Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization

Cheng Qin[a,b,c,1], Changshui Yu[b,1], Yaou Shen[a,1], Xiaodong Fang[d,e,1], Lang Chen[b,1], Jiumeng Min[d,1], Jiaowen Cheng[c], Shancen Zhao[d], Meng Xu[d], Yong Luo[b], Yulan Yang[d], Zhiming Wu[f], Likai Mao[d], Haiyang Wu[d], Changying Ling-Hu[b], Huangkai Zhou[d], Haijian Lin[a], Sandra González-Morales[g], Diana L. Trejo-Saavedra[h], Hao Tian[b], Xin Tang[c], Maojun Zhao[i], Zhiyong Huang[d], Anwei Zhou[b], Xiaoming Yao[d], Junjie Cui[c], Wenqi Li[d], Zhe Chen[a], Yongqiang Feng[b], Yongchao Niu[d], Shimin Bi[b], Xiuwei Yang[b], Weipeng Li[c], Huimin Cai[d], Xirong Luo[b], Salvador Montes-Hernández[j], Marco A. Leyva-González[g], Zhiqiang Xiong[d], Xiujing He[a], Lijun Bai[d], Shu Tan[c], Xiangqun Tang[b], Dan Liu[d], Jinwen Liu[d], Shangxing Zhang[b], Maoshan Chen[d], Lu Zhang[d,k], Li Zhang[c], Yinchao Zhang[a], Weiqin Liao[b], Yan Zhang[d], Min Wang[b], Xiaodan Lv[d], Bo Wen[d], Hongjun Liu[a], Hemi Luan[d], Yonggang Zhang[b], Shuang Yang[d], Xiaodian Wang[b], Jiaohui Xu[d], Xueqin Li[b], Shuaicheng Li[k], Junyi Wang[d], Alain Palloix[l], Paul W. Bosland[m], Yingrui Li[d], Anders Krogh[e], Rafael F. Rivera-Bustamante[h], Luis Herrera-Estrella[g,2], Ye Yin[d,2], Jiping Yu[b,2], Kailin Hu[c,2], and Zhiming Zhang[a,2]

[a]Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Ministry of Agriculture, Maize Research Institute of Sichuan Agricultural University, Wenjiang 611130, China; [b]Pepper Institute, Zunyi Academy of Agricultural Sciences, Zunyi 563102, China; [c]College of Horticulture, South China Agricultural University, Guangzhou 510642, China; [d]Beijing Genomics Institute-Shenzhen, Shenzhen 518083, China; [e]Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark; [f]College of Horticulture and Landscape Architecture, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China; [g]Laboratorio Nacional de Genómica para la Biodiversidad (Langebio) del Centro de Investigación y de Estudios Avanzados (Cinvestav), Irapuato, 36821, Mexico; [h]Departamento de Ingeniería Genética, Centro de Investigación y de Estudios Avanzados del IPN (Cinvestav)-Unidad Irapuato, Irapuato, 36821, México; [i]College of Biology and Science, Sichuan Agricultural University, Ya'an 625014, China; [j]Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Campo Experimental Bajío, Celaya, 38010, México; [k]Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China; [l]INRA Provence-Alpes-Côte d'Azur, UR1052, Unité de Génétique et Amélioration des Fruits et Légumes, CS 60094, F-84140 Montfavet Cedex, France; and [m]Chile Pepper Institute, New Mexico State University, Las Cruces, NM 88003

Contributed by Luis Herrera-Estrella, January 19, 2014 (sent for review December 12, 2013)

As an economic crop, pepper satisfies people's spicy taste and has medicinal uses worldwide. To gain a better understanding of *Capsicum* evolution, domestication, and specialization, we present here the genome sequence of the cultivated pepper *Zunla-1* (*C. annuum* L.) and its wild progenitor *Chiltepin* (*C. annuum* var. *glabriusculum*). We estimate that the pepper genome expanded ∼0.3 Mya (with respect to the genome of other Solanaceae) by a rapid amplification of retrotransposons elements, resulting in a genome comprised of ∼81% repetitive sequences. Approximately 79% of 3.48-Gb scaffolds containing 34,476 protein-coding genes were anchored to chromosomes by a high-density genetic map. Comparison of cultivated and wild pepper genomes with 20 resequencing accessions revealed molecular footprints of artificial selection, providing us with a list of candidate domestication genes. We also found that dosage compensation effect of tandem duplication genes probably contributed to the pungent diversification in pepper. The *Capsicum* reference genome provides crucial information for the study of not only the evolution of the pepper genome but also, the Solanaceae family, and it will facilitate the establishment of more effective pepper breeding programs.

de novo genome sequence | genome expansion | Solanaceae evolution

Pepper (*Capsicum*) is an economically important genus of the Solanaceae family, which also includes tomato and potato. The genus includes at least 32 species native to tropical America (1), of which *C. annuum* L., *C. baccatum* L., *C. chinense* Jacq., *C. frutescens* L., and *C. pubescens* (Ruiz & Pavon) were domesticated as far back as 6000 B.C. by Native Americans (2). Peppers have a wide diversity of fruit shape, size, and color. Pungent peppers are used as spices, and sweet peppers are used as vegetables. After the return of Columbus from America in 1492 and subsequent voyages of exploration, peppers spread around the world because of adaptation to different agroclimatic regions and rapid adoption of pepper in different cultures as food, medicine, and ornamentals (3, 4). Pepper global production in 2011 reached 34.6 million tons fresh fruit and 3.5 million tons dried pods harvested in 3.9 million hectares (www.fao.org). Despite the growing commercial importance of pepper, the molecular mechanisms that modulate fruit size, shape, and yield are mostly unknown.

Since the 1990s, genetic diversity and allelic shifts among cultivars, domesticated landraces, and wild accessions have been partially explored using restricted sets of anonymous or neutral

Significance

The two pepper genomes together with 20 resequencing accessions, including 3 accessions that are classified as semiwild/wild, provide a better understanding of the evolution, domestication, and divergence of various pepper species and ultimately, will enhance future genetic improvement of this important worldwide crop.

Author contributions: C.Q., R.F.R.-B., L.H.-E., Y. Yin, J.Y., K.H., and Z.Z. designed research; C.Q., C.Y., Y.S., X.F., L.C., J. Cheng, S. Zhao, Y. Luo, Z.W., C.L.-H., H. Lin, S.G.-M., D.L.T.-S., H.T., Xin Tang, M.Z., A.Z., J. Cui, Z.C., Y.F., Y.N., S.B., X. Yang, Weipeng Li, H.C., X. Luo, S.M.-H., M.A.L.-G., Z.X., S.T., Xiangqun Tang, J.L., S. Zhang, M.C., Li Zhang, W. Liao, Yan Zhang, M.W., B.W., H. Liu, H. Luan, Yonggang Zhang, X.W., X. Li, S.L., A.P., P.W.B., A.K., R.F.R.-B., L.H.-E., J.Y., K.H., and Z.Z. performed research; Wenqi Li, L.B., X. Lv, S.Y., J.X., J.W., and Y. Li contributed new reagents/analytic tools; C.Q., J.M., J. Cheng, M.X., Y. Yang, Z.W., L.M., H.W., H.Z., Z.H., A.Z., X. Yao, J. Cui, S.B., X. Luo, S.M.-H., M.A.L.-G., X.H., Xiangqun Tang, D.L., Lu Zhang, and Yinchao Zhang analyzed data; and C.Q., X.F., S. Zhao, and L.H.-E. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The *C. annuum* cv. *Zunla-1* and *C. annuum* var. *glabriusculum* whole-genome shotgun sequences reported in this paper have been deposited in the GenBank database (accession nos. ASJU00000000 and ASJV00000000, respectively). The RNA-sequence reads and small RNA-sequence reads data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE45037, GSE45040, and GSE45154). Additional information is accessible through the Pepper Genome Database website (http://peppersequence.genomics.cn).

See Commentary on page 5069.

[1]C.Q., C.Y., Y.S., X.F., L.C., and J.M. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: lherrera@langebio.cinvestav.mx, yinye@genomics.cn, yujiping62@163.com, hukailin@scau.edu.cn, or zzmmaize@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1400975111/-/DCSupplemental.

molecular markers (5–9) and annotated DNA sequences (10). These studies reported that the genetic variability among sweet and large-fruited *C. annuum* cultivars was very restricted and suggested that changes in the allelic frequencies and a subsequent loss of diversity during the transition from wild to cultivated populations occurred even in areas of species cohabitation. The relatively low levels of genetic diversity in the primary gene pool have constrained pepper genetic improvement. Another primary reason for limited applied and basic research in pepper has been lack of a reference genome sequence of ∼3.3 Gb (11). Recent work comparing two members of the Solanaceae family (pepper and tomato) has begun to shed light on the processes that influence the dynamics of genome size in angiosperms (12, 13).

To contribute to the understanding of pepper biology and evolution and accelerate agricultural applications, we generated and analyzed two reference genome sequences of cultivated *Zunla-1* and wild *Chiltepin* ($2n = 2x = 24$). The two pepper genomes together with 20 resequencing accessions, including 3 accessions that are classified as semiwild/wild, provide a better understanding of the evolution, domestication, and divergence of various pepper species and ultimately, will enhance future genetic improvement of this important worldwide crop.

Results and Discussion

Large Genome Assembly and Chromosome Anchoring. Because of their commercial and genetic advantages, we selected the widely cultivated *C. annuum* accession *Zunla-1* and it wild progenitor *Chiltepin* for genome sequencing (*SI Appendix*, *SI Text*). Using the whole-genome shotgun approach, we generated a total of 325- and 205-Gb high-quality reads from various Illumina sequencing libraries for *Zunla-1* and *Chiltepin*, respectively (*SI Appendix*, Tables S1 and S2). As expected, the genome size of *Zunla-1* was estimated to be 3.26 Gb, which is slightly larger than the 3.07-Gb size of *Chiltepin* by *K*-mer analysis (*SI Appendix*, Fig. S1 and Table S3); estimations are consistent with a previous report (11). Short sequencing reads, corresponding to 99- and 67-fold genomic depths (*SI Appendix*, Fig. S2), were hierarchically and iteratively assembled into contigs with N50 lengths (50% of the genome is in fragments of this length or longer) of 55 and 52 kb for *Zunla-1* and *Chiltepin*, respectively (Table 1). Pair-end information was used sequentially in assembler SOAPdenovo (14) to generate scaffolds comprising 3.48- and 3.35-Gb scaffolds with N50 lengths of 1.23 Mb and 445 kb, respectively (Table 1 and *SI Appendix*, Table S4). The smaller N50 scaffold length for *Chiltepin* was primarily caused by a lower sequencing depth and the lack of 40-kb libraries. In our analysis, we refer to the *Zunla-1* assembly as a reference for the *C. annuum* genome.

We assessed the quality and coverage of the two genomes using Sanger-derived BACs and ESTs from public databases. Of 1.7-Mb sequences from 15 BACs, ∼97% could be covered by the scaffolds with identity of 0.95 and E value of 1e-20, indicating reliable local assembly (*SI Appendix*, Table S5). More than 98% of 83,029 ESTs could be aligned to the genomes by the criteria of length >200 bp and hit >97%, which showed extensive genomic coverage (*SI Appendix*, Table S6). In addition, 23 and 18 large nuclear regions matching the chloroplast genome (>2 kb and >98% sequence identity) were identified in the reference and *Chiltepin* genomes, respectively (Dataset S1). This phenomenon is similar to that observed in tomato (15) and tobacco (16), suggesting active gene transfer from the choloroplast into the nuclear genome of the Solanaceae.

The scaffolds were then anchored to 12 linkage groups by 7,657 SNP markers in our newly developed high-density genetic map (*SI Appendix*, *SI Text*) (17), and they could be assigned as chromosomes 1–12 (Chr01–Chr12) according to the cytological analysis (1, 18) (Fig. 1, track A). The pseudochromosomes consist of 4,956 scaffolds with 31,201 genes located, corresponding to 79% of the reference (Table 1 and *SI Appendix*, Fig. S3 and Table S7). It has been reported that, during domestication, chromosome translocation events differentiate cultivars from wild progenitors (19), which helped us to precisely anchor 29,081 scaffolds (2.42 Gb; 30,123 genes) of *Chiltepin* to chromosomes (Table 1 and *SI Appendix*, Table S7). We also observed S shape when the genetic and physical distances were analyzed (*SI Appendix*, Fig. S3), reflecting extensive recombination suppression around the centromeres (Fig. 1, tracks A and B). Interestingly, Chr08 showed a short terminal arm (Fig. 1, track A and *SI Appendix*, Fig. S3), supporting the conclusion that the chromosome is acrocentric (19).

Repetitive Elements and Genome Expansion. Using a combination of homology-based searches and ab initio modeling, we found that more than 81% (∼2.7 Gb) of the pepper genomes were composed of different transposable elements (TEs), which is significantly higher than TEs (∼61%) in potato and tomato (Table 1 and Dataset S2). Most of the plant TE categories were identified in pepper, including 70.3% LTR retrotransposons and 4.5% DNA transposons (Table 1). Clearly, LTR retrotransposons contributed more to the genome expansion than those in potato (47.2%), tomato (50.3%), and grape (46.2%), which parallels the genomic topology of the maize genome (75%) (20).

Table 1. Comparison of features of pepper, tomato, and potato genomes

Genome features	Cultivated pepper	Wild pepper	Tomato*	Potato†
Assembled genome size (Mb)‡	3,349	3,480	760	727
Number of scaffolds§	967,017	1,973,483	NA	NA
Contig N50 (bp)¶	55,436	52,229	NA	NA
Scaffold N50 (bp)¶	1,226,833	445,585	NA	NA
GC content (%)	34.9	35.0	34.0	34.8
Repeat rate (%)	80.9	81.4	61.3	61.6
LTR rate (%)	70.3	70.1	50.3	47.2
Predicted protein-coding genes	35,336	34,476	33,726	38,492
Average gene length (bp)	3,363	3,235	3,006	2,476
Average CDS length (bp)	1,020	1,006	1,063	928
Average exon number per gene	4.27	4.04	4.6	3.49
Sequence anchored on chromosome (%)	78.95	69.68	NA	NA
Genes anchored on chromosome (%)	88.29	87.37	NA	NA

NA, not available; GC, guanine-cytosine; CDS, coding DNA sequence.
*Modified from ref. 15.
†Modified from ref. 27.
‡The fragments of the ungapped genome assembly.
§The length shorter than 100 bp was not included in the statistics.
¶N50 values of the genome assembly were calculated using the fragments longer than 100 bp.

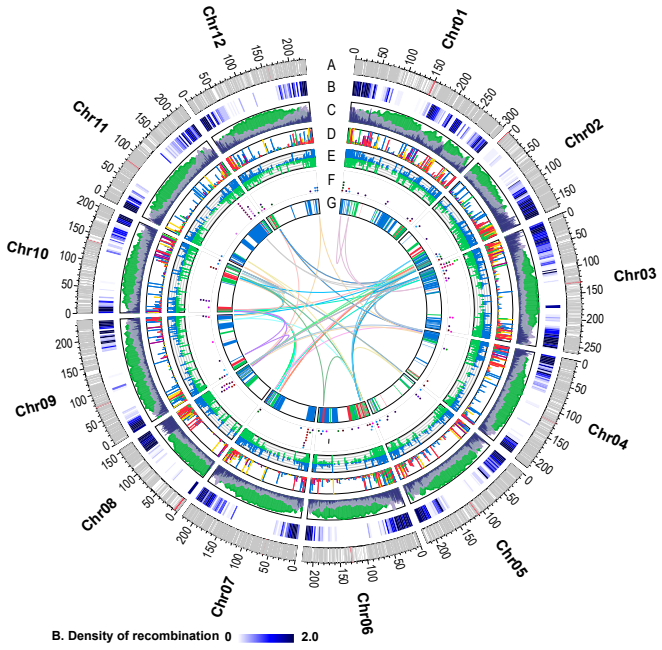

Fig. 1. Global view of the pepper genome. Track A denotes the 12 pseudochromosomes of pepper (megabases). The positions of the effective markers in the genetic map are shown as vertical gray lines. The loci of inferred centromeres are denoted by vertical red bars. Track B shows density of recombination. Track C shows density distribution of *Gypsy* (green), *Copia* (light blue), and protein-coding genes (navy). Track D shows distribution of tissue-specific expression genes, including root (red), stem (green), leaf (dark magenta), flower (blue), and fruit (gold). Track E shows genome-wide distribution of total small RNA loci (blue and green lines). The histograms plot small RNA reads from 20 to 25 nt, and they were normalized to account for the appearance of opposite strand inverse sequences. Track F shows distribution of the identified miRNA families denoted by different colors (Dataset S6). Track G shows connections of the triplicate loci denoted by different colors (Dataset S16).

The most abundant LTR retrotransposons were the *Gypsy* clade (54.5%) followed by *Copia* (8.6%) (Fig. 1, track C and Dataset S2). This scenario is quite different from some monocots, such as wheat (21, 22), in which the *Copia* clade is usually the predominant component of repetitive DNA. In the TEs identified, 23.1% and 16.2% are ancestral repeats that predate the divergence of pepper with tomato and potato, respectively (Dataset S2), whereas other lineage-specific TEs emerged during the genome expansion and account for 50.8% of the pepper genome (Dataset S3 and *SI Appendix*, Table S10).

To investigate the genome expansion event in pepper, we dated the insertion time of all LTRs based on divergence analysis (23). A peak of increased insertion activity was found ∼0.3 Mya (*SI Appendix*, Fig. S4*A*), suggesting that the expansion of the pepper genome was quite recent during the evolution of the Solanaceae family. Analysis of the insertion time and phylogenetic topology of *Copia* and *Gypsy* clades also supported this conclusion (*SI Appendix*, Figs. S4*B* and S5). Obviously, *Gypsy* had the highest insertion activity recently after Solanaceae species divergence, which made it the most abundant in pepper genome.

Gene Annotation and Transcription. To facilitate gene annotation, we generated 90.5-Gb RNA sequencing (RNA-Seq) data from 30 libraries representing all primary developmental stages and tissue types, including various fruits (Dataset S4). A combination of evidence-based and de novo approaches predicted 35,336 and 34,476 high-confidence protein-coding loci in the reference and *Chiltepin* genomes, respectively (*SI Appendix*, Table S9); over 90% of predicted genes were supported by ESTs, RNA-Seq entries, or homologous proteins (*SI Appendix*, Fig. S6). Gene density is relatively low surrounding centromeres where the TEs are inversely high, indicating that the repetitive sequences are unevenly scattered along chromosomes (Fig. 1, track C). For instance, the *Gypsy* clade filled in the gene-sparse deserts of the genome, but in contrast, the *Copia* elements usually accompanied genes in regions that exhibited high recombination rates.

We also obtained 2,717,180 unique tags by sequencing the flower buds and identified 6,527 long noncoding (lnc) RNAs by a self-developed program (Dataset S5). Among lnc-RNAs, 5,976 are intergenic, 222 are intron-overlapping, and the others are bidirectional. Sequencing of small RNAs from five different tissues allowed the identification of 5,581 phased siRNAs (Fig. 1, track E and *SI Appendix*, Table S10). Based on the plant microRNAs (miRNAs) miRBase database, 176 miRNAs were discovered in pepper and classified into 64 families (Dataset S6). Comparison with miRNAs of other Solanaceae members and plant species showed that 141 miRNAs are conserved and 35 miRNAs are specific to pepper (Fig. 1, track F and Dataset S6). We predicted 1,104 target genes for these miRNAs, of which 78% have putative functions (Dataset S7). Significantly, about one-half of the pepper miRNA families potentially plays an important role in posttranscriptional regulation by targeting mRNAs encoding transcription factors (TFs) (Dataset S8). In addition, target gene *Dihydrolipoamide dehydrogenase* (*Capana12g000245*) of can-miR5303 and *α-CT* (*Capana09g001602*), which are part of the capsaicinoid biosynthetic pathway, are potential targets of miRNAs (Dataset S7), suggesting the regulation of capsaicinoid biosynthesis by miRNAs. Overall, miRNA target genes are involved in a wide spectrum of regulatory functions and biological processes, including apoptosis, defense responses, and ATP binding (Dataset S9).

RNA-Seq expression profiles showed that over 31% of the protein-coding genes were constitutively expressed in the various tissues examined. We also identified 3,670 tissue-specific genes distributed in root (740), stem (113), leaf (197), fruit (835), and flower (1,785) (Fig. 1, track D). In blooming flowers, 599 tissue-specific genes were exclusively expressed ($P < 0.001$) and mainly involved in cell construction (enzyme regulator and inhibitor activity, pectinesterase activity, or cell wall and cytoskeleton modification) (Dataset S10).

Insights into Solanaceae Evolution. Sequence-based analysis of pepper gene families was conducted using OrthoMCL (24) and compared with those families in tomato, potato, and *Arabidopsis* (*SI Appendix*, Table S11). We identified 10,279 gene families shared among the four species and a total of 17,671 in pepper with more than one orthologous gene (*SI Appendix*, Fig. S7). Another 1,257 gene families, containing 3,143 genes, were specific to the pepper genome (Dataset S11). These pepper-specific genes have various biological functions; however, they are particularly overrepresented in the gene ontology category of biotic stimulus, indicating that the pepper has rapid and strong response to better face fluctuating environmental conditions (Dataset S12).

In total, 5,231 single copy orthologous genes identified in grape, papaya, pepper, tomato, potato, and *Arabidopsis* were used to construct a phylogenetic tree (Fig. 2 *A* and *B*). It showed that pepper separated from tomato and potato ∼36 Mya, during which time the *Capsicum* genus evolved in Solanaceae. We also observed that Solanaceae appeared nearly 156 Mya, very soon after the differentiation of monocots from dicots (15, 25). Approximately 38-Mb genomic sequences of pepper can be aligned to potato and tomato with 14% nucleotide divergence, whereas only 9.76% nucleotide divergence was detected within 106-Mb synteny regions between potato and tomato with the same approach previously described (15) (*SI Appendix*, Table S12).

In the pepper genome, we identified 1,052 and 799 large syntenic blocks, involving 12,601 and 10,596 genes compared with tomato and potato, respectively (Datasets S13–S15). However, 612 and 430 chromosomal translocation events occurred during the divergence of *Capsicum* relative to tomato and potato, respectively (Dataset S13). These translocations are distributed extensively

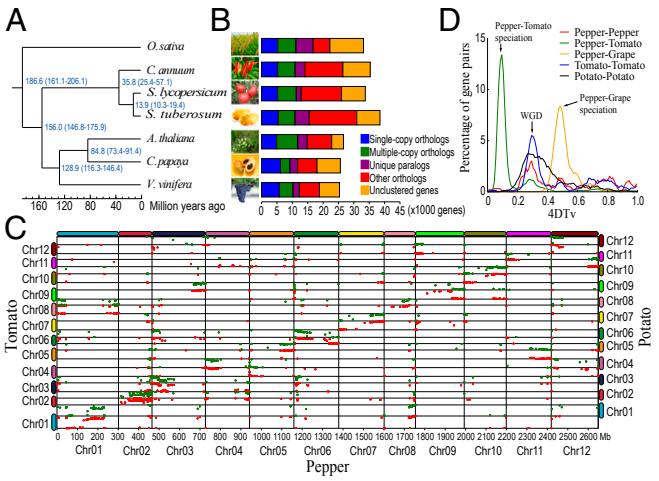

Fig. 2. Comparative analysis and evolution of the pepper genome. (*A*) Genomic differences among *C. annuum*, *Solanum lycopersicum*, *Solanum tuberosum*, *Arabidopsis thaliana*, *Carica papaya*, *Vitis vinifera*, and *Oryza sativa*. Neighbor-joining phylogenetic analysis was performed with orthologous genes and all coding DNA sequence (CDS) in *C. annuum* and the other six plants. (*B*) Clusters of orthologous and paralogous gene families in the seven plant species identified by OrthoMCL. (*C*) Syntenic blocks in the cultivated pepper, tomato, and potato show that genome rearrangements have occurred among these taxa. (*D*) Genome duplication in dicot genomes (pepper, tomato, potato, and grape) revealed by 4DTv analyses.

on all pepper chromosomes, providing evidence for generalized chromosomal rearrangements (Fig. 2*C*, Datasets S14 and S15, and *SI Appendix*, Fig. S8). The following translocations were proposed to happen between pepper and the common ancestor of tomato and potato: Chr01 vs. Chr01/Chr08, Chr03 vs.Chr03/Chr09, Chr04 vs. Chr02/Chr04, Chr05 vs. Chr04/Chr05, Chr08 vs. Chr01/Chr08, Chr09 vs. Chr09/Chr12, Chr11 vs. Chr05/Chr11, and Chr12 vs. Chr11/Chr12 [supporting previous reports (19, 26) with more precise details]. Meanwhile, 468 and 367 inversions were identified in pepper compared with tomato and potato, respectively (Datasets S14 and S15). In addition, comparison with the grape genomes revealed that a whole-genome triplication happened in the pepper genome, suggesting a common event among the Solanaceae (15) (Fig. 1, track G and *SI Appendix*, Table S13). Considerable gene loss of one or two copies of duplicated genes occurred after the triplication, resulting in few remaining triplicated genes in the pepper genome (Datasets S16 and *SI Appendix*, Table S14).

We then calculated the time of whole-genome duplication (WGD) events in Solanaceae lineages based on the distribution of distance–transversion rate at fourfold degenerate sites (4DTv methods) of paralogous gene pairs (Fig. 2*D*). Peaks at around 0.48 and 0.1 elaborated that the ancestral pepper–grape and pepper–tomato divergences occurred ∼89 and 20 Mya (15, 27), respectively; these findings are consistent with the phylogenetic analysis. The peak at ∼0.3 proved a recent WGD in the ancestral pepper–tomato lineage (15). As observed, there is no evidence of *Capsicum*-specific WGD after the pepper–tomato/pepper–potato divergence, again confirming the notion that proliferation of TEs primarily contributed to pepper genome expansion.

Molecular Footprints of Artificial Selection. Artificial selection, involved in two breeding processes of early domestication and modern intensive improvement (28), played an important role in the origin of cultivated peppers. We selected 18 cultivated accessions representing the major varieties of *C. annuum* and two semiwild/wild peppers for whole-genome resequencing (*SI Appendix*, Table S15). After alignment of the sequencing reads corresponding to 10- to 30-fold depth to the reference (*SI Appendix*, Table S16), we identified an average of 9,826,526 single nucleotide variations and 237,509 small insertions/deletions (*SI Appendix*, Table S17). As expected, the wild accessions possessed higher genetic diversity than the cultivars (*SI Appendix*, Table S17). The neighbor-joining tree and population structure further revealed that the wild and domesticated peppers are genetically distinguishable at an overall genomic level (*SI Appendix*, Figs. S9 and S10).

We next scanned the genome of these accessions to identify genome-wide signatures of artificial selection using the genetic bottleneck approach (29). To detect the reduction of genetic diversity of the pepper population caused by domestication, we used a sliding window strategy to estimate θ_π- and θ_w-values (Fig. 3*A* and *SI Appendix*, Fig. S11). The regions that showed significantly lower θ_π (Z test, $P < 0.005$) and θ_w (Z test, $P < 0.005$) in cultivars relative to the wild group were considered as potential artificial selection regions (Fig. 3*B*). We identified a total of 115 regions with strong selective sweep signals in the cultivated peppers (85.2 Mb or 2.6% of the genome and containing 511 genes) (Dataset S17 and *SI Appendix*, Fig. S12). The length of these selected regions ranged from 0.3 to 61.9 kb, and the polymorphism levels of these selected regions relative to the whole genome were relatively low (Fig. 3*B*), indicating that these regions seemed to have been affected by selection during domestication.

In total, 511 genes embedded in selected regions for domestic peppers were related mainly to transcription regulation, stress, and/or defense response, protein–DNA complex assembly, growth, and fruit development (Datasets S18 and S19). Of these genes, 34 TFs, including activating protein (AP2), ethylene-responsive-element-binding factor (ERF), and basic helix-loop-helix (bHLH) families, and 10 disease resistance protein containing the NB-ARC domain were identified (Dataset S20). This set of genes may contribute to the morphological and physiological differences between cultivated and wild peppers. For example, *Capana11g001329*, a homolog of the tomato gene (*Solyc05g005680*) encoding a Xyloglucan endotransglucosylase/hydrolase (XTH), was identified in our putative artificial selection genes (*SI Appendix*, Fig. S13). *Solyc05g005680* showed significantly differential expression during fruit ripening (15), whereas *Capana11g001329* was only expressed in early growing stages, suggesting that the gene may account for nonclimacteric fruits with a slower softening process (discussed below). The gene *Capana09g001426* is homologous to the rice *Rc* gene (30), which was a well-known domestication gene and thought to be associated with seed dormancy and pericarp color in rice (31). The region containing *Capana09g001426* showed a very strong selective sweep signal in the domesticated pepper genome (*SI Appendix*, Fig. S13). This finding suggested that the gene might play a role in shortening seed dormancy, a trait expected to be under strong artificial selection during domestication. We also identified the three genes *PepEST*, *CALTPI*, and *RGA15* (*Capana04g001148*, *Capana10g001225*, and *Capana01g004043*, respectively) that enhanced pepper resistance to pathogen and environmental stresses (32–34) (*SI Appendix*, Fig. S13).

Comparison of Fruit Development Between Pepper and Tomato. The ripening process greatly influences fruit quality and shelf life and differs significantly between climacteric fruits, such as tomato, and nonclimacteric fruits, such as pepper, which have a slower softening process and no response to ethylene (35) (*SI Appendix*, Fig. S14). We compared gene expression profiles between tomato and pepper during fruit ripening. Tomato had 2,281 differential genes, whereas pepper had 1,440 differential genes (Datasets S21 and S22), including in both cases, genes involved in cell wall remodeling, hormone signaling and metabolism, carbohydrate metabolism, protein degradation, and abiotic stress responses. However, important differences were identified. For instance, the number of genes involved in ethylene biosynthesis was lower in pepper (Datasets S21 and S22); zero of eight pepper genes encoding 1-aminocyclopropane-1-carboxylate synthase (the key enzyme in ethylene production) were up-regulated during ripening, whereas two 1-aminocyclopropane-1-carboxylate synthase genes were strongly induced in tomato, consistent with lower ethylene production in pepper (36). Similarly, the number of differentially expressed genes related to ethylene signaling and jasmonic acid

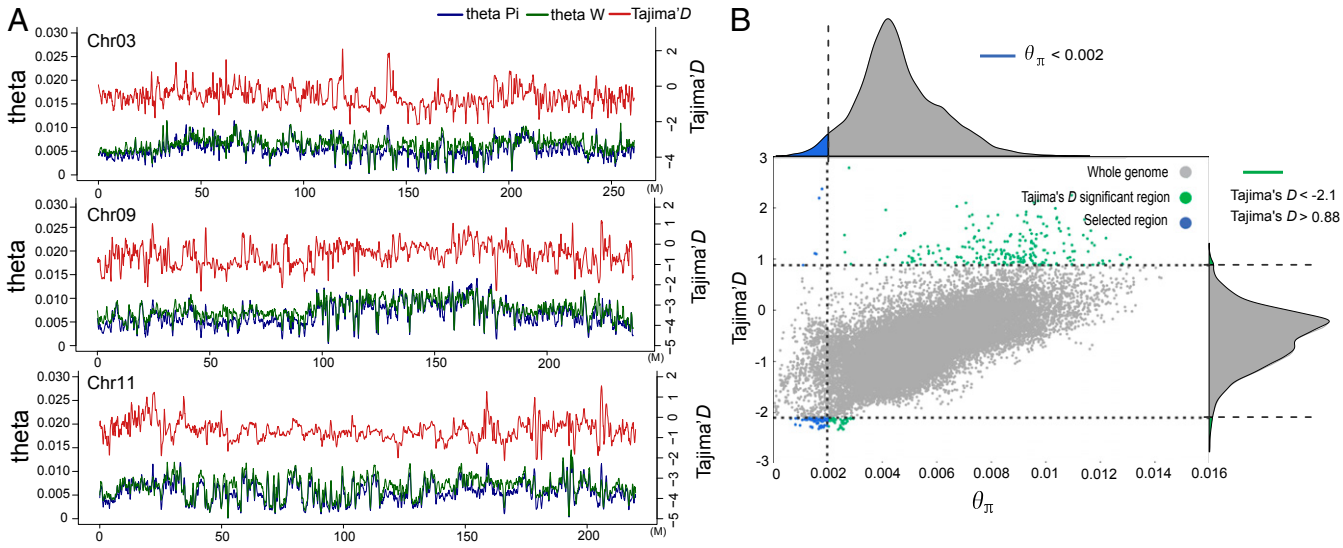

Fig. 3. Diversity in domesticated pepper population. (*A*) Diversity metrics (θ_π, θ_w, and Tajima *D*) are shown for 19 domesticated varieties across Chr03, -09, and -11. (*B*) Illustration of the strategy for candidate selection regions. The gray region above the *x* axis corresponds to regions with 0.5% significance level of diversity difference.

production and signaling was lower in pepper, whereas the number of differentially expressed auxin- and abscisic acid-related genes, including those involved in abiotic stress, was greater in pepper (Datasets S23 and S24), which is consistent with abscisic acid accumulation during ripening of strawberry, another nonclimacteric fruit (37). Interestingly, the negative regulators of leaf senescence, WRKY70 and ZAT10, suffer a stronger induction in pepper but not tomato (Datasets S23 and S24), suggesting that induction of these TFs might play an important role in the longer shelf life of peppers. Additionally, 15 of 39 tomato XTH genes showed differential expression during fruit ripening, whereas only 6 of 25 XTH genes in pepper had altered expression (Datasets S23 and S24). We suggest that the reduced level of XTH activity accounts for less softening of pepper fruit during ripening.

Evolution of Genes Involved in Capsaicin Synthesis. Capsaicinoid accumulation, which mainly consists of capsaicin and dihydrocapsaicin, is exclusive to *Capsicum* and responsible for the fruits' pungency (38). Based on previous studies on pepper pungency (39–41), we identified 51 gene families involved in capsaicinod biosynthesis in pepper and their orthologs in tomato, potato, and *Arabidopsis* (Datasets S25 and S26 and *SI Appendix*, Fig. S15). Phylogenetic analysis showed that pepper had independent pepper-specific duplications in 13 gene families compared with the other three species (such as ACLd, AT3, β-CT, C3H, CAD, CCR, Kas I, and PAL gene families) (Fig. 4*A* and *SI Appendix*, Fig. S16). The sequence divergence among gene duplications could have led to diverged functions or neofunctionalization (42), promoting the evolution of specialized capsaicinoid biosynthesis. Taking *AT3* as an example, we identified three tandem copies of At3 (Pun1) gene in pepper, which encodes a putative acyltransferase and acts as a regulator of pungency in certain *Capsicum spp*. (Fig. 4*A* and *SI Appendix*, Fig. S17*A*) (40, 41). Both AT3-D1 and AT3-D2 in wild and cultivated peppers have an amino acid substitution (K390R) in the conserved DFGWGKP motif (Fig. 4*B*). Analysis of *AT3-D1* indicated that the *pun1* allele (C locus) had a 2,724/2,930-bp deletion in nonpungent genotypes spanning the putative promoter and the first exon as reported previously (*SI Appendix*,

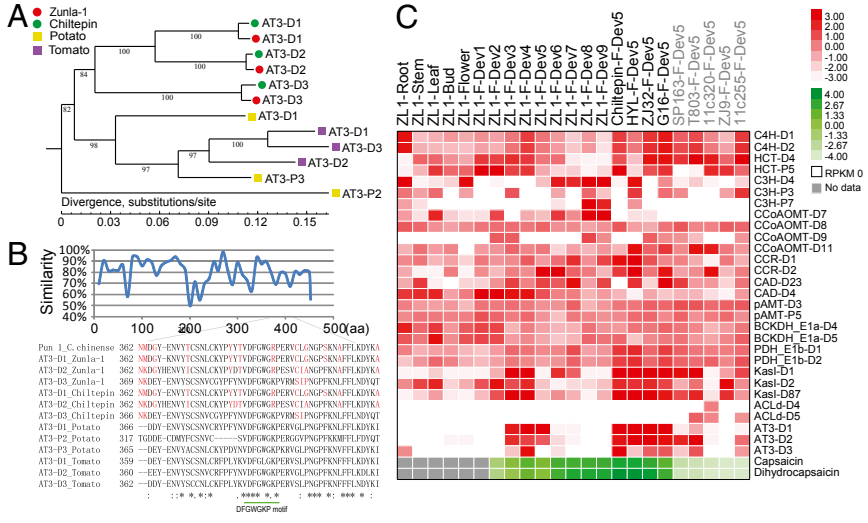

Fig. 4. Putative acyltransferase (AT3) genes for pungency and gene expression patterns of selected tissues and genes. (*A*) Phylogenetic analysis of the AT3 gene family among *Zunla-1*, *Chiltepin*, *Arabidopsis*, potato, and tomato. (*B*) The comparison of AT3 protein sequences among pepper, tomato, and potato. The nucleotide similarity of these proteins is shown in *Upper*, and the alignment is shown in *Lower*. The aligned codon sequences of an amino acid with zero, one, and two mutants among pepper, tomato and potato are marked by "*", ".", and ":", respectively. (*C*) The concentrations of capsaicin and dihydrocapsaicin, gene expression patterns in selected tissues of *Capsicum* species, and genes involved in capsaicin synthesis are shown. The green background denotes secondary metabolites (capsaicin and dihydrocapsaicin), and the red background indicates genes that are expressed in selected tissues of *Capsicum* species. RPKM, reads per kb per million mapped reads. Additionally, the gray background indicates missing data. *Zunla-1*, *Chiltepin*, HYL, JZ32, and G16 are pungent peppers; SP163, T803, 11c320, ZJ9, and 11c255 are nonpungent peppers. A model of capsaicin synthesis illustrating secondary metabolite and genes encoding enzymes is showed in *SI Appendix*, Fig. S15. Full names of genes encoding enzymes are shown in Dataset S26.

Fig. S17*B*) (40, 41). We also identified short insertions/deletions and nonsynonymous single base substitutions in both *AT3-D1* and *AT3-D2* in pungent domesticated peppers compared with *Chiltepin* (*SI Appendix*, Fig. S17 *B* and *C*).

When the tissue-specific and developmental expressions of genes involved in capsaicinod biosynthesis were examined, most gene families, except *ACL-D4* and *ACL-D5*, exhibited tissue- and stage-specific expressions accompanying gradual accumulation of capsaicinoids (Fig. 4*C*). However, *CCoAOMT-D9*, *AT3-D1*, and *AT3-D2* were only significantly expressed during the fruit developmental stages in which capsaicinoids were synthesized. We also carried out expression analysis of these expanded genes in five nonpungent peppers, which showed that the expression of *AT3-D1* was either undetectable or in trace amounts (Fig. 4*C*); this lack of expression may be caused by the large deletion in the *pun1* allele, which made it a pseudogene in nonpungent peppers. More interestingly, the expression of *AT3-D2* could probably keep the trace amount of capsaicin and dihydrocapsaicin detected in nonpungent peppers (Fig. 4*C*). We conclude that the dosage compensation effect by *AT3-D2* (*Capang02g002091*) and *AT3-D1* (*Capang02g002092*) in locus C (43) shaped the pungent diversification in peppers.

Conclusion

In this study, we sequenced, de novo assembled, and extensively annotated the genome of one of the most important vegetable crops (namely, the *Capsicum* genome). We characterized its genome structure and proposed that its large genome size is because of LTR expansion. We also annotated the genome with a wealth of transcriptome data, which include information on mRNAs, miRNAs, siRNAs, and lnc-RNAs. Importantly, RNA-Seq analysis facilitated annotation and allowed us to evaluate candidate genes for various traits. We also performed comparative analyses with other sequenced Solanaceae species and analyses of potentially key genes involved in pepper artificial selection, which will provide a resource for genetic improvement and breeding programs.

Materials and Methods

The inbred pepper cultivar *Zunla-1* (*C. annuum* L.) is an improved F_9 inbred line derived from a cross between two *C. annuum* cultivars grown by small farmers near the towns of Shanbao and Xinzhou (both in Zunyi County, Guizhou Province, China). *Chiltepin* (*C. annuum* var. *glabriusculum*) is a wild pepper landrace grown in northcentral Mexico near the El Patol municipality in the state of Queretaro. *SI Appendix* details the sequencing, assembly, annotation, and genome analysis.

ACKNOWLEDGMENTS. We appreciate the comments from Guangtang Pan, Manyuan Long, Dan Meckenstock, Doreen Ware, and Kevin T. Bilyk. We acknowledge financial support from Zunyi City and Zunyi Academy of Agricultural Sciences Natural Science Foundation of China Grant 201201 and Guangdong Natural Science Foundation of China Grant S2011030001410.

1. Moscone EA, et al. (2007) The evolution of chili peppers (*Capsicum* – Solanaceae): A cytogenetic perspective. *Acta Hortic* 745:137–170.
2. Perry L, et al. (2007) Starch fossils and the domestication and dispersal of chili peppers (*Capsicum spp*. L.) in the Americas. *Science* 315(5814):986–988.
3. Bosland PW, Votava EJ (2012) Peppers: Vegetable and spice *Capsicums*. *Crops Prod Sci Hortic* 12:1–11.
4. Hayman M, Kam PCA (2008) Capsaicin: A review of its pharmacology and clinical applications. *Curr Anaesth Crit Care* 19(5-6):338–343.
5. Votava E, Nabhan G, Bosland P (2002) Genetic diversity and similarity revealed via molecular analysis among and within an in situ population and ex situ accessions of chiltepín (*Capsicum annuum* var. *glabriusculum*). *Conserv Genet* 3(2):123–129.
6. Tam SM, et al. (2009) LTR-retrotransposons Tnt1 and T135 markers reveal genetic diversity and evolutionary relationships of domesticated peppers. *Theor Appl Genet* 119(6):973–989.
7. González-Jara P, Moreno-Letelier A, Fraile A, Piñero D, García-Arenal F (2011) Impact of human management on the genetic variation of wild pepper, *Capsicum annuum* var. *glabriusculum*. *PLoS One* 6(12):e28715.
8. Pacheco-Olvera A, Hernández-Verdugo S, Rocha-Ramírez V, González-Rodríguez A, Oyama K (2012) Genetic diversity and structure of pepper (*Capsicum Annuum* L.) from northwestern Mexico analyzed by microsatellite markers. *Crop Sci* 52(1):231–241.
9. Nicolaï M, Cantet M, Lefebvre V, Sage-Palloix A-M, Palloix A (2013) Genotyping a large collection of pepper (*Capsicum* spp.) with SSR loci brings new evidence for the wild origin of cultivated *C. annuum* and the structuring of genetic diversity by human selection of cultivar types. *Genet Resour Crop Evol* 60(8):2375–2390.
10. Aguilar-Meléndez A, Morrell PL, Roose ML, Kim S-C (2009) Genetic diversity and structure in semiwild and domesticated chiles (*Capsicum annuum*; Solanaceae) from Mexico. *Am J Bot* 96(6):1190–1202.
11. Moscone EA, et al. (2003) Analysis of nuclear DNA content in Capsicum (Solanaceae) by flow cytometry and Feulgen densitometry. *Ann Bot (Lond)* 92(1):21–29.
12. Park M, et al. (2011) Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. *BMC Genomics* 12(1):85.
13. Park M, et al. (2012) Evolution of the large genome in Capsicum annuum occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. *Plant J* 69(6):1018–1029.
14. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272.
15. Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–641.
16. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5(2):123–135.
17. Cheng J, et al. (2014) Genome-wide development and application of single nucleotide polymophism markers for genetic map construciton and assoation studies for yield-related traits in pepper (*Capsicum annuum* L.). *PLoS Genet*, in press.
18. Pochard E (1970) Description of trisomic individuals of *Capsicum annuum* L. obtained in progeny of a haploid plant. *Ann Amel Plantes* 20:233–256.
19. Wu F, et al. (2009) A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus Capsicum. *Theor Appl Genet* 118(7):1279–1293.
20. Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.
21. Jia J, et al. (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496(7443):91–95.
22. Ling HQ, et al. (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* 496(7443):87–90.
23. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20(1):43–45.
24. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178–2189.
25. Bolot S, et al. (2009) The 'inner circle' of the cereal genomes. *Curr Opin Plant Biol* 12(2):119–125.
26. Livingstone KD, Lackney VK, Blauth JR, van Wijk R, Jahn MK (1999) Genome mapping in capsicum and the evolution of genome structure in the solanaceae. *Genetics* 152(3):1183–1202.
27. Xu X, et al. (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355):189–195.
28. Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. *Nature* 457(7231):843–848.
29. Li YH, et al. (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* 14(1):579.
30. Sweeney MT, Thomson MJ, Pfeil BE, McCouch S (2006) Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* 18(2):283–294.
31. Gu X-Y, et al. (2011) Association between seed dormancy and pericarp color is controlled by a pleiotropic gene that regulates abscisic acid and flavonoid synthesis in weedy red rice. *Genetics* 189(4):1515–1524.
32. Wan H, et al. (2012) Analysis of TIR- and non-TIR-NBS-LRR disease resistance gene analogous in pepper: Characterization, genetic variation, functional divergence and expression patterns. *BMC Genomics* 13(1):502.
33. Ko MK, et al. (2005) A Colletotrichum gloeosporioides-induced esterase gene of nonclimacteric pepper (Capsicum annuum) fruit during ripening plays a role in resistance against fungal infection. *Plant Mol Biol* 58(4):529–541.
34. Jung HW, Kim W, Hwang BK (2003) Three pathogen-inducible genes encoding lipid transfer protein from pepper are differentially activated by pathogens, abiotic, and environmental stresses. *Plant Cell Environ* 26(6):915–928.
35. Giovannoni J (2001) Molecular biology of fruit maturation and ripening. *Annu Rev Plant Physiol Plant Mol Biol* 52:725–749.
36. Villavicencio L, Blankenship SM, Sanders DC, Swallow WH (1999) Ethylene and carbon dioxide production in detached fruit of selected pepper cultivars. *J Am Soc Hortic Sci* 124(4):402–406.
37. Symons GM, et al. (2012) Hormonal changes during non-climacteric ripening in strawberry. *J Exp Bot* 63(13):4741–4750.
38. Aza-González C, Núñez-Palenius HG, Ochoa-Alejo N (2011) Molecular biology of capsaicinoid biosynthesis in chili pepper (Capsicum spp.). *Plant Cell Rep* 30(5):695–706.
39. Mazourek M, et al. (2009) A dynamic interface for capsaicinoid systems biology. *Plant Physiol* 150(4):1806–1821.
40. Stewart CJ, Jr., et al. (2005) The Pun1 gene for pungency in pepper encodes a putative acyltransferase. *Plant J* 42(5):675–688.
41. Stewart C, Jr., Mazourek M, Stellari GM, O'Connell M, Jahn M (2007) Genetic control of pungency in C. chinense via the Pun1 locus. *J Exp Bot* 58(5):979–991.
42. Walsh JB (1995) How often do duplicated genes evolve new functions? *Genetics* 139(1):421–428.
43. Deshpande RB (1935) Studies in Indian chillies: 4. Inheritance of pungency in Capsicum annuum L. *Indian J Agric Sci* 5:513–516.