398	31 561 416..31 565 029	—	Cla008724	Ripening-related protein grip22	Intergenic	—	Lipoprotein A	—
	BVWS02343	32 923 309..32 924 254	+	Cla008579	Phospholipase A22	Exon	Lipid catabolic process	Calcium ion binding; phospholipase A2 activity	—
4	BVWS00405	5 942 099..5 945 562	+	Cla011671	GTP-binding family protein	Intergenic	Protein transport, regulation of growth	Protein binding, nucleotide binding	Cytosol, nucleus, ER
	BVWS01427	10 205 343..10 211 318	+	Cla006847	Nop25 superfamily protein	Intron	—	—	Nucleolus
5	BVWS02422	6 367 196..6 367 822	+	Cla021804	bHLH-like transcription factor	Exon	Transcriptional regulation	Transcriptional regulation	Nucleus
	BVWS01953	19 362 579..19 365 571	—	Cla001042	Shikimate kinase	Intergenic	Amino acid biosynthesis	Kinase activity, magnesium ion binding	Plastid, chloroplast stroma, cytoplasm
6	BVWS02256	6 687 405..6 694 182	—	Cla009404	Haloacid dehalogenase-like hydrolase	Intergenic	Metabolic process	Catalytic activity	—
	BVWS02455	9 730 451..9 732 498	+	Cla010544	Cytochrome P450 CYP714B3	Intergenic	fatty acid metabolic process, oxidation-reduction	electron carrier, heme binding, iron ion binding, monooxygenase, oxygen binding	Membrane, ER, Vacuole
9	BVWS00309	3 409 466..3 412 876	—	Cla015221	Cytochrome P450	Intron	Callose deposition, root development, oxidation/reduction	Electron carrier, iron ion binding, heme binding, protein binding	Membrane, ER, mitochondria
	BVWS02323	32 706 023..32 708 090	—	Cla003249	CYC1D1 (CYCLIN D1); cyclin-dependent protein kinase regulator	Exon	Response to sucrose, brassinosteroid and cytokinin stimulus; cell cycle	Protein binding	Nucleus

Table 2 Level of significance, marker effect and location of associated alleles are revealed by association mapping

Year	Trait	Locus	Chromosome number	Genetic distance (cM)	Physical distance (Mb)	GLM P value	R ²	Allelic sizes (bp)	Allelic effect	No. of genotypes with type of allele
2013	FL	BVWS00358	7	9.7	3.93	0.0303325	0.1018875	177:177	-7.371135503	16
2012		BVWS00358				0.0155809	0.1198053	150:150	-1.428738633	73
								177:177	-13.75719344	16
2012	FL	BVWS01708	4	0.0	0.329	0.0118073	0.155797	150:150	-6.478449225	73
2013		BVWS01708				3.06E-04	0.2166411	278:278	0.498701341	76
								329:329	9.653739726	18
2013	FW	BVWS00711	10	24.9	6.99	0.0290696	0.1067577	329:329	8.809532625	18
								175:175	2.41806641	66
								146:146	0	25
	FL	BVWS00711				0.0545139	0.0857415	175:175	4.274701446	66
2013	RT	BVWS00681	2	40.5	18.39	0.0189544	0.1123728	146:146	0	25
								210:210	0.172952407	20
								236:236	0	72

FL, fruit length; FW, fruit width; RT, rind thickness.

Moreover, with each generation during the breeding process, only a few seeds from the best plants formed the next generation. This winnowing effect can exacerbate a genetic bottleneck, and further reduce genetic diversity throughout the genome (Doebly et al. 2006). The extent of this loss of diversity depends on the initial population size during domestication (Doebly et al. 2006). However, this loss in diversity is not experienced equally by all genes in the genome (Gross and Olsen 2010). For instance, generally genetic bottlenecks will not influence the diversity levels of neutral genes that are not linked to agronomic traits in domesticated organisms. However, genes that influence important traits favored by humans will experience a greater loss of diversity, because plants carrying these alleles contribute the most next-generation progeny, as other alleles are eliminated from the population (Glaszmann et al. 2010). Another important factor is that balancing or diversifying selection can increase optimal phenotypes. A non-shattering phenotype in grain crops is an example of balancing selection for maintaining an adequate balance between harvesting and threshing (Zhang et al. 2009). Alternatively, conscious selection or gene flow from wild or weedy relatives can increase genetic diversity while the cultivated genome is under a selective sweep (Gross and Olsen 2010). Genomic scans for the signature of selection offer a means of identifying new genes of agronomic importance, even when the gene function and the phenotype of interest are yet unknown (Doebly et al. 2006).

Our current study generated genome-wide scans for locating selective sweep and balancing selection, and it identified important genes corresponding to these regions. By resequencing 20 watermelon accessions that belong to *vulgaris*, *mucospermus*, and wild subsp. *lanatus*, Guo et al. (2013) identified 108 regions (7.78 Mb in size) containing 741 candidate genes. The region contains gene categories for the recognition of pollen and for pollen–pistil interaction, a large cluster of 12 tandemly arrayed S-locus protein kinase genes, and genes involved in plant responses to abiotic and biotic stresses (Guo et al. (2013). Furthermore, Guo et al. (2013) identified genes related to several known selected traits including carbohydrate metabolism, fruit flavor (terpene metabolism), and seed oil content (fatty acid metabolism) in the regions of selective sweep.

Guo et al. (2013) studied 11 accessions of sweet watermelon to characterize selective sweep. Selective sweeps characterized using too small of sample size can cause bias because of narrow genetic diversity, limited population history, selection timing, phasing error, and false LD resolution (Tang et al. 2007; Granka et al. 2012; Qanbari et al. 2012). We used genome-wide microsatellites in 90 diverse sweet watermelon accessions in our current study, in contrast to the 11 accessions used by Guo et al. (2013). Furthermore, microsatellite markers are more highly polymorphic than biallelic SNPs. Therefore, microsatellites can resolve balancing selection more efficiently than SNPs, primarily because microsatellites amplify more alleles in diverse populations. We used a sliding window approach to analyze genome-wide population differentiation and fixation indices, along with heterozygosity levels estimated across the chromosomes. We corroborated

several genomic regions under selective sweep previously described by Guo et al. (2013). Vigouroux et al. (2002) identified 15 SSRs that exhibited evidence for selection in maize using a similar research approach as ours, and concluded that non-neutral SSRs are good candidates for agronomically important genes. Our study used microsatellite alleles and entire world collections to characterize thirty genomic regions under selective sweep. Because microsatellite alleles are mutational hot spots, using them for characterizing genomic regions undergoing selection may provide additional information, compared with characterizing genome-wide selection using SNPs.

A major challenge remaining is how to distinguish true signals from those due to genetic drift caused by narrow genetic diversity (Tang et al. 2007; Granka et al. 2012; Qanbari et al. 2012). A possible solution is to estimate selective sweeps by analyzing diverse populations collected from various geographic areas with diverse marker sets, hypothesizing that true signals generated by selection would overlap across the populations (Qanbari et al. 2012). We are currently undertaking this, along with the validation of results obtained from larger populations of sweet watermelons.

It is still unclear whether crop plant recombination rates are positively associated with genetic diversity, despite considerable debate (Bauer et al. 2013). Bauer et al. (2013) characterized the recombination rate across the maize genome as highly variable. Recombination hot spots have been hypothesized to positively associate with genetic diversity, at least as demonstrated in human genetic diversity (Spencer et al. 2006). In this study, we noted that the recombination rate varies among and within chromosomes. In LD blocks, the recombination rate is quite low compared with the rest of the chromosomes. Recombination rates observed in the current research are comparable to those previously estimated (Ren et al. 2012).

Individual marker allele frequencies are very critical in association mapping studies. All the linked markers in our current research have an allele frequency of 0.15 or more. Out of 96 accessions a causative allele needs to be present in at least 12 individuals or more to infer association. Our current research identified important genes containing microsatellite motifs in UTR and promoter regions exhibiting associations with fruit length. BVWS01708 and BVWS00358 both showed linkage to fruit length. BVWS01708 is located near the promoter TATA box of a DTW domain-containing protein. BVWS00358 is a GAGA type transcription factor in the 5' UTR of a gene that expresses a C₂H₂ Zinc finger motif. Interestingly, these microsatellites containing GA and TA type repeats are specific to GAGA and TATA motifs in the watermelon genome. Furthermore, the typical American ecotype is characterized by elongated fruit length (oval shape), compared with the typical African and Asian ecotypes (round shape), so these 2 microsatellites, BVWS00358 and BVWS01708, might have important roles in ecotype differentiation. None of the QTLs mapped in the current study could be validated with the previously mapped QTLs by Sandlin et al. (2012). This may be due to the mapping

population used in Sandlin et al. (2012) research, which were derived from wide crosses.

Analysis of sweet watermelon genome population structure, recombination rate, and selective sweeps will be immensely useful in the design of association mapping studies. The extensive resequencing of several sweet watermelon accessions around the world, along with wild-type counterparts, facilitated by the rapid progress in sequencing and analysis tools currently underway, will further facilitate generating genome scans for selective sweeps, and thereby, for mining novel alleles to improve horticulturally important traits.

Supplementary Material

Supplementary material can be found at <http://www.jhered.oxfordjournals.org/>.

Funding

USDA-NIFA (2013-38821-21453), NSF-EPSCOR#1003907, Gus R. Douglass Institute and NIH grant number P20RR016477.

Acknowledgments

We declare that we have no conflict of interest. The authors thank GRDI graduate assistantships for L. Abburi. The authors are grateful to R. Jarret, Plant Genetic Resources Conservation Unit, USDA-ARS (Griffin, GA) for providing the seeds of germplasm accessions.

References

- Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincet R, Schipprack W, et al. 2013. Intraspecific variation of recombination rate in maize. *Genome Biol.* 14:R103.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Series B (Methodological)*. 57:289–300.
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*. 23:2633–2635.
- Cavagnaro PF, Chung SM, Manin S, Yildiz M, Ali A, Alessandro MS, Iorizzo M, Senalik DA, Simon PW. 2011. Microsatellite isolation and marker development in carrot - genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across Apiaceae. *BMC Genomics*. 12:386.
- Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, Huang S, Weng Y. 2010. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). *BMC Genomics*. 11:569.
- Córdoba JM, Chavarro C, Schlueter JA, Jackson SA, Blair MW. 2010. Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. *BMC Genomics*. 11:436.
- Crossa J, Franco J. 2004. Statistical methods for classifying genotypes. *Euphytica*. 137:19–37.
- Dane F, Liu J. 2007. Diversity and origin of cultivated and citron type watermelon (*Citrullus lanatus*). *Genetic Resour Crop Evol.* 54:1255–1265.
- Doebley JF, Gaut BS, Smith BD. 2006. The molecular genetics of crop domestication. *Cell*. 127:1309–1321.

- Earl D, Vonholdt B. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genetics Resour.* 4:359–361.
- Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* 14:2611–2620.
- Fursa T. 1972. On the taxonomy of the genus *Citrullus* Schrad. *Bot Zh.* 57:31–34.
- Glazmann JC, Kilian B, Upadhyaya HD, Varshney RK. 2010. Accessing genetic diversity for crop improvement. *Curr Opin Plant Biol.* 13:167–173.
- Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. 2012. Limited evidence for classic selective sweeps in African populations. *Genetics.* 192:1049–1064.
- Gross BL, Olsen KM. 2010. Genetic perspectives on crop domestication. *Trends Plant Sci.* 15:529–537.
- Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet.* 45:51–58.
- Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet.* 45:51–58.
- Hamblin MT, Warburton ML, Buckler ES. 2007. Empirical comparison of Simple Sequence Repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One.* 2:e1367.
- Jarret RL, Merrick LC, Holms T, Evans J, Aradhya MK. 1997. Simple sequence repeats in watermelon (*Citrullus lanatus* (Thunb.) Matsum. & Nakai). *Genome.* 40:433–441.
- Jeffrey C. 2001. Cucurbitaceae. In: P. Hanelt, editor. *Mansfeld's encyclopedia of agricultural and horticultural crops*. Vol. 3. Berlin, Heidelberg: Springer. p. 1510–1557.
- Levi A, Thomas C, Keinath A, Wehner T. 2001. Genetic diversity among watermelon (*Citrullus lanatus* and *Citrullus colocynthis*) accessions. *Genetic Resour Crop Evol.* 48:559–566.
- Levi A, Newman M, Reddy UK, Zhang X, Xu Y. 2004. ISSR and AFLP markers differ among American watermelon cultivars with limited genetic diversity. *Am Soc Hort Sci.* 129:553–558.
- Manly BFJ. 1985. *The statistics of natural selection*. London: Chapman and Hall.
- Nimmakayala P, Abburi VL, Abburi L, Alaparthi SB, Cantrell R, Park M, Choi D, Hankins G, Malkaram S, Reddy UK. 2014a. Linkage disequilibrium and population-structure analysis among *Capsicum annuum* L. cultivars for use in association mapping. *Mol Genet Genomics.* 289:513–521.
- Nimmakayala P, Abburi VL, Bhandary A, Abburi L, Vajja VG, Reddy R, Malkaram S, Venkatramana P, Wijeratne A, Tomason YR, et al. 2014b. Use of VeraCode 384-plex assays for watermelon diversity analysis and integrated genetic map of watermelon with single nucleotide polymorphisms and simple sequence repeats. *Mol. Breed.* 34:537–548.
- Parida SK, Dalal V, Singh AK, Singh NK, Mohapatra T. 2009. Genic non-coding microsatellites in the rice genome: characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. *BMC Genomics.* 10:140.
- Paris HS, Daunay MC, Janick J. 2013. Medieval iconography of watermelons in Mediterranean Europe. *Ann Bot.* 112:867–879.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics.* 155:945–959.
- Qanbari S, Strom TM, Haberer G, Weigend S, Gheyas AA, Turner F, Burt DW, Preisinger R, Gianola D, Simianer H. 2012. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. *PLoS One.* 7:e49525.
- Ren Y, McGregor C, Zhang Y, Gong G, Zhang H, Guo S, Sun H, Cai W, Zhang J, Xu Y. 2014. An integrated genetic map based on four mapping populations and quantitative trait loci associated with economically important traits in watermelon (*Citrullus lanatus*). *BMC Plant Biol.* 14:33.
- Ren Y, Zhao H, Kou Q, Jiang J, Guo S, Zhang H, Hou W, Zou X, Sun H, Gong G, et al. 2012. A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS One.* 7:e29453.
- Romão R. 2000. Northeast Brazil: A secondary center of diversity for watermelon (*Citrullus lanatus*). *Genetic Resour Crop Evol.* 47:207–213.
- Sandlin K, Prothro J, Heesacker A, Khalilian N, Okashah R, Xiang W, Bachlava E, Caldwell DG, Taylor CA, Seymour DK, et al. 2012. Comparative mapping in watermelon [*Citrullus lanatus* (Thunb.) Matsum. et Nakai]. *Theor Appl Genet.* 125:1603–1618.
- Sandlin K, Prothro J, Heesacker A, Khalilian N, Okashah R, Xiang W, Bachlava E, Caldwell DG, Taylor CA, Seymour DK, et al. 2012. Comparative mapping in watermelon [*Citrullus lanatus* (Thunb.) Matsum. et Nakai]. *Theor Appl Genet.* 125:1603–1618.
- Schneider, Roessli D, Excoffier L. 2000. *Arlequin: a software for population genetics data analysis* User manual.
- Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet.* 2:e148.
- Stägel A, Portis E, Toppino L, Rotino GL, Lanteri S. 2008. Gene-based microsatellite development for mapping and phylogeny studies in eggplant. *BMC Genomics.* 9:357.
- Sun L, Yang W, Zhang Q, Cheng T, Pan H, Xu Z, Zhang J, Chen C. 2013. Genome-wide characterization and linkage mapping of simple sequence repeats in mei (*Prunus mume* Sieb. et Zucc.). *PLoS One.* 8:e59562.
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.
- Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5:e171.
- Zoltán T, Gyulai G, Szabó Z, Horváth L and Heszy L. 2007. Watermelon (*Citrullus l. lanatus*) production in Hungary from the Middle Ages (13th century). *Hungarian Agric Res.* 4:14–19.
- Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J. 2002. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc Natl Acad Sci USA.* 99:9650–9655.
- Wasylikowa K, Van Der Veen M. 2004. An archaeobotanical contribution to the history of watermelon, *Citrullus lanatus* (Thunb.) Matsum. & Nakai (syn. *C. vulgaris* Schrad.). *Veget Hist Archaeobot.* 3:213–217.
- Watterson GA. 1978. The homozygosity test of neutrality. *Genetics.* 88:405–417.
- Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution.* 38:1358–1370.
- Yeh YR, Boyle T. 1999. POPGENE Version 1.31. Microsoft Window-based Freeware for Population Genetic Analysis Quick User Guide.
- Yu J, Buckler ES. 2006. Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol.* 17:155–160.
- Zhang H, Wang H, Guo S, Ren Y, Gong G, Weng Y, Xu Y. 2012. Identification and validation of a core set of microsatellite markers for genetic diversity analysis in watermelon, *Citrullus lanatus* Thunb. Matsum. Nakai. *Euphytica.* 186:329–342.
- Zhang LB, Zhu Q, Wu ZQ, Ross-Ibarra J, Gaut BS, Ge S, Sang T. 2009. Selection on grain shattering genes and rates of rice domestication. *New Phytol.* 184:708–720.
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, et al. 2007. An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* 3:e4.

Received September 2, 2014; First decision September 10, 2014;
Accepted October 15, 2014

Corresponding Editor: John Stommel