Supplementary Information for

Long-read bitter gourd (*Momordica charantia*) genome and the genomic architecture of non-classic domestication

Hideo Matsumura[&*1], Min-Chien Hsiao[&2], Ya-Ping Lin[2], Atsushi Toyoda[3], Naoki Taniai[4], Kazuhiko Tarora[4], Naoya Urasaki[4], Shashi S. Anand[2], Narinder P. S. Dhillon[5], Roland Schafleitner[6], Cheng-Ruei Lee*[278]

1. Gene Research Center, Shinshu University, Ueda, Nagano, Japan
2. Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan
3. National Institute of Genetics, Mishima, Shizuoka, Japan
4. Okinawa Prefectural Agricultural Research Center, Itoman, Okinawa, Japan
5. World Vegetable Center East and Southeast Asia/Oceania, Kasetsart University, Kamphaeng Saen, Nakhon Pathom, Thailand
6. The World Vegetable Center, Tainan, Taiwan
7. Institute of Plant Biology, National Taiwan University, Taipei, Taiwan
8. Genome and Systems Biology Degree Program, National Taiwan University, Taipei, Taiwan

& These authors contribute equally
* Author of correspondence
Cheng-Ruei Lee (chengrueilee@ntu.edu.tw)
Hideo Matsumura (hideoma@shinshu-u.ac.jp)

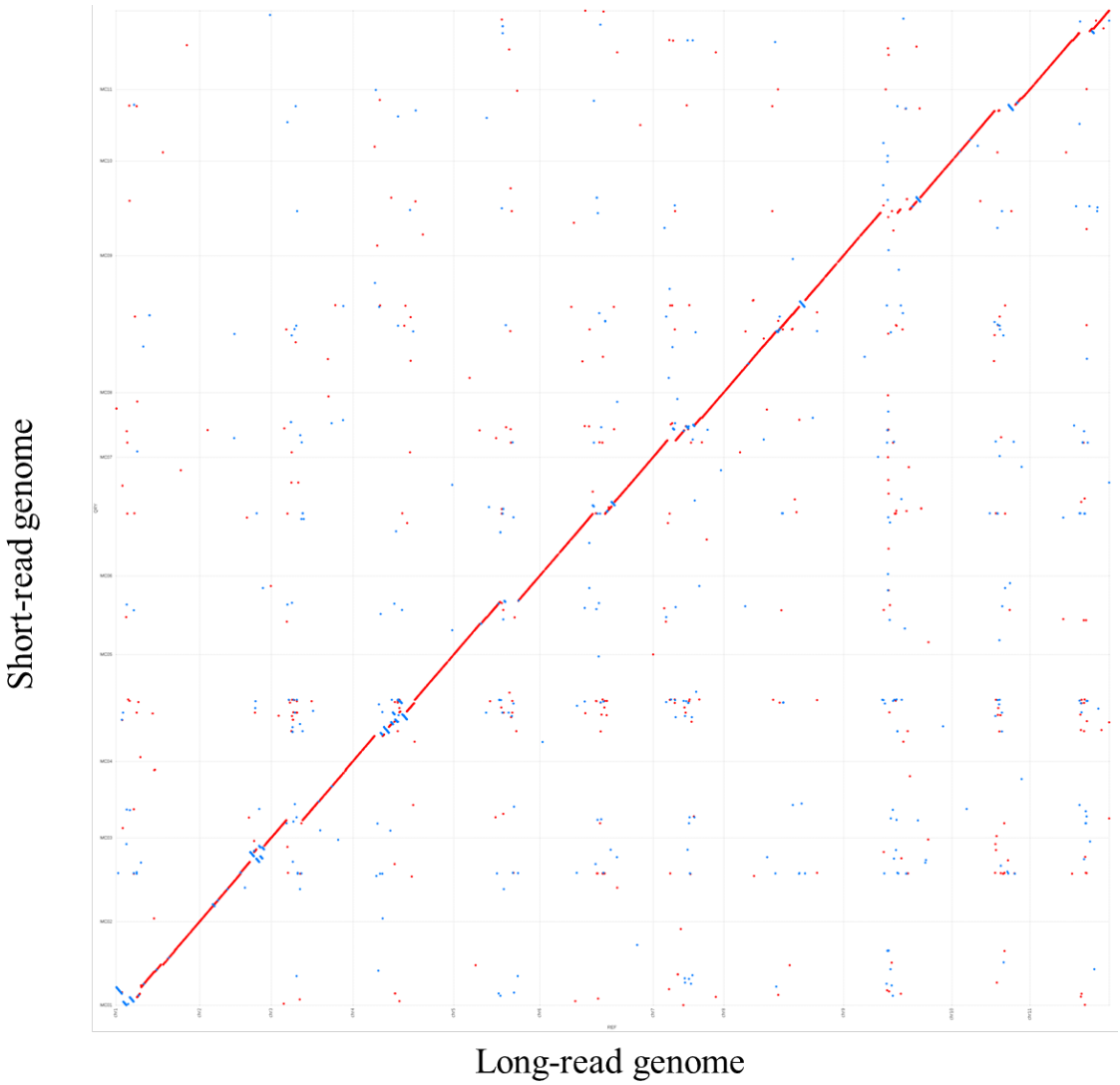
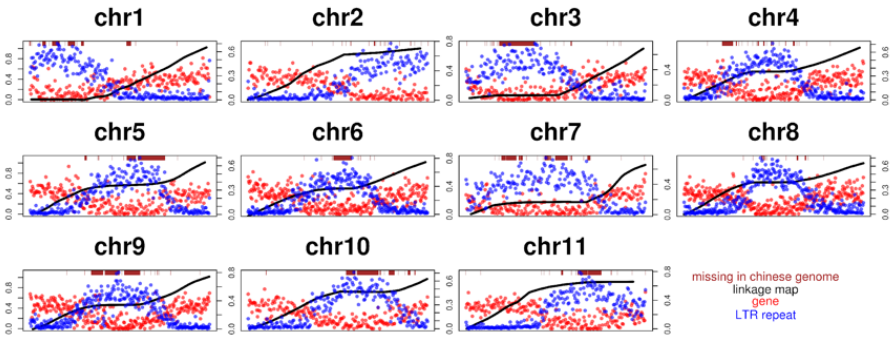
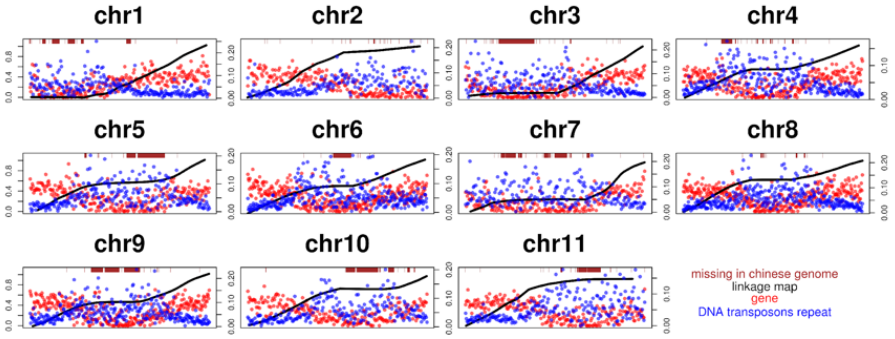


Fig. S1. Alignment of two versions of *M. charantia* genome. Short-read genome was the Dali-11 genome, and long-read genome was the PacBio assembly in this study. Red lines indicate alignments with identical orientation, whereas blue lines indicate the invert orientation between the two genomes.

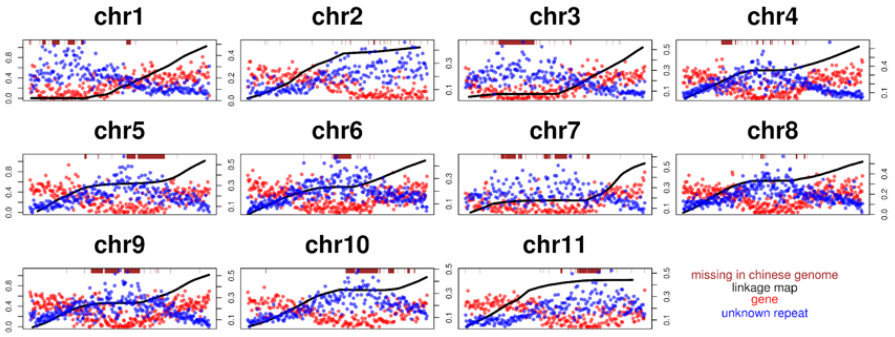
(A)


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
missing in chinese genome
linkage map
gene
LTR repeat

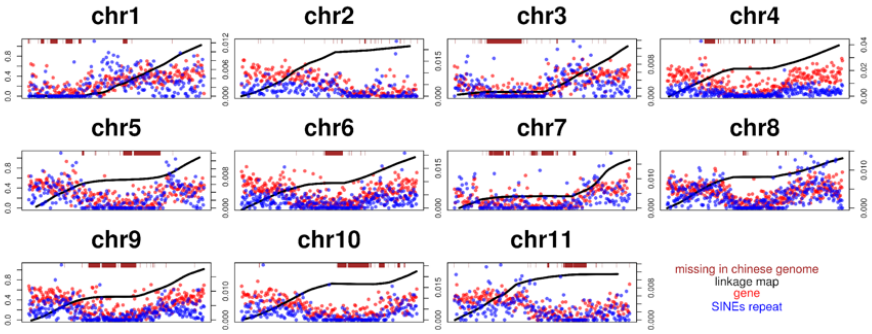
(B)


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
missing in chinese genome
linkage map
gene
DNA transposons repeat

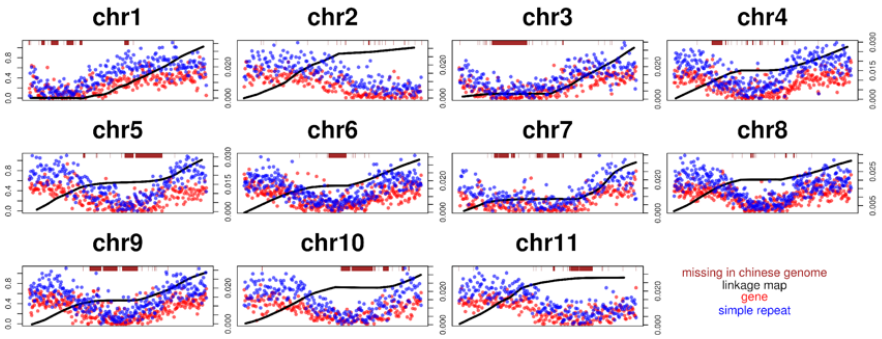
(C)


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
missing in chinese genome
linkage map
gene
unknown repeat

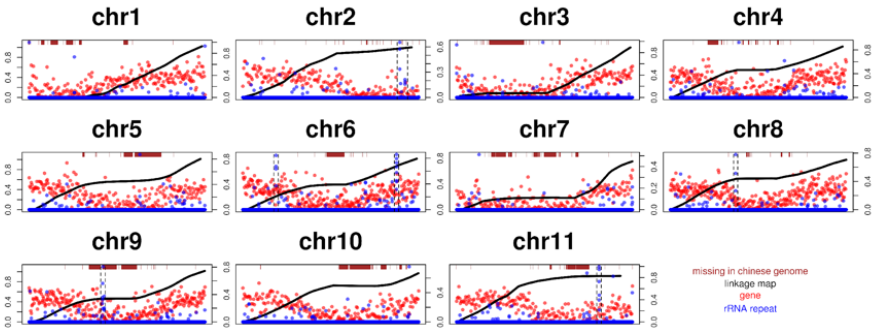
(D)


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
missing in chinese genome
linkage map
gene
SINEs repeat

(E)



(F)



(G)

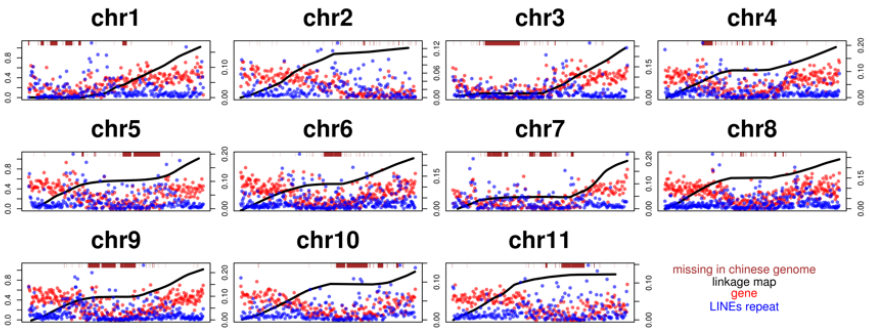


Fig. S2. Genomic distributions of different types of repetitive elements: (A) Long terminal repeats (LTR). (B) DNA transposons. (C) Unknown repeats. (D) Short interspersed nuclear elements (SINE). (E) Simple repeats. (F) rRNA. (G) Lone interspersed nuclear elements (LINE). Vertical dashed lines in (G) represent rRNA clusters.

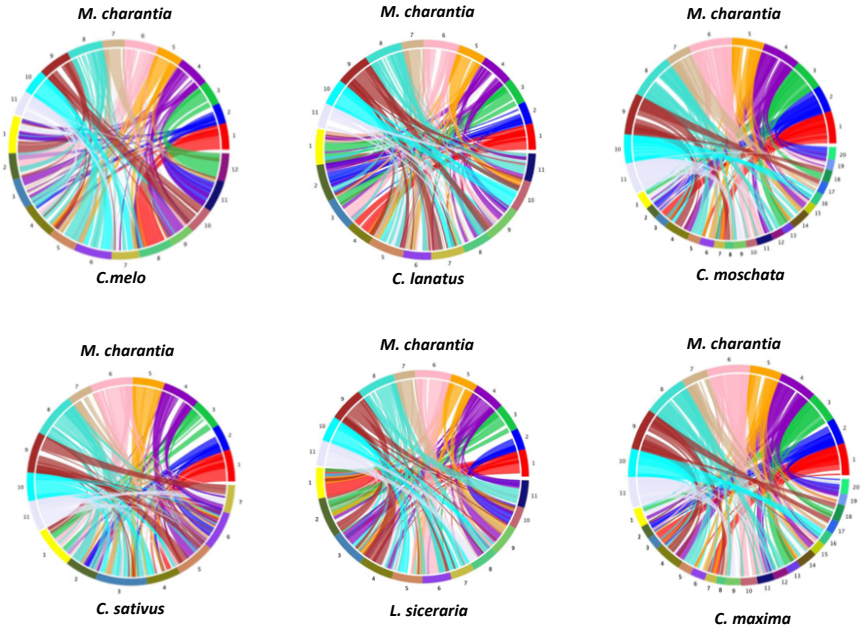


Fig. S3. Synteny comparison between our *M. charantia* assembly and six other Cucurbitaceae species, *Cucumis melo*, *Citrullus lanatus*, *Cucurbita moschata*, *Cucumis sativus*, *Lagenaria siceraria*, and *Cucurbita maxima*.

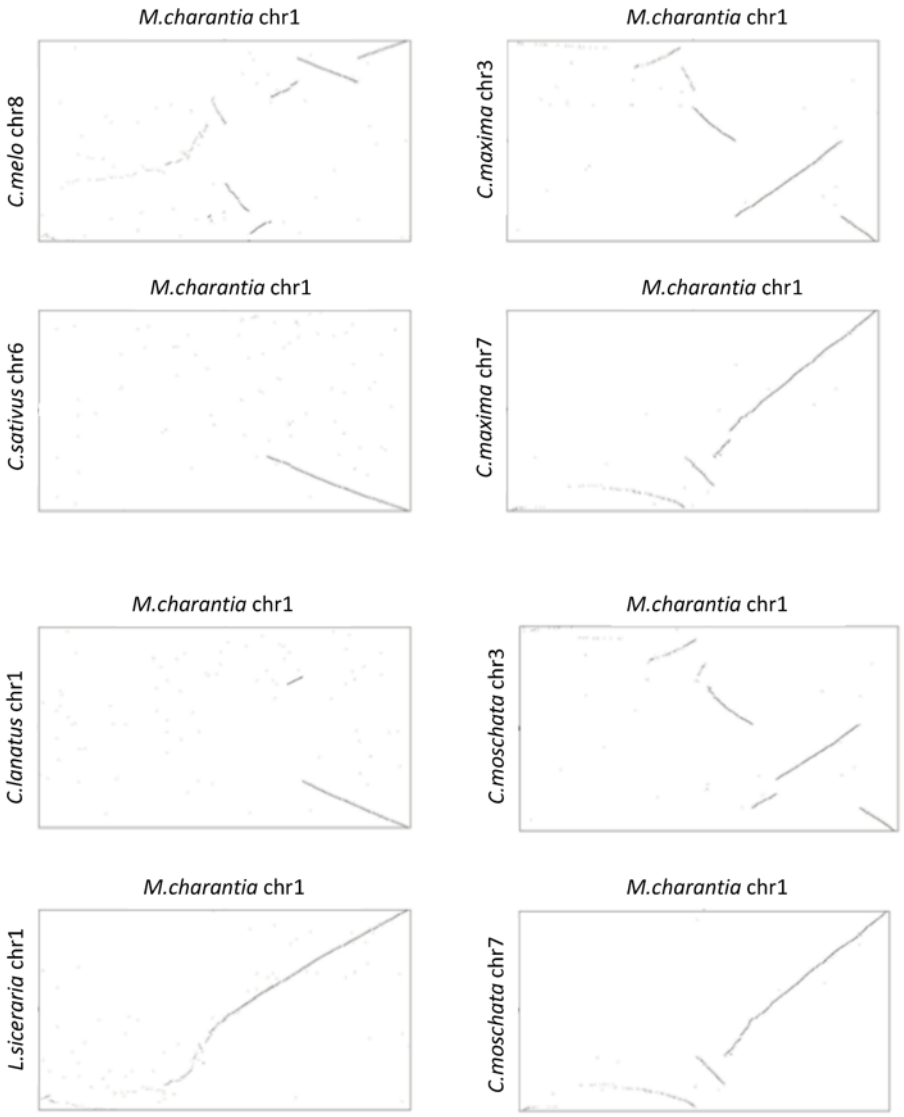


Fig. S4. Dotplots showing the synteny between *Momordica charantia* chromosome 1 and the corresponding chromosomes in six other Cucurbitaceae species, *Cucumis melo*, *Cucumis sativus*, *Citrullus lanatus*, *Lagenaria siceraria*, *Cucurbita moschata*, and *Cucurbita maxima*. The two *Cucurbita* species have two chromosomes matching *M. charantia* chromosome 1 due to whole-genome duplication.

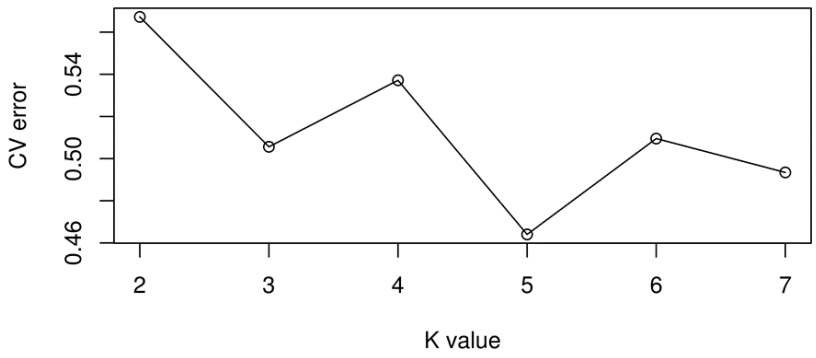


Fig. S5. ADMIXTURE cross validation errors for each K value.

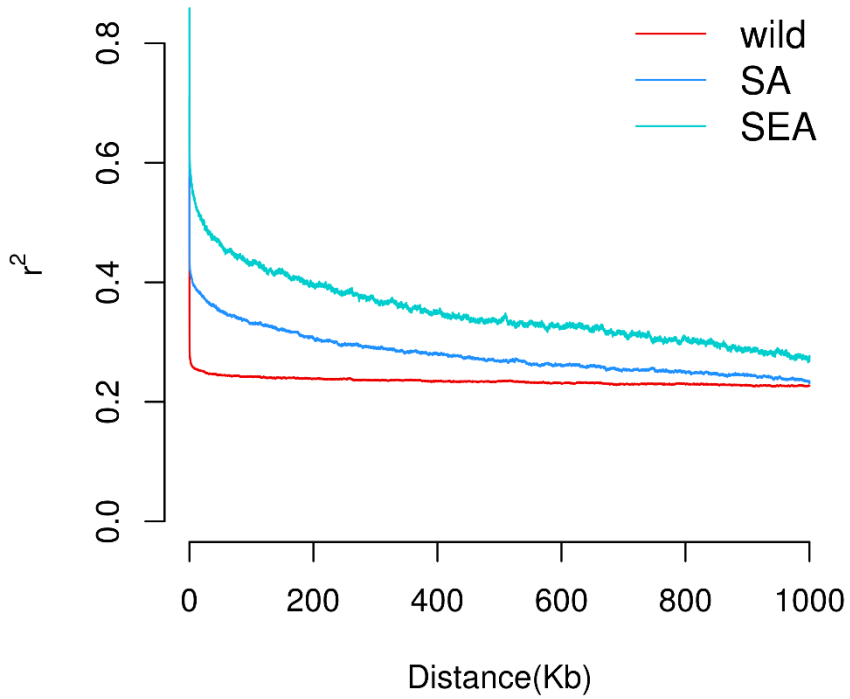


Fig. S6. Linkage disequilibrium **(**LD) decay of wild, South Asian (SA), and Southeast Asian (SEA) groups.

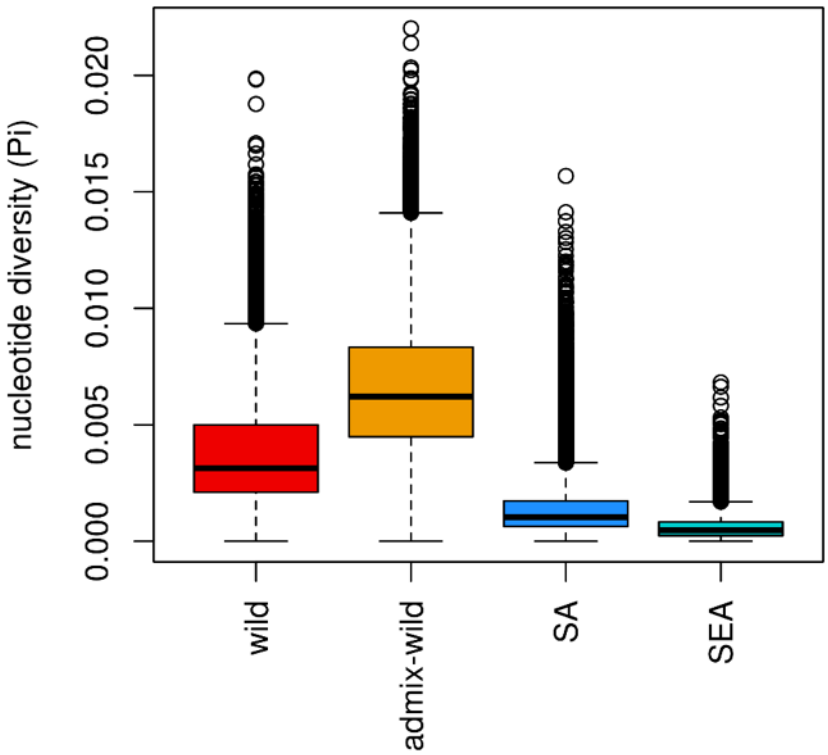


Fig. S7. Nucleotide diversity in 50-kb windows. The box represents median and the first and third quantiles of each distribution. The whiskers extend 1.5 interquartile range beyond the quartiles, with minimum value at zero. SA: South Asian cultivars. SEA: Southeast Asian cultivars.

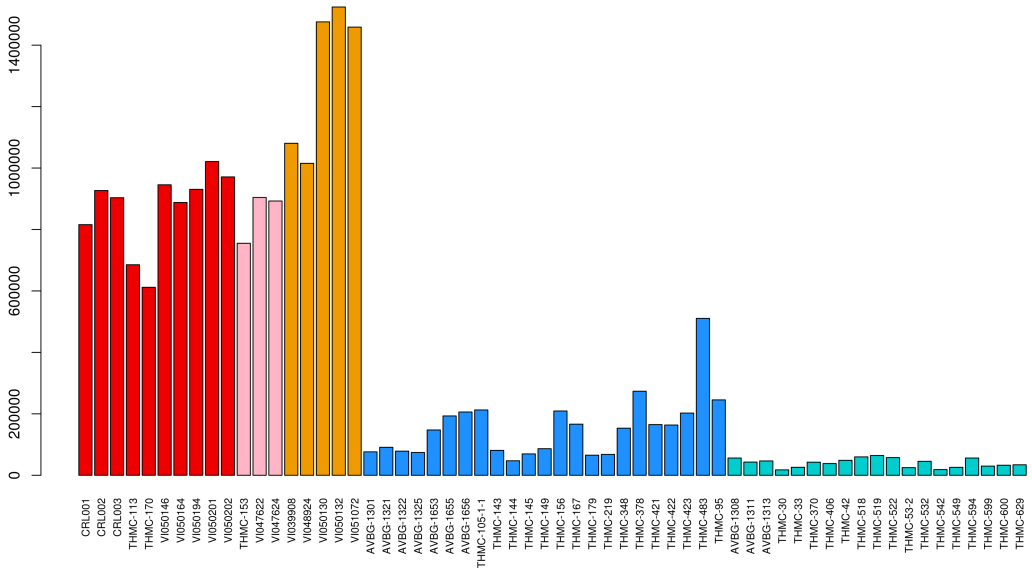


Fig. S8. Number of heterozygous sites of each individual. Red, pink, orange, blue, and light blue represent the Taiwan wild group (TAI), Thailand wild group (THAI), admix-wild group, South Asian cultivar group (SA), and Southeast Asian cultivar group (SEA).

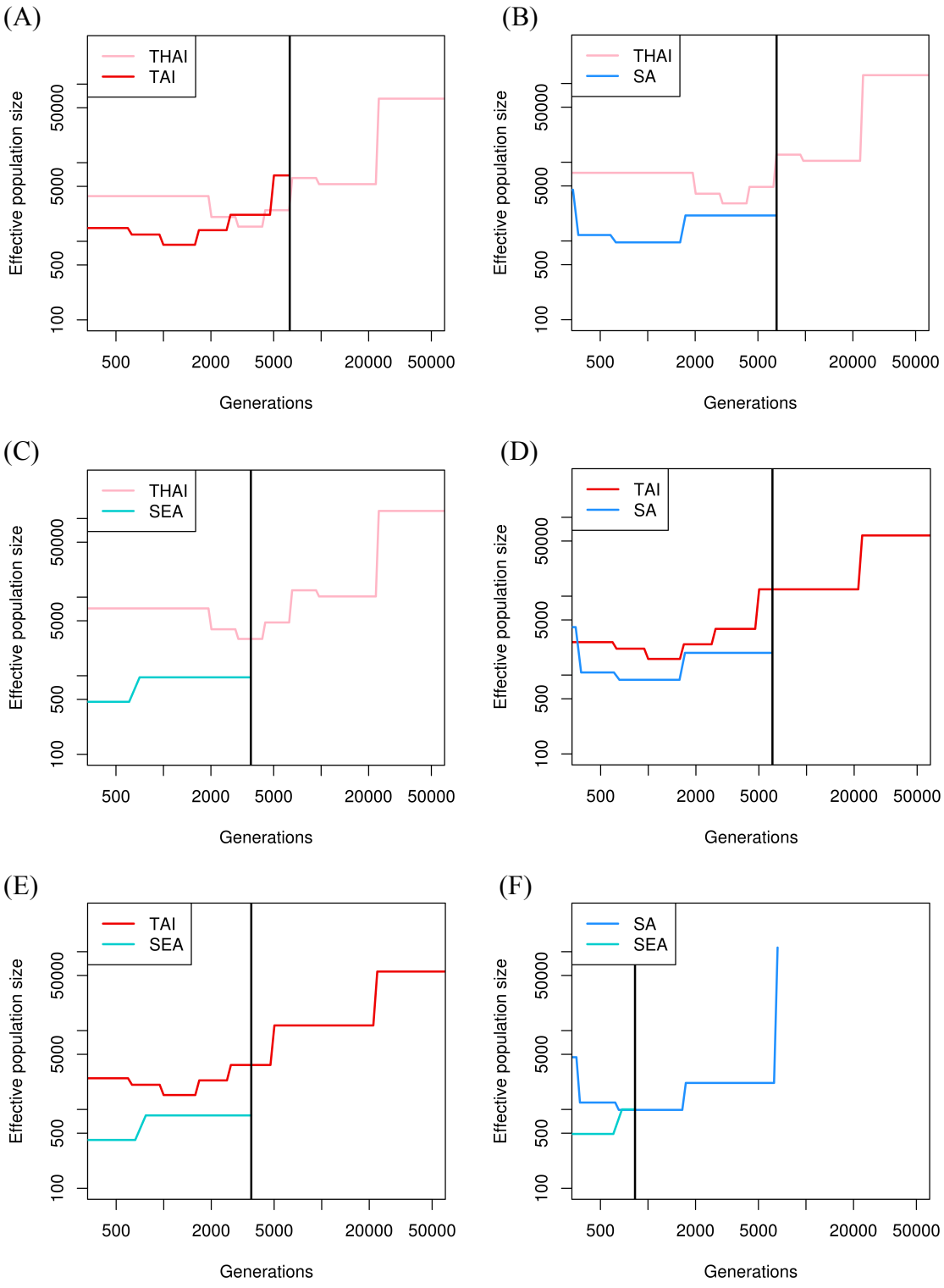


Fig. S9. Pairwise population divergence time estimation with SMC++: (A) THAI and TAI. (B) THAI and SA. (C) THAI and SEA. (D) TAI and SA. (E) TAI and SEA. (F) SA and SEA. Blacklines denote the estimated divergence time. THAI: the Thailand wild group. TAI: the Taiwan wild group. SA: the South Asian cultivar group. SEA: the Southeast Asian cultivar group.

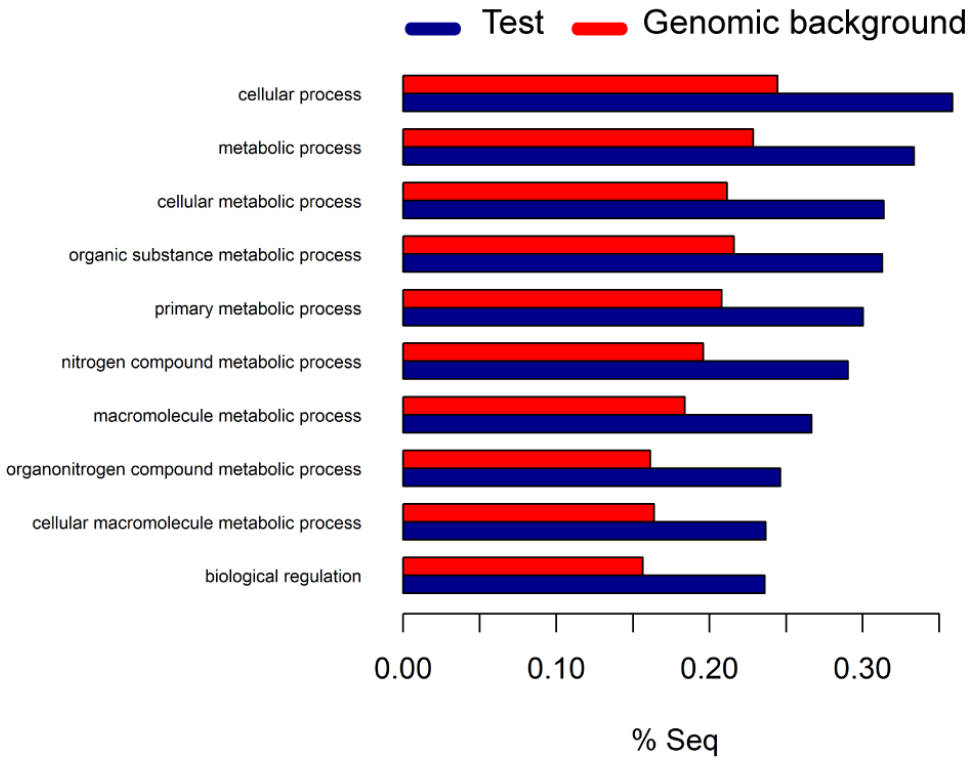


Fig. S10. Significant biological process terms in GO enrichments performed with the selection regions: the union of the top 1% windows of each method to test for signatures of selection. Red bars represent the genomic background for reference, and blue bars represent the selection regions. The horizontal axis is the percentage of a particular term in test or reference set.

Table S1. Comparison of genome assemblies in Cucurbitaceae

	Cucumis sativus	*Citrullus lanatus*	*Citrullus lanatus* (charleston gray)	*Citrullus lanatus* (97103_v2)	*Cucurbita argyrosperma*	*Cucurbita pepo*	*Lagenaria siceraria*	*Cucumis melo*	*Momordica charantia* (2017 assembly)	*Momordica charantia* (Dali-11 assembly)	*Momordica charantia* (this study)
assembly size (Mb)	243	353	396	365	229	263	313.4	375	285	294	303
Contig N50 (kb)	19.8	26.4	36.7	2,300	463.4	110.0	28.3	60.8	21.9	62.6	9,898.0
scaffold N50 (kb)	1,140	2,380	7,471	21,900	620	1,750	8,701	4,680	1,101	611	25,371
sequence on chromosome (Mb)	176.9	330	378.7	362.7	0	214.1	297.6	337.5	0	251	291.7
sequence on chromosome / assembly size (%)	72.80	93.48	95.63	99.3	0.00	81.41	94.96	90.00	0.00	85.37	96.27

Table S2. Repeat annotation of *Momordica charantia*

total length	302,992,168 bp (302,989,368 bp excluding N sites)		
GC level	35.56%		
bases masked	159,137,698 bp (52.52 %)		
	number of elements	length occupied	percentage of sequence
SINEs	7,730	880,637	0.29
tRNA	6,704	709,276	0.23
LINEs	28,173	8,192,849	2.7
L1	18,359	7,449,099	2.46
RTE-BovB	7,970	513,592	0.17
LTR elements	95,383	72,614,725	23.97
Gypsy	44,031	40,679,168	13.43
Copia	34,651	25,113,678	8.29
DNA elements	59,043	15,543,591	5.13
MULE-MuDR	25,621	6,698,338	2.21
CMC-EnSpm	10,434	3,779,991	1.25
Unclassified	215,486	57,350,078	18.93
small RNA	8,203	1,440,675	0.48
Satellites	1,552	292,321	0.1
Simple repeats	86,358	3,354,855	1.11
Low complexity	21,080	1,021,681	0.34

Table S3. Repeat comparison between two *Momordica charantia* assemblies

	Momordica charantia (Dali-11)	*Momordica charantia* (this study)
Genome size	251,380,684	302,989,368
Repeat percentage	45.43	52.52
DNA elements	12,216,336	15,543,591
DNA elements percentage	4.86	5.13
LINE	6,913,599	8,192,849
LINE percentage	2.75	2.7
LTR	49,059,921	72,614,725
LTR percentage	19.52	23.97
SINE	804,878	880,637
SINE percentage	0.32	0.29
simple repeats	2,982,683	4,668,857
simple repeats percentage	1.19	1.55
unknown	41,510,819	57,350,078
unknown percentage	16.51	18.93

Table S4. Top 15 genes enriched for the top 0.1% SNPs with highest trait association

Trait	Gene	Proportion SNPs with top 0.1% association	Annotation
Color	chr10.2901	0.94	BPA1_ARATH Binding partner of ACD11 1
Color	chr9.1019	0.71	PPR76_ARATH Pentatricopeptide repeat-containing protein At1g51965, mitochondrial
Color	chr10.2733	0.65	TPSGD_VITVI (-)-germacrene D synthase
Color	chr5.3365	0.56	AGT23_ARATH Alanine--glyoxylate aminotransferase 2 homolog 3, mitochondrial
Color	chr5.3363	0.46	TOL6_ARATH TOM1-like protein 6
Color	chr10.2734	0.46	RNH_HALSA Ribonuclease HI
Color	chr9.1489	0.45	PP268_ARATH Putative pentatricopeptide repeat-containing protein At3g47840
Color	chr8.1391	0.44	NBP35_ARATH Cytosolic Fe-S cluster assembly factor NBP35
Color	chr7.2532	0.35	AKR1_SOYBN Probable aldo-keto reductase 1
Color	chr5.2799	0.30	LPAT2_ARATH 1-acyl-sn-glycerol-3-phosphate acyltransferase 2
Color	chr5.3367	0.29	Y5370_ARATH PAN domain-containing protein At5g03700
Color	chr8.1392	0.29	G6PD_SOLTU Glucose-6-phosphate 1-dehydrogenase, cytoplasmic isoform
Color	chr1.2324	0.29	SAP4_ARATH Zinc finger A20 and AN1 domain-containing stress-associated protein 4
Color	chr6.3557	0.27	APRR2_ARATH Two-component response regulator-like APRR2
Color	chr8.3849	0.26	MYC4_ARATH Transcription factor MYC4
Length	chr7.839	0.74	Y3093_ARATH Uncharacterized protein At3g60930, chloroplastic
Length	chr4.3687	0.53	RS193_ARATH 40S ribosomal protein S19-3
Length	chr7.840	0.50	ADS3_ARATH Palmitoyl-monogalactosyldiacylglycerol delta-7 desaturase, chloroplastic
Length	chr6.2989	0.37	EIX2_SOLLC Receptor-like protein EIX2
Length	chr4.3689	0.27	U83A1_ARATH UDP-glycosyltransferase 83A1
Length	chr8.590	0.25	ANT_ARATH AP2-like ethylene-responsive transcription factor ANT
Length	chr3.267	0.21	PPP7L_ARATH Serine/threonine-protein phosphatase 7 long form homolog
Length	chr8.443	0.20	BRT1_ARATH Adenine nucleotide transporter BT1, chloroplastic/mitochondrial
Length	chr11.2845	0.19	DRP2B_ARATH Dynamin-2B OS=Arabidopsis thaliana

Length	chr4.2162	0.15	DXR_ORYSJ 1-deoxy-D-xylulose 5-phosphate reductoisomerase, chloroplastic
Length	chr9.408	0.14	PRP8A_ARATH Pre-mRNA-processing-splicing factor 8A
Length	chr5.3007	0.12	KN5C_TOBAC Kinesin-like protein KIN-5C
Length	chr8.413	0.12	NAA50_XENTR N-alpha-acetyltransferase 50
Length	chr4.3309	0.12	OPF13_ARATH Transcription repressor OFP13
Length	chr8.533	0.10	ACL5_ARATH Thermospermine synthase ACAULIS5
Spine	chr6.3621	0.47	RCD1_ARATH Inactive poly [ADP-ribose] polymerase RCD1
Spine	chr2.2718	0.33	PPP7L_ARATH Serine/threonine-protein phosphatase 7 long form homolog
Spine	chr2.2736	0.30	PPP7L_ARATH Serine/threonine-protein phosphatase 7 long form homolog
Spine	chr6.3605	0.30	ATL29_ARATH RING-H2 finger protein ATL29
Spine	chr5.3210	0.27	IDM1_ARATH Increased DNA methylation 1
Spine	chr2.2431	0.27	ATPAM_MAIZE ATP synthase subunit alpha, mitochondrial
Spine	chr2.2752	0.25	SBT4E_ARATH Subtilisin-like protease SBT4.14
Spine	chr6.3609	0.25	Y4958_ARATH Uncharacterized membrane protein At4g09580
Spine	chr2.2814	0.24	POL_AVIRE Gag-Pol polyprotein (Fragment)
Spine	chr6.3531	0.21	CAAT6_ARATH Cationic amino acid transporter 6, chloroplastic
Spine	chr6.3601	0.21	TRN2_ARATH Protein TORNADO 2
Spine	chr8.725	0.20	PX11A_ARATH Peroxisomal membrane protein 11A
Spine	chr2.2733	0.20	PPP7L_ARATH Serine/threonine-protein phosphatase 7 long form homolog
Spine	chr8.4288	0.19	IAA14_ARATH Auxin-responsive protein IAA14
Spine	chr2.2437	0.19	ATESY_VITVI (-)-alpha-terpineol synthase

Table S5. Accessions used in this study

Accession	Country of collection	Origin	Color	Length	Skin pattern	Genetic group
AVBG1308	Philippines	Cultivar	Medium green	Medium	Spiny	SEA
AVBG1311	Thailand	Cultivar	Light green	Medium	Smooth	SEA
AVBG1313	Thailand	Cultivar	Light green	Long	Smooth	SEA
AVBG1321	India	Cultivar	Green	Medium	Spiny	SA
AVBG1322	India	Cultivar	Green	Medium	Spiny	SA
AVBG1325	India	Cultivar	Green	Long	Spiny	SA
AVBG1653	Bangladesh	Cultivar	Green	Medium	Spiny	SA
AVBG1655	Bangladesh	Cultivar	Green	Medium	Spiny	SA
AVBG1656	Bangladesh	Cultivar	Green	Medium	Spiny	SA
THMC 105-1-1	Bangladesh	Cultivar	Green	Short	Smooth	SA
THMC113	Belize	Cultivar	Green	Short	Spiny	TAI
THMC143	India	Cultivar	Medium green	Medium	Spiny	SA
THMC144	India	Cultivar	Green	Medium	Smooth	SA
THMC145	India	Cultivar	Medium green	Medium	Spiny	SA
THMC149	India	Cultivar	Green	Long	Smooth	SA
THMC153	Thailand	Cultivar	Green	Short	Spiny	THAI
THMC156	India	Cultivar	Green	Short	Spiny	SA
THMC167	India	Cultivar	Dark green	Medium	Spiny	SA
THMC170	Taiwan	Cultivar	Green	Short	Spiny	TAI
THMC179	India	Cultivar	White	Long	Spiny	SA
THMC219	Pakistan	Cultivar	Green	Long	Spiny	SA
THMC30	Philippines	Cultivar	Green	Long	Smooth	SEA

THMC33	Philippines	Cultivar	Dark green	Medium	Smooth	SEA
THMC348	Bangladesh	Cultivar	Green	Long	Spiny	SA
THMC370	Vietnam	Cultivar	Light green	Short	Smooth	SEA
THMC378	Thailand	Cultivar	Light green	Short	Smooth	SA
THMC406	Vietnam	Cultivar	Light green	Medium	Smooth	SEA
THMC42	Philippines	Cultivar	Medium green	Long	Smooth	SEA
THMC421	Bangladesh	Cultivar	Green	Long	Spiny	SA
THMC422	Bangladesh	Cultivar	Green	Long	Spiny	SA
THMC423	Bangladesh	Cultivar	Green	Long	Spiny	SA
THMC483	Bangladesh	Cultivar	Green	Medium	Spiny	SA
THMC518	Vietnam	Cultivar	Light green	Long	Spiny	SEA
THMC519	Vietnam	Cultivar	Light green	Medium	Smooth	SEA
THMC522	China	Cultivar	Light green	Long	Smooth	SEA
THMC532	China	Cultivar	Light green	Long	Smooth	SEA
THMC53-2	Philippines	Cultivar	Green	Short	Smooth	SEA
THMC542	China	Cultivar	Dark green	NA	Smooth	SEA
THMC549	China	Cultivar	Green	Long	Spiny	SEA
THMC594	China	Cultivar	Light green	Long	Smooth	SEA
THMC599	China	Cultivar	Light green	Long	Smooth	SEA
THMC600	China	Cultivar	Light green	Long	Smooth	SEA
THMC629	China	Cultivar	Light green	Long	Smooth	SEA
THMC95	Bangladesh	Cultivar	Green	Long	Spiny	SA
CRL001	Taiwan	Wild	Green	Short	Spiny	TAI
CRL002	Taiwan	Wild	Green	Short	Spiny	TAI
CRL003	Taiwan	Wild	Green	Short	Spiny	TAI

VI039908	Thailand	Wild	Green	Short	Spiny	ADMIX
VI047622	Thailand	Wild	Green	Short	Spiny	THAI
VI047624	Thailand	Wild	Green	Short	Spiny	THAI
VI048924	Thailand	Wild	Green	Short	Spiny	ADMIX
VI050130	Taiwan	Wild	Green	Short	Spiny	ADMIX
VI050132	Taiwan	Wild	Green	Short	Spiny	ADMIX
VI050146	Taiwan	Wild	Green	Short	Spiny	TAI
VI050164	Taiwan	Wild	Green	Short	Spiny	TAI
VI051072	Philippines	Wild	Green	Short	Spiny	ADMIX
VI050194	Taiwan	Wild	Green	Short	Spiny	TAI
VI050201	Taiwan	Wild	Green	Short	Spiny	TAI
VI050202	Taiwan	Wild	Green	Short	Spiny	TAI

Table S6. *MatK* primer sequences for validating *Momordica cochinchinensis* samples

Primer	5' sequence 3'
matK-AF	CTA TAT CCA CTT ATC TTT CAG GAG T
matK-8R	AAA GTT CTA GCA CAA GAA AGT CGA