nature genetics

OPEN

The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions

Shaogui Guo[1,2,17], Jianguo Zhang[3,4,17], Honghe Sun[1,2,5,17], Jerome Salse[6,17], William J Lucas[7,17], Haiying Zhang[1], Yi Zheng[2], Linyong Mao[2], Yi Ren[1], Zhiwen Wang[3], Jiumeng Min[3], Xiaosen Guo[3], Florent Murat[6], Byung-Kook Ham[7], Zhaoliang Zhang[7], Shan Gao[2], Mingyun Huang[2], Yimin Xu[2], Silin Zhong[2], Aureliano Bombarely[2], Lukas A Mueller[2], Hong Zhao[1], Hongju He[1], Yan Zhang[1], Zhonghua Zhang[8], Sanwen Huang[8], Tao Tan[9], Erli Pang[9], Kui Lin[9], Qun Hu[10], Hanhui Kuang[10], Peixiang Ni[3,4], Bo Wang[3], Jingan Liu[1], Qinghe Kou[1], Wenju Hou[1], Xiaohua Zou[1], Jiao Jiang[1], Guoyi Gong[1], Kathrin Klee[11], Heiko Schoof[11], Ying Huang[3], Xuesong Hu[3], Shanshan Dong[3], Dequan Liang[3], Juan Wang[3], Kui Wu[3], Yang Xia[1], Xiang Zhao[3], Zequn Zheng[3], Miao Xing[3], Xinming Liang[3], Bangqing Huang[3], Tian Lv[3], Junyi Wang[3], Ye Yin[3], Hongping Yi[12], Ruiqiang Li[13], Mingzhu Wu[12], Amnon Levi[14], Xingping Zhang[1], James J Giovannoni[2,15], Jun Wang[3,16], Yunfu Li[1], Zhangjun Fei[2,15] & Yong Xu[1]

Watermelon, *Citrullus lanatus*, is an important cucurbit crop grown throughout the world. Here we report a high-quality draft genome sequence of the east Asia watermelon cultivar 97103 (2*n* = 2× = 22) containing 23,440 predicted protein-coding genes. Comparative genomics analysis provided an evolutionary scenario for the origin of the 11 watermelon chromosomes derived from a 7-chromosome paleohexaploid eudicot ancestor. Resequencing of 20 watermelon accessions representing three different *C. lanatus* subspecies produced numerous haplotypes and identified the extent of genetic diversity and population structure of watermelon germplasm. Genomic regions that were preferentially selected during domestication were identified. Many disease-resistance genes were also found to be lost during domestication. In addition, integrative genomic and transcriptomic analyses yielded important insights into aspects of phloem-based vascular signaling in common between watermelon and cucumber and identified genes crucial to valuable fruit-quality traits, including sugar accumulation and citrulline metabolism.

Watermelon (*C. lanatus*) is an important cucurbit crop, accounting for 7% of the worldwide area devoted to vegetable production. The annual world production of watermelon is about 90 million tons, making it among the top five most consumed fresh fruits (http://faostat.fao.org/). Watermelon belongs to the xerophytic genus *Citrullus* Schrad. ex Eckl. et Zeyh. of the botanical family *Cucurbitaceae*. The center of diversity and possible center of origin of *Citrullus* is southern Africa[1]. *C. lanatus* includes three subspecies: *C. lanatus* subsp. *lanatus*, which represents a group of ancient cultigens, the 'tsamma' or 'citron' watermelon, that naturally thrives in southern Africa; *C. lanatus* subsp. *mucosospermus*, which represents the egusi watermelon group that has large edible seeds with a fleshy pericarp[2]; and *C. lanatus* subsp. *vulgaris*, which represents the sweet (dessert) watermelon group that gave rise to the modern cultivated watermelon[3].

The large edible watermelon fruits contribute to the diets of consumers throughout the world. Although comprised mainly of water (often over 90%), watermelon also contains important nutritional compounds, including sugars, lycopene and cardiovascular health–promoting amino acids, such as citrulline, arginine and glutathione[4–6]. Watermelon and cucurbit species in general have unique developmental mechanisms that facilitate the rapid growth and formation of giant pepo fruits[7]. Fruits of modern watermelon varieties are diverse in shape, size, color, texture, flavor and nutrient composition. However, years of cultivation and selection targeting yield and desirable fruit

[1]National Engineering Research Center for Vegetables, Beijing Academy of Agriculture and Forestry Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China), Beijing, China. [2]Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, New York, USA. [3]Beijing Genomics Institute–Shenzhen, Shenzhen, China. [4]T-Life Research Center, Fudan University, Shanghai, China. [5]College of Plant Science and Technology, Beijing University of Agriculture, Beijing, China. [6]Institut National de la Recherche Agrinomique, Unités Mixtes de Recherche 1095, Genetics, Diversity and Ecophysiology of Cereals, Clermont-Ferrand, France. [7]Department of Plant Biology, College of Biological Sciences, University of California, Davis, California, USA. [8]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. [9]College of Life Sciences, Beijing Normal University, Beijing, China. [10]College of Horticulture and Forestry, Huazhong Agriculture University, Wuhan, China. [11]Institut für Nutzpflanzenwissenschaften und Ressourcenschutz Crop Bioinformatics, University of Bonn, Bonn, Germany. [12]Xinjiang Academy of Agricultural Sciences, Urumqi, China. [13]Beijing Novogene Bioinformation Technology Co. Ltd, Beijing, China. [14]US Department of Agriculture (USDA), Agricultural Research Service, US Vegetable Lab, Charleston, South Carolina, USA. [15]USDA Robert W. Holley Center for Agriculture and Health, Ithaca, New York, USA. [16]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [17]These authors contributed equally to this work. Correspondence should be addressed to Yong Xu (xuyong@nercv.org), Z.F. (zf25@cornell.edu), Y.L. (liyunfu@baafs.net.cn) or Jun Wang (wangj@genomics.org.cn).

Received 4 June; accepted 22 October; published online 25 November 2012; doi:10.1038/ng.2470

qualities have narrowed the genetic base of watermelon[8], resulting in a major bottleneck in watermelon improvement.

Knowledge of genome sequences is indispensable for basic biological research and crop improvement. Here we report a high-quality genome sequence of an east Asia watermelon cultivar, 97103 ($2n = 2\times = 22$), and resequencing of 20 watermelon accessions spanning the genetic diversity of *C. lanatus*. Our comprehensive genomic and transcriptome analyses provide insights into the structure and evolution of the watermelon genome, the genetic diversity and structure of watermelon populations and the molecular mechanisms of important biological processes such as fruit quality and phloem-based vascular signaling. Together, these results will assist in identifying and accessing the plethora of watermelon genetic diversity that remains to be tapped for biological discovery and crop improvement.

RESULTS

Genome sequencing and assembly

We selected the Chinese elite watermelon inbred line 97103 for genome sequencing. We generated a total of 46.18 Gb of high-quality genomic sequence using Illumina sequencing technology (**Supplementary Table 1**), representing 108.6-fold coverage of the entire watermelon genome, which has an estimated genome size of ~425 Mb on the basis of our 17-mer depth distribution analysis of the sequenced reads (**Supplementary Fig. 1**) and an earlier flow cytometry analysis[9]. *De novo* assembly of the Illumina reads resulted in a final assembly of 353.5 Mb, representing 83.2% of the watermelon genome. The assembly consists of 1,793 scaffolds (≥500 bp) with N50 lengths of 2.38 Mb and 26.38 kb for the scaffolds and contigs, respectively (**Supplementary Table 2**). A total of 234 scaffolds covering approximately 330 Mb (93.5% of the assembled genome) were anchored to the 11 watermelon chromosomes, among which 126 and 94 scaffolds accounting for 70% and 65% of the assembled genome were ordered and oriented, respectively[10].

We sought to determine why 16.8% of the genome was not covered by our genome assembly by aligning unassembled reads (17.4% of the total reads) to the assembled genome with less stringent criteria (**Supplementary Note** and **Supplementary Table 3**). We found that the unassembled genome regions are composed primarily of sequences that are similar to those of the assembled regions. Distribution of the unassembled reads on the watermelon chromosomes showed the same pattern as that for transposable elements (**Fig. 1a** and **Supplementary Fig. 2**). We identified three major repeat units from the unassembled sequences on the basis of their substantial read depths and sequence similarities to centromeres, telomeres and ribosomal DNA (rDNA) clusters. We further confirmed the nature of these repeats by FISH (**Fig. 1b–d**). Together these results support the notion that underestimation of the repeat proportion has an important role in the unassembled component of *de novo* genome assemblies, especially those generated using next-generation sequencing technologies[11–18].

We further evaluated the quality of the assembled watermelon genome using approximately one million ESTs, four completely sequenced BACs and paired-end sequences of 667 BAC clones. Our analyses supported the high quality of the watermelon genome assembly (**Supplementary Note**, **Supplementary Tables 4–6** and **Supplementary Figs. 3** and **4**), which is favorably comparable to several other recently published plant genomes[11–18] using next-generation sequencing technologies (**Table 1**).

Repeat sequence annotation and gene prediction

Transposable elements are major components of eukaryotic genomes. We identified a total of 159.8 Mb (45.2%) of the assembled watermelon genome as transposable element repeats. Among these repeats, 68.3% could be annotated with known repeat families. The long terminal repeat (LTR) retrotransposons, mainly Gypsy-type and Copia-type LTRs, are predominant. The distribution of transposable element divergence rates showed a peak at 32% (**Supplementary Fig. 5**). We further identified 920 (7.8 Mb) full-length LTR retrotransposons in the watermelon genome. We found that over the past 4.5 million years, LTR retrotransposons accumulated much faster in watermelon than in cucumber[14] (**Supplementary Fig. 6**) such that the overall difference in their genome sizes may reflect the differential LTR retrotransposon accumulation.

We predicted 23,440 high-confidence protein-coding genes in the watermelon genome (**Supplementary Table 7**), which is close to the number of genes predicted in the cucumber genome[19]. Approximately 85% of the watermelon predicted genes had either known homologs or could be functionally classified (**Supplementary Table 8**). In addition, we also identified 123 ribosomal RNA (rRNA), 789 transfer RNA, 335 small nuclear RNA and 141 microRNA genes (**Supplementary Table 9**).

In accordance with previously reported plant genomes, the watermelon protein-coding genes showed a clear enrichment pattern within subtelomeric regions. In contrast, the transposable element–related fraction of the genome was located primarily within the pericentromeric and centromeric regions. The short arms of chromosomes 4, 8 and 11 are highly enriched with repeat sequences (**Supplementary Fig. 7**). The 97103 genome contained one 5S and two 45S rDNA clusters on the short arm of chromosomes 4 and 8 (ref. 10). Using FISH, we further investigated rDNA patterns in genomes of 20 representative watermelon accessions (**Supplementary Table 10**). The number and location of 5S and 45S rDNA sites in the genomes of the ten modern cultivated (*C. lanatus* subsp. *vulgaris*) and six semiwild watermelon (*C. lanatus* subsp. *mucosospermus*) were identical to those in the 97103 genome, whereas the genomes of the four more distantly related wild watermelon (*C. lanatus* subsp. *lanatus*) contained one 45S and two 5S rDNA sites, with the additional 5S rDNA site on the short arm of chromosome 11 (**Supplementary Fig. 8**). These results indicate that chromosome fusion, fission and transposition of rDNA might occur during the evolution of *C. lanatus* species. Our analysis also confirmed the phylogenetic relationship of these three watermelon

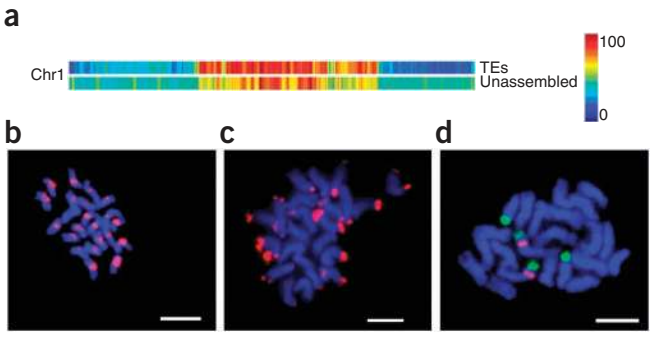

Figure 1 Distribution of unassembled reads on chromosome 1 and FISH patterns of probes from three repeat units related to the centromere, telomere and 45S rDNA clusters. (**a**) Distribution of unassembled reads on chromosome 1. The distribution of unassembled reads on the other ten chromosomes is shown in **Supplementary Figure 2**. TEs, transposable elements. (**b**) FISH of watermelon chromosomes stained with 4′-6-diamidino-2-phenylindole (DAPI, blue) using probes from repeat units similar to the centromere (pink). (**c**) FISH using probes from repeat units similar to the telomere (red). (**d**) FISH using probes from repeat units similar to the rDNAs (green, 45S; pink, 5S). Scale bars, 5 μm.

subspecies[20] and supported the hypothesis that *C. lanatus* subsp. *mucosospermus* is the recent ancestor of *C. lanatus* subsp. *vulgaris*.

Table 1 Comparison of watermelon genome assembly with other plant genomes

Species	Genome assembly size (Mb)	Estimated genome size (Mb)	Genome covered by assembly (%)	N50 scaffold (kb)	N50 contig (kb)	Sequencing technologies
Watermelon	353.3	425	83.2	2,378.2	26.4	Illumina
Date palm	381	658	57.9	30.5	6.4	Illumina
Pigeonpea	605.8	833.1	72.7	516.1	22	Illumina
Cucumber	243.5	367	66.3	226.5	19.8	Sanger+Illumina
Apple	603.9	742.3	81.3	NA	16.2	Sanger+454
Strawberry	201.9	240	84.1	1,360	NA	454+Illumina+SOLiD
Cacao	326.9	430	76	473.8	19.8	Sanger+Illumina+454
Chinese cabbage	283.8	529	53.6	1,971.1	27.3	Sanger+Illumina
Thellungiella parvula	137.1	160	85.7	5,290	NA	Illumina+454

NA, not applicable.

Cucurbit genome evolution

Genome-wide duplication in angiosperms is common and represents an important molecular mechanism that has shaped modern plant karyotypes. In the watermelon genome, we identified seven major triplications that corresponded to 302 paralogous relationships covering 29% of the genome (**Fig. 2a**). These ancestral triplicates corresponded to the shared paleohexaploidization event (referenced as γ) reported for eudicots[21] that dates back to 76–130 million years ago. This would be well in advance of the cucurbit genome speciation event that occurred 15–23 million years ago (**Supplementary Fig. 9**).

To access the nature of evolutionary events leading to modern cucurbit genome structures, we analyzed the syntenic relationships between watermelon, cucumber[19], melon[22] and grape[21]. We chose grape as the reference, as it is known to be the closest relative to the eudicot ancestor structured in seven protochromosomes[23]. We identified a total of 3,543 orthologous relationships covering 60% of the watermelon genome. We then investigated the detailed chromosome-to-chromosome relationships within the *Cucurbitaceae* family and identified orthologous chromosomes between watermelon, cucumber and melon (**Fig. 2b**). The complicated syntenic patterns illustrated as mosaic chromosome-to-chromosome orthologous relationships unveiled a high degree of complexity of chromosomal evolution and rearrangement among these three important crop species of the *Cucurbitaceae* family.

Integration of independent analyses of duplications within, and syntenies between, the four eudicot genomes (watermelon, cucumber, melon and grape) led to the precise characterization in watermelon of the seven paleotriplications identified recently as the basis for the definition of seven ancestral chromosomal groups in eudicots[24]. On the basis of the ancestral hexaploidization (γ) reported for the eudicots, we propose an evolutionary scenario that has shaped the 11 watermelon chromosomes from the 7-chromosome eudicot ancestors through the 21 paleohexaploid intermediates. We suggest that the transition from the 21-chromosome eudicot intermediate ancestors involved 81 fissions and 91 fusions to reach the modern 11-chromosome structure of watermelon, which is represented as a mosaic of 102 ancestral blocks (**Fig. 2c**).

Assessment of genetic diversity in watermelon germplasm

We selected 20 representative watermelon accessions for genome resequencing. These included ten cultivated accessions representing the major varieties of *C. lanatus* subsp. *vulgaris* (five east Asia and five America ecotypes), six semiwild *C. lanatus* subsp. *mucosospermus* and four wild *C. lanatus* subsp. *lanatus* (**Supplementary Table 10** and **Supplementary Fig. 10**). We sequenced these accessions to between 5× and 16× coverage and mapped the short reads to the genome of 97103 (**Supplementary Table 11**). We identified a total of 6,784,860 candidate SNPs and 965,006 small insertions/deletions (indels) among the 20 resequenced lines and 97103. The major variations existed between *C. lanatus* subsp. *lanatus* and the other two subspecies, whereas the variation within the cultivated watermelon, especially *C. lanatus* subsp. *vulgaris* America ecotype, was relatively low (**Supplementary Table 12**). The accuracies of our SNP and indel calling were 99.3% and 98%, respectively, as indicated by Sanger sequencing (**Supplementary Note** and **Supplementary Table 13**). This extensive watermelon genome variation dataset, covering a wide spectrum of watermelon genetic diversity, represents a valuable resource for biological discovery and germplasm improvement.

We evaluated the genetic diversity of the watermelon population using two common summary statistics, π and θw values[25]. The estimated amount of diversity in watermelon (**Supplementary Table 14**) was substantially lower than that found in maize[26], soybean[27] and rice[28]. Wild watermelon contains greater genetic diversity, indicating additional genetic opportunity for watermelon improvement. We also investigated the population structure and relationships among the watermelon accessions through construction of a neighbor-joining tree (**Fig. 3a**) and principal component analysis (PCA) (**Fig. 3b**). Both analyses indicated the close relationship between *C. lanatus* subsp. *vulgaris* and *C. lanatus* subsp. *mucosospermus* (**Supplementary Note**). Additional analysis of population structure using the FRAPPE program[29] with *K* (the number of populations) set from 2 to 5 identified a new subgroup within the *C. lanatus* subsp. *mucosospermus* group (when $K = 5$) and admixtures between *C. lanatus* subsp. *vulgaris* and *C. lanatus* subsp. *mucosospermus* (**Fig. 3c** and **Supplementary Note**). The new subgroup shows some characteristics of the cultivated watermelon, such as soft flesh texture, pink flesh color and relatively high sugar content (**Supplementary Table 10** and **Supplementary Fig. 10**). Together these results offer further support for our proposed evolutionary scenario of *C. lanatus* subsp. *mucosospermus* to *C. lanatus* subsp. *vulgaris* derived from the FISH analysis of chromosomal rDNA distribution.

We next scanned the genome for regions with the highest differences of genetic diversity ($\pi_{mucosospermus}/\pi_{vulgaris}$) between *C. lanatus* subsp. *mucosospermus* and *C. lanatus* subsp. *vulgaris*. These regions represent potential selective sweeps during watermelon domestication, as modern watermelon cultivars are thought to have been domesticated from *C. lanatus* subsp. *mucosospermus*. We identified a total of 108 regions (7.78 Mb in size) containing 741 candidate genes (**Fig. 4** and **Supplementary Table 15**). Although gene complements in these regions could have been affected by genetic hitchhiking, we identified biological processes significantly enriched in candidate genes that were related to important selected traits when compared to the whole genome, including regulation of carbohydrate use, sugar-mediated signaling, carbohydrate metabolism, response to sucrose stimulus, regulation of nitrogen-compound metabolism, cellular response to nitrogen starvation and growth (**Supplementary Note** and **Supplementary Tables 16–18**).

It is noteworthy that certain noncentromeric regions, especially a large region on chromosome 3 (from ~3.4 Mb to ~5.6 Mb), have particularly high nucleotide divergence only among *C. lanatus* subsp.

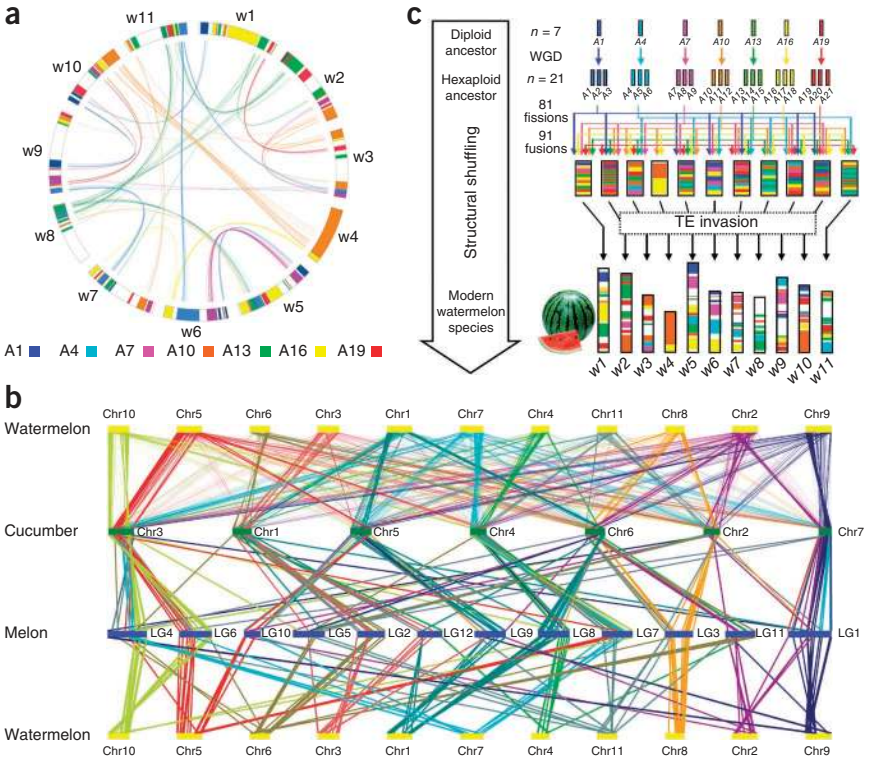

Figure 2 Genome synteny, duplication patterns and evolutionary history of watermelon, cucumber and melon. (**a**) Schematic representation of paralogous pairs identified within the watermelon genome (chromosomes w1–w11). Each line represents a syntenic region. Different colors reflect origin from the seven ancestral eudicot chromosome karyotype (A1, A4, A7, A10, A13, A16 and A19). (**b**) Schematic representation of syntenies among watermelon (chromosomes 1–11), cucumber (chromosomes 1–7) and melon genomes (linkage groups (LG) 1–12). Each line represents a syntenic region. Shared synteny between two of the three species is linked by a light gray line. (**c**) Evolution of the watermelon genome (w1–w11 at the bottom) from the common eudicot genome ancestors of seven chromosomes (A1, A4, A7, A10, A13, A16 and A19) and the derived paleohexaploid $n = 21$ (A1–A21) ancestor intermediate. Colored blocks represent the evolution of segments from the 7- or 21-chromosome ancestors to reach the modern watermelon genome structure. The 172 chromosomal fusions and fissions are highlighted with colored arrows. TE, transposable element; WGD, whole-genome duplication.

mucosospermus accessions (**Fig. 4**). A previous report described a similar finding in three different rice crosses, and it was suggested that these population-specific high-divergence regions were highly associated with genes involved in reproductive barriers[30]. We analyzed genes in the large high-diversity region on chromosome 3 and, indeed, found that the most significantly enriched gene categories were recognition of pollen and the pollen-pistil interaction; both of these gene categories are related to reproductive barriers (**Supplementary Table 19**). In addition, we determined that the region contained a large cluster of 12 tandemly arrayed S-locus protein kinase genes, which are involved in reproductive barriers[31]. The high nucleotide divergence of reproductive barrier genes in *C. lanatus* subsp. *mucosospermus*, the recent progenitor of modern cultivated watermelon, indicates that the domestication of watermelon could be a possible force responsible for the rapid evolution of reproductive barriers, as has been reported in rice[30]. Furthermore, genes involved in plant responses to abiotic and biotic stresses were also significantly enriched in this region, in addition to genes related to several known selected traits such as carbohydrate metabolism, fruit flavor (terpene metabolism) and seed oil content (fatty acid metabolism) (**Supplementary Table 19**).

Evolution of disease resistance genes in watermelon

The watermelon crop suffers major losses from numerous diseases. Therefore, improvement in pathogen resistance is an ongoing objective of watermelon breeding programs. To investigate the molecular basis for pathogen susceptibility, we searched for three major classes of resistance genes in the watermelon genome, namely the nucleotide-binding site and leucine-rich repeat (NBS-LRR), lipoxygenase (LOX)[32] and receptor-like gene families[33]. We identified a total of 44 NBS-LRR genes, including 18 Toll interleukin receptor (TIR)-NBS-LRR– and 26 coiled-coil (CC)-NBS-LRR–encoding genes (**Supplementary Table 20**). The watermelon NBS-LRR genes evolved independently, and we detected no sequence exchanges between different homologs. Such evolutionary patterns are similar to those of type II R genes in lettuce and *Arabidopsis*[34], indicating that watermelon has low diversity of NBS-LRR genes. The number of NBS-LRR genes in the watermelon genome is similar to that in cucumber[14] and papaya[35] but is considerably fewer than that in maize[36], rice[37] and apple[12]. In contrast, the LOX gene family has undergone an expansion in the watermelon genome with 26 members, 19 of which are arranged in two tandem gene arrays (**Supplementary Fig. 11**). Similar findings have been reported in cucumber, with expansion of the LOX gene family having been considered as a possible complementary mechanism to cope with pathogen invasion[14]. We further identified 197 receptor-like genes in the watermelon genome, among which 35 encode receptor-like proteins lacking a kinase domain and 162 encode receptor-like kinases that have an intracellular kinase domain in addition to the extracellular LRR and transmembrane domains (**Supplementary Table 20**). Many of these resistance genes are located on chromosomes in clusters (**Supplementary Fig. 11**), suggesting tandem duplications as their evolutionary basis.

It has been speculated that the lack of resistance to a wide range of diseases in modern watermelon cultivars is the result of the many years of cultivation and selection that have focused on desirable fruit qualities at the expense of disease resistance[8,38]. To test this notion, we performed *de novo* assemblies of unmapped reads pooled each from modern cultivated (*C. lanatus* subsp. *vulgaris*) and semiwild and wild (*C. lanatus* subsp. *mucosospermus* and *C. lanatus* subsp. *lanatus*, respectively) accessions. We identified 11 and 69 genes from the cultivated and the semiwild and wild groups, respectively, that are homologous to known plant proteins (**Supplementary Table 21**). It is worth mentioning here that the 69 new genes identified from the semiwild and wild group were highly enriched with disease-related genes including, 6 TIR-LRR-NBS genes, 1 PR-1 gene and 3 lipoxygenase genes, whereas none of the 11 genes identified in the cultivated group were disease related. In addition, all of the 44 NBS-LRR genes identified in the 97103 genome were also present in the semiwild and

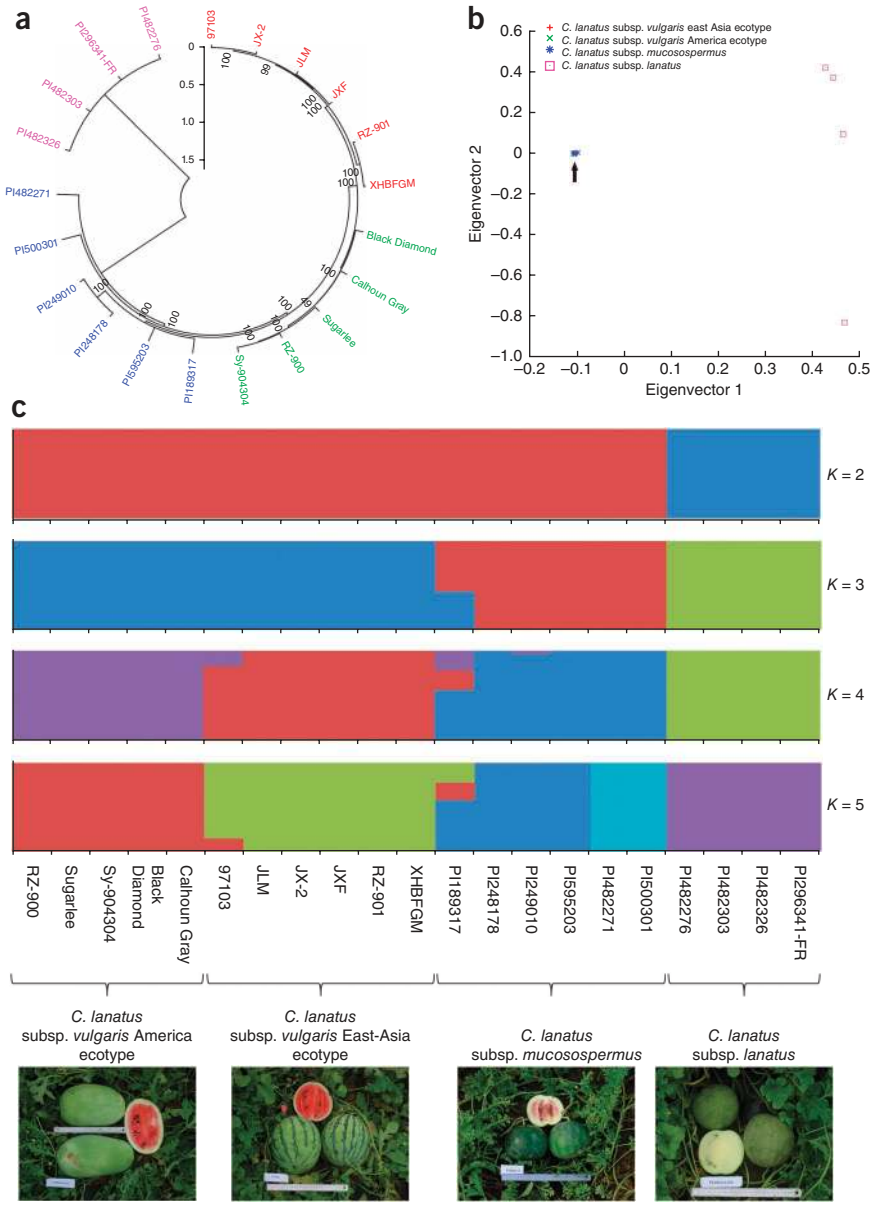

Figure 3 Population structure of watermelon accessions. (**a**) Neighbor-joining phylogenetic tree of watermelon accessions on the basis of SNPs. (**b**) PCA analysis of watermelon accessions using SNPs as markers. *C. lanatus* subsp. *vulgaris* east Asia and America ecotypes and *C. lanatus* subsp. *mucosospermus* accessions cluster together (arrow) and are almost indistinguishable. (**c**) Population structure of watermelon accessions. Each color represents one population, each accession is represented by a vertical bar, and the length of each colored segment in each vertical bar represents the proportion contributed by ancestral populations. Shown are representative watermelon pictures from each subspecies or ecotype.

wild accessions (**Supplementary Note**). These findings support the hypothesis that a large portion of disease resistance genes has been lost during watermelon domestication.

Analysis of cucurbit phloem sap and vascular transcriptomes

The angiosperm enucleate sieve tube system contains mRNA, some of which has been shown to function as a long-distance signaling agent[39,40]. Through deep transcriptome sequencing (**Supplementary Table 22**), we identified 13,775 and 14,242 mRNA species in watermelon and cucumber vascular bundles, respectively, and 1,519 and 1,012 transcripts in the watermelon and cucumber phloem sap, respectively (**Supplementary Tables 23–26**). Notably, we found that the gene sets in the vascular bundles between the two cucurbit species were almost identical, whereas only 50–60% of the transcripts detected in the phloem sap were common between the two species (**Supplementary Note** and **Supplementary Table 27**). Gene Ontology (GO) term enrichment analysis indicated that the major categories among the common phloem transcripts were response to stress or stimulus (**Supplementary Table 28**), which is fully consistent with the central role of the plant vascular system, and the phloem in particular, in the long-distance communication system that integrates abiotic and biotic stress signaling at the whole-plant level[41]. In contrast, analysis of the phloem transcripts that are unique to watermelon identified macromolecular biosynthesis process and protein metabolic process as the major GO categories (**Supplementary Table 29**). The unique phloem sap transcripts may reflect specialized functions that are unique to the role of the phloem in these species. It is noteworthy that the watermelon phloem contained 118 transcription factors, whereas we identified only 46 transcription factors in cucumber and 32 transcription factors that were common to both (**Supplementary Tables 30–32**).

Pumpkin (*Cucurbita maxima*) has been used as a model system for phloem studies[42,43]. We developed pumpkin vascular bundle and phloem sap transcript catalogs through generation and *de novo* assembly of the Illumina paired-end RNA sequencing (RNA-Seq) reads. Comparative analysis of the watermelon, cucumber and pumpkin phloem transcriptomes indicated that approximately 36% of their transcripts were in common (**Supplementary Fig. 12**). These conserved transcripts probably carry out functions that are central to the operation of the sieve tube system in most cucurbit and possibly additional species.

Regulation of watermelon fruit development and quality

Watermelon fruit development is a complex process involving major changes in size, color, texture, sugar content and nutritional components. To obtain a comprehensive characterization of the genes involved in the development and quality of watermelon fruit, we performed strand-specific RNA-Seq[44] of both the flesh and rind at four crucial stages of fruit development in the inbred line 97103 (**Supplementary Table 33**). We identified 3,046 and 558 genes that were differentially expressed in the flesh and rind, respectively, during fruit development and 5,352 genes that were differentially expressed between the flesh and rind in at least one of the four stages (**Supplementary Tables 34–36**). GO term enrichment analysis indicated that during fruit development in both the flesh and rind, biological processes such as cell-wall biogenesis, flavonoid metabolism and

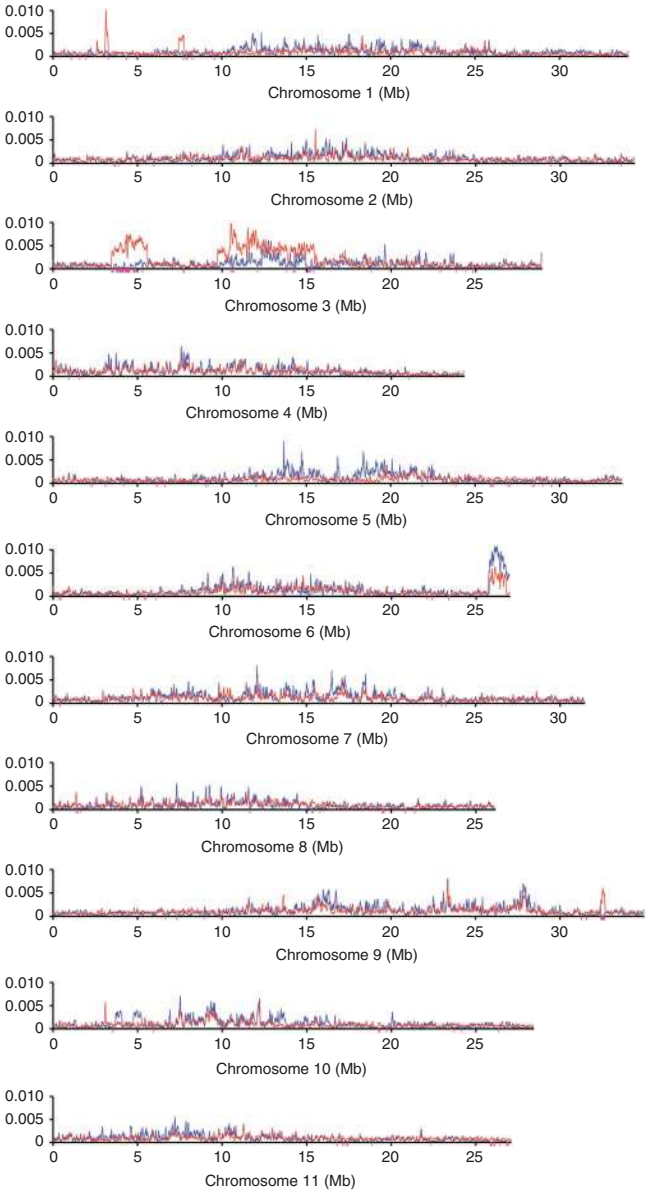

Figure 4 Diversity (π) distribution for *C. lanatus* subsp. *vulgaris* (blue) and *C. lanatus* subsp. *mucosospermus* (red) across the 11 watermelon chromosomes. Pink bars below the *x* axis correspond to regions with 1% significance level of diversity difference ($\pi_{mucosospermus}/\pi_{vulgaris}$).

defense responses were significantly altered (false discovery rates (FDR) < 0.01), whereas carotenoid, hexose and monosaccharide metabolic processes were only significantly altered in the flesh, supporting major physiological differences, including sugar content and fruit color, between the flesh and rind (**Supplementary Table 37**).

Sugar content is a key factor in determining watermelon fruit quality. The sweetness of a watermelon is determined by both the total sugar content and the ratios among the major accumulated sugars: glucose, fructose and sucrose[45]. In young 97103 fruit flesh, fructose and glucose are the predominant sugars, whereas in mature 97103 fruit flesh, both sucrose and total sugar content are substantially increased, with sucrose then becoming the dominant sugar; in the rind, the sugar content remains relatively lows (**Supplementary Table 38**). Final sugar accumulation in watermelon fruit is determined by sugar unloading from the phloem followed by uptake and metabolism within the fruit flesh. The annotated watermelon genome contains a total of 62 sugar metabolic enzyme genes and 76 sugar transporter genes, among which 13 sugar metabolic genes and 14 sugar transporter genes were differentially expressed during flesh development and between the flesh and rind tissues (**Supplementary Tables 39** and **40**). On the basis of these results and prior published work from other plant species[46,47], we propose a model for sugar metabolism in the cells of watermelon fruit flesh (**Supplementary Fig. 13**). Specifically, during watermelon flesh development, α-galactosidase, insoluble acid invertase, neutral invertase, sucrose phosphate synthase, UDP-glucose 4-epimerase, soluble acid invertase and UDP-galactose/glucose pyrophosphorylase function as key enzymes involved in regulating sugar unloading and metabolism. Furthermore, the 14 differentially expressed sugar transporters are probably responsible for sugar partitioning (**Supplementary Note**).

Transcription factors also have a role in sugar accumulation[48]. Of the 1,448 putative transcription factor genes identified in the watermelon genome, 193 showed significant expression changes (FDR < 0.01) during flesh development and also in flesh compared to rind at later stages, including transcription factors from families known to be involved in the regulation of sugar accumulation (**Supplementary Note** and **Supplementary Tables 41** and **42**). It is noteworthy that one bZIP gene, *Cla014572*, is downregulated during flesh development and contains the sucrose-controlled upstream open reading frame (SC-uORF) (**Supplementary Note** and **Supplementary Fig. 14**). It was recently reported that transgenic plants constitutively expressing the tobacco SC-uORF containing the bZIP gene *tbz17* but lacking its SC-uORF had increased sugar concentrations[49]. Therefore, our analysis is consistent with a role for *Cla014572* as a key regulator of sugar accumulation during fruit development.

MADS-box genes, such as *MADS-RIN* (also known as *LeMADS-RIN*)[50] and *TAGL1* (ref. 51) in tomato, have been reported to regulate the fruit expansion and ripening processes. Phylogenetic analysis of watermelon, cucumber and *Arabidopsis* MADS-box transcription factors, together with *MADS-RIN* and *TAGL1*, identified two MADS-box transcription factors from watermelon in each of the *RIN* and *AGL1* clades (**Supplementary Note** and **Supplementary Fig. 15**). These four genes (*Cla000691* and *Cla010815* in the *RIN* clade and *Cla009725* and *Cla019630* in the *AGL1* clade) are among the most highly expressed MADS-box transcription factors during fruit development (**Supplementary Table 43**). Notably, unlike *MADS-RIN*, which is highly expressed only in ripening fruits, both *Cla000691* and *Cla010815* are highly expressed throughout fruit development, indicating they could have evolved to participate in other functions in addition to ripening. It is noteworthy in this regard that close banana and strawberry homologs of *MADS-RIN* also show expression and/or functional activities that extend beyond the ripening fruit[52,53]. The expression profiles of *Cla009725* and *Cla019630* during fruit development are similar to that of *TAGL1*, which consistent with their potential roles in regulating fruit expansion and ripening[51].

Citrulline is a nonessential amino acid produced from glutamine and has various benefits to health and athletic performance. Its name is derived from *citrullus*, the Latin word for watermelon, from which it was first isolated[54]. Watermelon flesh and rind serve as a natural source of citrulline, and its abundance increases substantially during fruit maturation but then declines as the fruit becomes over-ripe (**Supplementary Fig. 16**). On the basis of our annotation of the watermelon genome, we identified 14 genes in the citrulline metabolic pathway (**Supplementary Fig. 17**). Compared to the *Arabidopsis* citrulline metabolic pathway, this pathway in watermelon has undergone expansion in the arginosuccinase and arginosuccinate synthase

families. Both are involved in converting citrulline to L-arginine. We found an arginosuccinase and two arginosuccinate synthase genes to be highly downregulated during watermelon flesh development (**Supplementary Table 44**). Thus, citrulline accumulation in the maturing fruit flesh is probably a result of decreased activities of citrulline degradation.

DISCUSSION

The draft watermelon genome sequence presented here represents an important resource for plant research and crop genetic improvement and also supports further evolutionary and comparative genomics studies of the *Cucurbitaceae*. The evolutionary scenario outlined for the cucurbits provides a clearly described series of genetic phenomena underlying a modern plant genome and yields new insights into the events that underlie the transition from ancestral chromosomes to modern chromosome architecture. Genome resequencing of representative watermelon accessions has provided a large source of haplotype data with great potential for genome manipulation, trait discovery and allele mining. Insights regarding the genetic diversity and population structure of watermelon accessions, as well as chromosome regions and genes under human selection, will shape future efforts in watermelon genetic research and breeding. The unique metabolic and regulatory networks in developing watermelon fruit identified from our functional genomics study represent an initial genomics-enabled milestone for the understanding and genetic improvement of crucial nutritional attributes, including sugar and amino acid contents. In addition, genomic resources that are available for both watermelon and cucumber greatly enhance the capacity to investigate at the whole-plant level the phloem-based vascular signaling systems that function to integrate developmental and physiological processes.

URLs. Watermelon genome database, http://www.icugi.org/ and http://www.iwgi.org/; FAO Statistics database, http://faostat.fao.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. This whole-genome shotgun sequencing project has been deposited at DDBJ/EMBL/GenBank under the accession AGCB00000000. Sequence reads of genome resequencing and transcriptome sequencing have been deposited into the NCBI sequence read archive (SRA) under accessions SRA052158, SRA052198 and SRA052519. The four completely sequenced BACs have been deposited into GenBank under accessions JN402338, JN402339, JX027061 and JX027062.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to E. Legg of Syngenta and J. de Wit and Z. Sun of Rijk Zwaan for management support of this project. This research was supported by grants from the Ministry of Science and Technology of the People's Republic of China (2010DFB33740, 2012AA020103, 2012AA100101, 2012AA100103 and 2012AA100105), the Ministry of Agriculture of the People's Republic of China (CARS-26), the Major Program of Beijing Natural Science Foundation of China (5100001), the Beijing Municipal Science and Technology Commission of China (D111100001311002) to Yong Xu, the National Natural Science Foundation of China (30972015 and 31171980) to H. Zhang, the National Natural Science Foundation of China (31272184) to S. Guo, the Agence Nationale de la Recherche (Program ANRjc-PaleoCereal, ANR-09-JCJC-0058-01) to J.S., the USDA National Institute of Food and Agriculture (NIFA 201015479) to W.J.L., the US National Science Foundation (IOS-0923312 and IOS-1025642) and US-Israel Binational Agricultural Research and Development Fund (IS-4223-09C) to Z.F. and the USDA Agricultural Research Service.

AUTHOR CONTRIBUTIONS

Yong Xu, Z.F., Y.L., Jun Wang, M.W., X. Zhang, S.H., W.J.L. and S. Guo designed and managed the project. J.Z., Z.W., R.L., Junyi Wang, Y.Y., P.N., X.G., X. Zhao, Z. Zheng, B.W., Juan Wang, K.W., B.H., X.L., T.L., M.X., J.M., A.B., L.A.M., K.K., H. Schoof, Y.R., H. Zhang, H. Zhao, Y. Zhang, Q.K., W.H., X. Zou, J.J., A.L. and Zhonghua Zhang conducted sequencing, assembly and annotation. Z.W., J.S., Y.H., M.X. and F.M. performed evolutionary analyses. H. Sun, L.M., Z.F., S. Guo. X.H., S.D. and D.L. were involved in genome resequencing analysis. S. Guo, Yong Xu, Y. Zheng, S. Gao, Z.F., W.J.L., B.-K.H., Zhaoliang Zhang, Yimin Xu, S.Z. and J.J.G. performed RNA-Seq analysis. S. Guo, Y.R., J.L., H. Sun, H.H., H.Y. and G.G. performed fruit quality trait measurement and analyses. H.K., Q.H. and Z.F. performed analysis of disease-resistance genes. Z.F., M.H., Y. Zheng, Y. Xia, K.L., T.T. and E.P. performed database development. Yong Xu, Z.F., S. Guo, W.J.L., J.S., J.J.G., Y.R., H. Sun, Z.W., J.M., H.K. and A.L. wrote and/or revised the paper. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2470.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Erickson, D.L., Smith, B.D., Clarke, A.C., Sandweiss, D.H. & Tuross, N. An Asian origin for a 10,000-year-old domesticated plant in the Americas. *Proc. Natl. Acad. Sci. USA* **102**, 18315–18320 (2005).
2. Fursa, T.B. On the taxonomy of the genus *Citrullus* Schad. *Bot. Zh.* **57**, 31–34 (1972).
3. Jeffrey, C. Cucurbitaceae. in *Mansfeld's Encyclopedia of Agricultural and Horticultural Crops* (ed. Hanelt, P.) Vol. 3, 1510–1557 (Springer, 2001).
4. Hayashi, T. *et al.* L-citrulline and L-arginine supplementation retards the progression of high-cholesterol-diet–induced atherosclerosis in rabbits. *Proc. Natl. Acad. Sci. USA* **102**, 13681–13686 (2005).
5. Collins, J.K. *et al.* Watermelon consumption increases plasma arginine concentrations in adults. *Nutrition* **23**, 261–266 (2007).
6. Perkins-Veazie, P., Collins, J.K., Davis, A.R. & Roberts, W. Carotenoid content of 50 watermelon cultivars. *J. Agric. Food Chem.* **54**, 2593–2597 (2006).
7. Schaefer, H. & Renner, S.S. Phylogenetic relationships in the order Cucurbitales and a new classification of the gourd family (*Cucurbitaceae*). *Taxon* **60**, 122–138 (2011).
8. Levi, A., Thomas, C.E., Wehner, T.C. & Zhang, X. Low genetic diversity indicated the need to broaden the genetic base of cultivated watermelon. *HortScience* **36**, 1096–1101 (2001).
9. Arumuganathan, K. & Earle, E.D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
10. Ren, Y. *et al.* A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS ONE* **7**, e29453 (2012).
11. Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011).
12. Velasco, R. *et al.* The genome of the domesticated apple (*Malus* x *domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
13. Al-Dous, E.K. *et al. De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* **29**, 521–527 (2011).
14. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
15. Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109–116 (2011).
16. Wang, X. *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
17. Varshney, R.K. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89 (2012).
18. Dassanayake, M. *et al.* The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* **43**, 913–918 (2011).
19. Li, Z. *et al.* RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* **12**, 540 (2011).
20. Dane, F. & Liu, J. Diversity and origin of cultivated and citron type watermelon (*Citrullus lanatus*). *Genet. Resour. Crop Evol.* **54**, 1255–1265 (2007).
21. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
22. Deleu, W. *et al.* A set of EST-SNPs for map saturation and cultivar identification in melon. *BMC Plant Biol.* **9**, 90 (2009).

23. Salse, J. *In silico* archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.* **15**, 122–130 (2012).
24. Abrouk, M. *et al.* Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479–487 (2010).
25. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
26. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030 (2010).
27. Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
28. Xu, X. *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2012).
29. Tang, H., Peng, J., Wang, P. & Risch, N.J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
30. Harushima, Y. *et al.* Diverse variation of reproductive barriers in three intraspecific rice crosses. *Genetics* **160**, 313–322 (2002).
31. Nasrallah, J.B. & Nasrallah, M.E. Pollen-stigma signaling in the sporophytic self-incompatibility response. *Plant Cell* **5**, 1325–1335 (1993).
32. Feussner, I. & Wasternack, C. The lipoxygenase pathway. *Annu. Rev. Plant Biol.* **53**, 275–297 (2002).
33. Fritz-Laylin, L.K., Krishnamurthy, N., Tor, M., Sjolander, K.V. & Jones, J.D. Phylogenomic analysis of the receptor-like proteins of rice and *Arabidopsis*. *Plant Physiol.* **138**, 611–623 (2005).
34. Kuang, H. *et al.* Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell* **16**, 2870–2894 (2004).
35. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
36. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
37. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **296**, 79–92 (2002).
38. Harris, K.R., Wechter, W.P. & Levi, A. Isolation, sequence analysis, and linkage mapping of NBS-LRR disease resistance gene analogs in Watermelon. *J. Am. Soc. Hortic. Sci.* **134**, 649–657 (2009).
39. Kim, M., Canio, W., Kessler, S. & Sinha, N. Developmental changes due to long-distance movement of a homeobox fusion transcript in tomato. *Science* **293**, 287–289 (2001).
40. Kehr, J. & Buhtz, A. Long distance transport and movement of RNA through the phloem. *J. Exp. Bot.* **59**, 85–92 (2008).
41. Lough, T.J. & Lucas, W.J. Integrative plant biology: role of phloem long-distance macromolecular trafficking. *Annu. Rev. Plant Biol.* **57**, 203–232 (2006).
42. Yoo, B.C. *et al.* A systemic small RNA signaling system in plants. *Plant Cell* **16**, 1979–2000 (2004).
43. Ham, B.K. *et al.* A polypyrimidine tract binding protein, pumpkin RBP50, forms the basis of a phloem-mobile ribonucleoprotein complex. *Plant Cell* **21**, 197–215 (2009).
44. Zhong, S. *et al.* High-throughput Illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* **8**, 940–949 (2011).
45. Brown, A.C. & Summers, W.L. Carbohydrate accumulation and color development in watermelon. *J. Am. Soc. Hortic. Sci.* **110**, 683–686 (1985).
46. Slewinski, T.L. Diverse functional roles of monosaccharide transporters and their homologs in vascular plants: a physiological perspective. *Mol. Plant* **4**, 641–662 (2011).
47. Dai, N. *et al.* Metabolism of soluble sugars in developing melon fruit: a global transcriptional view of the metabolic transition to sucrose accumulation. *Plant Mol. Biol.* **76**, 1–18 (2011).
48. Slewinski, T.L. & Braun, D.M. Current perspectives on the regulation of whole-plant carbohydrate partitioning. *Plant Sci.* **178**, 341–349 (2010).
49. Thalor, S.K. *et al.* Deregulation of sucrose-controlled translation of a bZIP-type transcription factor results in sucrose accumulation in leaves. *PLoS ONE* **7**, e33111 (2012).
50. Vrebalov, J. *et al.* A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (*rin*) locus. *Science* **296**, 343–346 (2002).
51. Vrebalov, J. *et al.* Fleshy fruit expansion and ripening are regulated by the tomato *SHATTERPROOF* gene *TAGL1*. *Plant Cell* **21**, 3041–3062 (2009).
52. Elitzur, T., Vrebalov, J., Giovannoni, J.J., Goldschmidt, E.E. & Friedman, H. The regulation of MADS-box gene expression during ripening of banana and their regulatory interaction with ethylene. *J. Exp. Bot.* **61**, 1523–1535 (2010).
53. Seymour, G.B. *et al.* A *SEPALLATA* gene is involved in the development and ripening of strawberry (*Fragaria* × *ananassa* Duch.) fruit, a non-climacteric tissue. *J. Exp. Bot.* **62**, 1179–1188 (2011).
54. Wada, M. Über citrullin, eine neue aminosäure im presssaft der wassermelone, *Citrullus vulgaris* Schrad. *Biochem. Z.* **224**, 420 (1930).

ONLINE METHODS

Sequencing and assembly of the 97103 genome. Paired-end and mate-pair Illumina libraries were prepared following the manufacturer's instructions. Additional steps, including DNA circularization, digestion of linear DNA, fragmentation of circularized DNA and purification of biotinylated DNA, were performed before adaptor ligation for mate-pair libraries with insert size ≥2 kb. All libraries were sequenced on the Illumina GAII system using standard Illumina protocols. Raw Illumina reads were first processed by removing low-quality reads, adaptor sequences and possible contaminated reads of bacterial and viral origin. The clean reads were then assembled using SOAP*denovo*[55] (**Supplementary Note**).

Transposable element annotation. Repeat sequences were first identified *de novo* from the genome assembly using PILER[56] and RepeatScout[57]. LTR retrotransposons were identified by LTR_FINDER[58] with default parameters. We then used RepeatMasker (http://www.repeatmasker.org/) and the known repbase library (http://www.girinst.org/repbase/index.html) to find transposable element repeats in the assembled genome. Transposable elements were then classified as previously described[14] (**Supplementary Note**).

Gene prediction and functional annotation. The repeat-masked watermelon genome was used for gene predictions. AUGUSTUS[59] and GlimmerHMM[60] were used for *ab initio* gene prediction. We also aligned watermelon EST and complementary DNA (cDNA) sequences onto the genome assembly using BLAT (identity ≥0.98 and coverage ≥0.95) to derive spliced alignments. Protein sequences of six plants (*Arabidopsis*, cucumber, poplar, rice, papaya and grape) were then aligned onto the watermelon genome using TBLASTN at an *E* value cutoff of 1×10^{-5}, and the homologous genome sequences were then aligned against the matching proteins using GeneWise[61] for accurate spliced alignments. Outputs of the three methods described above (*ab initio*, cDNA and EST mapping and homology mapping) were integrated using GLEAN[62] to produce the consensus gene models. Functions of the predicted watermelon genes were assigned using AHRD (Automated assignment of Human Readable Descriptions) as described previously[63] (**Supplementary Note**).

FISH analysis. FISH analysis of unassembled repeat units similar to centromeres, telomeres and rDNAs in the 97103 genome and 45S and 5S rDNAs in the genomes of the 20 watermelon accessions were performed according to Ren *et al.*[10].

Analysis of cucurbit genome evolution. A method for the identification of orthologous regions between plant genomes, on the basis of integrative sequence alignment criteria combined with a statistical validation[64], was used to unravel the *Cucurbitaceae* evolutionary paleohistory using watermelon genome information and results from a previous paleogenomics analysis[65]. The duplication event in the watermelon genome was dated following the method of Murat *et al.*[66]. The syntenic relationships between watermelon, cucumber, melon and grape were analyzed using alignment parameters and statistical tests following the method of Salse *et al.*[64] (**Supplementary Note**).

Genome resequencing, read mapping and SNP and small indel calling. Paired-end Illumina libraries were prepared following the manufacturer's instructions and sequenced on an Illumina GAII system. We sequenced 44, 75 or 90 bp at each end. The Illumina paired-end reads from each watermelon accession were aligned to the reference 97103 genome sequences using BWA[67]. For shorter reads (44 bp), we used the paired-end mapping mode of BWA and only kept reads with mapping quality >16. For longer reads (75 or 90 bp), we used the single-end mapping mode of BWA and only kept reads that were uniquely mapped to the reference genome. After mapping, SNPs and small indels (1–5 bp) were identified on the basis of the mpileup files generated by SAMtools[68]. Genotypes supported by at least two reads and with allele frequency ≥0.3 were assigned to each genomic position. Only homozygous SNPs and small indels were accepted.

Identification of new genes. Unmapped reads from the ten modern cultivated and ten semiwild and wild watermelon accessions were each pooled. The two pools of unmapped reads were assembled separately into contigs using SOAP*denovo*[55]. Contigs shorter than 2 kb were discarded. New genes were predicted from the assembled contigs using AUGUSTUS[59] and then compared against the NCBI nr database using BLASTP.

Nucleotide diversity and selective sweep detection. Two standard estimates of the scaled mutation rate, θw, the proportion of segregating sites, and π, the average pairwise nucleotide diversity[25], were used to characterize nucleotide diversity among the examined watermelon populations. A sliding window approach was used to calculate the θw and π along all 11 watermelon chromosomes with a window size of 50 kb and a step size of 10 kb. To identify potential selective sweeps, we compared nucleotide diversity between populations of modern cultivars (*C. lanatus* subsp. *vulgaris*) and semiwild accessions (*C. lanatus* subsp. *mucosospermus*). Genome regions with highest (top 1%) genetic diversity ($\pi_{mucosospermus}/\pi_{vulgaris}$) were identified as potential selective sweeps. We also identified regions with smallest (top 1%) $\pi_{mucosospermus}/\pi_{vulgaris}$ values to serve as negative controls. GO terms enriched in genes from the selective sweeps were identified with GO::TermFinder[69].

Phylogenetic relationship and population structure analyses. In these analyses, we used a subset of ~1.46 million SNPs that had information in all 21 watermelon accessions. A neighbor-joining tree was constructed using PHYLIP (http://evolution.genetics.washington.edu/phylip.html) and MEGA5 (ref. 70) was used to display the tree. PCA was performed with EIGENSOFT[71]. We further used FRAPPE[29] to investigate the population structure with 10,000 iterations and the number of clusters (K) of 2–5.

Comparative transcriptome analysis of watermelon, cucumber and pumpkin phloem sap and vascular tissues. Phloem sap and vascular tissues were collected from the main stem located in the central region of 6-week-old watermelon (cv. 97103), cucumber (cv. Chinese Long) and pumpkin (cv. Big max) plants. Paired-end and single-end strand-specific RNA-Seq libraries were prepared from these tissues according to Zhong *et al.*[44] and sequenced on the Illumina HiSeq 2000 system. Three independent biological replica samples were prepared. RNA-Seq reads were first aligned to rRNA sequences using Bowtie[72] to eliminate possible rRNA sequence contamination. The resulting watermelon and cucumber reads were aligned to the corresponding genomes using TopHat[73]. Pumpkin RNA-Seq reads were *de novo* assembled into contigs using Trinity[74], and then reads were aligned back to the assembled contigs using Bowtie[72]. The count of mapped reads for each gene from each sample was then derived and normalized to fragments per kilobase of exon model per million mapped reads (FPKM). We defined vascular transcripts as those with FPKM ≥ 2 in all three biological replica samples and phloem sap transcripts as those with FPKM ≥ 2 in all three biological replica samples and enriched with at least twofold higher levels in phloem sap compared to vascular bundle (**Supplementary Note**).

Fruit transcriptome sequencing and analysis. Fruit flesh and rind tissues of 97103 were collected at four developmental stages: 10, 18, 26 and 34 days after pollination[75]. RNA extraction and strand-specific RNA-Seq library preparations were performed as described[44], and RNA-Seq libraries were sequenced on the Illumina HiSeq 2000 system. Two independent biological replica samples were prepared. RNA-Seq reads were first processed to remove rRNA sequence contamination. The resulting reads were aligned to the watermelon 97103 genome sequences using TopHat[73]. After alignment, for each watermelon gene model, the count of mapped reads from each sample was derived and normalized to FPKM. Differentially expressed genes during flesh or rind development and between the flesh and rind at the same stages were identified using the LIMMA[76] and DESeq[77] packages, respectively. Raw P values of multiple tests were corrected using the FDR[78].

55. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
56. Edgar, R.C. & Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (suppl. 1), i152–158 (2005).
57. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–358 (2005).

58. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–268 (2007).
59. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** (suppl. 2), ii215–225 (2003).
60. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
61. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
62. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
63. The Tomato Genome Sequencing Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
64. Salse, J. *et al.* Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.* **10**, 619–630 (2009).
65. Salse, J. *et al.* Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. USA* **106**, 14908–14913 (2009).
66. Murat, F. *et al.* Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545–1557 (2010).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
68. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
69. Boyle, E.I. *et al.* GO:TermFinder: open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).
70. Tamura, K. *et al.* MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
71. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
72. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
73. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
74. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
75. Guo, S. *et al.* Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. *BMC Genomics* **12**, 454 (2011).
76. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Molec. Biol.* **3**, Article 3 (2004).
77. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
78. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

doi:10.1038/ng.2470