

Use of VeraCode 384-plex assays for watermelon diversity analysis and integrated genetic map of watermelon with single nucleotide polymorphisms and simple sequence repeats

Padma Nimmakayala · Venkata Lakshmi Abburi · Abhishek Bhandary · Lavanya Abburi · Venkata Gopinath Vajja · Rishi Reddy · Sridhar Malkaram · Pegadaraju Venkatramana · Asela Wijeratne · Yan R. Tomason · Amnon Levi · Todd C. Wehner · Umesh K. Reddy

Received: 14 November 2013 / Accepted: 20 February 2014
© Springer Science+Business Media Dordrecht 2014

Abstract Watermelon (*Citrullus lanatus* var. *lanatus*) is one of the most important vegetable crops in the world. Molecular markers have become the tools of choice for resolving watermelon taxonomic relationships and evolution. Increased numbers of single nucleotide polymorphism (SNP) markers together with simple sequence repeat (SSR) markers would be useful for phylogenetic analyses of germplasm accessions and for linkage mapping for marker-assisted breeding with quantitative trait loci and single genes. We aimed to

construct a genetic map based on SNPs (generated by Illumina Veracode multiplex assays for genotyping) and SSR markers and evaluate relationships inferred from SNP genotypes between 130 watermelon accessions collected throughout the world. We incorporated 282 markers (232 SNPs and 50 SSRs) into the linkage map. The genetic map consisted of 11 linkage groups spanning 924.72 cM with an average distance of 3.28 cM between markers. Because all of the SNP-containing sequences were assembled with the whole-genome sequence draft for watermelon, chromosome numbers could be readily assigned for all the linkage groups. We found that 134 SNPs were polymorphic in 130 watermelon accessions chosen for diversity studies. The current 384-plex SNP set is a powerful tool for characterizing genetic relatedness and for developing medium-resolution genetic maps.

Padma Nimmakayala and Umesh K. Reddy have contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-014-0056-9) contains supplementary material, which is available to authorized users.

P. Nimmakayala · V. L. Abburi · A. Bhandary · L. Abburi · V. G. Vajja · R. Reddy · S. Malkaram · Y. R. Tomason · U. K. Reddy (✉)
Department of Biology, Gus R. Douglass Institute,
West Virginia State University, Institute,
WV 25112-1000, USA
e-mail: uredy@wvstateu.edu

P. Venkatramana
BioDiagnostics, Inc, 507 Highland Drive, River Falls,
WI 54022, USA

Present Address:
P. Venkatramana
Douglas Scientific LLC, 3600 Minnesota Street,
Alexandria, MN 56308, USA

A. Wijeratne
Molecular and Cellular Imaging Center, Ohio
Agriculture Research and Development Center, Ohio
State University, Wooster, OH 44691, USA

A. Levi
U.S. Vegetable Laboratory, ARS, USDA, 2875 Savannah
Highway, Charleston, SC 29414, USA

T. C. Wehner
Department of Horticultural Science, North Carolina State
University, Raleigh, NC 27695-7609, USA

Keywords SNPs · High throughput genotyping · SSRs · Genetic mapping · Watermelon · Cultivar diversity

Introduction

Watermelon [*Citrullus lanatus* (Thunb.) Matsum. & Nakai] ($2n = 2x = 22$) belongs to the genus *Citrullus* Schrad. ex Eckl. & Zeyh (Robinson and Decker-Walters 1997) of the Cucurbitaceae family and has a genome size of 425 Mb (Guo et al. 2013). *Citrullus lanatus* includes two botanical varieties, *lanatus* and *citroides* (Bailey) Mansf. (Pitrat et al. 1999; Meeuse 1962). The variety *lanatus* is one of the most important vegetable crops in the world (Paris et al. 2013). The variety *citroides* is cultivated for pickles or medicinal purposes in South Africa and is also called Tsamma or citron melon (Whitaker and Davis 1962; Whitaker and Bemis 1976). Excavations in Egypt and Libya suggested that northern Africa was a primary center of origin and domestication of cultivated watermelon (Dane and Liu 2007). *Citrullus lanatus* subsp. *mucospermus*, representing the “egusi” watermelon group, has large edible seeds and a fleshy pericarp; *C. lanatus* subsp. *vulgaris* represents the red sweet type (Paris et al. 2013). Although the genetic diversity within subsp. *vulgaris* is extremely narrow, the species are phenotypically diverse in fruit shape, flesh pressure, fruit weight, soluble solids and rind thickness (Zhang et al. 2012).

Molecular markers are the tools of choice for determining relationships between cultivars from different breeding programs and from plant introduction accessions (Deschamps and Campbell 2010; Mammadov et al. 2012). Few single nucleotide polymorphism (SNP) markers are available for watermelon genome studies. SNP markers have been useful in genetic mapping and identification of quantitative trait loci (QTL) in other crops (Semagn et al. 2014; Thomson et al. 2012; Yan et al. 2010). Apart from SNPs, other sequence-specific markers are insertion and deletions (InDels), structural variations and simple sequence repeats (SSRs). SNPs are useful because they are easy to identify, inexpensive to use, and higher in frequency than other DNA markers and can be identified with several technologies (Deschamps and Campbell 2010).

The technologies can be used to genotype large sets of DNAs for mapping projects and for marker-assisted breeding (Wong et al. 2012; Li et al. 2012). SNPs are being used for several crops (Sim et al. 2012; Poland et al. 2012; Wong et al. 2012). Genotyping by sequencing (GBS) is an economical approach to finding SNPs and using them for mapping, diversity studies and association mapping (Aslam et al. 2012; Elshire et al. 2011; Ersoz et al. 2012). Once useful markers are identified from genome-based sequencing methods such as GBS, a suitable genotyping technique is required for routine use in multiple populations (Viquez-Zamora et al. 2013; Oliver et al. 2013).

SSRs are stretches of nucleotides consisting of a variable number of short tandem repeats that produce co-dominant, multi-allelic, reproducible bands on amplification (Stagel et al. 2008; Parida et al. 2009; Cavagnaro et al. 2010; Sun et al. 2013). Furthermore, SSR markers are transferrable across various *Citrullus* spp. and can be used with distantly related taxa (Jarret et al. 1997). The markers have been used for the construction of genetic maps of many plant species and provide dependable landmarks throughout the genome (Cordoba et al. 2010; Cavagnaro et al. 2011; Ren et al. 2012).

Three SNP maps constructed for watermelon consist of 378, 357 and 338 SNP markers (Sandlin et al. 2012) spanning 1,438, 1,514 and 1,144 cM, respectively. Probably the best genome-wide map for watermelon involved 698 SSRs, 219 InDels and 36 structural variants that covered 800 cM with a mean marker interval of 0.8 cM (Ren et al. 2012). This map positioned 234 watermelon genome-sequence scaffolds and accounted for 93.5 % of the assembled 353-Mb genome size (Ren et al. 2012). This map elucidated novel genomic features for locating recombination cold spots and the distribution of segregation distortion. However, SNP markers were not included in that mapping study. A genetic map based on SNPs together with SSR markers representing different linkage regions of the watermelon genome has not been constructed. More SNP markers and SSR markers would be useful for applications such as linkage disequilibrium (LD) mapping, phylogenetic analysis of germplasm accessions, and marker-assisted breeding to select and incorporate QTL into elite breeding lines (Berlin et al. 2010; Neves et al. 2011). Here, we aimed to construct a genetic map based on SNPs generated by Illumina

Veracode multiplex assays for genotyping as well as SSR markers. We hoped to evaluate relationships based on SNP genotypes between 130 watermelon cultivars chosen for their diversity of origin.

Materials and methods

Plant materials

We performed genetic mapping and cultivar diversity analysis using SNPs. For genetic mapping, we chose two plant introduction accessions as parents from the US Department of Agriculture germplasm collection (Griffin, GA). PI 244018 (*C. lanatus* var. *citroides*), a yellow-fleshed citron from Zimbabwe (Plant ID-TGR 98), was crossed with PI 270306 (*C. lanatus* var. *lanatus*), a white-fleshed watermelon from Zaire (Plant ID-Mangara). The F1 population was produced in summer 2006 at the West Virginia State University Agricultural Experiment Station. F2 populations of watermelon were generated from a single F1 plant that was self-pollinated. For the diversity study, we chose a set of 130 edible-type watermelon accessions from around the world (see Supplementary Information).

RAD library preparation

Genomic DNA from PI 244018 and PI 270306 was digested with the restriction endonuclease *Pst*I and processed as RAD (restriction site associated DNA) libraries as described by Baird et al. (2008) with modification. Briefly, ~300 ng genomic DNA was digested for 60 min at 37 °C in a 50- μ l reaction with 20 U *Pst*I (New England Biolabs [NEB]). Samples were heat-inactivated for 20 min at 65 °C. *Pst*I P1 adapters each contained a unique multiplex sequence index (barcode), which is read during the first 4 nucleotides of the Illumina sequence read. Then 100 nM P1 adaptor was added to each sample along with 1 μ l of 10 mM rATP (Promega), 1 μ l 10 \times NEB Buffer 4, 1.0 μ l (1,000 U) T4 DNA Ligase (high concentration; Enzymatics, Inc.) and 5 μ l H₂O, and incubated at room temperature for 20 min. Samples were then heat-inactivated for 20 min at 65 °C, pooled and randomly sheared with a Bioruptor (Diagenode) to an average size of 500 bp.

Samples were then run on 1.5 % agarose (Sigma) and 0.5 \times TBE gel, and DNA, 300–800 bp, was isolated

using a MiniElute Gel Extraction Kit (Qiagen). End-blunting enzymes (Enzymatics, Inc.) were used to remove single-strand overhangs of the DNA. Samples were purified on a Mini-elute column (Qiagen), and 15 U Klenow fragment exo- (Enzymatics) was used to add adenosine (Fermentas) overhangs on the 3' end of the DNA at 37 °C. After purification, 1 μ l of 10 μ M P2 adapter, a divergent modified Solexa adapter (2006 Illumina, Inc.), was ligated to the obtained DNA fragments at 18 °C. Samples were purified and eluted in 50 μ l. The elution was quantified by use of a Qubit fluorimeter and 20 ng of the product was used in PCR amplification with 20 μ l Phusion Master Mix (NEB), 5 μ l of 10 μ M modified Solexa Amplification primer mix (2006 Illumina, Inc., all rights reserved) and up to 100 μ l H₂O. Phusion PCR settings followed product guidelines (NEB) for a total of 18 cycles. Again, samples were gel-purified, excising DNA from 300 to 700 bp, and diluted to 1 nM. Two RAD libraries corresponding to PI 244018 and PI 270306 were run on an Illumina Genome Analyser at the University of Oregon Sequencing Facility (Eugene, OR, USA). Illumina/Solexa protocols were followed for paired-end (2 \times 54 bp) sequencing chemistry.

RAD long-read assembly and genotyping

To construct RAD long-read contigs, data from each accession were used to construct a reference assembly for SNP detection. First, sequences with >20 poor Illumina quality scores were discarded (typically < 5 % of all data). Remaining reads were then collapsed into RAD sequence “clusters,” which share 100 % sequence identity at the single-end Illumina read. We imposed a minimum of 25 \times and maximum 50 \times sequence coverage at RAD single-end reads to maximize the efficient assembly of sequences contributed from low-copy, single-dose genome positions. Single loci with coverage <25 \times often display short and fragmented contig assemblies due to insufficient sequence coverage, whereas loci with >500 identical single-end reads are often contributed from high-copy contaminant DNA (plastids) or can be contributed by dosing from multiple genomic loci (e.g., repetitive class sequences). The variable paired-end sequences for each common single-end locus were extracted from these filtered sequences and passed to the Velvet sequence assembler for contig assembly (Zerbino and Birney 2008). SNP-containing sequences were used

for mapping with the whole-genome sequence draft that is now publicly available (www.icugi.org), using Bowtie v1.0. For further analysis, we chose 522 SNPs. To annotate the sequences, we performed a similarity search against the non-redundant protein database with a BLASTx search (NCBI-BLAST v. 2.2.26) and an e-value cutoff of $10E-6$ and maximum of 20 highest scoring pairs per subject. The BLAST hits were then analyzed by using the command-line version of BLAST2GO (BLAST2GO PIPE 2.5) to find the corresponding gene ontology (GO) terms in biological process, molecular function and cellular component classes. The number of sequences mapped to each of the observed GO terms were computed and plotted using the graphical Blast2GO results. Mapping was performed with and without allowing mismatches to locate various chromosome positions. For the chosen SNPs, flanking regions were extracted by using a custom Python script, and a primer designability rank score (0–1) was calculated with Illumina's Veracode Assay Designer software. SNPs with the highest primer designability rank score, totaling 384, were selected for Illumina custom Oligo Pool analysis (OPA), a multiplexing procedure in a fluidics-based system that does not involve fixed arrays (for sequences, see Supplementary File). Parents of the mapping population and 94 F₂ progenies, along with 130 watermelon accessions, were genotyped using Illumina VeraCode technology for GoldenGate assays with the BeadXpress system (Illumina, San Diego, CA, USA) (Lin et al. 2009), according to the manufacturer's protocol. SNP data were analyzed with the Genotyping module (v1.6.3) of Illumina GenomeStudio (v2010.1) with GeneCall threshold 0.25.

To identify problem samples, a call rate was used to produce a scatter plot as a function of sample number, and samples with poor call rate were eliminated. In addition, GenTrain Scores and Gene Call Scores (GC Scores) calculated with the software were checked to refine the SNP calling. SNP clusters were manually adjusted when appropriate and recoded in the "AUX" module of the software according to the following key: 1 = robust; 2 = some manual editing; 3 = heterozygotes dispersed; 4 = heterozygotes similar to homozygous class; 5 = other not reliable; 6 = no amplification; 7 = monomorphic. These data were exported along with the genotypes to provide useful information on, for example, whether marker skewness was real or artificial because of a poorly performing assay.

SSR amplification and mapping

We obtained SSR primers (282 pairs) from the 'Charleston Grey' watermelon sequencing project (manuscript in preparation). In total, 22 bacterial artificial chromosome (BAC)-end sequences were provided by Dr. Hongbin Zhang (Texas A&M University, College Station, TX, USA). PCR was performed in a total volume of 10 μ l containing 10 ng DNA template, 1 \times Taq buffer, 2 mM MgCl₂, 0.2 mM dNTPs, 1 U *Taq* DNA polymerase (Fermentas) and 0.5 μ M each of forward and reverse primers. Amplification was performed in a GeneAmp PCR 9700 System thermocycler (Applied Biosystems) programmed at 94 °C for 2 min followed by 35 cycles of 94 °C for 30 s, 50–65 °C for 30 s and 72 °C for 1 min, and a final extension step at 72 °C for 10 min. Amplified products were separated on a high-throughput DNA fragment analyzer (*Advance* FS; Advanced Analytical Technologies, Inc. [AATI], Ames, IA, USA). Amplified PCR products were diluted 1:10 depending on the initial concentration of the products, the dilution and the injection voltage needed to adjust to prevent excessive PCR product on the fragment analyzer. PCR product of 2 μ l was pipetted into 22 μ l of 1 \times TE dilution buffer in respective wells of the sample plate. The samples were size-separated with a 96-capillary automated system with capillaries of 80 cm. Polymer and other required reagents were from a double-stranded DNA (dsDNA) DNF-900 kit (AATI). The DNF-900 dsDNA reagent kit can effectively separate the amplicon ranges between 35 and 500 bp and can differentiate a 1-bp difference between various alleles. Following the capillary electrophoresis, the data were processed by *PRO Size* 2.0 (AATI). The data were normalized to the 35-bp lower marker and 500-bp upper marker and calibrated to the 75- to 400-bp range.

Linkage analysis

The resulting genotypes for the mapping population were matched for respective parental genotypes and transformed into a locus file. Construction of a genetic linkage map involved the use of JoinMap 4.1 (Van Ooijen 2011) with regression mapping. Markers were grouped into linkage groups with a logarithm of odds (LOD) score of 10.0 as the initial threshold and groups were selected up to LOD 5.0. Default parameters were used with the maximum likelihood algorithm for map building with the exception of changing spatial

sampling thresholds. The Kosambi map function was used to estimate map distances.

Diversity analysis

The P matrix for five principal components was calculated from all the SNP genotypes using genotype principal component analysis (PCA) with the SNP & Variation Suite (SVS) v7.7.6 (Golden Helix, Bozeman, MT, USA; www.goldenhelix.com). A PCA chart is presented with the first two eigenvectors. Allele calls were also used for genetic diversity analysis with Tassel v3.0 (www.maizegenetics.net), and a neighbor-joining tree was built with MEGA 5 (www.megasoftware.net).

Results

SNP validation and sequence annotation

A total of 22.8 million reads were obtained from PI 244018 and PI 270306, representing ~1.2 Gb of sequence data. All reads were first coalesced into contigs by using the Velvet assembler. Initial de novo assembly produced ~2.9 Mb of watermelon genome sequence distributed over 12,105 individual contigs. Contig lengths ranged from 320 to 750 bp. The contig length distribution was in line with the fragment size range selected during RAD-Seq library preparation. Contigs were then evaluated for the presence of repetitive elements by using the RepeatMasker web server with the Arabidopsis Repbase library. The percentage of the RAD-Seq (RHA 464) assembly classified as repetitive by RepeatMasker was 0.9 %. This finding is consistent with a genome assembly principally from low-copy regions, because the 450-Mb watermelon genome is expected to contain more than 60 % repetitive nucleotide content. The GC dinucleotide content assembly was 32.2 %, which is consistent with results from paired-end RAD-Seq studies in other plant genomes (Pegadaraju et al. 2013). SNP-containing sequences were used to generate a map based on the publicly available draft sequence of the watermelon genome (www.icugi.org) with Bowtie v1.0. Mapping was performed with and without mismatches allowed to locate various chromosome positions. A total of 8,234 SNPs from both mapping parents were identified after submission to the Assay Design Tool of Illumina (<http://icom.illumina.com>)

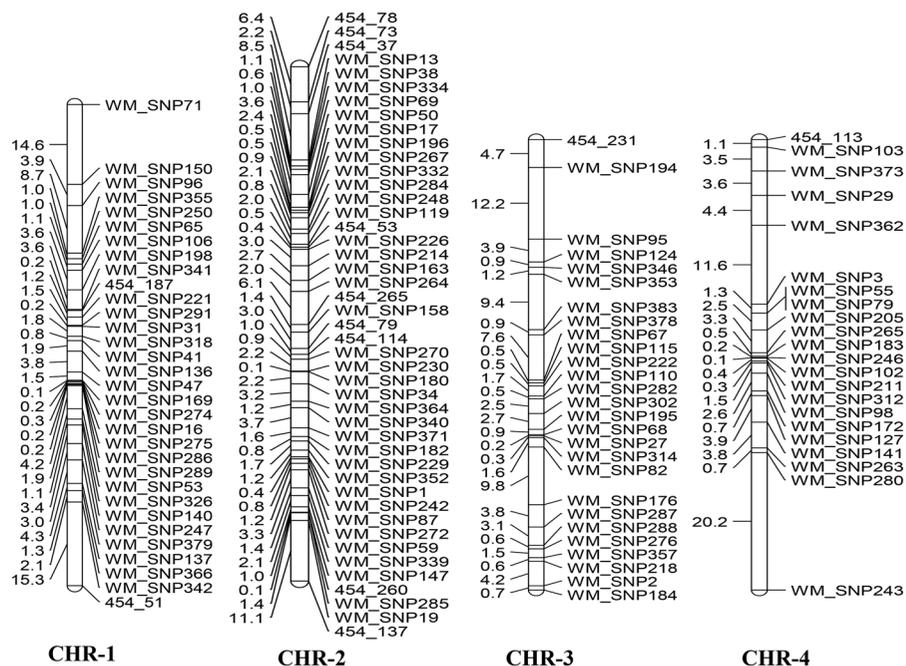
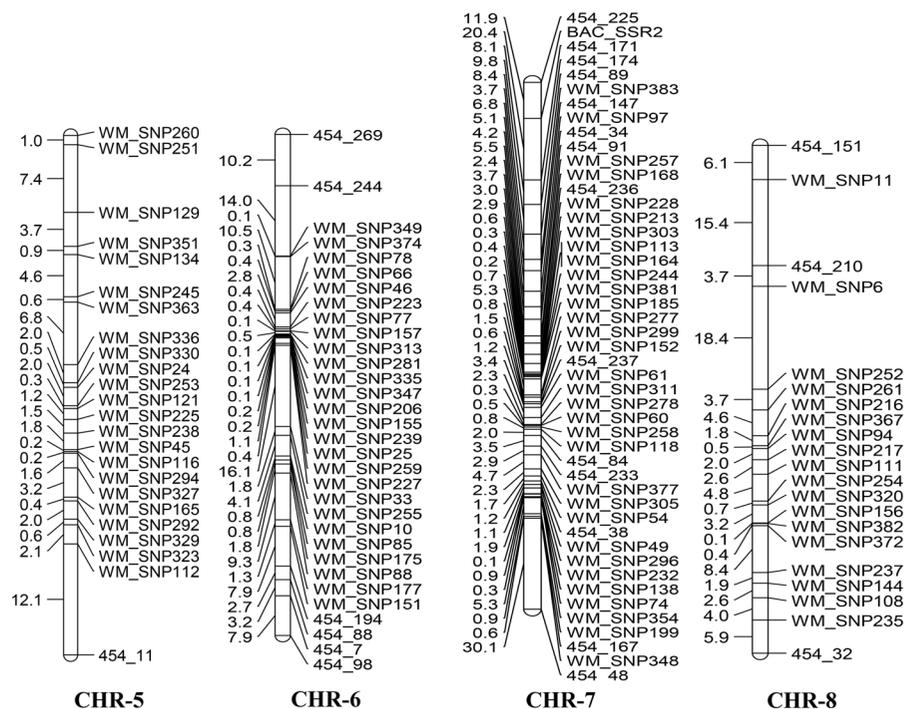
with a final design score of 0.7. We selected 522 of these SNPs by their uniform distribution across the chromosomes and their importance in biological and molecular functions as inferred by GO analysis. A set of 384 was assayed; 50 of the SNPs could not be reliably scored, mainly because of incorrect automatic clustering by the GenomeStudio software. We obtained a final set of 334 successfully genotyped, polymorphic and monomorphic SNPs across the watermelon accessions and mapping populations.

The GO annotations for SNPs showed a fairly consistent sampling of functional classes, which indicates that they represent various genes with known molecular functions involving important biological processes. Cellular, metabolic, biosynthetic and developmental processes were evenly represented. The annotation files for 384 SNP sequences and figures (Figs. 1S, 2S and 3S) pertaining to GO annotations for biological processes, molecular functions and cellular contents are in supplementary files.

Development of integrated genetic map

All 334 SNPs successfully assayed were polymorphic when used for genotyping the mapping population (parental and F2 progenies). A total of 288 SNPs and 104 SSRs (94 from 454-pyrosequencing and 10 from BAC-end sequencing, respectively) were used to construct the map. A total of 156 SNPs and 80 SSRs showed significant Chi squared deviation from monogenic, Mendelian segregation; 60 markers (46 SNPs and 14 SSRs) had significant deviations, with total Chi squared values >15 and were eliminated from the mapping data. Of 378 markers subjected to mapping, 282 (232 SNPs and 50 SSRs) were incorporated into the linkage map (Figs. 1, 2, 3).

Because most of the markers in linkage groups were assembled by using the whole-genome sequence for watermelon, chromosome numbers could be readily assigned to various linkage groups. Among the mapped SSRs, only two were from BAC-end sequences, which were mapped onto chromosomes (Chrs) 7 and 8. Various linkage groups were formed, with number of markers ranging from 16 to 47. SNPs ranged from 3 in Chr10 to 34 in Chr9. In all, 19 SNPs of Chr10 were present in the mapping data, with only three markers on the map because of their distorted segregation. Map distances of 11 chromosomes were 51.24, 94.33, 68.91, 66.22, 56.68, 99.66, 174.28, 90.84, 77.21, 76.60 and 68.75 cM.

Fig. 1 Linkage map of Chr1, 2, 3 and 4**Fig. 2** Linkage map of Chr5, 6, 7 and 8

Chromosome-wide physical locations of map positions for various SNPs are given in the Appendix table.

We examined the colinearity of the genetic and physical maps for various chromosomes (Fig. 4).

Markers on Chr2, 3, 4, 7, 8 and 9 were highly colinear with respect to their physical locations. The remaining chromosomes had undergone few to large rearrangements in terms of their physical location on the whole-

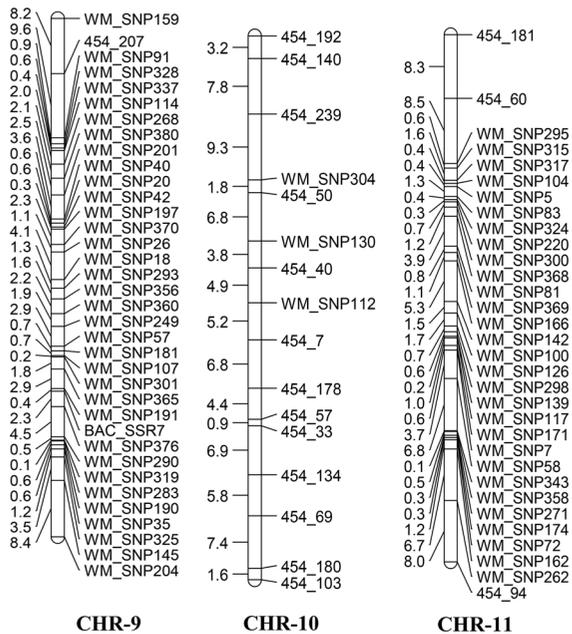


Fig. 3 Linkage map of Chr9, 10 and 11

genome sequence draft. Chr1 showed the highest disagreement between the genetic and physical maps, with only 13 markers across the length being colinear in terms of their position on the physical map. We observed 10 genome rearrangements in Chr1, one each on Chr2 and 3, none on Chr4, two large rearrangements each on Chr5 and 6, one each on Chr7, 8 and 9 and two on Chr11.

Genetic diversity and relationships between the cultivars

In this research, we determined the relationships between 130 watermelon cultivars across the world using SNP genotyping (Figs. 5 and 4S). A total of 134 SNPs were polymorphic in the set. Overall genetic similarity was 0.95 % among the cultivars, so edible watermelons of world collections maintain 5 % genetic diversity among them. PCA did not show any clear clustering pattern separating watermelons based on collection sites, but we found significant diversity among accessions. However, when African red-fleshed watermelons were analyzed separately, PI 32137, PI 525084 and PI 482248 were distinctly different and were separated by 7, 8.5 and 11 % genetic distance, respectively, from the remaining types (Fig. 5S). In particular, neighbor-joining analysis produced seven different clusters, with no clear distinction of

accessions based on the site of collection or geographical identity (Fig. 4S). Cluster 1 contained 29 cultivars: nine from America, three Asia, six Africa and the rest Europe. Cluster 2 had 16 cultivars: four from Africa and the rest from Europe or Asia. Cluster 3 had nine cultivars from Europe and Asia. Cluster 4 had 34 cultivars: nine from Africa and Europe and the rest mostly from Asia. Cluster 5 had 17 cultivars with a mix of collections from various parts of the world. Cluster 6 had 13 cultivars, with none from America. Cluster 7 had 15 cultivars: two from America and the rest from Africa, Asia and Europe.

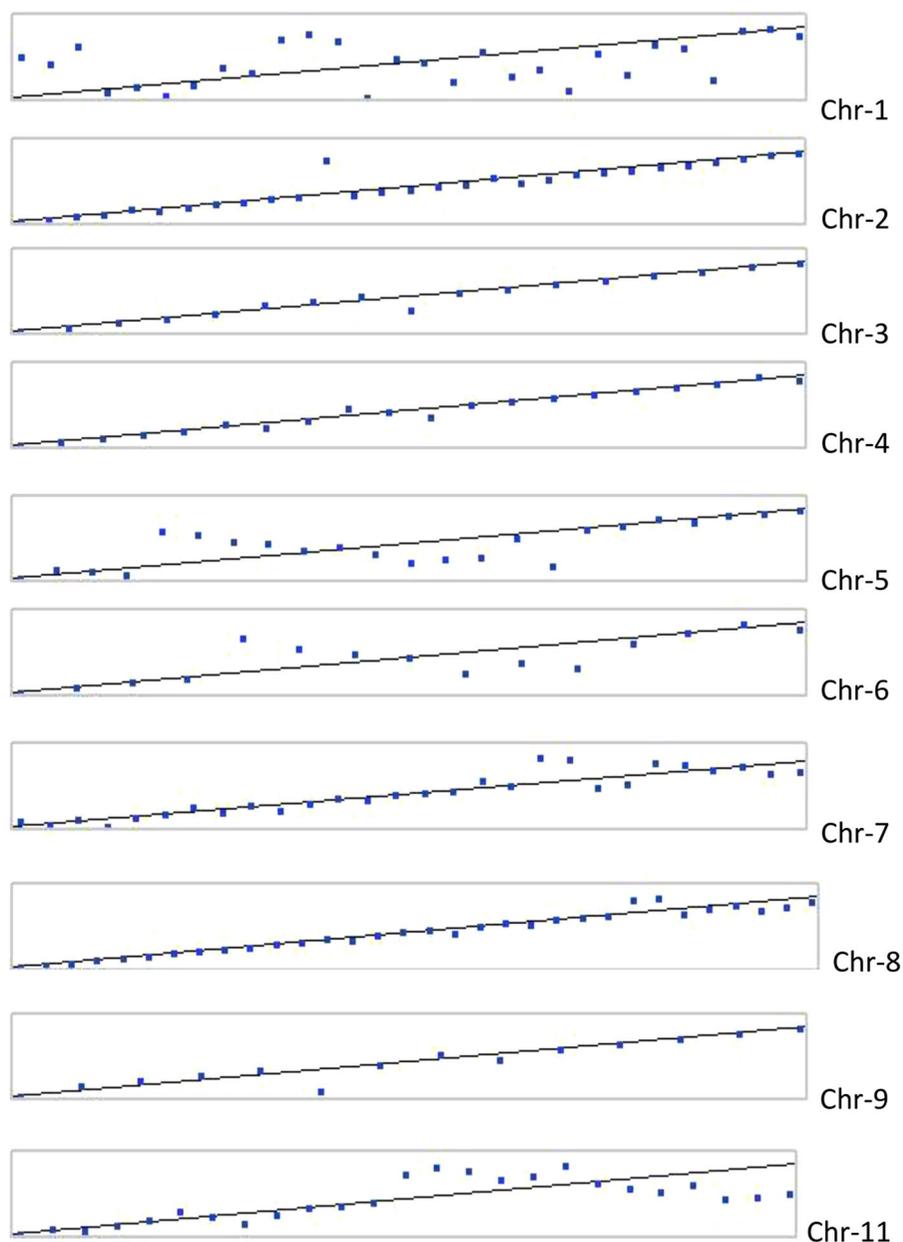
Discussion

Here we developed an SNP assay containing 384 markers that was suitable for genetic mapping and resolving genetic diversity among cultivated watermelon. Most SNP-containing sequences were found to have catalytic and binding activities and included a large number of hydrolases, kinases and transferases. Other abundant assignments were abiotic and biotic stress-response along with the other signal transduction, transport and transcriptional regulations. Platforms with the collections of functionally important SNPs and those with known annotations will be useful in future genetic studies (Esteras et al. 2012).

The genetic map we developed consisted of all 11 chromosomes spanning 924.72 cM. Map lengths of previously published SNP maps (Sandlin et al. 2012) are larger than that in our study. Genetic maps with sizes <800 cM would agree well with the small watermelon genome size of 450 Mb (Ren et al. 2012). Moreover, three SNP maps generated by Sandlin et al. (2012) had only 55 common markers, which indicates that most of their markers were not polymorphic. Moreover, linkage groups in their maps were not assigned to any chromosomes. In our study, the average distance between pairs of markers was 3.28 cM across the map, which is comparable with previous SNP maps (3.8, 4.2 and 3.4 cM) (Sandlin et al. 2012).

Previous mapping studies have shown the presence of distorted segregation in the wide cross of *lanatus* × *citroides* (Levi et al. 2002). Reddy et al. (2013) reported the presence of wide chromosomal structural differences among *lanatus* and *citroides*, which could explain the distorted segregation in the mapping populations derived from these subspecies. Compared with the physical and genetic map positions of various

Fig. 4 Colinearity between the genetic map and physical map. Markers that are distant from the ‘line of best fit’ are not colinear

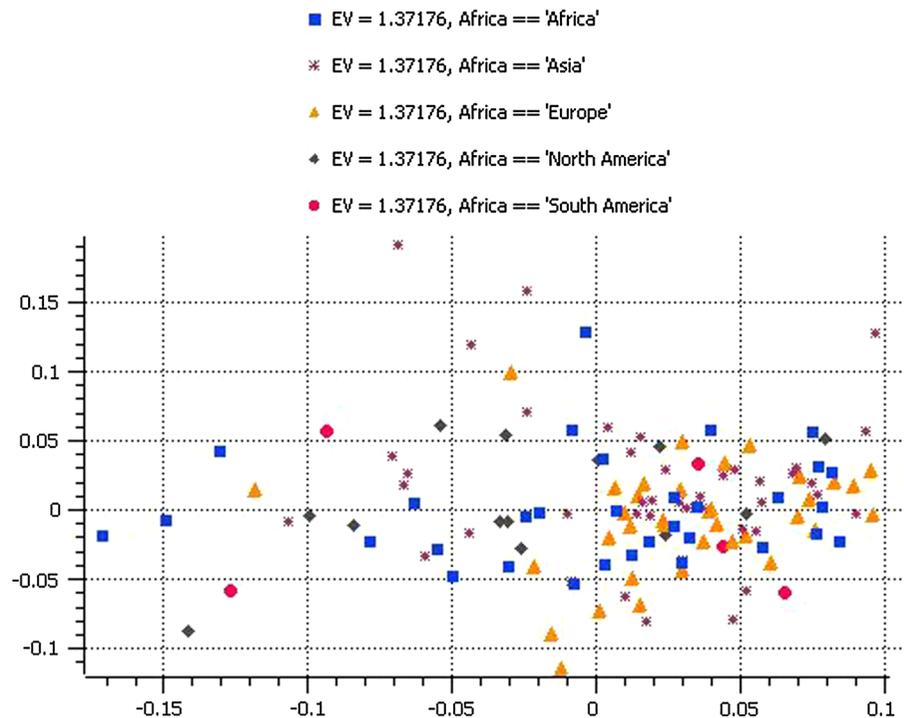


SNPs, we noted many disagreements, which may have occurred because our mapping population was derived from a cross of *C. lanatus* var. *lanatus* × *C. lanatus* var. *citroides*, which are known to produce genome-wide distortions.

Previous studies indicated that the molecular diversity in cultivated watermelon ranged from 2 to 4 % (Levi et al. 2013; Nimmakayala et al. 2011; Romão 2000; Levi et al. 2001; Zhang et al. 2012). From our analysis of cultivated watermelons only, we support

published findings of the narrow genetic diversity among American watermelon accessions. One explanation for the narrow genetic diversity in American and European germplasm could be the founder effect, whereby a small number of accessions are brought to a continent or region as people travel (Dane and Liu 2007; Tóth Zoltán et al. 2007). Watermelons may have entered Europe around 512 AD, when the Moors invaded the Iberian peninsula, or during the Crusades (Tóth Zoltán et al. 2007). In India and China,

Fig. 5 Principal component analysis (PCA) of the first two components of the world collections of watermelon accessions



watermelon was introduced around 800 and 1100 AD, respectively (Paris et al. 2013). The introduction of watermelon cultivars into the Americas occurred after the second voyage of Columbus and during the slave trade and colonization (Romão 2000; Paris et al. 2013; Tóth Zoltán et al. 2007). In contrast, when we analyzed cultivated types with edible flesh from Africa alone, the diversity of accessions ranged widely. Most importantly, our study disagrees with results obtained by Zhang et al. (2012), especially in showing distinct clustering of American and Chinese or East Asian types. The discrepancy may be due to the smaller set of Chinese and United States cultivars included in the Zhang et al. study. In our research, we had 130 cultivated forms of watermelon from the entire world, especially the ancestral cultivar forms from Africa. As we included several intermediate types between East Asian types and American types in the analysis, both the ecotypes became merged into a single group.

Most pioneering work on watermelon ecotypes was carried out by Fursa (1972). According to this, African collections are highly diverse and possess foundation types for global collections. Fursa (1972) divided rest of the world watermelon collections into various ecotypes based on their fruit characteristics. The

American ecotypes are oval-shaped fruits with red flesh and produce sweet to highly sweet flavor. In contrast, East Asian types are round, with flesh ranging from yellow to orange or red. The Russian ecotype is round, with pink or crimson flesh. Most Transcaucasian watermelons are oval-shaped with red flesh and are less sweet. Afghanistan and Iran have a wide variety of watermelons, with fruit shapes ranging from round to oval and a wide array of flesh colors: white, yellow, light pinkish and red. Because our collection contained fruit shapes ranging from round to oval with a wide range of colors, our study did not agree with that of Zhang et al. (2012), which was focused on resolving differences between East Asian and American ecotypes.

Watermelon fruit can be round, oval, blocky, oblong and elongate (Poole 1944; Guner and Wehner 2004). Weetman (1937) concluded that fruit shape dissimilarities in watermelons were shown to be fixed in the primordial ovary. One of the major factors known to alter shape of ovary in cucurbits is sex expression. If the sex expression is andromonoecious, the ovary is round in shape, with some exceptions. In contrast, most monoecious flowers produce an oval-shaped ovary. The sex expression is known to be altered by drought or

higher temperature. Cultivars with andromonoecious flowers can withstand drought or higher temperatures because the pollen is produced in the same bisexual flower. In monoecious flowers, pollen is produced in male flowers and could desiccate under stress. This situation may explain the altered genetic diversity across the geographical zones. However, critical research is further needed to understand the genes and genomic areas that are important for varietal differentiation and divergence with special reference to watermelon.

The 384-plex SNP set is a powerful tool for characterizing genetic relatedness and for developing high-resolution genetic maps. Robust allele calling and low-cost genotyping allows for analysis of large numbers of families in breeding populations or accessions in germplasm collections. High-density SNP arrays, with thousands of SNPs for crops such as maize and rice (Ganal et al. 2011; Hansey et al. 2012; Zhai et al. 2013), shows what is possible for the watermelon research community. The identification of SNPs for watermelon in this and previous studies will allow for genome-wide association mapping and marker-assisted selection to support breeding programs.

Acknowledgments This project was supported by the USDA-NIFA (no. 2013-38821-21453), National Science Foundation under Grant No. NSF 09-570 - EPS-1003907, Gus R. Douglass Institute and NIH Grant P20RR016477 to the West Virginia IDeA Network for Biomedical Research Funding. The authors thank GRDI graduate assistantships for A. Bhandary and 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. Conflict of interest We declare that we have no conflict of interest.

References

- Aslam M, Bastiaansen J, Elferink M, Megens H-J, Crooijmans R, Blomberg L, Fleischer R, Van Tassell C, Sonstegard T, Schroeder S, Groenen M, Long J (2012) Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). *BMC Genom* 13(1):391
- Baird NA, Eitter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3(10):e3376. doi:10.1371/journal.pone.0003376
- Berlin S, Lagercrantz U, von Arnold S, Ost T, Ronnberg-Wastl A (2010) High-density linkage mapping and evolution of paralogs and orthologs in *Salix* and *Populus*. *BMC Genom* 11(1):129
- Cavagnaro P, Senalik D, Yang L, Simon P, Harkins T, Kodira C, Huang S, Weng Y (2010) Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). *BMC Genom* 11(1):569
- Cavagnaro P, Chung S-M, Manin S, Yildiz M, Ali A, Alesandro M, Iorizzo M, Senalik D, Simon P (2011) Microsatellite isolation and marker development in carrot—genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across Apiaceae. *BMC Genom* 12(1):386
- Cordoba J, Chavarro C, Schlueter J, Jackson S, Blair M (2010) Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. *BMC Genom* 11(1):436
- Dane F, Liu J (2007) Diversity and origin of cultivated and citron type watermelon (*Citrullus lanatus*). *Genet Resour Crop Evol* 54(6):1255–1265. doi:10.1007/s10722-006-9107-3
- Deschamps S, Campbell M (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol Breed* 25(4):553–570. doi:10.1007/s11032-009-9357-9
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6(5):e19379. doi:10.1371/journal.pone.0019379
- Ersoz ES, Wright MH, Pangilinan JL, Sheehan MJ, Tobias C, Casler MD, Buckler ES, Costich DE (2012) SNP discovery with EST and NextGen sequencing in switchgrass (*Panicum virgatum* L.). *PLoS ONE* 7(9):e44112. doi:10.1371/journal.pone.0044112
- Esteras C, Gomez P, Monforte A, Blanca J, Vicente-Dolera N, Roig C, Nuez F, Pico B (2012) High-throughput SNP genotyping in *Cucurbita pepo* for map construction and quantitative trait loci mapping. *BMC Genom* 13(1):80
- Fursa T (1972) K Sistematičke roda *Citrullus* schrad. *Bot Z* 57(1):31–41
- Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner E-M, Hansen M, Joets J, Le Paslier M-C, McMullen MD, Montalent P, Rose M, Schön C-C, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* 6(12):e28334. doi:10.1371/journal.pone.0028334
- Guner N, Wehner TC (2004) The genes of watermelon. *Hort-Science* 39(6):1175–1182
- Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, Min J, Guo X, Murat F, Ham BK, Zhang Z, Gao S, Huang M, Xu Y, Zhong S, Bombarely A, Mueller LA, Zhao H, He H, Zhang Y, Zhang Z, Huang S, Tan T, Pang E, Lin K, Hu Q, Kuang H, Ni P, Wang B, Liu J, Kou Q, Hou W, Zou X, Jiang J, Gong G, Klee K, Schoof H, Huang Y, Hu X, Dong S, Liang D, Wang J, Wu K, Xia Y, Zhao X, Zheng Z, Xing M, Liang X, Huang B, Lv T, Wang J, Yin Y, Yi H, Li R, Wu M, Levi A, Zhang X, Giovannoni JJ, Wang J, Li Y, Fei Z, Xu Y (2013) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet* 45(1):51–58. doi:10.1038/ng.2470

- Hansey CN, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Buell CR (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS ONE 7(3):e33071. doi:[10.1371/journal.pone.0033071](https://doi.org/10.1371/journal.pone.0033071)
- Jarret RL, Merrick LC, Holms T, Evans J, Aradhya MK (1997) Simple sequence repeats in watermelon (*Citrullus lanatus* (Thunb.) Matsum. & Nakai). Genome/National Research Council Canada—Genome/Conseil national de recherches Canada 40(4):433–441
- Levi A, Thomas C, Keinath A, Wehner T (2001) Genetic diversity among watermelon (*Citrullus lanatus* and *Citrullus colocynthis*) accessions. Genet Resour Crop Evol 48(6):559–566. doi:[10.1023/a:1013888418442](https://doi.org/10.1023/a:1013888418442)
- Levi A, Thomas E, Joobeur T, Zhang X, Davis A (2002) A genetic linkage map for watermelon derived from a test-cross population: (*Citrullus lanatus* var. *citroides* × *C. lanatus* var. *lanatus*) × *Citrullus colocynthis*. Theor Appl Genet 105(4):555–563. doi:[10.1007/s00122-001-0860-6](https://doi.org/10.1007/s00122-001-0860-6)
- Levi A, Thies J, Wechter WP, Harrison H, Simmons A, Reddy U, Nimmakayala P, Fei Z (2013) High frequency oligonucleotides: targeting active gene (HFO-TAG) markers revealed wide genetic diversity among *Citrullus* spp. accessions useful for enhancing disease or pest resistance in watermelon cultivars. Genet Resour Crop Evol 60(2):427–440. doi:[10.1007/s10722-012-9845-3](https://doi.org/10.1007/s10722-012-9845-3)
- Li Q, Li P, Sun L, Wang Y, Ji K, Sun Y, Dai S, Chen P, Duan C, Leng P (2012) Expression analysis of beta-glucosidase genes that regulate abscisic acid homeostasis during watermelon (*Citrullus lanatus*) development and under stress conditions. J Plant Physiol 169(1):78–85. doi:[10.1016/j.jplph.2011.08.005](https://doi.org/10.1016/j.jplph.2011.08.005)
- Lin CH, Yeakley JM, McDaniel TK, Shen R (2009) Medium- to high-throughput SNP genotyping using VeraCode microbeads. Methods Mol Biol 496:129–142. doi:[10.1007/978-1-59745-553-4_10](https://doi.org/10.1007/978-1-59745-553-4_10)
- Mammadov J, Chen W, Mingus J, Thompson S, Kumpatla S (2012) Development of versatile gene-based SNP assays in maize (*Zea mays* L.). Mol Breed 29(3):779–790. doi:[10.1007/s11032-011-9589-3](https://doi.org/10.1007/s11032-011-9589-3)
- Meeuse ADJ (1962) The Cucurbitaceae of southern Africa. Bothalia 8(1):1–111
- Neves L, Mamani EMC, Alfenas A, Kirst M, Grattapaglia D (2011) A high-density transcript linkage map with 1,845 expressed genes positioned by microarray-based Single Feature Polymorphisms (SFP) in Eucalyptus. BMC Genom 12(1):189
- Nimmakayala P, Islam-Faridi N, Tomason Y, Lutz F, Levi A, Reddy U (2011) Citrullus. In: Wild crop relatives: genomic and breeding resources. Springer, New York, pp 59–66
- Oliver RE, Tinker NA, Lazo GR, Chao S, Jellen EN, Carson ML, Rines HW, Obert DE, Lutz JD, Shackelford I, Korol AB, Wight CP, Gardner KM, Hattori J, Beattie AD, Bjørnstad Å, Bonman JM, Jannink J-L, Sorrells ME, Brown-Guedira GL, Mitchell Fetch JW, Harrison SA, Howarth CJ, Ibrahim A, Kolb FL, McMullen MS, Murphy JP, Ohm HW, Rossnagel BG, Yan W, Mioclaus KJ, Hiller J, Maughan PJ, Redman Hulse RR, Anderson JM, Islamovic E, Jackson EW (2013) SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. PLoS ONE 8(3):e58068. doi:[10.1371/journal.pone.0058068](https://doi.org/10.1371/journal.pone.0058068)
- Parida S, Dalal V, Singh A, Singh N, 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 Genom 10(1):140
- Paris HS, Daunay M-C, Janick J (2013) Medieval iconography of watermelons in Mediterranean Europe. Ann Bot 112(5):867–879. doi:[10.1093/aob/mct151](https://doi.org/10.1093/aob/mct151)
- Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. BMC Genom 14(1):556. doi:[10.1186/1471-2164-14-556](https://doi.org/10.1186/1471-2164-14-556)
- Pitrat M, Chauvet M, Foury C (1999) Diversity, history and production of cultivated cucurbits. Acta Hort 492:21–28
- Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2):e32253. doi:[10.1371/journal.pone.0032253](https://doi.org/10.1371/journal.pone.0032253)
- Poole C (1944) Genetics of cultivated cucurbits. J Hered 35:122–128
- Reddy U, Aryal N, Islam-Faridi N, Tomason Y, Levi A, Nimmakayala P (2013) Cytomolecular characterization of rDNA distribution in various *Citrullus* species using fluorescent in situ hybridization. Genet Resour Crop Evol 60(7):2091–2100. doi:[10.1007/s10722-013-9976-1](https://doi.org/10.1007/s10722-013-9976-1)
- Ren Y, Zhao H, Kou Q, Jiang J, Guo S, Zhang H, Hou W, Zou X, Sun H, Gong G (2012) A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. PLoS ONE 7(1):e29453
- Robinson RW, Decker-Walters DS (1997) Cucurbits. Crop production science in horticulture, vol 6. CAB International, Wallingford, Oxon
- Romão R (2000) Northeast Brazil: a secondary center of diversity for watermelon (*Citrullus lanatus*). Genet Resour Crop Evol 47(2):207–213. doi:[10.1023/a:1008723706557](https://doi.org/10.1023/a:1008723706557)
- Sandlin K, Prothro J, Heesacker A, Khalilian N, Okashah R, Xiang W, Bachlava E, Caldwell DG, Taylor CA, Seymour DK (2012) Comparative mapping in watermelon [*Citrullus lanatus* (Thunb.) Matsum. et Nakai]. Theor Appl Genet 125(8):1603–1618
- Semagn K, Babu R, Hearne S, Olsen M (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. Mol Breed 33:1–14. doi:[10.1007/s11032-013-9917-x](https://doi.org/10.1007/s11032-013-9917-x)
- Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganai MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S, Francis DM (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS ONE 7(7):e40563. doi:[10.1371/journal.pone.0040563](https://doi.org/10.1371/journal.pone.0040563)
- Stagel A, Portis E, Toppino L, Rotino G, Lanteri S (2008) Gene-based microsatellite development for mapping and phylogeny studies in eggplant. BMC Genom 9(1):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(3):e59562. doi:[10.1371/journal.pone.0059562](https://doi.org/10.1371/journal.pone.0059562)

- Thomson M, Zhao K, Wright M, McNally K, Rey J, Tung C-W, Reynolds A, Scheffler B, Eizenga G, McClung A, Kim H, Ismail A, Ocampo M, Mojica C, Reveche MY, Dilla-Ermita C, Mauleon R, Leung H, Bustamante C, McCouch S (2012) High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. *Mol Breed* 29(4):875–886. doi:[10.1007/s11032-011-9663-x](https://doi.org/10.1007/s11032-011-9663-x)
- Tóth Zoltán GG, Szabó Z, Horváth L, Heszky L (2007) Watermelon (*Citrullus l. lanatus*) production in Hungary from the Middle Ages (13th century). *Hung Agric Res* 4:14–19
- van Ooijen JW (2011) Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet Res* 93(05):343–349. doi:[10.1017/S0016672311000279](https://doi.org/10.1017/S0016672311000279)
- Viquez-Zamora M, Vosman B, van de Geest H, Bovy A, Visser R, Finkers R, van Heusden A (2013) Tomato breeding in the genomics era: insights from a SNP array. *BMC Genom* 14(1):354
- Weetman L (1937) Inheritance and correlation of shape, size, and color of watermelon. *Iowa Agric Exp Sta Res Bull* 228:222–256
- Whitaker TW, Bemis WP (1976) Cucurbits. In: Simmonds NW (ed) *Evolution of crop plants*. Longman, London, pp 64–69
- Whitaker TW, Davis GN (1962) *Cucurbits. Botany, cultivation, and utilization*. Interscience Publishers, New York
- Wong M, Cannon C, Wickneswari R (2012) Development of high-throughput SNP-based genotyping in *Acacia auriculiformis* x *A. mangium* hybrids using short-read transcriptome data. *BMC Genom* 13(1):726
- Yan J, Yang X, Shah T, Sánchez-Villeda H, Li J, Warburton M, Zhou Y, Crouch J, Xu Y (2010) High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol Breed* 25(3):441–451. doi:[10.1007/s11032-009-9343-2](https://doi.org/10.1007/s11032-009-9343-2)
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome res* 18:821–829
- Zhai R, Feng Y, Zhan X, Shen X, Wu W, Yu P, Zhang Y, Chen D, Wang H, Lin Z, Cao L, Cheng S (2013) Identification of transcriptome SNPs for assessing allele-specific gene expression in a super-hybrid rice Xieyou9308. *PLoS ONE* 8(4):e60668. doi:[10.1371/journal.pone.0060668](https://doi.org/10.1371/journal.pone.0060668)
- 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(2):329–342. doi:[10.1007/s10681-011-0574-z](https://doi.org/10.1007/s10681-011-0574-z)