nature genetics

A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity

Jianjian Qi[1,12], Xin Liu[2,12], Di Shen[1,12], Han Miao[1,12], Bingyan Xie[1,12], Xixiang Li[1,12], Peng Zeng[2], Shenhao Wang[1], Yi Shang[1], Xingfang Gu[1], Yongchen Du[1], Ying Li[1], Tao Lin[1], Jinhong Yuan[1], Xueyong Yang[1], Jinfeng Chen[3], Huiming Chen[4], Xingyao Xiong[1,5], Ke Huang[5], Zhangjun Fei[6], Linyong Mao[6], Li Tian[7], Thomas Städler[8], Susanne S Renner[9], Sophien Kamoun[10], William J Lucas[11], Zhonghua Zhang[1] & Sanwen Huang[1]

Most fruits in our daily diet are the products of domestication and breeding. Here we report a map of genome variation for a major fruit that encompasses ~3.6 million variants, generated by deep resequencing of 115 cucumber lines sampled from 3,342 accessions worldwide. Comparative analysis suggests that fruit crops underwent narrower bottlenecks during domestication than grain crops. We identified 112 putative domestication sweeps; 1 of these regions contains a gene involved in the loss of bitterness in fruits, an essential domestication trait of cucumber. We also investigated the genomic basis of divergence among the cultivated populations and discovered a natural genetic variant in a β-carotene hydroxylase gene that could be used to breed cucumbers with enhanced nutritional value. The genomic history of cucumber evolution uncovered here provides the basis for future genomics-enabled breeding.

Ensuring an adequate and high-quality food supply for the expanding worldwide human population requires more effective plant breeding. To this end, it is critical to obtain a comprehensive understanding of the genetic variation within crop germplasm—the raw material of plant breeding[1]. Next-generation DNA sequencing technologies now permit cost-effective genome sequencing at a population scale, which has resulted in the construction of genome-wide variation maps for several major crops[2–8] as well as for the model plant *Arabidopsis thaliana*[9,10]. The cucumber (*Cucumis sativus* L.) is indigenous to India[11], where its wild form *Cucumis sativus* var. *hardwickii* still exists. To characterize patterns of genetic variation in cucumber, we previously sampled a core collection of 115 cucumber lines that capture 77.2% of the total genetic diversity estimated for 3,342 accessions from a wide geographic distribution[12] (**Fig. 1a** and **Supplementary Table 1**). For the present study, we generated a variation map of the cucumber genome at single-base resolution by performing deep resequencing of all 115 lines, and we also sequenced the wild cucumber genome *de novo* and compared it to the genome of cultivated cucumber[13]. These genomic resources generate new insights into the genetic basis of domestication and diversity for this important crop.

Resequencing of the 115 lines generated a total of 7.275 billion paired-end reads (632 Gb of sequence), with average depth of 18.3× and coverage of 95.2% (**Supplementary Table 1**). Through comparison with the reference genome of the inbred cucumber line 9930 (refs. 13,14), we detected a total of 3,305,010 SNPs, 336,081 small insertions and deletions (indels; shorter than 5 bp) and 594 presence-absence variations (PAVs) (**Table 1**, **Supplementary Figs. 1–4**, **Supplementary Tables 2–4**, **Supplementary Note** and **Supplementary Data Set**). We also carried out *de novo* sequencing and assembly of the wild cucumber accession PI183967 (CG0002) (**Supplementary Tables 5** and **6** and **Supplementary Note**), a well-studied *C. sativus* var. *hardwickii* accession[15,16]. The total length of the assembly was 204.8 Mb, and the N50 lengths of the contigs and scaffolds were 119 kb and 4.2 Mb, respectively. We predicted a total of 23,836 genes in the wild genome. By aligning the assembly against the genome for the cultivated accession 9930, we identified 21,021 orthologous genes. The accuracy of SNPs and genotyping inference was estimated to be 98.9% by PCR and Sanger sequencing of 400 randomly selected SNPs in 4 individual lines (**Supplementary Table 7** and **Supplementary Note**). We identified 74,166 nonsynonymous SNPs in 19,087 genes, including 1,713 nonsense SNPs in 1,516 genes causing start codon changes, premature stop codons or elongated transcripts. These variants are likely important in the functional evolution of cucumber genes and deserve further investigation.

The 115 cucumber lines we resequenced can be divided into 4 geographic groups (**Fig. 1a**,**b**). The Indian group consists of 30 lines mainly from India, including 13 lines identified morphologically

[1] Institute of Vegetables and Flowers of the Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Beijing, China. [2] BGI-Shenzhen, Shenzhen, China. [3] State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, Nanjing, China. [4] Hunan Vegetable Research Institute, Hunan Academy of Agricultural Sciences, Changsha, China. [5] Hunan Provincial Key Laboratory for Germplasm Innovation and Utilization of Crop, Horticulture & Landscape College, Hunan Agricultural University, Changsha, China. [6] Boyce Thompson Institute for Plant Research, US Department of Agriculture (USDA) Robert W. Holley Center for Agriculture and Health, Ithaca, New York, USA. [7] Department of Plant Sciences, University of California, Davis, Davis, California, USA. [8] Plant Ecological Genetics, Institute of Integrative Biology, Eidgenössische Technische Hochschule (ETH) Zurich, Zurich, Switzerland. [9] Department of Biology, University of Munich, Munich, Germany. [10] The Sainsbury Laboratory, Norwich Research Park, Norwich, UK. [11] Department of Plant Biology, University of California, Davis, Davis, California, USA. [12] These authors contributed equally to this work. Correspondence should be addressed to S.H. (huangsanwen@caas.cn) or Z.Z. (zhangzhonghua@caas.cn).

Received 7 April; accepted 20 September; published online 20 October 2013; doi:10.1038/ng.2801

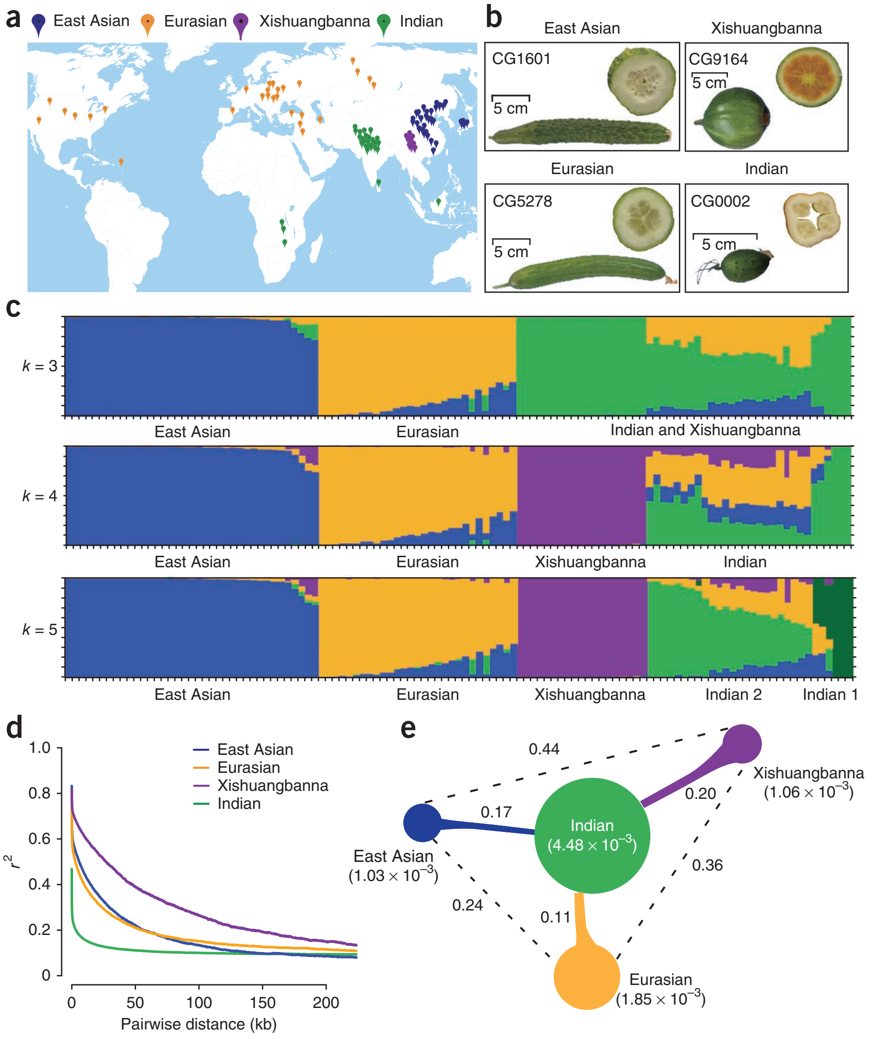

Figure 1 Cucumber populations. (**a**) The core collection of 115 lines sequenced in this study has a wide geographic distribution. Color codes indicate geographic groups. (**b**) Fruit morphology of the four groups. The cucumber line CG1601 (East Asian) bears fruits with dense, white spines and an elongated stalk. Fruits of cucumber line CG5278 (Eurasian) lack spines and have a short fruit stalk. Cucumber line CG9164 (Xishuangbanna) bears melon-like fruits with a low fruit shape index (length/width) and a unique orange endocarp. Cucumber line CG0002 (Indian) bears small, oval fruits with sparse, black spines. Note that the images differ in scale. (**c**) Model-based clustering analysis of the core set, given different number of groups (K = 3, 4 or 5), using STRUCTURE. The y axis quantifies subgroup membership, and the x axis shows the different accessions. (**d**) Decay of LD, measured by r^2, in the four groups. (**e**) Summary of nucleotide diversity and population divergence across the four groups. Values in parentheses represent measures of nucleotide diversity for the group, and values between pairs indicate population divergence (F_{ST}).

as the wild form *C. sativus* var. *hardwickii* (exclusively Indian in origin; **Supplementary Table 1**). The other 3 groups contain only lines from the cultivated form *Cucumis sativus* var. *sativus*: (i) Eurasian group: 29 lines primarily from central and western Asia, Europe and the United States, (ii) East Asian group: 37 lines mainly from China, Korea and Japan, and (iii) Xishuangbanna group: 19 landraces cultivated in the Xishuangbanna region of tropical southwestern China[17]. The Xishuangbanna group uniquely accumulates β-carotene in its fruit (**Fig. 1b**). Model-based analyses of population structure and phylogenetic reconstruction support the notion that the three cultivated groups are monophyletic and genetically quite homogeneous, whereas the Indian group shows clear evidence of substructure and genetic heterogeneity (**Fig. 1c**, **Supplementary Figs. 5–7**, **Supplementary Note** and **Supplementary Data Set**). The basal nature of the Indian group compared to the other three groups is supported by its significantly higher nucleotide diversity π (ref. 18) (**Fig. 1e**) and by its large numbers of private variants that account for 39.0% and 46.7% of the total SNPs and indels, respectively (**Table 1**). In addition, the decay of linkage disequilibrium (LD) with physical distance between SNPs occurs at only 3.2 kb in the Indian group (decaying to r^2 of 0.2), whereas the equivalent distances are 55.2–140.5 kb for the three domesticated groups (**Fig. 1d** and **Table 1**). We also defined LD blocks in the four groups (**Table 1** and **Supplementary Fig. 8** and **Supplementary Data Set**) and subsequently identified 317,323 tag SNPs that can be applied in cucumber breeding.

All crop species went through population bottlenecks during domestication[19–21]. As one could imagine that the number of rice plants necessary to feed an ancient household or community would be more than that of cucumber plants, the population size of founder plants in rice domestication would be larger than that in cucumber domestication. We thus hypothesize that the severity of domestication bottlenecks may systematically differ between fruit and grain food crops. To test this conjecture, we estimated the magnitude of the bottlenecks experienced by six crops by calculating the genome-wide reduction in genetic diversity in the cultivated groups compared to the wild groups (**Table 2**). The three fruit crops (cucumber, watermelon and tomato) exhibited larger reductions in genetic diversity than did grain crops (rice, maize and soybean). Demographic modeling using *δaδi*[22] suggests an effective population size of ~500 for cucumber at domestication; this estimate is markedly lower than corresponding estimates in maize[23] (~150,000) and also lower than that in rice[24] (~1,300) and soybean[7] (~1,000). Our comparison indicates that these fruit crops probably underwent narrower bottleneck events during domestication than the grain food crops, providing additional impetus for the use of wild germplasm in future fruit breeding.

Besides the genome-wide reduction in genetic diversity caused by population bottlenecks, domestication often resulted in a drastic loss of diversity in genomic regions carrying genes conferring favorable phenotypes, such as the non-shattering genes in rice[25,26] and the seed casing gene in maize[27]. To identify potential domestication sweeps in the cucumber genome, we scanned genomic regions with the largest reduction in diversity in cultivated groups and extreme divergence in allele frequency between wild and cultivated groups (**Supplementary Note**). In total, we identified 112 potential selective sweeps ranging from 50 kb to 780 kb in length (138 kb on average). These candidate

Table 1 General information on genetic variation in the cucumber genome

				Groups			
			Core set	Indian	Xishuangbanna	Eurasian	East Asian
Number of accessions			115	30	19	29	37
Variation[a]	SNPs		3,305,010	3,119,482	745,422	1,475,643	1,105,147
	Private SNPs		–	1,287,655	47,572	29,953	33,786
	Indels		336,081	288,377	84,844	120,661	83,544
	Private indels		–	157,068	10,195	16,615	13,000
	π (10^{-3})		3.17	4.48	1.06	1.85	1.03
LD[b]	LD decay (kb; r^2 = 0.2)			3.2	140.5	55.2	56.4
	LD blocks		–	79,545	14,582	34,734	24,057
	<1 kb			62,132	10,218	24,223	15,800
	1–10 kb		–	17,235	4,179	10,165	7,805
	>10 kb		–	178	185	346	452
Sweeps[c]	Regions		112	–	156	126	143
	Genes		2,054	–	2,133	2,104	2,387
Demography[d]	Population size at bottleneck				7.5 (4.5–10.5)	315.0 (270.0–375.0)	6.0 (1.5–9.0)
	Present population size				16,500 (12,000–19,500)	81,600 (48,450–91,800)	38,250 (31,950–44,700)
	Duration of bottleneck (years)				130.5 (117.0–136.5)	180.0 (120.0–315.0)	150.0 (75.0–315.0)
	Time after bottleneck to present (years)				3,450 (3,150–3,750)	3,450 (3,000–4,650)	2,550 (1,800–3,450)
					East Asian vs. Eurasian	Eurasian vs. Xishuangbanna	Eurasian vs. Xishuangbanna
Population divergence[e]	Regions	F_{ST}, top 5% threshold			0.57	0.88	0.76
		Regions			160	170	184
		Genes			2,319	2,207	2,466
	Nonsynonymous	F_{ST}, top 5% threshold			0.70	0.97	1.0
		SNPs			1,465	1,242	1,041
		Genes			869 (363)	759 (376)	616 (283)

[a]Private SNPs and indels are variations specific to each group. π, nucleotide diversity within each group. [b]LD blocks were defined using Haploview software. LD decay, the distance at which LD (measured as r^2) decayed to 0.2. [c]The regions and genes affected by sweeps were identified using the ratio of diversity of the wild group to that of the cultivated groups (π_w/π_c) in 50-kb sliding windows with a step size of 5 kb. The regions with the top 5% of π_w/π_c values and the top 5% of XP-CLR scores were considered to have undergone selective sweeps. [d]Demographic parameters (**Supplementary Fig. 11**) were inferred by $\delta a \delta i$[23]. Using the best-fitting parameters, the simulation was carried out 20 times on 500,000 randomly selected SNPs each time, and the mean value was calculated. The estimated 95% confidence intervals are given in parentheses. [e]F_{ST} values were calculated for each 50-kb sliding window with a step size of 5 kb for each SNP among the cultivated populations. The regions and nonsynonymous SNPs with the top 5% of F_{ST} values were considered. The number of genes containing nonsynonymous SNPs within the top 5% of F_{ST} values and not contained within the selected regions of high divergence is given in parentheses.

sweep regions occupied 7.8% of the assembled genome (15.4 Mb) and involved 8.5% of the annotated genes (2,054 genes) (**Fig. 2a** and **Supplementary Tables 8** and **9**). These regions also showed significantly extended LD (**Supplementary Note**).

To identify the potential trait-associated genes selected during cucumber domestication, we performed quantitative trait locus (QTL) mapping of domestication-related traits using three segregating populations (**Supplementary Note**). In total, we mapped five QTLs for fruit length (*fl1.1*, *fl3.1*, *fl4.1*, *fl4.2* and *fl6.1*) and three QTLs for leaf size (*ls1.1*, *ls2.1* and *ls2.2*). Except for one QTL (*ls2.1*), all others overlapped with a putative selective sweep region, and their peak signals resided within 1,000 kb (about 500 kb on average) of the nearby sweeps (**Fig. 2b**,**c** and **Supplementary Fig. 9** and **Supplementary Data Set**). For instance, within the physical interval of the *fl3.1* QTL for fruit length, there was a single candidate selective sweep (**Fig. 2b**) that contained 19 genes, including *Csa3G199660* that encodes cyclin and is likely involved in cell proliferation[28].

Wild cucumber plants bear extremely bitter fruit. Clearly, an essential step in the domestication of the wild cucumber into a fruit must have involved the (partial) loss of fruit bitterness. Two genetic loci, *Bi* and *Bt*, are known to confer bitterness in cucumber[29,30]. The recessive *bi* allele confers bitter-free foliage, and the dominant *Bt* allele renders the fruit extremely bitter. Using simple sequence repeat (SSR) markers[16] and SNP markers developed in this study, we subsequently mapped the

Table 2 Estimation of domestication bottlenecks for six crop species

	Species	Population[a]	Lines	Total genome coverage	Diversity (π)	Diversity ratio (π_w/π_c)
Fruit crops[b]	Cucumber	W	30	501×	0.0045	1.96
		C	85	1,606×	0.0023	
	Watermelon	W	10	41×	0.0076	5.43
		C	10	104×	0.0014	
	Tomato	W	16	88×	0.0042	2.63
		C	23	131×	0.0016	
Grain crops[c]	Rice	W	446	879×	0.003	1.25
		C	1,083	973×	0.0024	
	Maize	W	17	80×	0.0059	1.20
		C	23	123×	0.0048	
	Soybean	W	17	82×	0.0030	1.58
		C	14	7×	0.0019	

[a]W, wild population; C, cultivated population. [b]For cucumber, the Eurasian, East Asian and Xishuangbanna groups are combined as a single cultivated gene pool for comparison. For watermelon, π_w/π_c values are calculated on the basis of the data published in ref. 34. For tomato, our unpublished data were used (T.L., Y.L., Z.Z., Y.D., S.H. *et al.*, unpublished data). The wild population includes 16 lines mainly from the red-fruited wild species *Solanum pimpinellifolium* commonly regarded as the ancestor of cultivated tomato *Solanum lycopersicum*. [c]For rice, the cultivated rice population includes both *japonica* and *indica* rice[3]. For maize, the cultivated population in this comparison refers to the landrace population used for a domestication study[5]. Soybean data are from ref. 7.

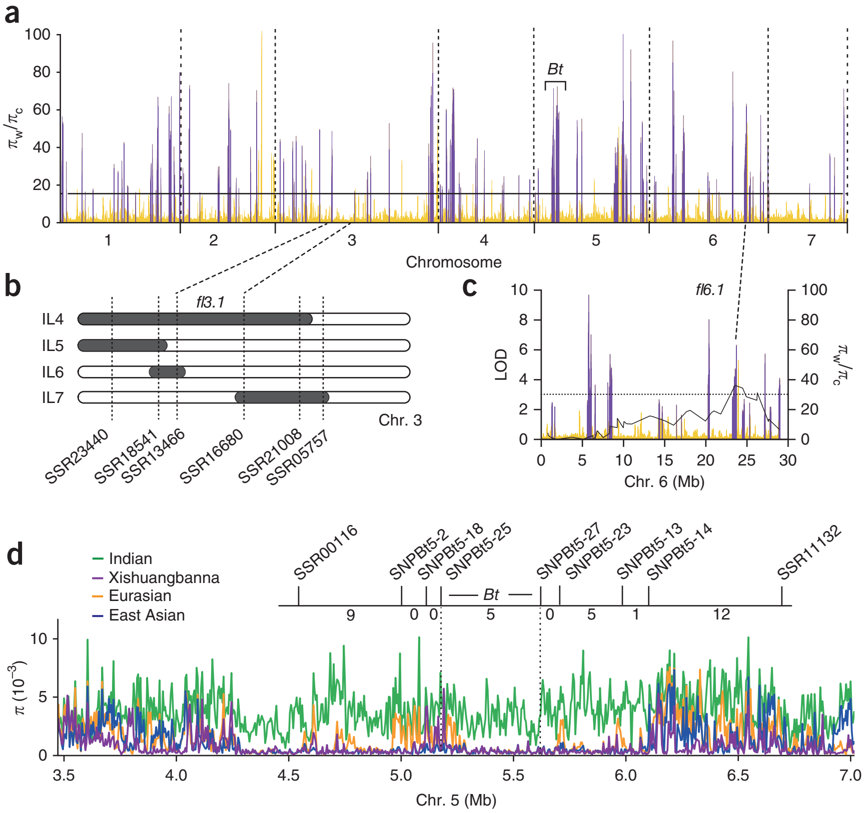

Figure 2 Detection and functional annotation of domestication sweeps. (**a**) Detected domestication sweeps on the seven chromosomes. A total of 112 regions with both the top 5% of π_w/π_c values (genetic diversity in the cultivated groups compared to the wild group) and the top 5% of XP-CLR scores were considered to be candidate sweeps (purple bars). The horizontal dashed line indicates the threshold (15.4) defining the top 5% of π_w/π_c values. Gold bars represent windows that are not considered to be candidate sweeps. Note that some gold bars above the π_w/π_c threshold are excluded as candidate sweeps because they do not pass the threshold defining the top 5% of XP-CLR scores. (**b**) A sweep within the physical interval of the *fl3.1* QTL for fruit length. The QTL was mapped in the interval defined by two markers (SSR13466 and SSR16680) on chromosome 3 by analysis of four single-fragment introgression lines (IL4–IL7) with the wild cucumber accession CG0002 (PI183967) as the donor parent and the cultivated cucumber line 931 as the recurrent parent. (**c**) The *fl6.1* QTL for fruit length was mapped by genetic analysis of the F_2:F_3 population from the cross of CG0002 (PI183967) and CG1601 (179 F_3 families). The peak of the QTL (SSR23284) is located within the sweep region (23.205–23.755 Mb) on chromosome 6. (**d**) Signal for a domestication sweep at the *Bt* locus that confers fruit bitterness. *Bt* resides between two SNP markers (SNPBt5-25 and SNPBt5-27), corresponding to a 442-kb region on chromosome 5. Numbers below the horizontal line indicate the numbers of recombinants between two neighboring markers in a large segregating population containing 1,822 F_2 individuals. The genetic order of all markers is consistent with their physical order. The mapped region for *Bt* overlaps with a large sweep region (5.290–6.070 Mb on chromosome 5) that shows almost no nucleotide diversity in the three cultivated groups (Xishuangbanna, Eurasian and East Asian).

Bi and *Bt* loci to chromosomes 6 and 5, respectively (**Supplementary Fig. 10**, **Supplementary Note** and **Supplementary Data Set**). We detected no signature of a selective sweep at the genomic region spanning the *Bi* gene (**Supplementary Fig. 10**). Intriguingly, we found almost no diversity in the cultivated groups in the region to which *Bt* was mapped, a strong signature of a selective sweep during domestication (**Fig. 2a,d**). Subsequently, we generated a high-resolution genetic map of the *Bt* locus using a large segregating population consisting of 1,822 F_2 individuals from a cross between the line 9110Gt and the sequenced Chinese Long line 9930 (**Supplementary Note**). On the basis of this analysis, the *Bt* locus was delimited to a 442-kb region on chromosome 5 (**Fig. 2d**) that harbors 67 predicted genes (**Supplementary Table 10**).

Newly identified alleles from another cultivated group represent low-hanging fruits for crop improvement, as their linkage with most unfavorable genes should have been eliminated by human selection. Therefore, it is important to study the extent and nature of genome-wide divergence among cultivated groups. In cucumber, the three cultivated groups diverge substantially, as shown by F_{ST} values and demographic changes (**Fig. 1e**, **Supplementary Figs. 11** and **12**, and **Supplementary Note**). We then searched for genomic regions showing the highest level of (near) fixation for alternative variants in cultivated groups (**Fig. 3a**, **Table 1**, **Supplementary Tables 11** and **12**, and **Supplementary Note**). Taking the Eurasian and East Asian groups as an example, there are 160 such regions (50–665 kb in length, 17.7 Mb in total) with an F_{ST} value greater than 0.57 (0.24 at the whole-genome level) that contain 2,319 genes. We also scanned nonsynonymous mutations with extremely high F_{ST} values (top 5%) between cultivated groups (**Fig. 3a**, **Table 1**, **Supplementary Table 13** and **Supplementary Note**). Between the Eurasian and East Asian groups, we detected 1,465 nonsynonymous polymorphisms with an F_{ST} value greater than 0.70 that encoded amino acid changes in 869 genes. These genes included several disease resistance (*R*) genes on chromosome 2 that were reported to be associated with resistance against fungal diseases in the Eurasian cucumber accessions (not discovered in the East Asian germplasm)[31]. Adaptation to the local microbial environment likely caused the differential selection of these *R* genes.

We sought to further explore the genomic landscape of population divergence to identify genes controlling important traits. The most obvious trait that distinguishes the Xishuangbanna group from the other three groups is the orange endocarp of the fruit (**Fig. 1b**), which is due to the accumulation of large amounts of β-carotene; a single recessive gene, *ore*, was reported to control this phenotype[32]. All cucumber lines in the Xishuangbanna group in our core collection (n = 19) showed the 'orange' phenotype, and all other lines (n = 96) did not. We therefore searched for nonsynonymous SNPs fixed between the Xishuangbanna group and the other 3 groups (F_{ST} = 1) and identified 43 such SNPs. Only 1 of the 43 SNPs resided within the physical interval that spans the *ore* gene (**Fig. 3b**). This SNP codes for an amino acid change at residue 257 (p.Ala257Asp) in *Csa3G183920*, a gene encoding a putative β-carotene hydroxylase[33] (**Supplementary Fig. 13**), designated *CsaBCH1* in this study (**Fig. 3c**).

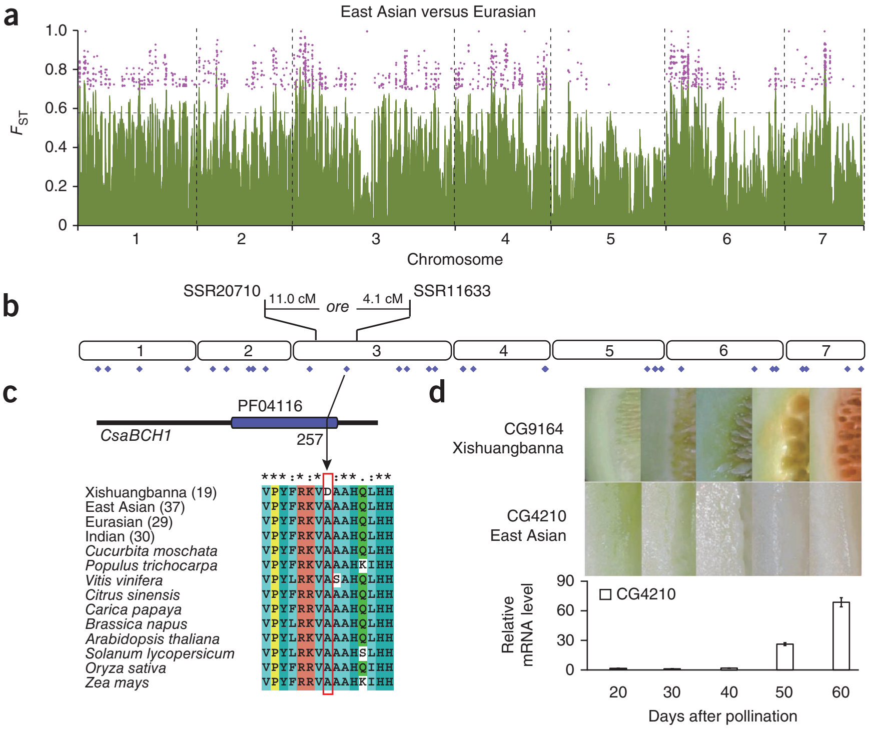

Figure 3 Population divergence and identification of a key mutation responsible for the accumulation of β-carotene in the fruit of the Xishuangbanna cucumbers. (**a**) Highly divergent regions (top 5%; $F_{ST} \geq 0.57$) and nonsynonymous SNPs (top 5%; $F_{ST} \geq 0.70$) between the East Asian and Eurasian groups. Green vertical bars higher than the dashed line ($F_{ST} = 0.70$) indicate highly divergent regions; purple dots indicate highly divergent nonsynonymous SNPs. (**b**) Physical positions of the genetically mapped *ore* gene and the 43 nonsynonymous SNPs with $F_{ST} = 1$ between the Xishuangbanna group ($n = 19$) and all other cucumbers ($n = 96$). Blue diamonds below the seven chromosomes indicate the positions of the SNPs. (**c**) A key mutation changed the conserved amino acid of a putative β-carotene hydroxylase (*CsaBCH1*). Residue 257 is located in the conserved PF04116 domain (the fatty acid hydroxylase domain). Xishuangbanna group cucumbers carry asparagine, whereas all other cucumbers and homologous proteins from ten other species carry alanine. Proteins used in the alignment refer to the corresponding accessions in GenBank: AEK86567 (*C. moschata*), XP_002327604 (*P. trichocarpa*), AAM77007.1 (*V. vinifera*), ABB49053 (*C. sinensis*), ADZ14893 (*C. papaya*), ABM54182 (*B. napus*), NP_194300 (*A. thaliana*), NP_001234348 (*S. lycopersicum*), ABF93742 (*O. sativa*) and ADC96676 (*Z. mays*). (**d**) *CsaBCH1* mRNA levels in the East Asian cucumber line CG4210 are significantly elevated during the period 40–60 d after pollination when the Xishuangbanna cucumber line CG9164 rapidly accumulates β-carotene. The orange endocarps of Xishuangbanna cucumber fruits represent the accumulation of large amounts of β-carotene. Three replicate RT-PCR assays were performed. The value obtained from the sample at 20 d was taken as 100%, and the values for other samples were normalized to that for the 20-d sample. Data are represented as average values with s.d.

This nonsynonymous polymorphism occurs within the region of *CsaBCH1* encoding the conserved domain (PF04116, the fatty acid hydroxylase domain). We aligned *CsaBCH1* orthologs from several other species and found that the Xishuangbanna group carries a rare variant that likely disrupts the function of the gene (**Fig. 3c**). To test this hypothesis, we cloned the *CsaBCH1* allele from the Chinese Long cucumber line 9930 and the *CsaBCH1*^{A257D} allele from the Xishuangbanna line CG9164 into an *Escherichia coli* expression vector, pET32a. Only one SNP in the coding region of the two alleles was identified, and it caused the polymorphism encoding p.Ala257Asp. Coexpression of *CsaBCH1*^{A257D} and pAC-BETA, which harbors all the genes for β-carotene biosynthesis, in *E. coli* did not convert β-carotene into its downstream hydroxylated products, whereas functional *CsaBCH1* did (**Supplementary Fig. 14**). Furthermore, we discovered that *CsaBCH1* was greatly upregulated during the maturation of fruit in the East Asian cucumber CG4210 (40–60 d after pollination) when the Xishuangbanna group cucumbers rapidly accumulate β-carotene (**Fig. 3d**). The expression pattern of this gene is in line with its role in the turnover of β-carotene in fruit (**Supplementary Fig. 13**). These results strongly indicate that *CsaBCH1* defines the *ore* locus. This study demonstrates the powerful synergy between genomics, population genetics and experimental biology for trait gene mining.

In this study, we provide a comprehensive description and analyses of genome-wide variation in a fruit species. Preliminary comparative analyses suggest that fruit crops may have undergone narrower bottlenecks during domestication than grain crops did and that they have more strongly reduced genetic diversity in their cultivated gene pools, underlining the importance of wild germplasm in fruit breeding. The identification of the *ore* gene indicates that the genomic landscape of population divergence should be further exploited to clone trait-related genes. The knowledge and resources generated herein will provide a basis for designing diagnostic tools for practical breeding.

URLs. Cucumber genome database, http://cmb.bnu.edu.cn/Cucumis_sativus_v20/ and http://www.icugi.org/; SOAP software, http://soap.genomics.org.cn/; Novoalign software, http://novocraft.com/; PCO software, https://www.stat.auckland.ac.nz/~mja/Programs.htm.

METHODS

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

Accession codes. Sequence data have been deposited in the NCBI Short Read Archive (SRA) under accession SRA056480. SNPs, indels, new assembly sequences, the genome assembly and annotation for wild cucumber and related scripts can be found in ftp://www.icugi.org/pub/reseq/cucumber/ and http://cmb.bnu.edu.cn/Cucumis_sativus_v20/resequence/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Ross-Ibarra (University of California, Davis) and J.-Y. Li (Chinese Academy of Agricultural Sciences (CAAS)) for critical comments, D. DellaPenna (Michigan State University) for assistance in β-carotene pathway analysis and the three anonymous referees for their constructive comments. We thank F. Cunningham (University of Maryland) for providing the pAC-BETA plasmid. This work was supported by funding from the National Program on Key Basic Research Projects in China (the 973 Program; 2012CB113900), the National

Natural Science Foundation of China (NSFC; 31225025) the National High-Tech Research Development Program in China (the 863 Program; 2010AA10A108 and 2012AA100101), and other NSFC grants (31030057, 30972011, 31101550, 31071797, 31171961, 31272161 and 31322047), the Chinese Ministry of Agriculture (the 948 program; 2008-Z42), the Chinese Ministry of Finance (1251610601001) and CAAS (seed grant to S.H.).

AUTHOR CONTRIBUTIONS

S.H. and Z.Z. conceived and designed the experiments. J.Q., D.S., H.M., X.G., S.W., Y.L., T.L., Y.S., X.Y., H.C., X.X., K.H., J.C. and L.T. performed the experiments. Z.Z., J.Q., X. Liu, B.X., X. Li, P.Z., J.Y., Y.D., Z.F., L.M., T.S., S.S.R., W.J.L., S.K. and S.H. analyzed the data. S.H., Z.Z., X. Liu and J.Q. wrote the manuscript. Z.F., T.S., S.S.R., W.J.L. and S.K. revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Morrell, P.L., Buckler, E.S. & Ross-Ibarra, J. Crop genomics: advances and applications. *Nat. Rev. Genet.* **13**, 85–96 (2011).
2. Chia, J.M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
3. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
4. Huang, X. *et al.* Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44**, 32–39 (2012).
5. Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
6. Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).
7. 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).
8. 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).
9. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
10. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
11. Sebastian, P., Schaefer, H., Telford, I.R. & Renner, S.S. Cucumber (*Cucumis sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proc. Natl. Acad. Sci. USA* **107**, 14269–14273 (2010).
12. Lv, J. *et al.* Genetic diversity and population structure of cucumber (*Cucumis sativus* L.). *PLoS ONE* **7**, e46919 (2012).
13. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
14. Li, Z. *et al.* RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* **12**, 540 (2011).
15. Li, X. *et al.* Construction of wild cucumber substitution lines. *Acta Horticulturae Sinica* **38**, 886–892 (2011).
16. Ren, Y. *et al.* An integrated genetic and cytogenetic map of the cucumber genome. *PLoS ONE* **4**, e5795 (2009).
17. Qi, C., Yuan, Z. & Li, Y. A new type of cucumber—*Cucumis sativus* L. var. *Xishuangbannanesis*. *Acta Horticulturae Sinica* **10**, 259–264 (1983).
18. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
19. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
20. Tang, H., Sezen, U. & Paterson, A.H. Domestication and plant genomes. *Curr. Opin. Plant Biol.* **13**, 160–166 (2010).
21. Gross, B.L. & Olsen, K.M. Genetic perspectives on crop domestication. *Trends Plant Sci.* **15**, 529–537 (2010).
22. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
23. Ross-Ibarra, J., Tenaillon, M. & Gaut, B.S. Historical divergence and gene flow in the genus *Zea*. *Genetics* **181**, 1399–1413 (2009).
24. Caicedo, A.L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
25. Konishi, S. *et al.* An SNP caused loss of seed shattering during rice domestication. *Science* **312**, 1392–1396 (2006).
26. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
27. Wang, H. *et al.* The origin of the naked grains of maize. *Nature* **436**, 714–719 (2005).
28. Horiguchi, G., Gonzalez, N., Beemster, G.T., Inze, D. & Tsukaya, H. Impact of segmental chromosomal duplications on leaf size in the *grandifolia-D* mutants of *Arabidopsis thaliana*. *Plant J.* **60**, 122–133 (2009).
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. Barham, W.S. The inheritance of a bitter principle in cucumbers. *Proc. Amer. Soc. Hort. Sci.* **62**, 441–442 (1953).
31. Kang, H. *et al.* Fine genetic mapping localizes cucumber scab resistance gene *Ccu* into an *R* gene cluster. *Theor. Appl. Genet.* **122**, 795–803 (2011).
32. Bo, K. *et al.* Inheritance and mapping of the *ore* gene controlling the quantity of β-carotene in cucumber (*Cucumis sativus* L.) endocarp. *Mol. Breed.* **30**, 335–344 (2012).
33. Walter, M.H. & Strack, D. Carotenoids and their cleavage products: biosynthesis and functions. *Nat. Prod. Rep.* **28**, 663–692 (2011).
34. Guo, S. *et al.* The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51–58 (2013).

ONLINE METHODS

Sample collection. A total of 3,342 cucumber lines collected worldwide were fingerprinted using 23 highly polymorphic SSR markers[12]. On the basis of fingerprinting results, a core germplasm set consisting of 115 cucumber lines was determined (**Supplementary Table 1**). The core set represents 77.2% of the total diversity. Population structure, geographic distribution and known traits of interest were also taken into account when determining the content of the core set.

Sequencing and mapping. Genomic DNA was extracted from etiolated seedlings from the 115 cucumber lines using the CTAB method[35]. At least 5 μg of genomic DNA was used for each line to construct sequencing libraries. Paired-end sequencing libraries with insert sizes of around 500 bp were constructed for all these lines, according to the manufacturer's instructions (Illumina). We sequenced 44 or 90 bp at each end by Illumina Genome Analyzer IIx or Illumina HiSeq 2000.

The cucumber reference genome and its annotation[13,14] were downloaded online (see URLs). Scaffolds not anchored on the chromosomes were joined together by 800 N nucleotides to generate one pseudochromosome named chromosome U. We used SOAP2 (ref. 36) to map all reads from each line against the reference genome with mismatches of 3 bp being allowed in a single read (parameters: −m 100 −x 888 −s 35 −l 32 −p 4 −v 3 for 90-bp reads and −v 2 for 44-bp reads). Reads from PCR duplication were removed. The mapping result was then split into chromosomes and sorted according to mapping coordinates. Both paired-end and single-end mapped reads (detailed information is provided in **Supplementary Table 1**) were then used in the following procedures for variation detection.

***De novo* assembly and comparative analysis of the wild cucumber genome.** *C. sativus* var. *hardwickii* is considered to be the progenitor of *C. sativus* var. *sativus*[37]. The wild cucumber accession PI183967 (CG0002), which was assembled *de novo* in this study, belongs to *C. sativus* var. *hardwickii*. In total, six sequencing libraries with insert sizes ranging from 170 bp to 10 kb were generated (**Supplementary Table 6**). Libraries were constructed according to the standard Illumina protocol. Contaminating reads were filtered out first. Following the SOAPdenovo[38] assembly process, contigs and scaffolds were constructed. In addition, we used fosmid and BAC-end sequences generated for the reference genome in a previous study[13] to construct superscaffold sequences. Only end sequences without conflicting paired-end relationships were used. In the final assembly, chloroplast and mitochondrial sequences were excluded by alignment against the reference sequences. The wild cucumber genome assembly was annotated independently using a previous annotation pipeline[14] integrating transcriptome sequences generated from wild cucumber. We used LASTZ to align the wild cucumber genome to the reference genome with the parameters T = 2 C = 2 H = 2,000 Y = 3,400 L = 6,000 K = 2,200 −format = axt. Only alignments >1 kb in length were considered. Alignments within annotated repetitive regions were excluded.

Variation calling. *SNPs*. SNPs were called in four steps.

(1) Possible SNPs between each cucumber line and the reference. Likelihoods of genotypes of each line at every genomic site were calculated by SOAPsnp[39] with parameters −L 50 −u −F 1. Then, SNPs were filtered using the criteria that the SNP quality value given by SOAPsnp had to be greater than 20 and the base quality had to be greater than 15. On the basis of these criteria, the likelihood of each individual's genotype in 'glf' format data was generated for each chromosome of each line.

(2) SNPs among the core set. We called SNPs in the core set by GLFmulti, a software based on the maximum-likelihood estimation of site frequency at each site. This software integrates the likelihoods of genotypes of each line at each site generated by SOAPsnp. The core set of SNPs was then obtained by filtering according to the site frequency and quality score given by GLFmulti, which takes sequencing quality and mapping quality into consideration.

(3) Putative SNP set. The core set of SNPs was further filtered using the following criteria: (i) two alleles had to exist in the population, confirming the respective position as polymorphic in the population; (ii) the total sequencing depth had to be >15× and < 2,400× (excluding possible errors resulting from low-quality mapping and repetitive sequences); (iii) the average mapping rate of reads mapped to the position had to be lower than 1.5, ruling out effects caused by duplications; and (iv) the nearest SNPs had to be more than 1 bp away. In addition, positions with a depth of less than 5× in each line were treated as 'genotype missing' instead of using the reference genotype. These SNPs were used as the putative SNPs in the core set.

(4) Further filtering according to segregation test and homozygosity proportion. For SNPs in the population, the sequencing depth of the two putative alleles in different individuals should be different. This test can distinguish any segregation pattern from random sequencing errors[2]. We thus applied the segregation test to the contingency table of read depth for SNP alleles by line. Permutations were used to determine the significance of allele depth in the population. Only sites with *P* value <0.01 were retained.

Because the cucumber lines were substantially inbred (expected ~10% residual heterozygosity), similar to in a previous study in maize[2], we filtered out sites at which less than 85% of the lines appeared to be homozygous and sites with a proportion of heterozygous genotypes more than three times that of the homozygous minor allele genotypes.

Identified SNPs were further classified on the basis of the gene annotation of the reference genome. SNPs were categorized as SNPs in intergenic regions, in 5′ UTRs, in coding sequences, in introns and in 3′ UTRs. SNPs in coding sequences were further grouped as synonymous SNPs not causing amino acid changes or nonsynonymous SNPs causing amino acid changes (nonsense mutations were also defined as mutations causing premature stops, elongated transcripts or introduction of false start codons).

Indels. To detect small insertions and deletions (shorter than 5 bp), all reads were mapped with a gap of less than 5 bp allowed (parameter −g 5) using SOAP2 (ref. 36). Indels (1–5 bp) were called by the SOAPindel pipeline (see URLs), as described previously[40]. Indels were identified in each line and then combined on the basis of the position and length of each insertion or deletion. To evaluate the accuracy of indels, we also used the Needleman-Wunsch global alignment algorithm (Novoalign software; see URLs) to align reads and identify indels using mpileup in SAMtools with parameters similar to those used in a previous study[41]. Detailed parameters are listed in the command line novoindex −k 13 −s 4 Novoalign −F ILMFQ −g 40 −x 5 −o SAM and samtools(v0.1.18) mpileup −h 300 −Augf (h: Coefficient for modeling homopolymer errors.[100]). More than 97% of the indels identified by the SOAPindel pipeline were also identified by the second pipeline.

PAVs. Reads that had both ends unmapped in all lines were pooled and assembled using SOAPdenovo[38] (parameters pregraph −s se_fa.lib −K 39 −R −o cucumber; contig −g cucumber −R −M 0; map −s se_fa.lib −g cucumber; scaff −g cucumber −F). Only scaffolds with lengths longer than 1 kb were retained. These scaffolds were then annotated by a homology-based method, Genewise[42], using all available plant proteins. Annotated genes were further filtered out if homologous genes could be found in the cucumber genome with over 80% identity and over 80% coverage at the nucleotide level by BLAST. As we used plant proteins for this annotation, possible bacterial contamination was excluded. Finally, we identified 594 genes as PAV genes, and these were further subjected to gene ontology annotation. To avoid false positives of identified PAVs due to the incompleteness of the reference genome, we mapped reads from the reference genome (30-fold) to the PAVs and found that no PAV could be mapped to >1× depth and coverage of >80%. We thus believe that all 594 identified PAV genes represent actual variation between different lines rather than false positives.

Population analysis. *Phylogeny*. To build a phylogenetic tree, we screened a subset of 25,228 4-fold degenerate SNPs without missing genotypes in all 115 lines from the entire SNP data set (3,305,010 SNPs). The rationale for choosing this subset is that these SNPs do not cause amino acid changes and, thus, should be under less selective pressure, more reliably reflecting population structure and demography. Using these SNPs, a phylogenetic tree was constructed using PhyML (version 3.0)[43] with the HKY85 model. A nonparametric bootstrap analysis was performed, and the number of bootstrap replicates was 100.

doi:10.1038/ng.2801

STRUCTURE. Population structure was investigated using the program STRUCTURE 2.3.1 (ref. 44), which is a model-based clustering method to infer the population structure assuming different numbers of clusters (K). A total of 25,228 4-fold degenerate SNPs without missing genotypes was used. First, we carried out analysis to determine the most likely group number, as previously described[45]. For each K value from 2 to 19, STRUCTURE was run 20 times on 1,000 randomly selected 4-fold degenerate SNPs. The statistic 'delta K', indicating the change in likelihood assuming different numbers of groups, was calculated and used to assess the most likely number of populations. Then, the subgroup membership of each accession was determined by 10,000 iterations for each K value from 3 to 5.

PCO. To perform principal-coordinate analysis, PCO[46] software (see URLs) was used. Because a maximum of 2,000,000 SNPs can be input into PCO software, we used 1,982,222 SNPs, including all 660,933 non-missing and 1,321,289 randomly selected SNPs.

$\delta a \delta i$. The site frequency spectrum in each group was first estimated by GLFmulti, the software we used for SNP detection. For each simulation, 500,000 SNPs were randomly selected and used in the $\delta a \delta i$ analysis. We fitted the two-group model (**Supplementary Fig. 11**) for all three cultivated groups by comparison to the Indian group. We fitted for 20 SNP data sets (500,000 random SNPs for each) and determined the 95% confidence interval of each parameter. As $\delta a \delta i$ estimates the time in units of $2N_e$ (the ancestral population size) generations, we further estimated $2N_e$ according to the formula $4N_e \times \mu \times L = \theta$, where μ is the mutation rate, L is the generation time and θ is the genetic diversity. As θ_w in the Indian group was estimated to be about 3×10^{-3} per base pair and assuming a mutation rate of 1×10^{-7} mutations per generation. $2N_e$ was estimated to be 150,000, and the time was scaled by $2N_e$ generations. To estimate the effective population size at domestication, we combined the three cultivated groups (East Asian, Eurasian and Xishuangbanna) into a single cultivated gene pool to compare with the ancestral Indian group.

π and Tajima's D. Nucleotide diversity (π) is used to measure the degree of variability within a population or species[18]. π and Tajima's D were calculated on the basis of the genotypes of each line at the SNP positions using BioPerl.

LD analysis. LD was calculated on the basis of SNPs with minor allele frequency (MAF) greater than 0.05 using Haploview software[47]. Four populations were separated, and SNPs in each population were extracted to perform the analysis. The parameters were −n −pedfile −info −log −minMAF 0.05 −hwcutoff 0.001 −dprime −memory 2096. Then, values for the r^2 and D' statistics were obtained. LD decay was calculated based on r^2 between two SNPs and the distance between the two SNPs.

LD blocks were also defined by Haploview. As block definition is time-consuming and requires considerable computer memory, we divided the genome into small regions of 500 kb (larger than the distance when LD decays to $r^2 = 0.2$). The parameters were −n −pedfile −info -minMAF 0 −hwcutoff 0.001 −log −blockoutput GAB −memory 30720 (or 14240) −pairwiseTagging. LD block definition was compared across the four groups, and tag SNPs in the four groups were merged into the final tag SNP set. TAGster[48] was used to identify the tag SNPs in all 115 lines, setting the parameters as −task: 1 −selection_method: 1 −LD_method: 2 −Vgenopair: 5 −cutoff_LD: 0.8 −max_distance: 250000 −selection_maf: [0.05,0.5] −evaluation_maf: [0.05,0.5] −maxtry: 1000000 −figure: 0.

Identification of domestication sweeps. To avoid the effect of genetic drift, we combined the three cultivated groups into a single cultivated gene pool for the analysis. The regions and genes under domestication sweeps should have significantly lower diversity in cultivated compared to wild cucumbers. Thus, the ratio of genetic diversity by comparing the Indian group to cultivated groups (π_w/π_c) in 50-kb sliding windows with a step size of 5 kb was used to identify regions with significantly lower levels of polymorphisms in cultivated groups. Windows with the top 5% of π_w/π_c values (15.4) were considered as candidate regions with significantly lower diversity in cultivated cucumbers. If the π_w value of a window was lower than 0.002, it was further removed. Finally, those windows with distance of ≤50 kb were merged into a single selected locus.

We then used updated XP-CLR[49] (acquired from the author, it uses the allele frequencies as input and is more useful for resequencing) to scan for the regions under selection in the cultivated groups. We used a 0.05-cM sliding window with 100-bp steps across the whole genome. The maximum number of SNPs assayed in each window was 200, and the command line was XPCLR −c freqInput outputFile −w1 gWin(Morgan) snpWin gridSize(bp) chrN. All SNPs were assigned to genetic positions using the previous genetic map[37] by assuming uniform recombination between mapped markers. Finally, we calculated the mean likelihood score in 50-kb sliding windows with a step size of 5 kb across the genome. Windows with the top 5% of XP-CLR values (206) were selected and merged into regions, as in the above analysis.

Candidate regions from the analysis of the ratio of genetic diversity were compared to those from XP-CLR analysis. Over 80% of those regions overlapped between the two methods. Finally, these shared regions were considered to be domestication sweeps. For these regions, we also calculated LD (r^2 calculated by Haploview[47]).

Identification of highly differentiated regions and nonsynonymous SNPs. The population fixation statistics F_{ST} were estimated for 50-kb sliding windows with a step size of 5 kb and each SNP using a variance component approach implemented in the HIERFSTAT[50] R package. The average F_{ST} of all sliding windows was considered as the value at the whole-genome level across different groups.

Sliding windows with the top 5% of F_{ST} values were selected initially. Neighboring windows were then merged into one fragment. If the distance between two fragments was <50 kb, fragments were merged into one region. These regions were regarded as highly diverged across groups. To identify the genes that are putatively under selective pressure between different groups, the nonsynonymous SNPs within the SNP list with the top 5% of F_{ST} value were selected, and the corresponding genes were then determined.

Cloning and functional analysis of cucumber β-carotene hydroxylase in *E. coli*. Full-length cDNA for *CsaBCH1*and *CsaBCH1*^{A257D} were isolated from leaf cDNA libraries using the primer BCH1_P1 (**Supplementary Table 14**), cloned into the TA cloning vector pMD19-T (TAKARA) and verified by Sanger sequencing. The two genes were amplified from the corresponding cloning constructs using the primer vect_1 (**Supplementary Table 14**) and were subcloned into the expression vector pET32a (the resulting plasmids were pET32a-*CsaBCH1* and pET32a-*CsaBCH1*^{A257D}). The plasmid pAC-BETA was kindly provided by F. Cunningham and contained all the genes for β-carotene formation[51]. Plasmid pAC-BETA was cotransformed with plasmid pET32a-*CsaBCH1* or pET32a-*CsaBCH1*^{A257D} into *E. coli* DH5α. Double transformants were selected on LB agar plates supplemented with ampicillin (100 μg/ml) (for selection of pET32a-*CsaBCH1*/pET32a-*CsaBCH1*^{A257D}) and chloramphenicol (50 μg ml−1) (for selection of pAC-BETA). Agar plates were incubated at 37 °C for 16 h and then at room temperature for 2–3 d in the dark to allow maximum color development (carotenoid accumulation). Single colonies of the double transformants were used to inoculate 50-ml aliquots of LB medium. Liquid cultures were grown at 37 °C until OD_{600} reached 0.8, and protein expression was induced with the addition of isopropyl-β-D-thiogalactopyranoside (IPTG) to the growth medium to a final concentration of 0.1 mM. Bacterial cultures were then grown at 30 °C for an additional 24 h. *E. coli* cells were harvested by centrifugation, and total carotenoids were extracted from the bacterial cell pellets according to procedures described in a previous report[52]. Chromatography was performed on a Waters ACQUITY UPLC system using an ACQUITY UPLC BEH C18 column (1.7 μm, 2.1 × 100 mm; Waters Corporation) placed in a column oven set at 30 °C. Chromatograms were collected at 450 nm using a photodiode array (PDA) detector. Mobile phase consisted of (A) acetonitrile and (B) acetone. Sample elution followed a linear gradient with the following proportion (v/v) of phase B (t, B%) used: (0, 5), (10, 40), (10.1, 90%), (12,90) and (12.1, 5). The flow rate was maintained at 0.25 ml/min, and the injection volume was 2 μl. β-carotene was identified by comparison of absorption spectrum and retention time to an authentic standard (Sigma). Multiple colonies for each coexpression experiment (pET32a-*CsaBCH1* and pAC-BETA or pET32a-*CsaBCH1*^{A257D} and pAC-BETA) were analyzed for carotenoid content (biological replicates of each experiment), and representative results are presented (**Supplementary Fig. 14**).

Expression analysis of *CsaBCH1* in cucumber fruits. The fruits of CG9164 (Xishuangbanna) and CG4210 (East Asian) accessions were harvested at

20, 30, 40, 50 and 60 d after pollination, respectively. Total RNA (10 µg) for each fruit sample was extracted. The TransScript SuperMix kit from Transgene was used for first-strand cDNA synthesis. We designed the RT–quantitative PCR (RT-qPCR) primer BCH_exp (**Supplementary Table 14**) for the *CsaBCH1* gene using Primer3. RT-qPCR reactions were conducted using the TransStart Green qPCR SuperMix kit from TransGene, and three replicates were performed. Relative expression levels were estimated by the $2^{-\Delta\Delta C_{\mathrm{T}}}$ method.

35. Murray, M.G. & Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
36. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
37. Yang, L. *et al.* Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. *Plant J.* **71**, 895–906 (2012).
38. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
39. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
40. Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 433–436 (2009).
41. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030 (2010).
42. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
43. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
44. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
45. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
46. Anderson, M.J. *PCO: A FORTRAN Computer Program for Principal Coordinate Analysis* (Department of Statistics, University of Auckland, Auckland, New Zealand, 2003).
47. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
48. Xu, Z., Kaplan, N.L. & Taylor, J.A. TAGster: efficient selection of LD tag SNPs in single or multiple populations. *Bioinformatics* **23**, 3254–3255 (2007).
49. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
50. de Meeûs, T. & Goudet, J. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infect. Genet. Evol.* **7**, 731–735 (2007).
51. Cunningham, F.X. Jr. & Gantt, E. A portfolio of plasmids for identification and analysis of carotenoid pathway enzymes: *Adonis aestivalis* as a case study. *Photosynth. Res.* **92**, 245–259 (2007).
52. Tian, L. & DellaPenna, D. Characterization of a second carotenoid β-hydroxylase gene from *Arabidopsis* and its relationship to the *LUT1* locus. *Plant Mol. Biol.* **47**, 379–388 (2001).