nature genetics

ARTICLES

The genome of the cucumber, *Cucumis sativus* L.

Sanwen Huang[1,19], Ruiqiang Li[2,3,19], Zhonghua Zhang[1,19], Li Li[2,19], Xingfang Gu[1,19], Wei Fan[2,19], William J Lucas[4,19], Xiaowu Wang[1], Bingyan Xie[1], Peixiang Ni[2], Yuanyuan Ren[2], Hongmei Zhu[2], Jun Li[2], Kui Lin[5], Weiwei Jin[6], Zhangjun Fei[7], Guangcun Li[8], Jack Staub[9], Andrzej Kilian[10], Edwin A G van der Vossen[11], Yang Wu[5], Jie Guo[5], Jun He[1], Zhiqi Jia[1], Yi Ren[1], Geng Tian[2], Yao Lu[2], Jue Ruan[2,12], Wubin Qian[2], Mingwei Wang[2], Quanfei Huang[2], Bo Li[2], Zhaoling Xuan[2], Jianjun Cao[2], Asan[2], Zhigang Wu[2], Juanbin Zhang[2], Qingle Cai[2], Yinqi Bai[2], Bowen Zhao[13], Yonghua Han[6], Ying Li[1], Xuefeng Li[1], Shenhao Wang[1], Qiuxiang Shi[1], Shiqiang Liu[1], Won Kyong Cho[14], Jae-Yean Kim[14], Yong Xu[15], Katarzyna Heller-Uszynska[10], Han Miao[1], Zhouchao Cheng[1], Shengping Zhang[1], Jian Wu[1], Yuhong Yang[1], Houxiang Kang[1], Man Li[1], Huiqing Liang[2], Xiaoli Ren[2], Zhongbin Shi[2], Ming Wen[2], Min Jian[2], Hailong Yang[2], Guojie Zhang[2,12], Zhentao Yang[2], Rui Chen[2], Shifang Liu[2], Jianwen Li[2], Lijia Ma[2,12], Hui Liu[2], Yan Zhou[2], Jing Zhao[2], Xiaodong Fang[2], Guoqing Li[2], Lin Fang[2], Yingrui Li[2,12], Dongyuan Liu[2], Hongkun Zheng[2,3], Yong Zhang[2], Nan Qin[2], Zhuo Li[2], Guohua Yang[2], Shuang Yang[2], Lars Bolund[2,16], Karsten Kristiansen[17], Hancheng Zheng[2,18], Shaochuan Li[2,18], Xiuqing Zhang[2], Huanming Yang[2], Jian Wang[2], Rifei Sun[1], Baoxi Zhang[1], Shuzhi Jiang[1], Jun Wang[2,17], Yongchen Du[1] & Songgang Li[2]

Cucumber is an economically important crop as well as a model system for sex determination studies and plant vascular biology. Here we report the draft genome sequence of *Cucumis sativus* var. *sativus* L., assembled using a novel combination of traditional Sanger and next-generation Illumina GA sequencing technologies to obtain 72.2-fold genome coverage. The absence of recent whole-genome duplication, along with the presence of few tandem duplications, explains the small number of genes in the cucumber. Our study establishes that five of the cucumber's seven chromosomes arose from fusions of ten ancestral chromosomes after divergence from *Cucumis melo*. The sequenced cucumber genome affords insight into traits such as its sex expression, disease resistance, biosynthesis of cucurbitacin and 'fresh green' odor. We also identify 686 gene clusters related to phloem function. The cucumber genome provides a valuable resource for developing elite cultivars and for studying the evolution and function of the plant vascular system.

The botanical family Cucurbitaceae, commonly known as cucurbits and gourds, includes several economically important cultivated plants, such as cucumber (*C. sativus* L.), melon (*C. melo* L.), watermelon (*Citrullus lanatus* (Thunb.) Matsum. & Nakai) and squash and pumpkin (*Cucurbita* spp.). Agricultural production of cucurbits uses 9 million hectares of land and yields 184 million tons of vegetables, fruits and seeds annually (http://faostat.fao.org). The cucurbit family also displays a rich diversity of sex expression, and the cucumber has served as a primary model system for sex determination studies[1]. The cucurbits are also model plants for the study of vascular biology, as both xylem and phloem sap can be readily collected for studies of long-distance signaling events[2,3].

Despite the agricultural and biological importance of cucurbits, knowledge of their genetics and genome is currently very limited. We have therefore sequenced and assembled the genome of the domestic cucumber, *C. sativus* var. *sativus* L.

All previous plant genome sequences have been derived using traditional Sanger technology[4–9]. The recent development of

[1]Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. [2]BGI-Shenzhen, Shenzhen, China. [3]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark. [4]Department of Plant Biology, College of Biological Sciences, University of California, Davis, California, USA. [5]College of Life Sciences, Beijing Normal University, Beijing, China. [6]National Maize Improvement Center of China, Key Laboratory of Crop Genetic Improvement and Genome of Ministry of Agriculture, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China. [7]Boyce Thompson Institute and USDA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York, USA. [8]High-Tech Research Center, Shandong Academy of Agricultural Sciences, Jinan, China. [9]US Department of Agriculture, Agricultural Research Service, Vegetable Crops Research Unit, Department of Horticulture, University of Wisconsin, Madison, Wisconsin, USA. [10]Diversity Arrays Technology, Canberra, Australia. [11]Wageningen UR Plant Breeding, Wageningen, The Netherlands. [12]The Graduate University of Chinese Academy of Sciences, Beijing, China. [13]High School Affiliated to Renmin University of China, Beijing, China. [14]Division of Applied Life Science (BK21 and WCU program), PMBBRC and EB-NCRC, Gyeongsang National University, Jinju, Republic of Korea. [15]National Engineering Research Center for Vegetables, Beijing, China. [16]Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. [17]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [18]South China University of Technology, Guangzhou, China. [19]These authors contributed equally to this work. Correspondence should be addressed to Y.D. (yongchen.du@mail.caas.net.cn), S.H. (huangsanwen@caas.net.cn), Jun Wang (wangj@genomics.org.cn) or Songgang Li (lisg@genomics.org.cn).

Received 6 May; accepted 28 September; published online 1 November 2009; doi:10.1038/ng.475

next-generation sequencing technologies has significantly improved sequencing throughput at a markedly reduced cost[10]. However, an intrinsic characteristic of next-generation technologies is their short read length (~50 bp), which prevents their direct application for *de novo* assembly of large genomes. When using these new technologies, assembly is typically carried out by mapping these short reads onto a known reference genome[11,12]. For the cucumber genome, we carried out a novel combination *de novo* sequencing strategy, taking advantage of the long read and clone length of Sanger technology and, for the first time, the high sequencing depth and low unit cost of Illumina GA technology.

Table 1 Cucumber genome assembly statistics

Assembly	Contig N50[a] (kb)	Contig total (Mb)	Scaffold N50 (kb)	Scaffold total (Mb)	% sequence anchored on chromosome
Sanger	2.6	204	19	238	—
Illumina GA	12.5	190	172	200	—
Sanger + Illumina GA	19.8	226.5	1,140	243.5	72.8%

[a]N50 refers to the size above which half of the total length of the sequence set can be found.

RESULTS

Sequencing and assembly

We selected the 'Chinese long' inbred line 9930, which is commonly used in modern cucumber breeding[13], for our genome sequencing project. We generated a total of 26.5 billion high-quality base pairs, or 72.2-fold genome coverage, of which the Sanger reads provided 3.9-fold coverage and the Illumina GA reads provided 68.3-fold coverage (**Supplementary Table 1**). The GA reads ranged in length from 42 to 53 bp.

We compared the assemblies obtained by Sanger reads only, Illumina GA reads only and Sanger plus Illumina reads. The 'hybrid' approach achieved markedly longer N50 (the size above which half of the total length of the sequence set can be found) in both contigs and scaffolds, so we used this assembly for further analyses (**Table 1** and **Supplementary Table 2**). The total length of the assembled genome was 243.5 Mb, about 30% smaller than the genome size estimated by flow cytometry of isolated nuclei stained with propidium iodide (367 Mb)[14] and by *K*-mer depth distribution of sequenced reads (350 Mb; **Supplementary Fig. 1**). Several types of satellite sequences were present in the data set, comprising 23.2% of all Sanger reads and 76.2% of unassembled reads (**Supplementary Table 3**). FISH analysis indicated that these are primarily located in the centromeric and telomeric regions[15]. The cucumber genome also contains a large number of rRNA sequences, and about 3.3% of the Sanger reads matched 45S rRNA. These results indicated that the majority of the remaining 30% of unassembled regions of the genome are likely to be heterochromatic satellite or rRNA sequences.

The high coverage of the cucumber genome by this assembly was also confirmed using the available EST, fosmid and BAC sequences. The assembly contains 96.8% of the 63,312 cucumber unigenes assembled from ~350,000 Roche 454–sequenced ESTs, 99.3% of the 6,952 NCBI-deposited ESTs of cucumber, 91.2% of the 50,441 NCBI-deposited ESTs of melon and 98.7% of the six finished fosmid and BAC sequences (**Supplementary Table 4**).

A genetic map was developed using 77 recombinant inbred lines from the intersubspecific cross between Gy14 (a North American processing market–type cucumber cultivar) and PI183967 (an accession of *C. sativus* var. *hardwickii* originating from India). The map spans 581 cM and contains 1,885 markers, including 995 microsatellite markers[16] and 890 Diversity Arrays Technology markers (marker sequences can be accessed at http://cucumber.genomics.org.cn). Using this map, we were able to anchor 72.8% of the assembled sequences onto the seven chromosomes. Among the 1,885 markers, 1,763 (93.5%) were uniquely aligned and used for constructing the pseudochromosomes. The majority (98.7%) of the markers were collinear with the sequence assembly (**Fig. 1a**). Comparison of the genetic and physical distances between markers revealed recombination suppression of two 10-Mb regions at either end of chromosome 4, a 20-Mb region on chromosome 5 and an 8-Mb region on chromosome 7. Using high-resolution FISH, we confirmed previously identified segmental inversion[16] within the suppression region on chromosome 5 between Gy14 and PI183967 (**Fig. 1b**), which provides an explanation for recombination suppression in these regions. These regions of recombination suppression are additionally useful for studying cucumber evolution during domestication.

After excluding 16 markers whose genetic positions were ambiguous, we examined the six remaining regions that had conflicts between the genetic map and our assembly. Upon inspection, we found that clone mate-pair information supported our assembly in all of these regions (**Supplementary Fig. 2**). We also identified no misassembly within the regions covered by the six finished fosmid or BAC sequences (**Supplementary Fig. 3**). The conflicts may be a result of chromosomal rearrangement that occurred between the sequenced genotype 9930 and the genotypes used to create the mapping population; alternatively, these markers may have been placed incorrectly on the genetic map. Sequencing depth distribution showed that we obtained more than 10× coverage on more than 97.5% of the assembly (**Supplementary Fig. 4**).

Repetitive sequences and transposons

The cucumber genome contains a large number of transposable elements, but only a few have previously been identified. We therefore constructed repeat libraries using multiple *de novo* methods and then derived a combined repeat library that contained 1,566 sequences (**Supplementary Table 5**), of which 469 (29.9%) were manually classified (**Supplementary Table 6**). We then used this library for repeat annotation of the cucumber genome. We identified a total of 54.4 Mb, which represents ~24% of the genome, as repeats. Among them, 51.5% could be classified based on known repeats. The long terminal repeat (LTR) retrotransposons (*gypsy* and *copia*) made up the majority of the transposable element classes and comprised 10.4% of the genome (**Supplementary Table 7**). The repeats divergence rate (percentage of substitutions in the matching region compared with consensus repeats in constructed libraries) distribution showed a peak at 20%. A fraction of LTR retrotransposons, long interspersed nuclear elements and DNA transposons (composing 2.3%, 0.4% and 0.2% of the genome, respectively) are of relatively recent origin, having a sequence divergence rate of less than 5% (**Supplementary Fig. 5**).

Gene annotation

We used three gene-prediction methods (cDNA-EST, homology based and *ab initio*) to identify protein-coding genes and then built a consensus gene set by merging all of the results (**Supplementary Fig. 6**). We predicted 26,682 genes, with a mean coding sequence size of 1,046 bp and an average of 4.39 exons per gene (**Supplementary Table 8**). Under an 80% sequence overlap threshold, we found that 26.7% of the genes were supported by models from all three gene prediction methods, 25% had both *ab initio* prediction and homology-based evidence, and 7.4% had *ab initio* prediction and cDNA-EST expression evidence; the remaining genes were primarily derived from pure

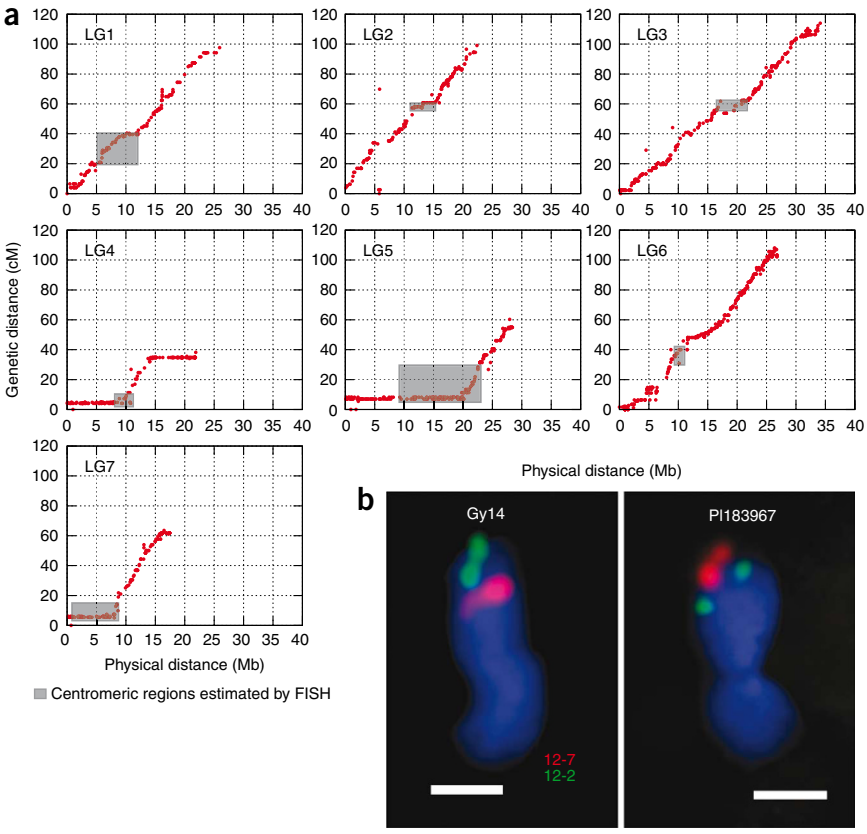

Figure 1 Integrated genetic and physical map of cucumber. (**a**) Genetic versus physical distance map of the seven cucumber chromosomes. The genetic map was constructed using a recombinant inbred line mapping population from the intersubspecific cross between Gy14 (domestic cucumber) and PI183967 (wild cucumber). (**b**) Segmental inversion between Gy14 and PI183967 on cucumber chromosome 5 detected by high-resolution FISH (12-2 and 12-7 denote individual fosmid clones). A low-resolution FISH analysis was also recently reported[16]. Scale bars represent 1 μm.

ab initio prediction, but the majority of these were supported by multiple gene finders (**Supplementary Table 9**). About 81% of the genes have homologs in the TrEMBL protein database, and 66% can be classified by InterPro. In sum, 82% of the genes have either known homologs or can be functionally classified (**Supplementary Table 10**). In addition to protein-coding genes, we identified 292 rRNA fragments and 699 tRNA, 238 small nucleolar RNA, 192 small nuclear RNA and 171 miRNA genes in the cucumber genome (**Supplementary Table 11**).

On the basis of pairwise protein sequence similarities, we carried out a gene family clustering analysis on all genes in sequenced plants, using rice as an outgroup. The cucumber genes consist of 15,669 families. Of these, 4,362 are cucumber unique families, among which 3,784 are single-gene families (**Supplementary Table 12**). The EST confirmation rate of these unique single-copy genes was much lower than the average of all predicted genes (33.4% vs. 72.3%, respectively). This category may therefore contain a number of false-positive predictions. In papaya, there are 4,622 unique families, but the actual number of genes is estimated to be 24,746, which is lower than the 28,629 predicted genes[7]. Thus, the actual number in cucumber should be lower than 26,682 and similar to that in papaya. The smaller average gene family size in cucumber (1.71) and papaya (1.77) supports this conclusion (**Fig. 2a**).

The cucumber genome contains the smallest number of tandem gene duplications (479) among all the plants we compared, whereas grapevine has the largest number (5,382; **Fig. 2a**). This may contribute in part to the small number of genes in cucumber.

Absence of recent whole-genome duplication

Whole-genome duplication (WGD) is common in angiosperm plants and produces a tremendous source of raw material for gene genesis. Previous research has revealed a paleohexaploidy (γ) event in the common ancestor of *Arabidopsis thaliana* and grapevine after the divergence of monocotyledons and dicotyledons[6]. Subsequently, two WGDs (α and β) occurred in *Arabidopsis*[17] and one (p) in poplar[8], whereas no recent WGD occurred in grapevine and papaya. Evidence indicates that rice underwent an ancient WGD[18]. We carried out a collinear gene-order analysis on the cucumber genome and observed no recent WGD and only a few segmental duplication events (**Supplementary Fig. 7**). We also used the distance-transversion rate at fourfold degenerate sites (4DTv method) to analyze paralogous gene pairs between syntenic blocks in *Arabidopsis* and cucumber, respectively. Two peaks (~0.06 and ~0.25) in *Arabidopsis* support the two recent WGDs (**Fig. 2b**). In cucumber, the analysis showed ancient duplication events (peak at ~0.60) but did not reveal recent WGD. This lack of recurrent WGD in the small cucumber genome provides an important complement to the grapevine and papaya genomes to study ancestral forms and arrangements of plant genes.

Synteny with flowering plant genomes

Given the similar gene arrangements between cucumber and other plant genomes, we defined syntenic blocks that contained 5,473, 6,525, 9,842, 8,439 and 3,992 cucumber genes collinear to *Arabidopsis*, papaya, poplar, grapevine and rice, respectively (**Supplementary Table 13** and **Supplementary Figs. 8–12**). The numbers of collinear genes were consistent with the phylogenetic distances of the other plants to cucumber. Within the syntenic blocks, we observed the highest density of collinear genes between cucumber and grapevine (90.5 genes per Mb), followed by papaya (76.1; the low contiguity of genome assembly may have, in part, decreased this value), poplar (68.8), rice (55.6) and *Arabidopsis* (43.5; **Supplementary Table 13**). This indicates that *Arabidopsis* has the most reshuffled or rearranged genome, whereas the genomes of grapevine and papaya are more conserved, probably because they have not undergone WGD since the ancestral paleohexaploidy.

Substantial fusion events involved in chromosomal evolution

Melon and cucumber belong to the same genus, although cucumber has seven chromosomes and melon has 12. Watermelon, their common distant relative, has 11 chromosomes. To investigate cucurbit chromosomal evolution, we compared the melon[19] and watermelon genetic maps to the cucumber genome (**Fig. 3a**). In total, 348 (66.7%) of the 522 melon markers and 136 (58.6%) of the 232 watermelon markers were aligned on the cucumber chromosomes

(**Supplementary Table 14**). The comparison revealed that there has been no substantial rearrangement of cucumber chromosome 7, which corresponds to melon chromosome 1 and watermelon group 7.

Using watermelon as an outgroup, we found that cucumber chromosomes 1, 2, 3, 5 and 6 were collinear to melon chromosomes 2 and 12, 3 and 5, 4 and 6, 9 and 10, and 8 and 11, respectively, indicating that after speciation these cucumber chromosomes each resulted from a fusion of two ancestral chromosomes. We also found that cucumber chromosome 6 and melon chromosome 3 have a syntenic segment, indicating that interchromosome rearrangement occurred in one of the two genomes after speciation. Cucumber chromosome 4 largely corresponds to melon chromosome 7, although a segment of melon chromosome 8 is syntenic with cucumber chromosome 4 (crossing the centromere). These data indicate that the rearrangement is most likely to have occurred before the divergence of cucumber and melon. In addition to chromosome fusion and interchromosome rearrangements, the comparison revealed the occurrence of several intrachromosome rearrangements (**Fig. 3a**).

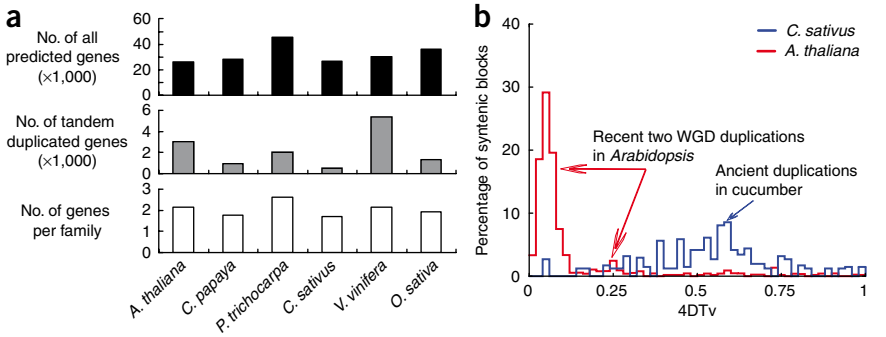

Figure 2 Comparison of cucumber genome with other sequenced plant genomes. (**a**) Numbers of predicted genes, numbers of tandem duplicated genes and gene family sizes of the six sequenced plant genomes. (**b**) The 4DTv distribution of duplicate gene pairs in cucumber and *Arabidopsis*, calculated based on alignment of codons with HKY substitution model.

Cucumber-melon microsynteny

To estimate the sequence divergence rate, we compared the four sequenced melon BACs to the cucumber genome (**Fig. 3b** and **Supplementary Fig. 13**). There are 56 genes on the melon BACs, 52 of which are collinear with the cucumber genome. The mean sequence similarity over coding regions is 95%. Although the gene region similarity is very high, the repeat content between the two genomes is quite different. New transposable elements were frequently inserted in the intergenic regions of both genomes. Hence, only 54% of the BAC sequences could be aligned onto the cucumber genome, with an average of 88% sequence identity. Nonetheless, the highly conserved gene content and order between the two species make the cucumber genome useful for genetic analysis of melon.

Using the annotated genes in the four melon BACs, we obtained and manually curated eight orthologous families among rice, cucumber, melon, *Arabidopsis* and papaya. Extrapolating from the age of divergence between *Arabidopsis* and papaya (54–90 million years ago), we estimated that cucumber and melon diverged about 4–7 million years ago, which is consistent with a previous estimate of 9 ± 3 million years ago[20].

Pathogen resistance genes

Only 61 nucleotide-binding site (NBS)-containing resistance (NBS-R) genes have been identified in cucumber, similar to papaya (55)[7] but only a fraction of what is found in *Arabidopsis* (200), poplar (398) and rice (600)[8]. Distribution of NBS genes on chromosomes

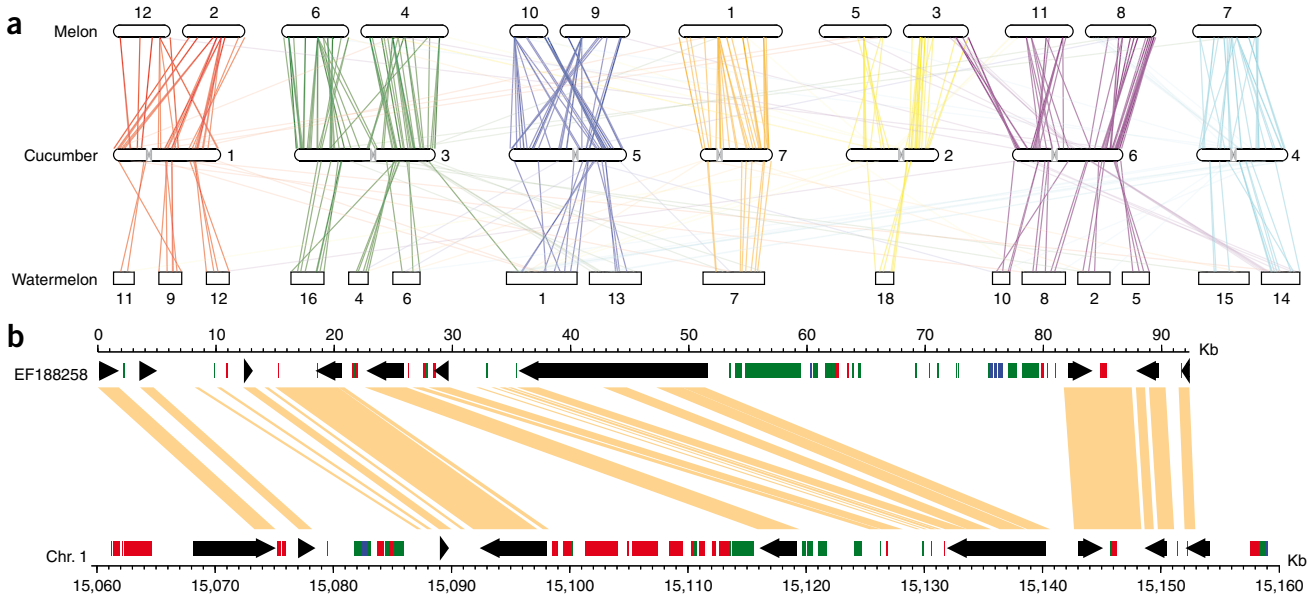

Figure 3 Comparative genomic analysis of cucurbits. (**a**) Comparative analysis of the melon and watermelon genetic maps with the cucumber sequence map. Cucumber, melon and watermelon have 7, 12 and 11 pairs of chromosomes, respectively. The current version of the watermelon genetic map is organized into 18 genetic groups. (**b**) Syntenic blocks between the cucumber genome and a melon BAC sequence (GenBank accession code EF188258.1). Genes are indicated by black arrows with the orientation indicated on the sequence. Rectangles, transposable elements; red, retrotransposable elements; blue, DNA transposons; green, unclassified transposable elements. Orthologous sequence regions between the two genomes are shown.

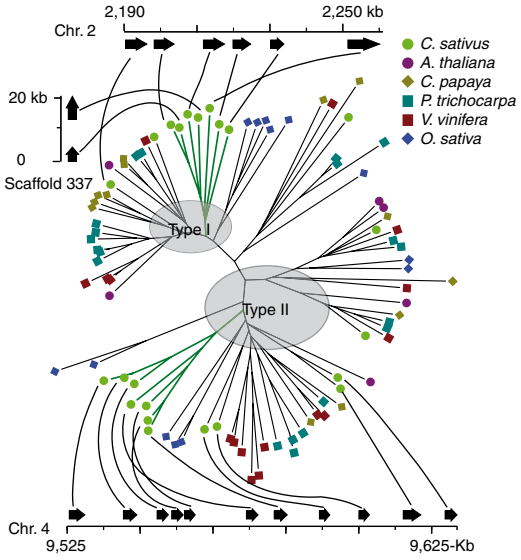

Figure 4 Lineage-specific expansion of the *LOX* gene family in the five sequenced dicot genomes and rice genome. The *LOX* family is divided into two groups, type I and type II. The two tandem duplicated gene clusters are ordered and shown on chromosomes 2 and 4, as well as one unmapped scaffold of the cucumber genome.

is nonrandom, with only five genes located on chromosomes 1, 6 and 7 and 20 genes located on chromosome 2 (**Supplementary Fig. 14**). Three-quarters of the NBS genes are located within 11 clusters, indicating that they evolved through tandem duplications, similar to other known plant genomes.

The lipoxygenase (LOX) pathway has an important role in developmentally and environmentally regulated processes in plants[21] and generates short-chain aldehydes and alcohols that are involved in plant defense and pest resistance[22]. The *LOX* gene family has been notably expanded in the cucumber genome (23 *LOX* genes in cucumber, 6 in *Arabidopsis*, 15 in papaya, 21 in poplar, 18 in grapevine and 15 in rice). Fourteen of the *LOX* genes are specific to the cucumber lineage. The majority of cucumber *LOX* genes (19 of 23) are distributed in three clusters, the largest of which contains 11 members that are arranged in tandem (**Fig. 4**). The other sequenced plant genomes show no obvious *LOX* clustering, with the exception of grapevine, which has one cluster harboring six copies.

Given that the cucumber has only 61 NBS-R genes, the expanded lipoxygenase pathway might be a complementary mechanism to cope with biotic stress. In support of this hypothesis, *Arabidopsis* has more NBS-R genes and fewer *LOX* genes than does papaya. The volatile (*E*,*Z*)-2,6-nonadienal (NDE) gives cucumber its 'fresh green' flavor[23] and confers resistance to some bacteria and fungi[24]. Lipoxygenase and one type of hydroperoxide lyase, 9-HPL, synthesize NDE from linolenic acid precursors. Genes encoding enzymes with 9-HPL activity are rarely found in other plants[25]. However, cucumber contains two tandem *HPL* genes, one of which has been experimentally confirmed as encoding an enzyme with 9-HPL activity[25]. The expansion of the *LOX* gene family and the duplicated *HPL* genes may be related to the high level of NDE synthesis in cucumber.

Eukaryotic translation initiation factors, particularly the eIF4E and eIF4G families, confer recessive resistance to plant RNA virus infections. An *EIF4E* gene in melon was found to mediate recessive resistance against melon necrotic spot virus[26]. In the cucumber genome, three *EIF4E* and three *EIF4G* genes have been identified, providing candidates for known recessive resistance genes against RNA viruses such as zucchini yellow mosaic virus and watermelon mosaic virus[27]. In some wild melon genotypes, enhanced expression of two glyoxylate aminotransferase genes (*At1* and *At2*) controls the resistance to downy mildew, a devastating foliar disease of cucurbits[28]. We identified two *At* homologs in cucumber that could be candidate genes for downy mildew resistance.

Novel biosynthetic pathways

Cucurbitacins are bitter cucurbit triterpenoid compounds that are toxic to most organisms but can attract specialized insects[29,30]. The presence of cucurbitacin in the cucumber is controlled by a mendelian gene, *Bi*[30]. Oxidosqualene cyclase catalyzes the formation of the triterpene carbon framework in plants[31]. An *OSC* gene, *CPQ*, in squash (*Cucurbita pepo* L.) is the first committed enzyme in the cucurbitacin biosynthesis pathway[32]. In cucumber, we identified four *OSC* genes; the *CPQ* ortholog *Csa008595* resides in a genetic interval that defines the *Bi* gene (**Supplementary Fig. 15**). Notably, *Csa008595* forms a cluster that contains an acyltransferase-encoding gene (*Csa008594*) and two cytochrome P450–encoding genes (*Csa008596* and *Csa008597*). Three of these (*Csa008594*, *Csa008595* and *Csa008597*) are coexpressed strongly in cucumber leaf tissue (**Supplementary Fig. 16**) in a pattern similar to that of the operon-like gene cluster involved in thalianol biosynthesis in *Arabidopsis*[33]. This gene cluster may therefore catalyze the stepwise formation of cucurbitacin in cucumber.

Cucumber is a model system for studying sex expression in plants[1]. Ethylene stimulates femaleness and is considered the sex hormone of cucumber[34]. We identified 137 cucumber genes that are related to the biosynthetic and signaling pathways of ethylene[35,36], but we found no gene family expansion in these pathways compared with other sequenced plant genomes (**Supplementary Table 15**). Thus, the origin of monoecy in cucumber might involve other evolutionary mechanisms.

The melon gene *Cm-ACS7* (ref. 37) and its cucumber ortholog *Cs-ACS2* (ref. 38) encode 1-aminocyclopropane-1-carboxylate synthase (ACS), a key regulatory enzyme in the ethylene biosynthetic pathway. Both genes are crucial to the inhibition of male organs and development of the female flower. *In situ* mRNA hybridization experiments revealed that both *Cm-ACS7* and *Cs-ACS2* transcripts accumulate only in the pistil and ovule, whereas their *Arabidopsis* ortholog, *AT4G26200* (**Supplementary Fig. 17**), is expressed only in the roots[39]. We also identified two ethylene-responsive elements (AWTTCAAA) and one flower meristem identity gene *LEAFY*-responsive element (CCAATGT) within the *Cs-ACS2* and *Cm-ACS7* promoter sequences, but these were absent from the promoter of *AT4G26200*. These findings indicate that the evolution of unisexual flowers in cucurbits may have involved the acquisition of new *cis* elements of the *ACS* genes.

To better understand the mechanism of sex determination in cucumber, we sequenced 359,105 EST sequences from near-isogenic unisexual and bisexual flower buds using the 454 pyrosequencing technology. Our analysis revealed that six auxin-related genes (auxin can regulate sex expression by stimulating ethylene production[40]) and three short-chain dehydrogenase or reductase genes (homologs to the sex determination gene *ts2* in maize[41]) are more highly expressed in unisexual flowers (**Supplementary Table 16**). This analysis provides an important resource for further study of sex determination in cucumber.

Novel developmental programs

The tendril is a specific climbing tool of vines, such as Vitaceae and all Cucurbitaceae. Darwin considered tendrils a key innovation in plant

evolution[42]. In cucumber and grapevine, gibberellic acid regulates tendril formation[43,44]. In most plants, the transition of GA_{12}-aldehyde to GA_{12} is catalyzed by cytochrome P450 monooxygenase. In cucurbits, it is also catalyzed by specific GA-7-oxidase genes, which are absent from *Arabidopsis*[45]. Cucumber has two GA-7-oxidase genes (**Supplementary Table 17**). GA-20-oxidase controls key steps leading to bioactive GA_1 and GA_4, and our data show that the cucumber has three lineage-specific clades (three copies; **Supplementary Fig. 18**). These specific genes might be associated with the role of gibberellic acid in the regulation of tendril formation. Tendril coiling involves rapid cell wall modification[46], and expansins are cell wall–loosening proteins in plants[47]. We found that, in cucumber, the expansin subfamily EXLA has undergone marked expansion through tandem duplication (eight genes in cucumber, compared with one to three genes in other genomes; **Supplementary Fig. 19**); this event may have contributed to the development of tendril coiling in cucumber.

Use in plant vascular biology studies

The evolution of the plant vascular system, comprising xylem and phloem tissues, had a pivotal role in the emergence of land plants. The sieve tube system of phloem, the equivalent of the animal arterial system, delivers nutrients and signaling molecules to developing organs[2]. A BLASTP analysis of 1,209 protein fragments from pumpkin phloem[48] identified 800 phloem proteins in the cucumber genome (**Supplementary Table 18**). Using these cucumber proteins, we conducted orthologous gene family (cluster) analysis (**Supplementary Table 19**) with their homologs in other vascular plants as well as the nonvascular moss *Physcomitrella patens*[49]. In total, we constructed 686 clusters (**Table 2**). About two-thirds (49 of 75) of the *Arabidopsis* and half (57 of 120) of the rice phloem proteins identified in previous studies[50,51] were included in this data set, indicating the effectiveness of these analyses and the value of this resource for vascular biology studies in plants.

The vascular and nonvascular plants shared 596 clusters; between monocots and eudicots, there are 648 clusters in common. Phloem protein II (*PP2*; cluster 2432) are present in angiosperms but absent from the moss genome. *PP2*-like genes are also present in gymnosperm[52], indicating their association with the advent of vascular plants. In cucurbits, these genes can increase the size-exclusion limit of plasmodesmata and facilitate cell-to-cell traffic of macromolecules[52] and thus are likely to have an essential role in vascular function. The sieve element occlusion proteins (gene cluster 4754), present in all eudicots but absent from mosses and monocots, represent a novel mechanism that evolved for sealing the sieve tube system after wounding[53].

The average number of genes in each cluster ranges from 2.9 to 5.1 in the vascular plants, compared to 1.7 in moss (**Table 2**). The increase of gene numbers per cluster may be associated with the evolution of the plant vascular system. The 16-kDa PP16 cluster (cluster 2599) has an average of 3.7 genes in the vascular plants compared to 2 in moss. The *CmPP16* gene in pumpkin is involved in transport of mRNA into the phloem[3]. The increase of the number of *PP16* genes in vascular plants indicates these new members may be involved in long-distance trafficking of mRNA.

To better understand xylem formation, we compared gene families related to lignin and cellulose biosynthesis between woody and herbaceous plants. The perennial woody plants, poplar and grapevine, have a large number of lignin biosynthesis–related genes (48 and 49, respectively), whereas the semiwoody plant papaya has an intermediate number (39). In contrast, the herbaceous plants *Arabidopsis* and cucumber have smaller numbers (28 and 26, respectively; **Supplementary Table 20**). Among these gene families, the number of genes in the cadmium-sensitive *CAD* family was consistent with this trend. In poplar and grapevine, homologs for *AT4G37980* and *AT4G37990* in *Arabidopsis*, which have low cadmium-sensitive enzymatic activity *in vitro* and may have only a minor role in lignin formation in this species[54], were expanded markedly. In papaya, there is an expansion of homologs for *AT1G37970*, which lack detectable cadmium-sensitive catalytic activities *in vitro* but are expressed predominantly in lignin-forming tissues[54] (**Supplementary Fig. 20**). Thus, the expansion of *CAD* genes may be associated with wood formation. It is also notable that grapevine has the largest *PAL* gene family, with 15 members, and that poplar and papaya have the largest number of *HCT* genes, with 7 members. Of the cellulose biosynthesis–related genes, poplar has more *CESA* and *COB* genes (18 of each) than do any of the other sequenced dicots (**Supplementary Table 20**).

Table 2 Summary of orthologous gene families (clusters) established using cucumber genes homologous to pumpkin phloem proteins

	Genes	Gene clusters	Average genes per cluster
P. patens[a]	1,072	622	1.7
O. sativa	2,458	676	3.6
S. bicolor	2,780	679	4.1
A. thaliana	2,351	682	3.5
C. papaya	1,944	672	2.9
P. trichocarpa	3,454	684	5.1
C. sativus[b]	1,986	686	2.9
V. vinifera	2,535	668	3.8

[a]Moss (*P. patens*) was used as the only outgroup. [b]For each cluster, at least one cucumber phloem protein was included.

DISCUSSION

The sequence of the cucumber genome provides an invaluable new resource for biological research and breeding of cucurbits. The high collinearity between cucumber and melon genomes enables cucumber to serve as a model system in the Cucurbitaceae family for comparative genomics studies in plants. The cucumber genome and related transcriptome analysis can provide insights into the mechanisms underlying sex determination, an important biological process that has been well characterized in cucumber at the phenotypic level. The genome can also advance our knowledge of the evolution and function of the plant vascular system.

We have also shown that, in combination with traditional Sanger sequencing, next-generation DNA sequencing technologies can be used effectively for *de novo* sequencing of plant genomes, making it possible to carry out rapid and low-cost sequencing for other important plant species.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The cucumber genome sequence has been deposited in GenBank with accession code ACHR00000000 (the version described here is the first version, with accession code ACHR01000000).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank L. Goodman for assistance in editing the manuscript and R. Quatrano, L. Kochian, L. Comai, V. Sundaresan, S. Kamoun and S. Renner for critical readings of the manuscript. This work was funded by the Chinese Ministry of Agriculture (948 program), Ministry of Science and Technology (2006DFA32140, 2007CB815701, 2007CB815703 and 2007CB815705) and Ministry of Finance

(1251610601001); the National Natural Science Foundation of China (30871707 and 30725008); the Chinese Academy of Agricultural Sciences (seed grant to S.H.); the Chinese Academy of Science (GJHZ0701-6 and KSCX2-YWN-023); the US Department of Agriculture (National Research Initiative grant 2006-35304-17346 to W.J.L.); the National Science Foundation (grant IOS-07-15513 to W.J.L.); and the Korea Science and Engineering Foundation–Ministry of Education, Science and Technology (WCU R33-10002 and BK21 grants to J.-Y.K.). WKC was partly supported by grants from the Environmental Biotechnology National Core Research Center (R15-2003-012-01003-0) and National Research Laboratory (2009-0066339). This work was also supported by the Shenzhen Municipal and Yantian District Governments and the Society of Entrepreneurs & Ecology. D. Qu and Z. Fang of the Chinese Academy of Agricultural Sciences provided management support for this work.

AUTHOR CONTRIBUTIONS

S.H., Y.D., Jun Wang and Songgang Li managed the project. S.H., Z.Z., W.J.L., X.G. and R.L. designed the analyses. X.G., H.M., L.L., Yuanyuan Ren, G.T., Y. Lu, Z.X., J.C., A., Z.W., J. Zhang, H. Liang, X.R., M.J., Hailong Yang, R.C., Shifang Liu and X.Z. conducted DNA preparation and sequencing. X.W., B.X., K.L., W.J., Guangcun Li, Z.F., J.S., A.K., E.A.G.v.d.V. and Y.X. contributed new reagents and analytic tools. S.H., Z.Z., W.J.L., X.G., R.L., X.W., B.X., K.L., W.J., J.H., Z.J., Yi Ren, Ying Li, X.L., S.W., Q.S., W.K.C., J.-Y.K., K.H.-U., H.M., Z.C., S.Z., J. Wu, Y.Y., H.K., Y.W., J.G., Y.H., M.L., B. Zhao, Shiqiang Liu, W.F., P.N., H. Zhu, Jun Li, J.R., W.Q., M. Wang, Q.H., B.L., Q.C., Y.B., Z.S., M. Wen, G.Z., Z.Y., Jianwen Li, L.M., H. Liu., Y. Zhou, J. Zhao, X.F., Guoqing Li, L.F., Yingrui Li, D.L., Hancheng Zheng and Shaochuan Li conducted the data analyses. S.H., R.L., Z.Z. and W.J.L. wrote the paper. Y.D., R.S., B. Zhang., S.J., G.Y., S.Y., Hongkun Zheng, Y. Zhang, N.Q., Z.L., L.B., K.K., Huanming Yang and Jian Wang revised the paper.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Tanurdzic, M. & Banks, J.A. Sex-determining mechanisms in land plants. *Plant Cell* **16**, S61–S71 (2004).
2. Lough, T.J. & Lucas, W.J. Integrative plant biology: role of phloem long-distance macromolecular trafficking. *Annu. Rev. Plant Biol.* **57**, 203–232 (2006).
3. Xoconostle-Cázares, B. *et al.* Plant paralog to viral movement protein that potentiates rransport of mRNA into the phloem. *Science* **283**, 94–98 (1999).
4. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
5. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
6. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
7. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
8. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
9. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
10. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
11. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
12. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
13. Staub, J.E., Serquen, F.C., Horejsi, T. & Chen, J.-f. Genetic diversity in cucumber (*Cucumis sativus* L.): IV. An evaluation of Chinese germplasm1. *Genet. Resour. Crop Evol.* **46**, 297–310 (1999).
14. Arumuganathan, K. & Earle, E. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
15. Han, Y.H. *et al.* Distribution of the tandem repeat sequences and karyotyping in cucumber (*Cucumis sativus* L.) by fluorescence in situ hybridization. *Cytogenet. Genome Res.* **122**, 80–88 (2008).
16. Ren, Y. *et al.* An integrated genetic and cytogenetic map of the cucumber genome. *PLoS One* **4**, e5795 (2009).
17. Bowers, J.E., Chapman, B.A., Rong, J. & Paterson, A.H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
18. Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38 (2005).
19. Fernandez-Silva, I. *et al.* Bin mapping of genomic and EST-derived SSRs in melon (*Cucumis melo* L.). *Theor. Appl. Genet.* **118**, 139 (2008).
20. Schaefer, H., Heibl, C. & Renner, S.S. Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc. Biol. Sci.* **276**, 843–851 (2009).
21. Liavonchanka, A. & Feussner, I. Lipoxygenases: occurrence, functions and catalysis. *J. Plant Physiol.* **163**, 348–357 (2006).
22. Schwab, W., Davidovich-Rikanati, R. & Lewinsohn, E. Biosynthesis of plant-derived flavor compounds. *Plant J.* **54**, 712–732 (2008).
23. Buescher, R.H. & Buescher, R.W. Production and stability of (*E, Z*)-2, 6-nonadienal, the major flavor volatile of cucumbers. *J. Food Sci.* **66**, 357–361 (2001).
24. Cho, M.J., Buescher, R.W., Johnson, M. & Janes, M. Inactivation of pathogenic bacteria by cucumber volatiles (*E,Z*)-2,6-nonadienal and (*E*)-2-nonenal. *J. Food Prot.* **67**, 1014–1016 (2004).
25. Matsui, K. *et al.* Fatty acid 9- and 13-hydroperoxide lyases from cucumber. *FEBS Lett.* **481**, 183–188 (2000).
26. Nieto, C. *et al.* An eIF4E allele confers resistance to an uncapped and non-polyadenylated RNA virus in melon. *Plant J.* **48**, 452–462 (2006).
27. Wai, T. & Grumet, R. Inheritance of resistance to watermelon mosaic virus in the cucumber line TMG-1: tissue-specific expression and relationship to zucchini yellow mosaic virus resistance. *Theor. Appl. Genet.* **91**, 699–706 (1995).
28. Taler, D., Galperin, M., Benjamin, I., Cohen, Y. & Kenigsbuch, D. Plant eR genes that encode photorespiratory enzymes confer resistance against disease. *Plant Cell* **16**, 172–184 (2004).
29. Balkema-Boomstra, A.G. *et al.* Role of cucurbitacin C in resistance to spider mite *Tetranychus urticae* in cucumber *Cucumis sativus* L. *J. Chem. Ecol.* **29**, 225–235 (2003).
30. Da Costa, C.P. & Jones, C.M. Cucumber beetle resistance and mite susceptibility controlled by the bitter gene in *Cucumis sativus* L. *Science* **172**, 1145–1146 (1971).
31. Phillips, D.R., Rasbery, J.M., Bartel, B. & Matsuda, S.P. Biosynthetic diversity in plant triterpene cyclization. *Curr. Opin. Plant Biol.* **9**, 305–314 (2006).
32. Shibuya, M., Adachi, S. & Ebizuka, Y. Cucurbitadienol synthase, the first committed enzyme for cucurbitacin biosynthesis, is a distinct enzyme from cycloartenol synthase for phytosterol biosynthesis. *Tetrahedron* **60**, 6995–7003 (2004).
33. Field, B. & Osbourn, A.E. Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008).
34. Rudich, J., Halevy, A.H. & Kedar, N. Ethylene evolution from cucumber plants as related to sex expression. *Plant Physiol.* **49**, 998–999 (1972).
35. Pirrung, M.C. Ethylene biosynthesis from 1-aminocyclopropanecarboxylic acid. *Acc. Chem. Res.* **32**, 711–718 (1999).
36. Stepanova, A.N. & Alonso, J.M. Ethylene signaling pathway. *Sci. STKE* **2005**, cm3 (2005).
37. Boualem, A. *et al.* A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons. *Science* **321**, 836–838 (2008).
38. Li, Z. *et al.* Molecular isolation of the M gene suggests that a conserved-residue conversion induces the formation of bisexual flowers in cucumber plants. *Genetics* **182**, 1381–1385 (2009).
39. Yamagami, T. *et al.* Biochemical diversity among the 1-amino-cyclopropane-1-carboxylate synthase isozymes encoded by the *Arabidopsis* gene family. *J. Biol. Chem.* **278**, 49102–49112 (2003).
40. Takahashi, H. & Jaffe, M.J. Further studies of auxin and ACC induced feminization in the cucumber plant using ethylene inhibitors. *Phyton (Buenos Aires)* **44**, 81–86 (1984).
41. DeLong, A., Calderon-Urrea, A. & Dellaporta, S.L. Sex determination gene TASSELSEED2 of maize encodes a short-chain alcohol dehydrogenase required for stage-specific floral organ abortion. *Cell* **74**, 757–768 (1993).
42. Darwin, C.R. *The Movements and Habits of Climbing Plants* (Murray, London, 1875).
43. Boss, P.K. & Thomas, M.R. Association of dwarfism and floral induction with a grape ('green revolution/' mutation. *Nature* **416**, 847–850 (2002).
44. Galun, E. The cucumber tendril—a new test organ for gibberellic acid. *Cell. Mol. Life Sci.* **15**, 184–185 (1959).
45. Lange, T. Cloning gibberellin dioxygenase genes from pumpkin endosperm by heterologous expression of enzyme activities in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **94**, 6553–6558 (1997).
46. Braam, J. In touch: plant responses to mechanical stimuli. *New Phytol.* **165**, 373–389 (2005).
47. Cosgrove, D.J. Loosening of plant cell walls by expansins. *Nature* **407**, 321–326 (2000).
48. Lin, M.-K., Lee, Y.-J., Lough, T.J., Phinney, B.S. & Lucas, W.J. Analysis of the pumpkin phloem proteome provides insights into angiosperm sieve tube function. *Mol. Cell. Proteomics* **8**, 343–356 (2009).
49. Rensing, S.A. *et al.* The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**, 64–69 (2008).
50. Aki, T., Shigyo, M., Nakano, R., Yoneyama, T. & Yanagisawa, S. Nano scale proteomics revealed the presence of regulatory proteins including rhree FT-Like proteins in phloem and xylem saps from rice. *Plant Cell Physiol.* **49**, 767–790 (2008).
51. Giavalisco, P., Kapitza, K., Kolasa, A., Buhtz, A. & Kehr, J. Towards the proteome of *Brassica napus* phloem sap. *Proteomics* **6**, 896–909 (2006).
52. Dinant, S. *et al.* Diversity of the superfamily of phloem lectins (phloem protein 2) in angiosperms. *Plant Physiol.* **131**, 114–128 (2003).
53. Pélissier, H.C., Peters, W.S., Collier, R., van Bel, A.J. & Knoblauch, M. GFP tagging of sieve element occlusion (SEO) proteins results in green fluorescent forisomes. *Plant Cell Physiol.* **49**, 1699–1710 (2008).
54. Kim, S.J. *et al.* Expression of cinnamyl alcohol dehydrogenases and their putative homologues during *Arabidopsis thaliana* growth and development: lessons for database annotations. *Phytochemistry* **68**, 1957–1974 (2007).

ONLINE METHODS

Removal of contamination for Sanger reads. Sanger reads were aligned against mitochondrion (assembled by us based on the gene sequences of mitochondria of rice and *Arabidopsis*), chloroplast (GenBank accession code AJ970307) and satellite (GenBank X03768, X03769, X03770, X69163, AY424361 and AY424362) sequences. Reads with identity >95% were filtered.

***De novo* assembly of Solexa data.** The De Bruijn graph method was used to represent all possible sequences assembled by Solexa reads, with a *K*-mer as a node and the $(K-1)$ base overlap between two *K*-mers as an edge. Some tips and low-coverage *K*-mers in the graph were removed to reduce sequencing errors and eliminate branches. The De Bruijn graph was then converted to a contiging graph by turning a series of linearly connected *K*-mers into a precontig node. Dijkstra's algorithm was implemented to detect bubbles, which were then straightforwardly merged into a single path if sequences of the branches were sufficiently similar. By this approach, the repeat regions could be assembled into consensus sequences.

Contigs were next connected by paired reads to form a scaffolding graph. Edges in this graph were connections between contigs, and the edge length was estimated from the insert size of the paired reads. The paired-end information was used step by step, from insert sizes around 200 bp and 500 bp to 2 kb. At each step, two procedures were applied: the repeat-masking method masked the complicated connections around repeat contigs, and the subgraph linearization turned the interleaving contigs into linear structure. This process yielded the final set of Solexa contigs and scaffolds.

Combination of Sanger reads and Solexa scaffolds. RePS2 (ref. 55) software was used to assemble the Solexa scaffolds and Sanger reads. We counted the depth of each 17-mer in the 3.9× plasmid and fosmid ends to create the 17-mer database, which contained all the depth information of the 17-mers. This database was then used to check all the contigs to identify repeated ones. A contig was defined as a repeat if over 80% of the 17-mers it contained were with higher depth than the threshold. After removing the repeat contigs, the scaffolds were divided into fake paired reads with read length of 600 bp and insert size of 1,700 bp. All segments over 200 bp were put into the second data set, which was then assembled as a unique region. In the same way as the construction of Solexa scaffolds, the plasmid, fosmid and BAC ends were used, step by step, to construct a 'superscaffold'.

Misassembly checking and gap filling. In the final stage, we used the repeat sequences to fill the gaps in the scaffolds using the following steps. First, we mapped all of the reads that contained paired-end information (Solexa and plasmid reads, as well as fosmid and BAC ends) to the scaffolds, and we used the unique contigs to establish the paired-end relationship between the contigs. Second, we identified repeat contigs with paired ends that uniquely connected two other scaffolded contigs. If the length of the repeat contig and the estimated size of the gap were similar, the gap was filled by this repeat. Any remaining repeat contigs that were not used for gap filling were added into the final set of scaffolds.

Chromosome anchoring along the cucumber genetic map. The marker sequences in the cucumber genetic map were aligned against the scaffold sequences using BLASTN at an *E*-value cutoff of 1×10^{-20}. Hits with coverage >30% and identity >90% were considered mapped markers. Based on the mapped markers, the scaffold sequences were anchored on the cucumber chromosomes. During this process, the scaffolds with mapped markers that showed inconsistent genetic positions were manually checked by paired-end relationships; the incorrect scaffold was then split.

FISH analysis. The FISH protocol was described in a previous study[16]. To better visualize the segmental inversion, we chose chromosome spreads where chromosome 5 appeared in a straight form. Instead of showing all chromosomes[16], only chromosome 5 is shown in **Figure 1b** of this study. In addition, the image was taken in a higher resolution. Scale bars represent 1 μm, as compared to 3 μm previously[16]. Red and green signals were detected with anti-digoxigenin antibody coupled to rhodamine (Roche) and by anti-avidin antibody conjugated with FITC (Vector Laboratories), respectively.

Identification of repetitive elements in the cucumber genome. Four *de novo* software packages, ReAS[56], PILER-DF[57], RepeatScout[58] and LTR_Finder[59], were used to search for repeat sequences within the cucumber genome. All repeat sequences with lengths >100 bp and gap 'N' <5% constituted the raw transposable element library.

The repeat elements belonging to rRNA and satellite sequences were first filtered using BLASTN (E value $\leq 1 \times 10^{-10}$, identity ≥ 80%, coverage ≥ 50% and minimal matching length ≥ 100 bp). All-versus-all BLASTN (E value $\leq 1 \times 10^{-10}$) searches were then conducted iteratively, and the shorter sequences were filtered when two repeats aligned with identity ≥ 80%, coverage ≥ 80% and minimal matching length ≥ 100 bp; this yielded a nonredundant transposable element library. The nonredundant repeats were then searched against the Swiss-Prot protein database to filter the protein-coding genes by BLASTX (E value $\leq 1 \times 10^{-4}$, identity ≥ 30%, coverage ≥ 30% and minimal matching length ≥ 30 amino acids). After manual curation, a *de novo* transposable element library for cucumber was obtained.

Transposable elements in the cucumber genome assembly were identified both at the DNA and protein level. RepeatMasker was applied for DNA-level identification using a custom library (a combination of Repbase, plant repeat database and our cucumber *de novo* transposable element library). At the protein level, RepeatProteinMask was used to conduct WU-BLASTX searches against the transposable element protein database. Overlapping transposable elements belonging to the same type of repeats were integrated together, whereas those with low scores were removed if they overlapped >80% and belonged to different types.

Gene prediction. Our strategy for gene prediction was to conduct *de novo* predictions on the repeat-masked genome and then integrate them with spliced alignments of proteins and transcripts to genome sequences using GLEAN[60]. Cucumber genome sequences were masked by identified repeat sequences with length >500 bp, except for miniature inverted-repeat transposable elements, which are usually found near genes or inside introns. The EST and full-length cDNA sequences of cucumber were processed by PASA[61] to train gene prediction software BGF[62], GlimmerHMM[63] and SNAP[64]. Augustus[65] and Genscan[66] software used gene model parameters trained for *Arabidopsis*. We aligned the protein sequences of five sequenced plants (*Arabidopsis*, papaya, poplar, grapevine and rice) onto the cucumber genome using TBLASTN, at an *E*-value cutoff of 1×10^{-5}, and the homologous genome sequences were aligned against the matching proteins using GeneWise[67] for accurate spliced alignments. The cDNA and EST sequences of cucumber and melon were aligned against the cucumber genome using BLAT (identity ≥ 0.95, coverage ≥ 0.90) to generate spliced alignments. We also aligned TIGR unigenes[68] from Cucurbitales, Fabales and Fagales to the cucumber genome by ATT_gap2 (ref. 69). All of these resources were combined by GLEAN[60] to produce the consensus gene sets.

Identification of noncoding RNA genes in the cucumber genome. The tRNA genes were identified by tRNAscan-SE[70] with default parameters. The C/D-box small nucleolar RNAs were identified by Snoscan[71] using yeast rRNA and yeast methylation sites. Other noncoding RNAs, including miRNA, small nuclear RNA and H/ACA-box small nucleolar RNA, were identified using INFERNAL software by searching against the Rfam[72] database with default parameters.

Construction of gene families. We adapted the Treefam[73] method to construct gene families for the genes in cucumber, *Arabidopsis*, papaya, poplar, grapevine and rice (outgroup).

Construction of syntenic blocks. We identified syntenic blocks between two species (A and B) by an automatic clustering algorithm on a dot plot graph, which included five steps. First, markers (gene pairs) were generated between A and B. All protein sequences of A were aligned to all proteins of B using BLASTP (E value $< 1 \times 10^{-10}$ and identity > 20%). The fragmental alignments were conjoined for each gene pair. Those gene pairs with aligned regions covering <50% were filtered. The remaining gene pairs were plotted on the dot graph as markers (points). Second, the Euclidean distance was calculated for each pair. Distances were calculated based on the gene order in each chromosome rather than the genomic position. Third, hierarchical clustering was

doi:10.1038/ng.475

determined for all of the points. If the distance between two points was less than the distance cutoff, a link was assigned. The distance cutoff was adapted in accordance with the selected species. Fourth, the quality was estimated for each cluster by calculating the point number (N), average point distance (D) and correlation coefficient (R). A score (S) was calculated to show the overall quality, defined as $S = N \times \mathrm{sqrt}(2)/D \times R$. Finally, problematic clusters were filtered. Clusters with $N < 8$ or $|R| < 0.5$ were filtered out. The clusters caused by tandem duplication were further filtered by determining the slope (L) of the regression line within a range of $0.1 < |L| < 10$. This algorithm can also be used to study intraspecies synteny.

4DTv calculation. After the identification of syntenic blocks, the pairwise protein alignments for each gene pair were first constructed with MUSCLE[74]. The nucleotide alignment was then created according to the protein alignment. 4DTv was then calculated on concatenated nucleotide alignments with HKY substitution models[75].

Comparative analysis between cucumber and melon. Cucumber genome sequences were aligned with melon BAC sequences using NUCmer, a program in the MUMmer package[76]. The delta-filter program was then run with the −1 option to remove complex alignments. Orthologous gene pairs were identified by the reciprocal best method.

The Bayesian relaxed molecular clock approach was used to estimate divergence time using the program MULTIDIVTIME, which was implemented using the Thornian Time Traveler (T3) package. The calibration time (fossil record time) interval (54–90 million years ago) of Capparales was obtained from previous results[77,78].

URLs. *Arabidopsis thaliana* (TIGR Release 5.0), ftp://ftp.tigr.org/pub/data/a_thaliana/ath1; *Carica papaya* (assembly v1.0, EVidence Modeler genes), http://www.life.uiuc.edu/ming; *Populus trichocarpa* (assembly release v1.0, annotation v1.1), http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html; *Vitis vinifera* (published assembly, annotation v1), http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/; *Oryza sativa* (assembly International Rice Genome Sequencing Project build 3), http://rgp.dna.affrc.go.jp/IRGSP/download.html; *Oryza sativa* (GLEAN genes annotated by Beijing Genomics Institute), ftp.genomics.org.cn/pub/ricedb/rice_update_data/GLEAN_genes/IRGSP_japonica/; *Physcomitrella patens* (assembly release v1.0, annotation v1.1), http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html; *Sorghum bicolor* (assembly release v1.0, annotation v1.4), http://www.phytozome.net/sorghum; UniGene sequences of *Cucurbitales*, *Fabales* and *Fagales*, http://plantta.jcvi.org/; cucumber marker sequences, http://cucumber.genomics.org.cn; UniProt (Swiss-Prot/TrEMBL) release 14.1, http://www.uniprot.org/downloads; InterPro v18.0, http://www.ebi.ac.uk/interpro/; KEGG release 47, ftp://ftp.genome.jp/pub/kegg/pathway/; Repbase release 13.07, http://www.girinst.org/repbase/index.html; Plant Repeat Databases (TIGR), http://plantrepeats.plantbiology.msu.edu/index.html; Rfam release 9.0, http://rfam.sanger.ac.uk/; Thornian Time Traveler (T3) package, http://abacus.gene.ucl.ac.uk/software.html; RepeatMasker, http://www.repeatmasker.org.

55. Wang, J. *et al.* RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res.* **12**, 824–831 (2002).
56. Li, R. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLOS Comput. Biol.* **1**, e43 (2005).
57. Edgar, R.C. & Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** Suppl 1, i152–i158 (2005).
58. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21** Suppl 1, i351–i358 (2005).
59. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
60. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
61. Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. & Buell, C.R. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* **7**, 327 (2006).
62. Li, H. *et al.* Test data sets and evaluation of gene prediction programs on the rice genome. *J Comp Sci Tech* **20**, 446–453 (2005).
63. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
64. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
65. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** Suppl 2, ii215–ii225 (2003).
66. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
67. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
68. Childs, K.L. *et al.* The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res.* **35**, D846–D851 (2007).
69. Huang, X., Adams, M.D., Zhou, H. & Kerlavage, A.R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37–45 (1997).
70. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
71. Lowe, T.M. & Eddy, S.R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–1171 (1999).
72. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
73. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006).
74. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
75. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
76. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
77. Crepet, W.L., Nixon, K.C. & Gandolfo, M.A. Fossil evidence and phylogeny: the age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits. *Am. J. Botany* **91**, 1666–1682 (2004).
78. Wikström, N., Savolainen, V. & Chase, M.W. Evolution of the angiosperms: calibrating the family tree. *Proc. Biol. Sci.* **268**, 2211–2220 (2001).