



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

# Draft genome sequence of *Solanum aethiopicum* provides insights into disease resistance, drought tolerance, and the evolution of the genome

### Citation for published version:

Song, B, Song, Y, Fu, Y, Kizito, EB, Kamenya, SN, Kabod, PN, Liu, H, Muthemba, S, Kariba, R, Njuguna, J, Maina, S, Stomeo, F, Djikeng, A, Hendre, PS, Chen, X, Chen, W, Li, X, Sun, W, Wang, S, Cheng, S, Muchugi, A, Jamnadass, R, Shapiro, HY, Van Deynze, A, Yang, H, Wang, J, Xu, X, Odeny, DA & Liu, X 2019, 'Draft genome sequence of *Solanum aethiopicum* provides insights into disease resistance, drought tolerance, and the evolution of the genome', *GigaScience*, vol. 8, no. 10, giz115.  
<https://doi.org/10.1093/gigascience/giz115>

### Digital Object Identifier (DOI):

[10.1093/gigascience/giz115](https://doi.org/10.1093/gigascience/giz115)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

GigaScience

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

















### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.





## RESEARCH

# Draft genome sequence of *Solanum aethiopicum* provides insights into disease resistance, drought tolerance, and the evolution of the genome

Bo Song <sup>1,2,3,†</sup>, Yue Song <sup>1,3,4,†</sup>, Yuan Fu <sup>1,2,†</sup>, Elizabeth Balyejusa Kizito <sup>5</sup>, Sandra Ndagire Kamenya<sup>5,6</sup>, Pamela Nahamya Kabod<sup>5</sup>, Huan Liu <sup>1,2,3</sup>, Samuel Muthemba <sup>7</sup>, Robert Kariba <sup>7</sup>, Joyce Njuguna<sup>6</sup>, Solomon Maina <sup>6</sup>, Francesca Stomeo <sup>6,8</sup>, Appolinaire Djikeng<sup>6,9</sup>, Prasad S. Hendre <sup>7</sup>, Xiaoli Chen <sup>1,2</sup>, Wenbin Chen<sup>1,2</sup>, Xiuli Li<sup>1,2</sup>, Wenjing Sun<sup>1,2</sup>, Sibow Wang<sup>1,3</sup>, Shifeng Cheng<sup>1,2</sup>, Alice Muchugi<sup>7</sup>, Ramni Jamnadass <sup>7</sup>, Howard-Yana Shapiro<sup>7,10,11</sup>, Allen Van Deynze <sup>10</sup>, Huanming Yang<sup>1,2</sup>, Jian Wang<sup>1,2</sup>, Xun Xu <sup>1,2,3</sup>, Damaris Achieng Odeny <sup>12,\*</sup> and Xin Liu <sup>1,2,3,\*</sup>

<sup>1</sup>BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China; <sup>2</sup>China National GeneBank, BGI-Shenzhen, Jinsha Road, Shenzhen 518120, China; <sup>3</sup>State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China; <sup>4</sup>BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China; <sup>5</sup>Uganda Christian University, Bishop Tucker Road, Box 4, Mukono, Uganda; <sup>6</sup>Biosciences Eastern and Central Africa (BeCA) – International Livestock Research Institute (ILRI) Hub, P.O. Box 30709, Nairobi 00100, Kenya; <sup>7</sup>African Orphan Crops Consortium, World Agroforestry Centre (ICRAF), United Nations Avenue, Nairobi 00100, Kenya; <sup>8</sup>Present address: European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; <sup>9</sup>Present address: Centre for Tropical Livestock Genetics and Health (CTLGH), University of Edinburgh, Edinburgh EH25 9RG, UK; <sup>10</sup>University of California, 1 Shields Ave, Davis, CA, USA; <sup>11</sup>Mars, Incorporated, 6885 Elm Street, McLean, VA 22101, USA and <sup>12</sup>International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) – Eastern and Southern Africa, P.O. Box 39063, Nairobi 00623, Kenya

\*Correspondence address. Xin Liu, BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. Tel: +86-180-2546-0332; E-mail: [liuxin@genomics.cn](mailto:liuxin@genomics.cn)  <http://orcid.org/0000-0002-3629-3752>; Damaris Achieng Odeny, P.O. Box 39063–00623, Nairobi, Kenya. Tel: +254 20 7224559; E-mail: [D.Odeny@cigar.org](mailto:D.Odeny@cigar.org)  <http://orcid.org/0000-0003-3256-2940>

<sup>†</sup>Equal contribution.

## Abstract

**Background:** The African eggplant (*Solanum aethiopicum*) is a nutritious traditional vegetable used in many African countries, including Uganda and Nigeria. It is thought to have been domesticated in Africa from its wild relative, *Solanum anguivi*. S.

Received: 26 January 2019; Revised: 14 July 2019; Accepted: 24 August 2019

© The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

*aethiopicum* has been routinely used as a source of disease resistance genes for several Solanaceae crops, including *Solanum melongena*. A lack of genomic resources has meant that breeding of *S. aethiopicum* has lagged behind other vegetable crops. **Results:** We assembled a 1.02-Gb draft genome of *S. aethiopicum*, which contained predominantly repetitive sequences (78.9%). We annotated 37,681 gene models, including 34,906 protein-coding genes. Expansion of disease resistance genes was observed via 2 rounds of amplification of long terminal repeat retrotransposons, which may have occurred ~1.25 and 3.5 million years ago, respectively. By resequencing 65 *S. aethiopicum* and *S. anguivi* genotypes, 18,614,838 single-nucleotide polymorphisms were identified, of which 34,171 were located within disease resistance genes. Analysis of domestication and demographic history revealed active selection for genes involved in drought tolerance in both “Gilo” and “Shum” groups. A pan-genome of *S. aethiopicum* was assembled, containing 51,351 protein-coding genes; 7,069 of these genes were missing from the reference genome. **Conclusions:** The genome sequence of *S. aethiopicum* enhances our understanding of its biotic and abiotic resistance. The single-nucleotide polymorphisms identified are immediately available for use by breeders. The information provided here will accelerate selection and breeding of the African eggplant, as well as other crops within the Solanaceae family.

**Keywords:** *Solanum aethiopicum*; African eggplant; *Solanum anguivi*; LTR-Rs; biotic stress; drought tolerance

## Background

The African eggplant, *Solanum aethiopicum* (NCBI:txid205524), is an indigenous non-tuberiferous Solanaceae crop that is mainly grown in tropical Africa [1], especially in Central and West Africa. *S. aethiopicum* is hypervariable [2, 3] and is generally classified into 4 groups: Gilo, Shum, Kumba, and Aculeatum. Gilo is the most important group and has edible fruits, while Shum has small and bitter fruits. Kumba is used as a leafy vegetable, while Aculeatum is used as an ornamental [3, 4] or as rootstock because of its excellent disease resistance [5]. The African eggplant is reported to have anti-inflammatory activity [6], and its roots and fruits have been used to treat colic, high blood pressure, and uterine complications in Africa [6].

Although *S. aethiopicum* is one of the most important cultivated eggplants in Africa [7, 8], it remains an “orphan crop” because research and breeding investments are lagging behind other Solanaceae relatives, such as *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), and *Solanum melongena* (edible eggplant). Consequently, there have been few robust genomic resources, such as a well-annotated reference genome. Genomics-assisted breeding is an effective approach that would facilitate the breeding of orphan crops such as the African eggplant. Previous attempts to develop molecular markers for *S. aethiopicum*, using the *S. melongena* genome as a reference, have been unsuccessful because of compromised accuracy [9]. An alternative approach that uses genome editing has been successfully deployed in other Solanaceae crops, including *Physalis pruinosa* [11, 12], but cannot be implemented in *S. aethiopicum* because of its lack of well-annotated reference genome and gene sequences.

The African eggplant serves as a gene reservoir for other economically important crops within the Solanaceae family. Thanks to its cross-compatibility with *S. melongena* [4, 10] and its outstanding resistance to various pathogens, including *Fusarium*, *Ralstonia*, and *Verticillium* [5, 11–13], *S. aethiopicum* has been used to develop rootstocks [13] or improve the disease resistance of *S. melongena* [14]. Because the genomic basis of resistance in *S. aethiopicum* is poorly understood, it can be time-consuming to use it as a donor in such interspecific crosses. Mapping resistance genes and then developing markers associated with these genes might resolve this challenge. The development and expansion of resistance genes is usually accompanied by the amplification of long terminal repeat retrotransposons (LTR-Rs). A typical example is shown in the Solanaceous hot pepper (*Capiscum annuum*), in which a burst of LTR-Rs substantially mediated the retrotransposition of nucleotide-binding, leucine-rich

repeat-related (NLR) genes, leading to the expansion of resistance genes [15]. LTR-Rs are abundant in plant genomes, including Solanaceae crops such as *Nicotiana sylvestris* (~38.16%) [16], pepper (>70.0%) [17], potato (62.2%) [18], tomato (50.3%) [19], and petunia (>60%) [20]. The role of LTR-Rs in the *S. aethiopicum* genome remains unknown, and whether the resistance seen in *S. aethiopicum* is a result of LTR-R amplification remains to be investigated. The generation of a reference genome for *S. aethiopicum*, as well as for other orphan crops, is urgently needed to advance their research and breeding.

Here, we report a draft whole-genome assembly and annotation for *S. aethiopicum*. We found 2 amplifications of LTR-Rs that occurred ~1.25 and 3.5 million years ago (MYA), resulting in the expansion of resistance genes. We also resequenced 2 *S. aethiopicum* groups, “Gilo” and “Shum,” and *S. anguivi* at a high depth (~60×) and identified 18,614,838 single-nucleotide polymorphisms (SNPs), 34,171 of which are located within resistance genes. Subsequently, we generated a pan-genome of *S. aethiopicum*. The genomic data provided in this study will greatly advance research and breeding activities of the African eggplant.

## Data Description

We sequenced the genome of *S. aethiopicum* using a whole-genome shotgun approach. A total of 242.61 Gb raw reads were generated by sequencing the libraries with insert sizes of 250 and 500 bp, and mate-pair libraries with sizes ranging between 2,000 and 20,000 bp, on an Illumina HiSeq 2000 platform. The filtered reads used for downstream analysis are shown in Supplementary Table 1. *k*-mer (*k* = 17) analysis [21] revealed the *S. aethiopicum* genome to be diploid and homozygous, with an estimated genome size of 1.17 Gb (Supplementary Fig. 1). “Clean reads” amounting to 127.83 Gb (~109×) were used to assemble the genome using Platanus [22] (see Methods). A final assembly of 1.02 Gb in size was obtained, containing 162,187 scaffolds with N50 contig and scaffold values of 25.2 and 516.15 kb (Table 1 and Supplementary Table 2), respectively. Our results reveal that the *S. aethiopicum* genome is larger than that of other *Solanum* genomes, including tomato (0.76 Gb) and potato (0.73 Gb) [18, 19], but it has a comparable guanine-cytosine (GC) ratio (33.12%) (Supplementary Table 3).

Repetitive elements, predominantly transposable elements (TEs) (Supplementary Table 4), occupied 811 Mb (78.9%) of the sequenced genome. Most annotated TEs were retrotransposon elements, including long terminal repeats (LTRs), short interspersed nuclear elements, and long interspersed nuclear

**Table 1:** Statistical data for the *Solanum aethiopicum* genome and gene annotation

Parameter	Value
<b>Scaffolds</b>	
Number	162,187
Total length	1.02 Gb
N50	516.1 kb
Longest	2.94 Mb
<b>Contigs</b>	
Number	231,821
Total length	936 Mb
N50	25.2 kb
Longest	366.2 kb
GC content	33.13%
Number of genes	34,906
Average/total coding sequence length	1104.3 bp/38.5 Mb
Average exon/intron length	265.8 bp/613.1 bp
Total length of transposable elements	805.7 Mb (78.23%)

elements. Together these retrotransposons made up 75.42% of the assembly. DNA transposons accounting for 2.87% of the genome were also annotated (Supplementary Table 4).

Protein-coding gene models were predicted by a combination of homologous search and *ab initio* prediction. The resulting models were pooled to generate a final set of 34,906 protein-coding genes. Predicted gene models were, on average, 3,038 bp in length, with a mean of 3.15 introns. The mean length of coding sequences, exons, and introns was 1,104, 265, and 613 bp, respectively (Table 1, Supplementary Table 5, Supplementary Fig. 2). As expected, these gene features were similar to those of other released genomes, including *Arabidopsis thaliana* [23] and other Solanaceae crops including *S. lycopersicum*, *S. tuberosum*, *C. annuum*, and *N. sylvestris* [16, 18, 19, 24] (Supplementary Table 5). We further assessed the annotation completeness of this assembly by searching for 1,440 core embryophyta genes (CEGs) with BUSCO, version 3.0 [25]. We found 80.4% CEGs in this assembly, with 77.8% being single copies and 2.6% being duplicates (Supplementary Table 6). We also annotated the non-coding genes by homologous search, leading to the identification of 128 microRNA, 960 transfer RNA (tRNA), 1,185 ribosomal RNA (rRNA), and 503 small nuclear RNA (snRNA) genes (Supplementary Table 7).

We annotated 31,863 (91.28%) proteins for their homologous function in several databases. Homologs of 31,099 (89.09%), 26,319 (75.4%), and 20,932 (59.97%) proteins were found in TrEMBL, InterPro, and SwissProt databases, respectively (Supplementary Table 8). The remaining 3,043 (8.72%) genes encoded putative proteins with unknown functions.

## Analyses

### Genome evolution and phylogenetic analysis

By comparing with 4 other sequenced Solanaceae genomes (*S. melongena*, *S. lycopersicum*, *S. tuberosum*, and *C. annuum*), 25,751 of the *S. aethiopicum* genes were clustered into 19,310 families using OrthoMCL (version 2.0) [26], with an average of 1.33 genes each. Single-copy genes shared by these 5 genomes were concatenated as a supergene representing each genome and were used to build a phylogenetic tree (Fig. 1A). The split time between *S. aethiopicum* and *S. melongena* was estimated to be ~2.6 MYA. McScanX [27] identified 182 syntenic blocks. We detected evidence of whole-genome duplication (WGD) events in

this genome by calculating the pairwise synonymous mutation rates and the rate of 4-fold degenerative third-codon transversion (4DTV) of 1,686 paralogous genes in these blocks. The 4DTV distribution plot displayed 2 peaks, at ~0.25 and 1, indicating 2 WGDs (Fig. 1B). The first 1 (peak at 1) represents the ancient WGD event shared by asterids and rosids [28], while the second WGD event is shared by Solanaceae plants. This suggests that its occurrence predates the split of Solanaceae.

### Evolution of gene families

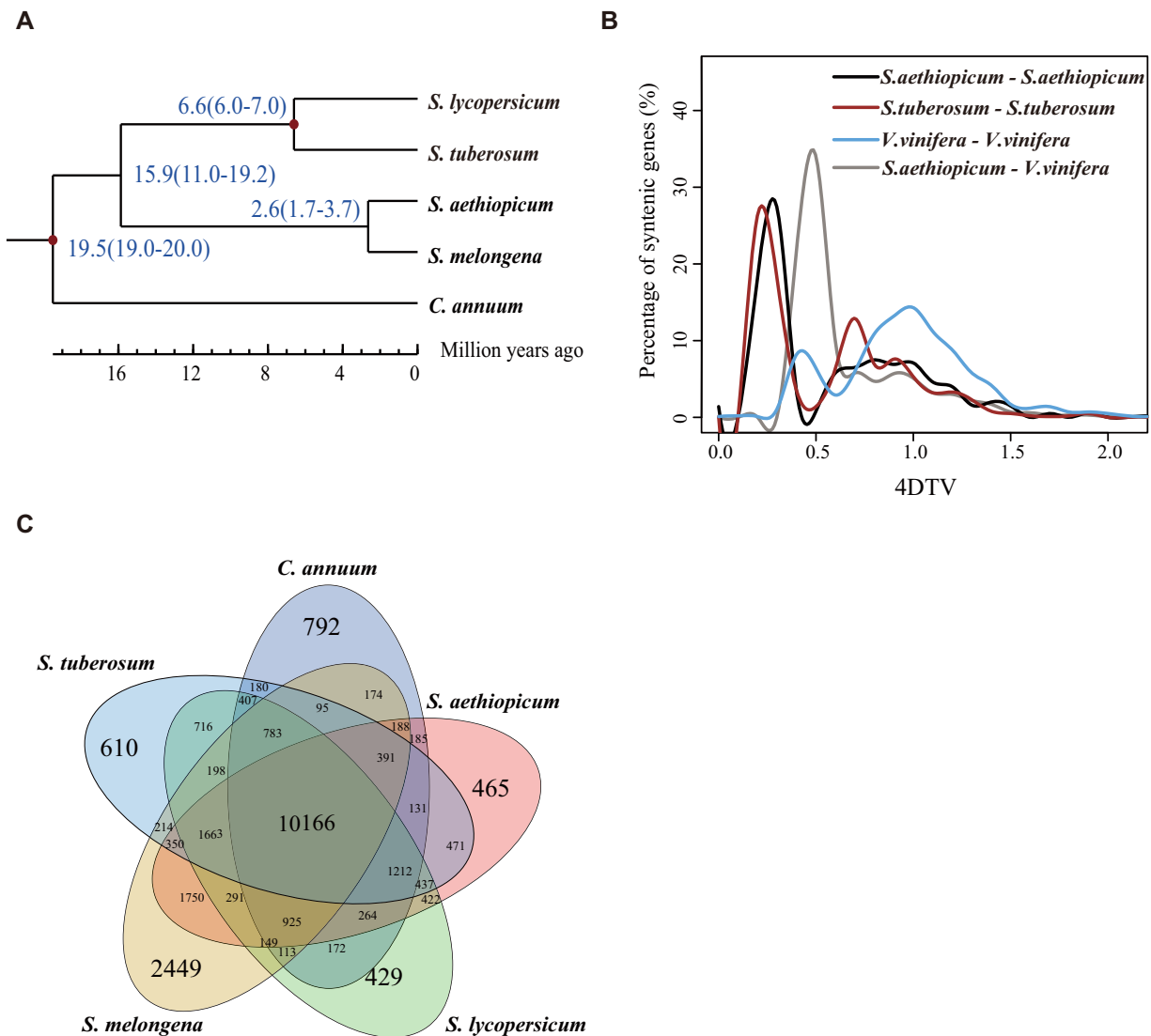
OrthoMCL [26] clustering of genes from *S. aethiopicum*, *S. melongena*, *S. lycopersicum*, *S. tuberosum*, and *C. annuum* identified 25,751 gene families. Among these, 465 gene families were unique to *S. aethiopicum* and 10,166 were common (Supplementary Table 9, Fig. 1C). As expected, the number of shared gene families decreased as a function of evolutionary distance between *S. aethiopicum* and the selected species (Supplementary Table 10). For example, *S. aethiopicum* shared 15,723 gene families with *S. melongena*, compared with only 13,461 genes shared with *C. annuum*. To further investigate the evolution of gene families, we identified expanded and contracted gene families. Compared with *S. melongena*, 437 gene families were expanded; most expanded gene families were found to be involved in biological processes related to drought or salinity tolerance or disease resistance, including defense response (GO:0006952), response to oxidative stress (GO:0006979), glutamate biosynthetic processes (GO:0006537), and response to metal ions (GO:0010038) (Supplementary Table 11). No gene families were contracted when comparing with *S. melongena*.

### Amplification of LTR-Rs

LTR-Rs made up ~70% of the genome and accounted for 89.31% of the total TEs in *S. aethiopicum* (Supplementary Table 4). Consistent with previous studies of LTR-Rs, most LTR-Rs were classified as being in Ty3/Gypsy (82.36% of total LTR-Rs) and Ty1/Copia (14.90% of total LTR-Rs) subfamilies. The proportion of Ty3/Gypsy in *S. aethiopicum* is comparable to that reported in the hot pepper genome (87.7% of Ty3/Gypsy) [24]. To investigate the roles of LTR-Rs in the evolution of *S. aethiopicum*, we detected 36,599 full-length LTR-Rs using LTRharvest [29] with the parameters “-maxlenltr 2000, -similar 75” and LTRdigest software [30]. We further analyzed their evolution, activity, and potential biological functions.

The age of each LTR-R was inferred by comparing the divergence between the 5' and 3' LTR-R, using a substitution rate of  $1.3e-8 \text{ year}^{-1} \text{ site}^{-1}$  [31]. Two amplifications of LTR-Rs were found in *S. aethiopicum*, while only 1 was detected in tomato and hot pepper (Fig. 2A). The early amplification occurred at ~3.5 MYA, coincident with the LTR-R burst found in *C. annuum* [15] (Fig. 2A). The second amplification was at 1.25 MYA, coinciding with the LTR-R burst in the tomato genome [19] (Fig. 2A). Although the time of LTR-Rs amplification is vertically coincident between different species, they occurred separately in each genome since the ancestor of *S. aethiopicum* diverged from that of hot pepper and tomato ~20 and 4 MYA, respectively (Fig. 1A). These results imply that environmental stimulators shared between these species during their evolution could have triggered the amplifications observed. We also estimated the amplification time of Ty3/Gypsy and Ty1/Copia LTR-Rs and found 2 peaks at ~1.25 and 3.5 MYA for Gypsy LTR-Rs (Fig. 2B) but only 1 peak (~1.25 MYA) for Ty1/Copia LTR-Rs (Fig. 2C). Compared with the amplification time of Ty3/Gypsy and Ty1/Copia LTR-Rs in differ-





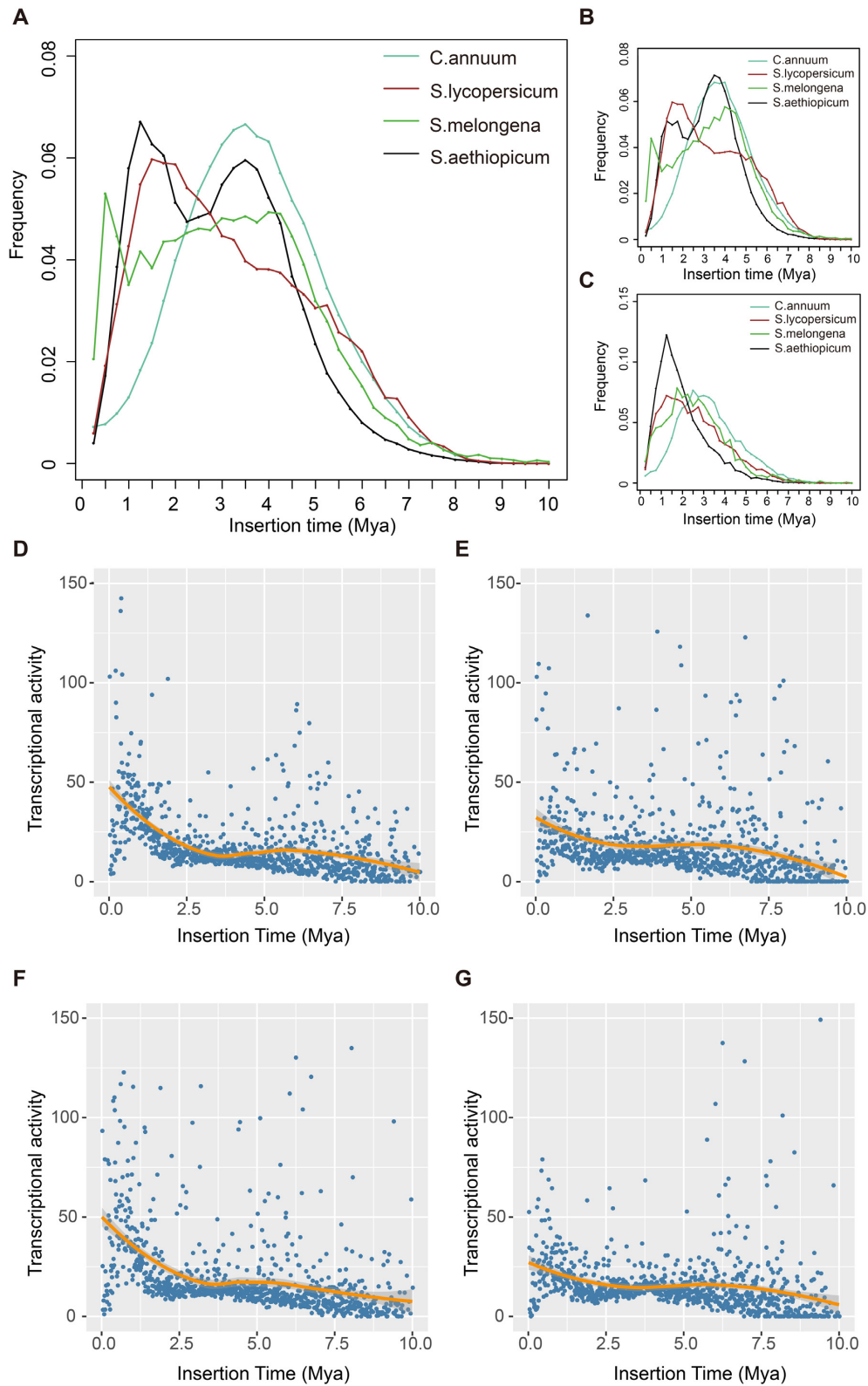
**Figure 1:** Comparative analysis of the *Solanum aethiopicum* genome. (A) Phylogenetic analysis of *Solanum melongena*, *S. lycopersicum*, *S. tuberosum*, *S. aethiopicum*, and *Capsicum annuum* using single-copy gene families. The species differentiation time between *S. aethiopicum* and *S. melongena* was 2.6 million years. (B) Distribution of 4DTV distance, which showed 2 peaks ~0.25 and 1 (black line), representing 2 whole-genome duplication events. (C) Venn diagram showing overlaps of gene families between *S. melongena*, *S. lycopersicum*, *S. tuberosum*, *S. aethiopicum*, and *C. annuum*. A total of 465 gene families were unique to *S. aethiopicum* and 10,166 were common to the genomes of the 5 species. *V. vinifera*: *Vitis vinifera*.

ent species, we observed that the insertion time of Ty1/Copia LTR-RTs in *S. aethiopicum* and tomato were earlier than that of *S. melongena* and hot pepper. On the contrary, the insertion time of Ty3/Gypsy LTR-RTs (~3.5 MYA) in *S. aethiopicum* was consistent with the insertion time of hot pepper (Fig. 2B and C).

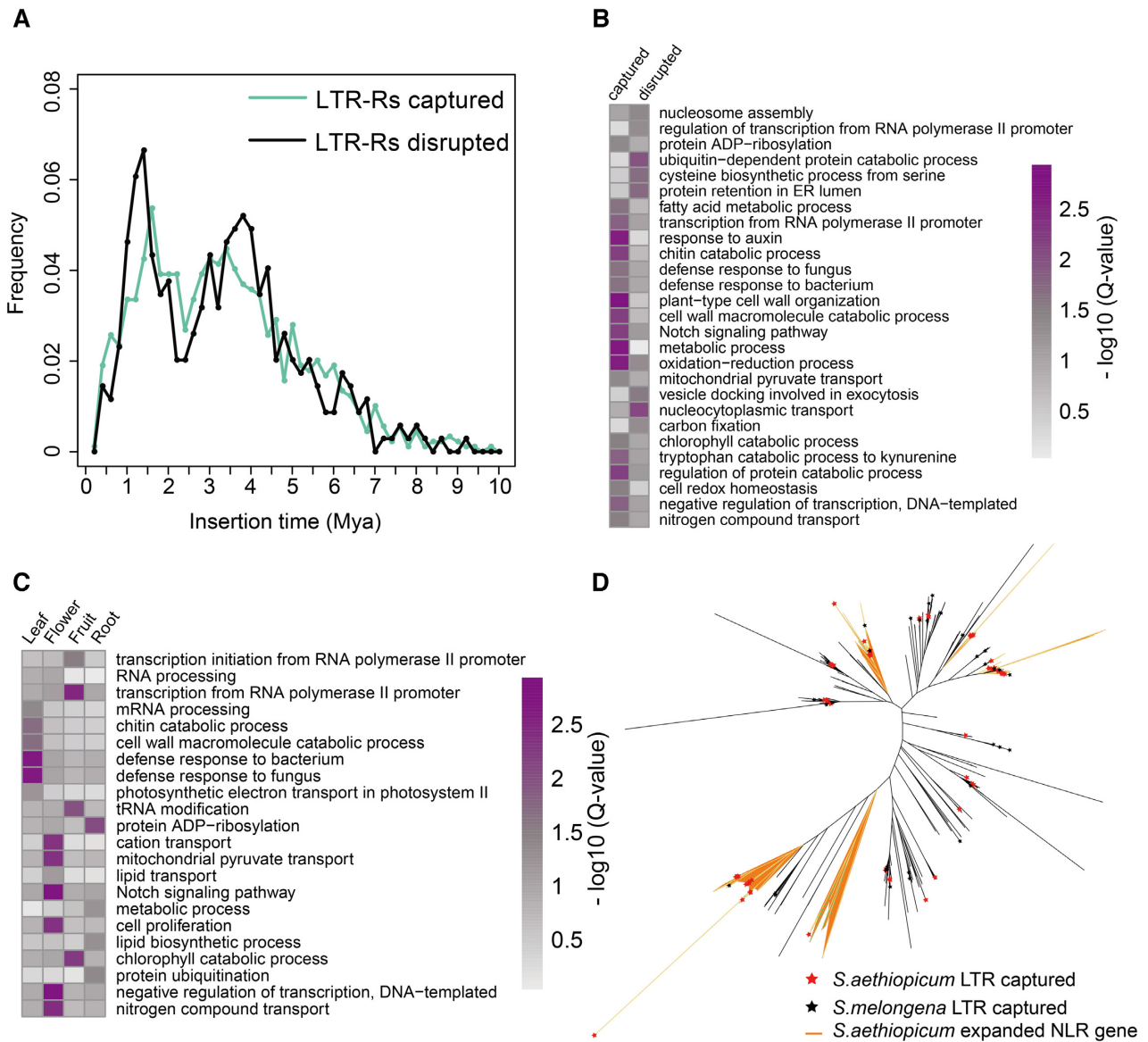
To investigate the activities of these LTR-Rs, we measured their expression levels by using RNA-sequencing (RNA-seq) data from different tissues (see Methods). Younger LTR-Rs were expressed at higher levels than those of older LTR-Rs. We detected 2 peaks of LTR-R activity, at positions corresponding to the 2 rounds of LTR-R insertions (Fig. 2D–G). The slight shift of the former peaks indicates that the activities degenerated more slowly than the LTR-R sequences (Fig. 2D–G). The LTR-R activities varied across these tissues. The degeneration of LTR-R activities was slower in fruits and roots than those in flowers and leaves (Fig. 2D). This pattern was also confirmed by the varied activity of each LTR-R across these tissues (Supplementary Fig. 3A), implying that these LTR-Rs have different roles in development.

### Increased resistance is facilitated by LTR-R amplification

We identified 1,156 LTR-R captured genes and 491 LTR-R disrupted genes. The insertion time of LTR-R captured and LTR-R disrupted genes both ranged between 1.5 and 3.5 MYA (Fig. 3A), showing a pattern similar to the insertions of whole LTR-Rs (Fig. 2A). These results suggest that LTR-R-mediated gene disruption and capture occurred simultaneously. We further classified the LTR-R captured genes into Gene Ontology (GO) categories and performed GO enrichment analysis. GO terms related to disease resistance including “defense response to fungus (GO:0006952),” “chitin catabolic process (GO:0006032),” “chitinase activity (GO:0004568),” “chitin binding (GO:0008061),” “cell wall macromolecule catabolic process (GO:0016998),” and “defense response to bacterium (GO:0042742)” were overrepresented in the LTR-R captured genes (Fig. 3B, Supplementary Table 12),



**Figure 2:** Long terminal repeat retrotransposon (LTR-R) insertion time distribution and the expression level of LTR-Rs in different tissues. Insertion time distribution of total LTR-Rs (A), Ty3/Gypsy LTR-Rs (B), and Ty1/Copia LTR-Rs (C) of *Capsicum annuum*, *Solanum melongena*, *S. lycopersicum*, and *S. aethiopicum*. The x- and y-axes, respectively, indicate the insertion time and the frequency of inserted LTR-Rs. Expression levels of LTR-Rs in flower (D), fruit (E), leaf (F), and root (G) tissues.



**Figure 3:** LTR-R captured and disrupted genes. (A) The distribution of ages of LTR-R captured and disrupted genes. (B) GO enrichment analysis between the LTR-R captured and disrupted gene set. (C) GO terms enriched in LTR-R captured genes that are specifically and highly expressed in various tissues, including leaf, flower, root, and fruit. (D) Phylogenetic tree of the nucleotide-binding, leucine-rich repeat-related (NLR) gene in *Solanum aethiopicum* and *S. melongena*. ADP: adenosine diphosphate; ER: endoplasmic reticulum.

suggesting that they may be involved in enhancing disease resistance.

We also analyzed the expression of genes captured by LTR-Rs. It was intriguing to find that most of these genes were active in only 1 tissue (Supplementary Fig. 3B). Among these genes, 159 (13.75%), 105 (9.08%), 106 (9.16%), and 129 (11.15%) were specifically and highly expressed in root, leaf, flower, and fruit, respectively. The genes captured by LTR-Rs that were specifically active in leaf tissues were significantly enriched in functions relating to disease resistance (Supplementary Table 13). The biological processes and molecular activities related to disease resistance mentioned above were overrepresented in these genes (Fig. 3C). The high expression level of resistance genes in leaves would arm the plant with stronger resistance to pathogens. On the contrary, these GO terms were not enriched in the genes that were specifically and highly expressed in leaves. Instead, as ex-

pected, “photosynthesis” and “photosystem I” were significantly overrepresented (Supplementary Table 14). The discrepancy between these 2 gene sets highlights the contribution to resistance of LTR-R captured genes.

Proteins containing nucleotide-binding, leucine-rich repeat domains (NB-LRRs) are major components that are responsible for defense against various phytopathogens [32]. The NB-LRR family is highly expanded in plants, with numbers ranging from <100 to >1,000 [33, 34]. As NB-LRR genes are often co-localized with LTR-Rs [35], we inspected their genomic locations in the *S. aethiopicum* genome. Because proteins containing the nucleotide-binding (NB) site can also confer disease resistance, we searched for all the NB-containing genes in the genome. As a result, we identified 447 NB-containing genes in the genome, among which 62 (13.9%) NB-containing genes co-localized with LTR-Rs were identified as LTR-R captured genes. The

**Table 2:** Statistical data for single-nucleotide polymorphisms and indels of 65 accessions

Type	Class	No. (%)
SNPs	Exon	392,160 (2.11)
	Intron	669,855 (3.60)
	Intergenic	17,552,823 (94.29)
	Synonymous	126,172 (0.68)
	Nonsynonymous	267,710 (1.44)
	<b>Total</b>	<b>18,614,838</b>
Indels	Exon	32,349 (1.62)
	Intron	145,362 (7.27)
	Intergenic	1,821,530 (91.11)
	Frame shift	2,977 (0.15)
	<b>Total</b>	<b>1,999,241</b>

phylogenetic tree shows a substantial expansion of NB-containing genes after the amplification of LTRs in *S. aethiopicum* (Fig. 3D). A similar expansion was also observed in *S. melongena*. However, the number was remarkably fewer than in *S. aethiopicum*, probably because of the limited number of LTR-Rs in the *S. melongena* genome (Supplementary Table 15).

### Polymorphisms in different *S. aethiopicum* groups

We resequenced 60 *S. aethiopicum* genotypes in 2 major groups, Gilo and Shum, and 5 accessions of *S. anguivi*, the progenitor of *S. aethiopicum* [36]. We generated ~60 Gb raw data (60×) (Supplementary Table 20) and identified 18,614,838 SNPs and 1,999,241 indels, with an average of 3,530,488 SNPs for each accession (Supplementary Table 16). On average, there were 18,090 SNPs and 1,943 indels per megabase. Among them, 424,509 (2.06%), 815,217 (3.95%), and 19,374,353 (93.99%) were located in exons, introns, and intergenic regions, respectively (Table 2). There were 267,710 SNPs that resulted in amino acid sequence changes by introducing new start codons, premature stop codons, or nonsynonymous substitutions (Table 2). We also identified 1,255,302 structural variations (SVs). Of the detected indels, 177,711 (8.89%) were located in genic regions, among which 2,977 caused frameshift changes and, therefore, resulted in amino acid sequence changes that may have led to gene malfunctions. Furthermore, 106,377 SVs were identified in genic regions, including 53,736 (50.51%) deletions, 34,368 (32.31%) insertions, and 8,872 (8.34%) duplications.

On counting the SNPs and indels in each group, we found 12,777,811, 15,165,053 and 8,557,818 SNPs in “Gilo,” “Shum,” and “*S. anguivi*,” respectively, accounting for 68.64%, 81.47%, and 45.97% of the total SNPs, respectively. There were 2,019,539 (10.85%), 4,747,418 (25.50%), and 587,885 (3.16%) SNPs unique to Gilo, Shum, and *S. anguivi*, respectively (Fig. 4A). Most (93.13%) SNPs in *S. anguivi* were shared with either Gilo or Shum (Fig. 4A), which is in line with the fact that *S. anguivi* is the ancestor [36]. Similarly, 92.62% of the indels identified in *S. anguivi* were also shared with Gilo or Shum (Fig. 4B).

Nucleotide diversity ( $\pi$ ) of all the genotypes was determined to be  $3.58 \times 10^{-3}$  for whole genomes,  $2.06 \times 10^{-3}$  for genic regions, and  $3.75 \times 10^{-3}$  for intergenic regions. Nucleotide diversity for each genotype revealed lower diversity for Gilo (*S. anguivi*:  $3.16 \times 10^{-3}$ , Shum:  $3.65 \times 10^{-3}$ , and Gilo:  $2.55 \times 10^{-3}$ , respectively). Linkage disequilibrium (LD) estimation using Haploview (version 4.2) [37] revealed that  $r^2$  reached the half maximum value at ~150 kb (Fig. 4C), which is smaller than in

other Solanaceae crops, e.g., tomato (2,000 kb) [38]. Because *S. aethiopicum* has been routinely used to improve disease resistance in eggplant and other Solanaceae crops [14], we further identified SNPs that were strongly associated with resistance genes by selecting those lying within resistance genes. A total of 34,171 SNPs were finally selected, which could be used in the selection of Solanaceae plants with disease resistance (Supplementary Table 16).

### Population structure and demography of *S. aethiopicum*

To investigate the evolution and population demography of *S. aethiopicum*, we first built a maximum-likelihood (Fig. 5A, Supplementary Fig. 4) phylogenetic tree using the full set of SNPs. We observed population structure in the genome-wide diversity. As anticipated, the accessions from Gilo and Shum were clearly separated in the tree, with only 1 exception in each group, probably caused by labelling errors. On the other hand, accessions of *S. anguivi*, the known ancestor of *S. aethiopicum*, did not cluster separately, but grouped with either Gilo or Shum. This structure was also supported by principal component analysis (PCA), which clearly separated these accessions into 2 clusters (Fig. 5B, Supplementary Fig. 5).

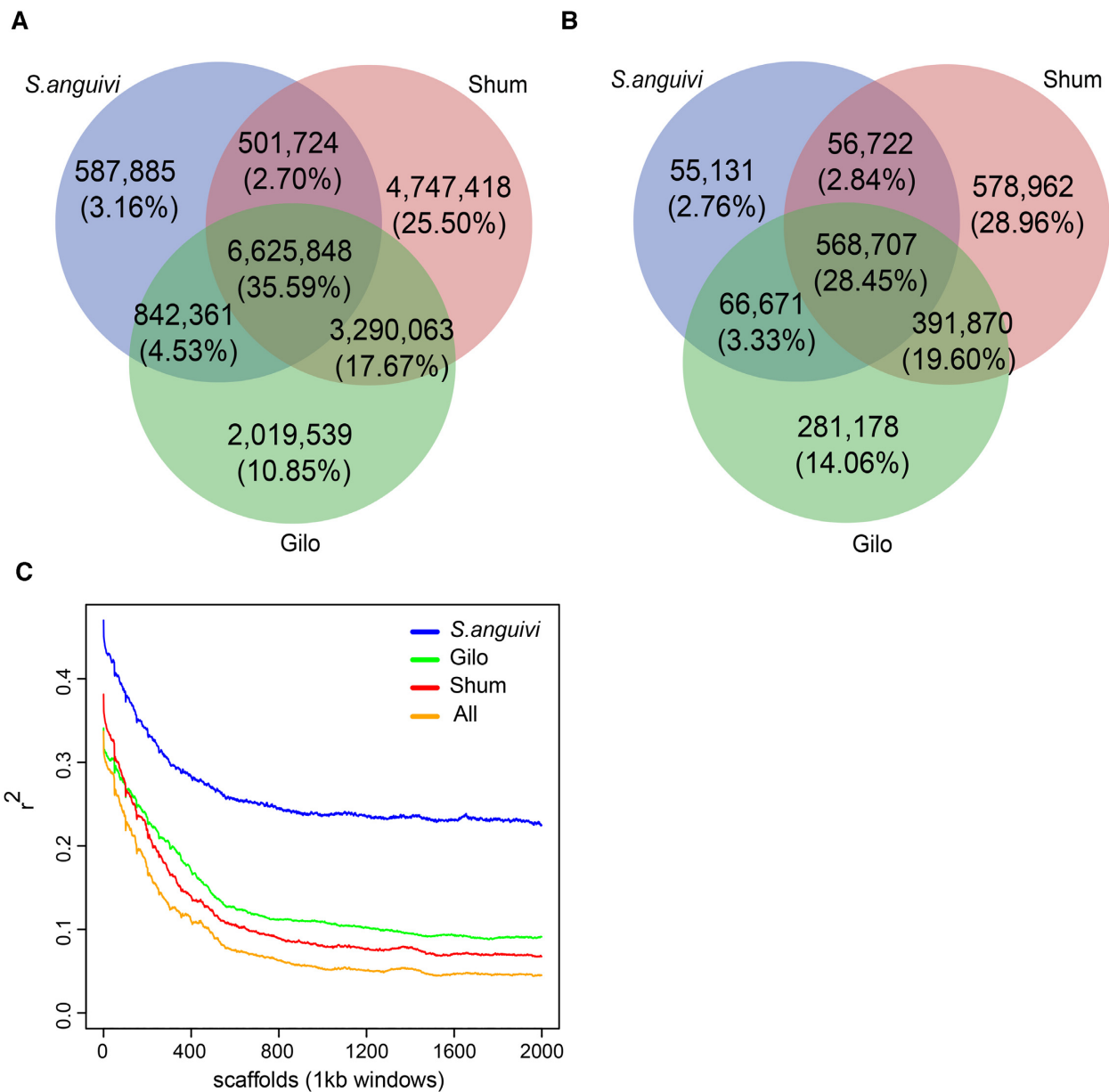
The domestication history of *S. aethiopicum* was inferred by constructing a multilevel population structure using ADMIXTURE [39]. This enabled us to estimate the maximum-likelihood ancestry (Fig. 5A). The parameter K, representing the number of subgroups to be divided, was set in the range of 2–9, and the cross-validation (CV) error was calculated individually. The CV error converged to 0.4375 when K = 6, suggesting the division of the resequenced accessions into 6 subgroups: I–VI (Fig. 5A). The structure changes with increasing K-value from 2 to 6, showing a time-lapse domestication history of *S. aethiopicum* that was first split into 2 groups, Gilo and Shum. The former was subsequently divided into subgroups I and II. Two groups emerged in Shum when K = 3, each of which was then divided into 2 subgroups when K = 6. In summary, Gilo was divided into 2 subgroups (I and II) and Shum was divided into 4 subgroups (III–VI).

The demographic history of *S. aethiopicum* was inferred using the pairwise sequential Markovian coalescent model [40]. By doing this, we inferred changes in the effective population sizes of *S. aethiopicum* (Fig. 5C). Our data revealed distinct demographic trends from 10,000 to 100 years ago, in which a bottleneck was shown ~4,000–5,000 years ago, followed by an immediate expansion of population size. The great population expansion might be associated with the early domestication of *S. aethiopicum* in Africa because it coincides with human population growth in western Africa, also occurring 4,000–5,000 years ago [41].

### Artificially selected genes in *S. aethiopicum*

We used reduce of diversity (ROD) and fixation index-statistics (Fst) measures to detect artificially selected regions along the genome. Briefly, ROD and Fst were calculated in a sliding non-overlap 10-kb window. Regions with ROD > 0.75 and Fst > 0.15 were identified as candidate regions under selection. As a result, genomic regions of 3,238 and 1,062 windows were found to be under selection during the domestication of Gilo and Shum, respectively (Supplementary Table 17). Among them, 161 windows were common between these 2 groups, while 3,077 and 901 windows were unique to Gilo and Shum, respectively. Genes located within these regions were identified as selected genes. Thirty-six and 1,406 selected genes were identified in Shum and Gilo, respectively, and 12 of these genes were

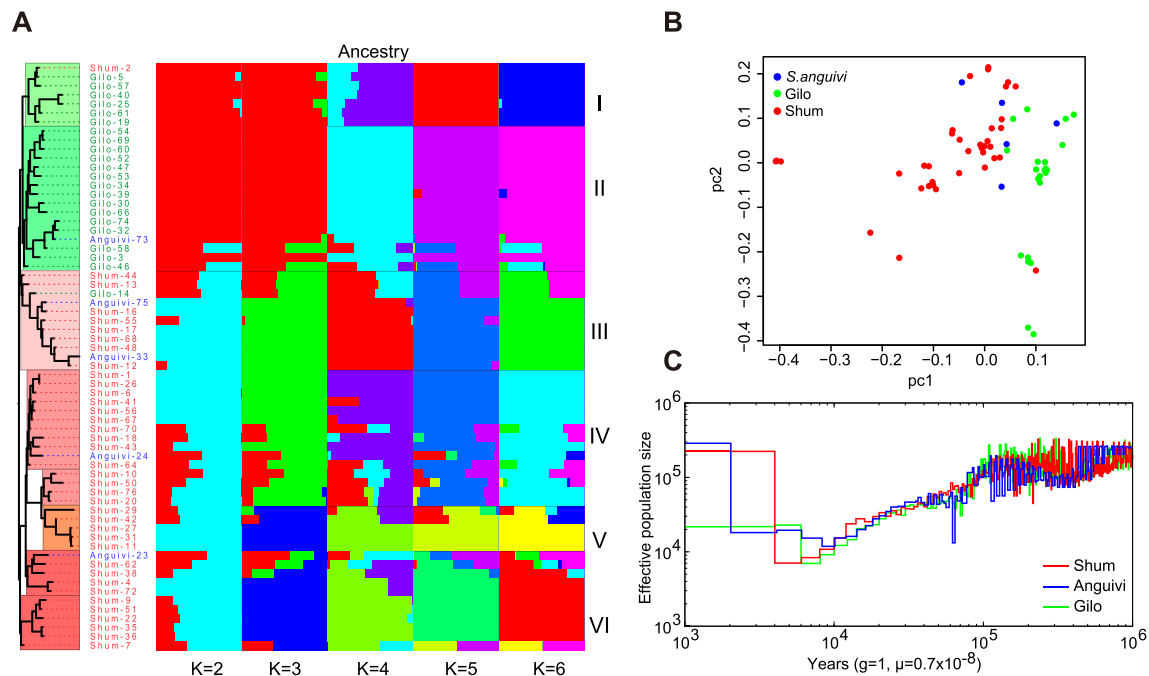




**Figure 4:** Single-nucleotide polymorphisms (SNPs), indel, and linkage disequilibrium (LD) decay for "Gilo," "Shum," and "*S. anguivi*" groups. (A) SNPs numbering 2,019,539 (10.85%), 4,747,418 (25.50%), and 587,885 (3.16%) were unique to Gilo, Shum, and *S. anguivi*, respectively. Most (93.13%) of SNPs in *S. anguivi* were shared with either Gilo or Shum. (B) Indels amounting to 14.06%, 28.96%, and 2.76% were unique to Gilo, Shum, and *S. anguivi*, respectively, and, like the SNP statistics in these groups, 92.62% of indels in *S. anguivi* were shared with either Gilo or Shum. (C) LD estimation revealed that  $r^2$  reaches the half maximum value at ~150 kb.

selected in both. Ten of the 12 genes were annotated in the SwissProt database with known functions and included many genes known to be involved in tolerance to unfavorable environmental stresses, such as autophagy-related gene 18f (*ATG18f*), ATP-binding cassette transporter B (*ABCB18*), lysine-tRNA ligase (*LYSRS*), acyl-coenzyme A oxidase 4 (*ACX4*), inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase (*VIP2*) (Supplementary Table 18). For example, *ATG18* is reported to be involved in defense response to powdery mildew fungus through autophagy in *Arabidopsis* [42]; it is also involved in response to nutrition starvation by serving as an accessory component to *ATG1/13* kinase complex [43]. *ABCB* is reported to be associated with lipid transport and confers tolerance to heavy

metal ions, such as aluminium [44], cadmium, and lead [45]. The expression of *LYSRS* has been shown to be specifically induced in tomato root during the unusual accumulation of metal ions [46]. *VIP2* is reported to be critical in myo-inositol phosphate signalling pathways, and is known to be involved in responses to drought and salt stresses [47]. Furthermore, 2 genes encoding pentatricopeptide repeat-containing protein were also found among these genes, suggesting that RNA editing may have played a crucial role in the domestication of *S. aethiopicum* [48]. GO enrichment analysis showed that genes selected in both the Gilo and Shum groups were enriched in "transport" (Supplementary Table 19). GO terms for "response to auxin," "response to hormone," "response to salt stress" and "response to water"



**Figure 5:** Population structure and demographic history of *Solanum aethiopicum*. (A) A maximum-likelihood phylogenetic tree and population structure constructed using the full set of single-nucleotide polymorphisms (SNPs). (B) Principal component analysis (PCA). (C) Pairwise sequential Markovian coalescent model analysis indicated a distinct demographic history of *S. aethiopicum* from 10,000 to 100 years ago, in which a bottleneck was shown ~4,000–5,000 years ago, followed by an immediate expansion of population size.

were also overrepresented in genes selected either in Gilo or Shum only. This result could explain the enhanced tolerance to drought and salinity in *S. aethiopicum*.

We also focused on the diversity of genes co-localized with LTR-Rs. A total of 24,682 SNPs were located within these co-localized genes, corresponding to 0.133% of the total number of SNPs (18,614,838). This is substantially fewer than would be expected if SNPs were evenly distributed across all genes, particularly because the LTR-R co-localized genes comprise 3.31% of the total gene set. The repellent of SNPs in these genes suggests purifying selection, which was also supported by the large amount (9,728; 39.41%) of rare SNPs (minor allele frequency <5%) found among the co-localized genes. We also observed that nonsynonymous SNPs (9,544) were much more abundant than synonymous ones (5310) among the co-localized genes. These variations led to amino acid changes in the encoded proteins, which may have contributed to the diversification of resistance genes.

### Pan- and core-genome of *S. aethiopicum*

Gene content varies across different accessions. A single reference assembly is insufficient to include all *S. aethiopicum* genes. Therefore, we assembled contigs for individual accessions using pair-end reads, with coverages ranging from 30 to 60× (Supplementary Table 20).

We assembled the genomes individually using SOAPdenovo2 [49] and filtered out contigs smaller than 2 kb. As a result, 753,084 contigs were retained, among which 432,785 were from Shum, 260,119 were Gilo and 60,180 were from *S. anguivi*. These contigs were further pooled separately and cleaned by removing duplicates using CD-HIT [50]. This led to the retention of 97,429, 76,638, and 36,915 contigs for Shum, Gilo and *S. anguivi*, respectively. The annotation of these contigs resulted in 41,626, 33,194,

and 17,662 protein-coding genes, among which we identified accessory gene sets of 29,389, 23,726, and 12,829 for Shum, Gilo, and *S. anguivi*, respectively, by comparing against the reference genome sequence. We generated a pan-genome of *S. aethiopicum* (including Shum, Gilo and *S. anguivi* groups) of 51,351 genes (Supplementary Table 21). These genes were further clustered together with those annotated in the reference using CD-HIT. Overall, we identified 7,069 genes unique to the pan-genome gene set, suggesting that they had been missed from the reference. The average length of accessory genes was 1.62 kb with 2.22 introns. This is comparable to gene models in the reference genome, providing further evidence of accurate annotation. We further assigned their putative functions by querying against protein databases. A total of 48,572 (94.59%) genes were fully annotated and functional descriptions (Supplementary Table 22) provided. Among the identified gene models, 10,409 (20.27%) were common to these 3 groups and were thus defined as “core” genes. As expected, they were mainly composed of housekeeping genes (Supplementary Table 23). However, it is important to note that the number of core genes may have been underestimated because *S. anguivi* was underrepresented, while the other 2 *S. aethiopicum* groups, Kumba and Aculeatum, were not included in the present study.

### Discussion

*Solanum aethiopicum* is cross-compatible with *S. melongena* and is routinely used as a donor of disease resistance genes to its close relative [14]. Genomic analysis of *S. aethiopicum* revealed higher LTR-mediated expansion of resistance gene families than its other close relatives, including tomato, potato, eggplant, and hot pepper. LTR amplification is one of the major forces driving genome evolution. It shapes the genome by capturing, interrupting, or flanking genes [51]. The consequences of LTR insertions

depend on the genomic position of insertion. For example, inserting into protein-coding sequences results in pseudogenization. LTR-Rs adjacent to protein-coding genes can downregulate or silence the expression of flanking genes by extending methylation regions or by producing antisense transcripts [52–55]. LTR-Rs also mediate gene retroposition, capturing genes back into the genome [51]. In the present study, LTRs preferentially captured genes related to disease resistance, resulting in the overrepresentation of GO terms related to disease resistance in the LTR-captured genes. Enrichment of the GO terms “chitin binding (GO:0008061)” and “chitinase activity (GO:0006032)” (Fig. 3B, Supplementary Table 12) implies that these genes may have been selected to resist infection by fungal pathogens, such as *Fusarium oxysporum* [56]. On the contrary, no GO term enrichment was seen in genes that were disrupted by LTR-Rs. This suggests that gene disruption by LTR-Rs may be a random event in terms of gene function. The age distribution of LTR-R captured genes coincidentally fit with that of the LTR-R-disrupted genes, suggesting that these 2 events may have occurred simultaneously (Fig. 3A). It is not clear why genes related to disease resistance were favoured by LTR-Rs, but one explanation is that the disease resistance genes may have been more active than other genes at the time of LTR retrotransposition. The expression pattern of LTR-R captured genes also varied between tissues. Those related to resistance were specifically active in the leaf, while those engaged in the transport of cations, nitrogen, and cell proliferation were active in flowers. This outcome suggests low abundance of transcripts for disease resistance genes, resulting in a relatively low chance to adequately capture the genes in flowers under normal conditions. Another possible scenario is that LTR retrotransposition occurred under stress conditions, which resulted in the simultaneous induction of the expression of resistance genes in gametes and the activity of LTR retrotransposition. Such possible stresses might be extreme environmental conditions or pathogen infection. A “reinforcement model” has been proposed to explain the simultaneous accumulation of stress-responsive genes and the activity of retrotransposons in genomes under environmental stress [57, 58].

There are 4 major groups of *S. aethiopicum*: “Gilo,” “Shum,” “Kumba,” and “Aculeatum.” We resequenced accessions from the Gilo and Shum groups, which are widely consumed as vegetables. The accessions resequenced in this study were clustered into 6 subgroups (4 for Shum and 2 for Gilo). By scanning for regions with lower genomic diversity, we identified regions and several genes involved in responses to salt, water, and drought tolerance that were under selection during the domestication of *S. aethiopicum*. Furthermore, purification selection was also found among disease resistance genes.

In the present study, resequencing *S. aethiopicum* and *S. anguivi* genomes at a high depth (30–60×) (Supplementary Table 20) enabled us to assemble draft genomes for these individuals. Despite resequencing only a few genotypes from the 2 groups, we intend to supplement the reference gene set with accessory genes by pooling the resequenced contigs for gene prediction and annotation. This “pan-genome” is expected to provide a more comprehensive understanding of *S. aethiopicum* in the future.

We report a reference genome for African eggplant, which will provide a basic data resource for further genomic research and breeding activities for *S. aethiopicum*. The gene sequences annotated in the genome will be essential for developing genome editing vectors to create mutants to further understand the functions of genes within the genome and develop superior genotypes. Molecular markers developed using

the genome sequences will also enable more efficient and precise selection of superior accessions by breeders.

## Methods

### DNA extraction, library construction and sequencing, and genome assembly

High molecular weight genomic DNA was extracted from young leaves of 14-day-old seedlings of *Solanum aethiopicum* “Shum” accession 303, which had been previously and repeatedly selfed to ensure homozygosity. Shum 303 is a selection of African eggplant from Uganda, with green fruits and pigmented stem and leaf veins. DNA was extracted using a modified cetyl trimethylammonium bromide (CTAB) protocol, as previously described [59]. Briefly, 2.5 g fresh leaf tissue was flash-frozen in liquid nitrogen and ground to a fine powder, before adding 15 mL of 2× extraction buffer (100 mM Tris-HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2% w/v CTAB, 10 μL/mL β-mercaptoethanol), then incubated at 65°C. One volume of chloroform: isoamyl alcohol (24:1) was added and mixed and the sample was centrifuged twice. The aqueous phase was precipitated overnight and the washed pellet was treated with RNaseA. A repeat chloroform extraction was performed, as above, to remove RNaseA and any other contaminants. The aqueous phase was collected and DNA was precipitated and washed with ethanol. DNA was allowed to dry, then was resuspended in 100 μL elution buffer.

High molecular weight DNA was fragmented and used to construct paired-end libraries with insert sizes of 250 bp, 500 bp, 2 kb, 6 kb, 10 kb, and 20 kb, following standard Illumina protocols. The libraries were sequenced on an Illumina HiSeq 2000 platform, resulting in a total of 242.61 Gb raw reads. Filtering of duplicated, low-quality reads and reads with adaptors was done using SOAPfilter (version 2.2, an application included in the SOAPdenovo2 package, [RRID:SCR.014986](https://doi.org/10.1093/bioinformatics/btu088)) [49] with the parameters “-M 2, -f 0, -p”. Reads with ≥40% low-quality bases or with ≥10% uncalled bases (“N”) were filtered. We used 17 k-mer counts [21] of high-quality reads from small insert libraries to evaluate the genome size and heterozygosity using GCE [60] and Kmergenie [61]. We assembled the genome using Platanus ([RRID:SCR.015531](https://doi.org/10.1093/bioinformatics/btu088)) [22].

Genomic DNA used for resequencing was extracted from young leaves of 65 accessions. DNA was sheared into small fragments of ~200 bp and used to construct paired-end libraries, following standard BGI protocols as previously described [62], and subsequently sequenced on a BGI-500 sequencer. Briefly, the DNA fragments were ligated to BGISEQ-500 compatible adaptors, followed by an index PCR amplification, the products of which were then pooled and circularized for sequencing on the BGISEQ-500 (BGI, Shenzhen, China). Ultra-deep data were produced for each accession, with coverage ranging from ~45 to ~75× (Supplementary Table 20).

### RNA extraction, library construction, and sequencing

For RNA extraction, seeds of Gilo and Shum inbred lines were obtained from Uganda Christian University. The seeds were planted in a screenhouse at the Beca-ILRI Hub (Nairobi, Kenya) in polyvinylchloride pots (13 cm height and 11.5 cm diameter) containing sterile forest soil and farmyard manure (2:1). The seedlings were later transplanted into larger polyvinylchloride pots of 21 cm height and 14 cm diameter. Plants were raised in a screenhouse at 21–23°C and 11–13°C day and night temperatures, respectively (average 12 light hours per day). The plants

were regularly watered to maintain moisture at required capacity.

Two plants were selected randomly from each of Gilo and Shum accessions and were tagged at the seedling stage for tissue sampling. Fresh tissues were sampled from each of the tagged plants and flash-frozen in liquid nitrogen immediately. Total RNA was extracted from the frozen tissues using the ZR Plant RNA Miniprep™ Kit (Zymo Research, Irvine, CA, USA), according to the manufacturer's instructions. RNA integrity was evaluated by electrophoresis in denaturing agarose gel (1% agarose, 5% formamide, 1× TAE) stained with 3× GelRed (Biotium, Fremont, CA, USA). RNA was quantified using the Qubit RNA Assay Kit (Thermo Fisher Scientific, Carlsbad, CA, USA). Ribosomal RNA (rRNA) was removed from 4 μL of total RNA from each sample using the Epicentre Ribo-zero™ rRNA Removal Kit (Epicentre, Madison, WI, USA). The rRNA-depleted RNA was then used to generate strand-specific RNA-seq libraries using TruSeq® Stranded mRNA Kit (Illumina, San Diego, CA, USA). Twenty mRNA libraries were prepared, multiplexed (10 samples at a time), and sequenced as paired-end reads on the MiSeq (Illumina, San Diego, USA) platform at the BeCA-ILRI Hub. Similar to the process of filtering genomic reads, SOAPfilter software [49] was used, with the parameters “-M 2, -f 0, -p” to filter low-quality reads and adaptor sequences. Reads with ≥40% low-quality bases or with ≥10% uncalled bases (“N”) were filtered out.

### Repeat annotation

Tandem repeats were searched in the genome using TRF, version 4.04 [63]. Transposable elements (TEs) were identified by a combination of homology-based and *de novo* approaches. Briefly, the assembly was aligned to a known repeats database (Repeatbase16.02) using RepeatMasker (RRID:SCR\_012954) and Repeat-ProteinMask (version 3.2.9) [64] at both the DNA and protein level. In the *de novo* approach, RepeatModeler (version 1.1.0.4, RRID:SCR\_015027) [65] was used to build a *de novo* repeat library using the *S. aethiopicum* assembly, in which redundancies were filtered out. TEs in the genome were then identified by RepeatMasker [64]. LTRs were identified using LTRharvest [29], with the criterion of 75% similarity on both sides. LTRdigest [30] was used to identify the internal elements of LTR-Rs with the eukaryotic tRNA library [66]. Identified LTR-Rs including intact poly purine tracts and primer binding sites with LTR-Rs on both sides were considered to be the final intact LTR-Rs. These were then classified into superfamilies, *Gypsy* and *Copia*, by querying against Repeatbase 16.02 [67].

### Annotation of gene models and ncRNA

Gene models were predicted using a combination of *de novo* prediction, homology search, and RNA-aided annotation. Augustus software (RRID:SCR\_008417) [68] was used to perform *de novo* prediction after the annotated repeats were masked in the assembly. To search for homologous sequences, protein sequences of 4 closely related species (*S. lycopersicum*, *S. tuberosum*, *Cap-sisium annuum*, and *Nicotiana glauca*), together with *Arabidopsis thaliana*, were used as query sequences to search the reference genome sequence using TBLASTN (RRID:SCR\_011822) [69] with the e-value ≤1e−5. Regions mapped by these query sequences were subjected to GeneWise (RRID:SCR\_015054) [70], together with their flanking sequences (1,000 bp) to identify the positions of start/stop codons and splicing. For RNA-aided annotation, RNA-seq data from different tissues of *S. aethiopicum* were

mapped to the genome assembly of *S. aethiopicum* using HISAT (RRID:SCR\_015530) [71]. Mapped reads were then assembled using StringTie (RRID:SCR\_016323) [72]. GLEAN software [73] was used to integrate mapped transcripts from different sources to produce a consensus gene set. tRNAscan-SE (RRID:SCR\_010835) [74] was performed to search for reliable tRNA positions. snRNA and miRNA were detected by searching the reference sequence against the Rfam database (RRID:SCR\_007891) [75] using BLAST [69]. rRNAs were detected by aligning with BLASTN (RRID:SCR\_004870) [69] against known plant rRNA sequences [76]. For functional annotation, protein sequences were searched against Swissprot, TrEMBL, KEGG (release 88.2), InterPro, Gene Ontology, COG, and Non-redundant protein NCBI databases [77–82].

### Gene family analysis

Proteins of *S. aethiopicum*, *S. tuberosum* (PGSC v3.4) [18], *S. lycopersicum* (v2.3) [19], *C. annuum* (PGA v1.6) [24], and *S. melongena* (Sme2.5.1) [83] were selected to perform all-against-all comparisons using BLASTP (RRID:SCR\_001010) [69], with an e-value cutoff of ≤1e−5. OrthoMCL (RRID:SCR\_007839) [26] and the default MCL inflation parameter of 1.5 were used to define the gene families. Single-copy families were selected to perform multiple sequence alignment using MAFFT (RRID:SCR\_011811) [84]. Four-fold degenerate sites were picked and used to construct a phylogenetic tree based on the maximum-likelihood method by PhyML (RRID:SCR\_014629) [85], with *C. annuum* as the outgroup. WGD analysis was achieved by identifying colinearity blocks by paralog gene pairs in MCscanX, with default parameters [27]. Each aligned paralog gene pair was concatenated to a super-sequence in 1 colinearity block and 4dTv (transversion of 4-fold degenerate site) values of each block were calculated. We also determined the distribution of 4dTv values to estimate the speciation between species or WGD events. The divergence time of *S. aethiopicum* was estimated using the MCMCTree program [86], with the constructed phylogenetic trees and the divergence time of *C. annuum* [24] and *S. tuberosum* [18].

### Analysis of LTR-Rs

Insertion times of identified, intact LTR-Rs were estimated on the basis of the sequence divergence between the 5′ and 3′ LTR of each element. The nucleotide distance *K* between 1 pair of LTR-Rs was calculated using the Kimura 2-parameter method in Distmat (EMBOSS package) [87]. An average base substitution rate of 1.3e−8 [31] was used to estimate the insertion time, based on the following formula:

$$T = K/2r [15].$$

Transcriptomic data were used to analyse the activity of intact LTR-Rs. After filtering and removing low-quality reads, high-quality reads from each were mapped against the full-length LTR-R sequence using BWA-MEM software [88], with default parameters. Expression levels of intact LTR-Rs were calculated using EdgeR [89] and visually presented using heatmap in R [90].

### Analysis of NB-containing genes

NB domain-containing genes in the *S. aethiopicum* genome were identified using a method previously described [15, 91]. Briefly, the hidden Markov model (HMM) profile of the NB-ARC domain (PF00931) was used as a query to perform an HMMER search (version 3.2.1, RRID:SCR\_005305 [92]) against protein sequences of tomato, potato, hot pepper [18, 19, 24], and annotated sequences



of *S. aethiopicum*, with an e-value cut-off of  $\leq 1e-60$ . Aligned NB-ARC domain sequences of *S. aethiopicum* were extracted and used to build the *S. aethiopicum*-specific HMM model. NB-ARC domain sequences of tomato, potato, and hot pepper were mapped as the query sequences against the *S. aethiopicum* genome using TBLASTN [69], with an e-value cut-off of  $\leq 1e-4$  using GeneWise software [70] to identify candidate NB-containing genes at the whole-genome level. Final NB-containing genes were confirmed by searching the genome with an *S. aethiopicum*-specific NB-ARC HMM model, constructed with an e-value cut-off of  $\leq 1e-4$ . Retroduplicated NLRs were identified according to the method described by Kim et al. (2017) [15]. Phylogenetic trees for *S. aethiopicum* and *S. melongena* NB-containing genes were constructed using FastTree (RRID:SCR.015501) [93], with default parameters.

### SNP calling

The GATK pipeline (RRID:SCR.001876) [94] was used to call SNPs and indels. Briefly, low-quality, duplicated, and adaptor-contaminated reads were filtered using SOAPfilter (version 2.2) [49] before further processing. To reduce the compute time, scaffolds in the assembly were sequentially linked into 24 pseudo-chromosomes, in which the original scaffolds were separated by 100 Ns, before mapping reads using BWA (RRID:SCR.010910) [88], with default parameters. Picard Tools [95] and SAMtools (RRID:SCR.002105) [96] were used to further process the alignment outputs, including sorting and marking of duplicates. After alignment and sorting, the GATK pipeline (version 4.0.11.0) was used to call SNPs by sequentially implementing the following modules: RealignerTargetCreator, IndelRealigner, UnifiedGenotyper, samtools mpileup, VariantFiltration, BaseRecalibrator, AnalyzeCovariates, PrintReads, and HaplotypeCaller, with default parameters. This pipeline produced a file in gvcf format, which displayed the called SNPs and indels filtered according to genotype information. The file was then analysed using PLINK software [97] for quality control, with “GENO>0.05, MAF<0.1, HWE test p-value  $\leq 0.0001$ ” parameters (GENO: maximum per-SNP missing; MAF: minor allele frequency; HWE: Hardy-Weinberg disequilibrium P-value). The loci of these SNPs and indels were anchored back to the original scaffolds and annotated using SnpEff [98]. To identify structural variations (SVs), sample information was added using AddOrReplaceReadGroups, a module of Picard-tools, and SVs were detected using DiscoverVariantsFromContigAlignmentsSAMSpark, a GATK module.

### Population analysis

A maximum-likelihood phylogenetic tree was constructed, based on the genotypes at all the SNP loci using FastTree [93], with default parameters. To perform PCA, Beagle4.1 [99] was used to impute the unphased genotypes. All imputed and identified genotypes at SNP loci were pooled and finalized using PLINK [97] and ReSeqTools [100], which were then subjected to PCA using GCTA software [101]. The population was clustered using ADMIXTURE software [39], with K (the expected number of clusters) increasing from 2 to 9. The K value with the minimum cross-validation error was eventually selected.

Genome-wide linkage disequilibrium (LD) was calculated for populations of different groups using Haploview [102] in windows of 2,000 kb. Briefly, the correlation coefficient ( $r^2$ ) between SNP pairs in a non-overlapping sliding 1-kb bin was calculated and then averaged within bins.

Candidate regions under selection were identified by comparing polymorphism levels—measured by ROD, as well as by  $F_{ST}$ —between Gilo, Shum, and *Solanum anguivi* groups. ROD was calculated using the formula

$$ROD = 1 - \pi_{cul}/\pi_{wild},$$

where  $\pi_{cul}$  and  $\pi_{wild}$  denote the nucleotide diversity within the cultivated and wild populations, respectively.

$F_{ST}$  measurement was calculated according to the formula

$$F_{ST} = (\pi_{between} - \pi_{within})/\pi_{between},$$

where  $\pi_{between}$  and  $\pi_{within}$  represent the average number of pairwise differences between 2 individuals sampled from different or the same population.

### Construction of pan- and core-genome

To build a gene set including as many *S. aethiopicum* genes as possible, we assembled contigs of all 65 resequenced accessions individually using SOAPdenovo2 [49]. The assembled contigs from each group (Gilo, Shum, and *S. anguivi*) were then merged. CD-HIT-EST [50] was used to eliminate redundancy and generate the final dataset of pan-genomes for each group. Similarly, all these contigs were merged into a pan-genome of *S. aethiopicum*. Gene models were predicted from these contigs as described above and their functions were also annotated.

### Availability of supporting data and materials

The raw sequence data from our genome project were deposited in the NCBI SRA with BioProject number PRJNA523664 and in the CNGB Nucleotide Sequence Archive database under project accession number CNP0000317. Assembly and annotation of the *S. aethiopicum* genome are available in GigaDB [103].

### Additional files

Supplementary Table 1. Summary of the library types and data generated in this study  
 Supplementary Table 2. Statistics of the *S. aethiopicum* assembly  
 Supplementary Table 3. Comparison of the genomic characteristics in different genomes  
 Supplementary Table 4. Statistics of repeat annotation, transposable elements  
 Supplementary Table 5. Statistics of gene model in different species  
 Supplementary Table 6. Evaluation of predicted gene models using 1,440 CEGs  
 Supplementary Table 7. Statistics of predicted ncRNA  
 Supplementary Table 8. Result of function annotation  
 Supplementary Table 9. Statistics of gene families  
 Supplementary Table 10. Common shared gene families between *S. aethiopicum* and other species  
 Supplementary Table 11. GO annotation of expansion gene family of *S. aethiopicum*  
 Supplementary Table 12. GO classification and enrichment analysis of LTR-captured and LTR disrupted genes in *S. aethiopicum*  
 Supplementary Table 13. GO classification and enrichment analysis of LTR capture genes with specific activity in different tissues

Supplementary Table 14. GO classification and enrichment analysis of total specific active gene in 4 tissues  
 Supplementary Table 15. Number of annotated NLR genes in *S. aethiopicum*, *S. melongena* and other species  
 Supplementary Table 16. SNPs within resistance genes  
 Supplementary Table 17. Under selection genomic regions of Gilo and Shum  
 Supplementary Table 18. Gene position and functional annotation description for 12 Artificially selected genes (both in Gilo and Shum)  
 Supplementary Table 19. GO enrich for Artificially selected genes (Gilo and Shum)  
 Supplementary Table 20. Sequencing data of Pan-genome  
 Supplementary Table 21. Pan-genome annotation  
 Supplementary Table 22. Statistics of functional annotation for the protein-coding genes from the three groups (*S. anguivi*, Shum and Gilo) and pan-genome (All)  
 Supplementary Table 23. Statistics of genes which commended shared by three groups (*S. anguivi*, Gilo and Shum)  
 Supplementary Figure 1. 17-mer analysis for estimating the *S. aethiopicum* genome size  
 Supplementary Figure 2. Distributions of the gene model for four categories in the relative species  
 Supplementary Figure 3. Distinct expression pattern of LTR-Rs and their captured genes in different tissues  
 Supplementary Figure 4. Maximum-likelihood phylogenetic tree of 65 samples using the full-set of SNPs  
 Supplementary Figure 5. Principal-component analysis

## Abbreviations

4DTV: 4-fold degenerative third-codon transversion; BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; CEG: core embryophyta gene; COG: Clusters of Orthologous Groups; CV: cross-validation; Fst: fixation distance; GATK: Genome Analysis Toolkit; Gb: gigabase pairs; GC: guanine-cytosine; GCE: Genomic Character Estimator; GCTA: Genome-wide Complex Trait Analysis; GO: gene ontology; HMM: hidden Markov model; kb: kilobase pairs; KEGG: Kyoto Encyclopedia of Genes and Genomes; LTR: long terminal repeat; LD: linkage disequilibrium; LTR-R: long terminal repeat retrotransposon; MAFFT: Multiple Alignment using Fast Fourier Transform; Mb: megabase pairs; MYA: million years ago; NB-LRR: nucleotide-binding, leucine-rich repeat domain; NCBI: National Center for Biotechnology Information; NLR: nucleotide-binding, leucine-rich repeat-related; PCA: principal component analysis; RNA-seq: RNA-sequencing; rRNA: ribosomal RNA; SNP: single-nucleotide polymorphism; snRNA: small nuclear RNA; SRA: Sequence Read Archive; TE: transposable element; TrEMBL: Translation of European Molecular Biology Laboratory; TRF: Tandem Repeats Finder; tRNA: transfer RNA; WGD: whole-genome duplication.

## Competing interests

The authors declare that they have no competing interests.

## Funding

This work was supported by the National Natural Science Foundation of China (grant number 31601042), the Science, Technology and Innovation Commission of Shen-

zhen Municipality (grant numbers JCYJ20151015162041454 and JCYJ20160331150739027), and by the Guangdong Provincial Key Laboratory of Genome Read and Write (grant number 2017B030301011).

## Authors' contributions

D.A.O., X.X., A.V., X.Liu., H.S., R.J., A.M., J.W., and H.Y. conceived the project; D.A.O., F.S., E.B.K., A.V., S.C., and H.L. managed and supervised the work; B.S., Y.F. and X.C. managed the samples at BGI; B.S., Y.F. and W.C. assembled the whole genome; and Y.F., Y.S., and W.S. annotated the genome. S.N.K., S.M., and R.K. extracted high molecular weight DNA. H.L. and S.P. constructed DNA libraries and sequenced the genome. S.N.K. and S.M. prepared RNA libraries and sequenced the transcriptome. J.N. and S.N.K. assembled and analysed the transcriptome. Y.S., X.Li. and B.S. performed the analysis of gene families, LTR evolution, and transcriptomic data; P.N.K. extracted DNA for resequencing samples. Y.F. and B.S. analysed the resequencing data; Y.S., Y.F., S.W. and B.S. collected datasets required for the genome annotation and analyses. B.S., X.Liu., Y.S., D.A.O., P.S.H. and Y.F. wrote and revised the manuscript.

## Acknowledgements

We acknowledge Uganda Christian University for providing seeds of the African eggplant.

## References

1. Sunseri F, Polignano GB, Alba V, et al. Genetic diversity and characterization of African eggplant germplasm collection. *Afr J Plant Sci* 2010;**4**:231–41.
2. Adeniji O, Kusolwa P, Reuben S. Genetic diversity among accessions of *Solanum aethiopicum* L. groups based on morpho-agronomic traits. *Plant Genet Resour* 2012;**10**(3): 177–85.
3. Plazas M, Andújar I, Vilanova S, et al. Conventional and phenomics characterization provides insight into the diversity and relationships of hypervariable scarlet (*Solanum aethiopicum* L.) and gboma (*S. macrocarpon* L.) eggplant complexes. *Front Plant Sci* 2014;**5**:318.
4. Prohens J, Plazas M, Raigón MD, et al. Characterization of interspecific hybrids and first backcross generations from crosses between two cultivated eggplants (*Solanum melongena* and *S. aethiopicum* Kumba group) and implications for eggplant breeding. *Euphytica* 2012;**186**(2):517–38.
5. Toppino L, Valè G, Rotino GL. Inheritance of *Fusarium* wilt resistance introgressed from *Solanum aethiopicum* Gilo and *Aculeatum* groups into cultivated eggplant (*S. melongena*) and development of associated PCR-based markers. *Mol Breed* 2008;**22**(2):237–50.
6. African garden eggplant|FAO|Food and Agriculture Organization of the United Nations. African garden eggplant. <http://www.fao.org/traditional-crops/africangardenegg/en/>. Accessed on 19 September 2018.
7. Schippers RR. African indigenous vegetables: an overview of the cultivated species, Chatham, UK:NRI/CTA. 2000, pp. 213.
8. Maundu P, Achigan-Dako E, Morimoto Y. Biodiversity of African vegetables. In: Shackleton CM, Pasquini MW, Drescher AW, eds. *African Indigenous Vegetables in Urban Agriculture*. Routledge; 2009:65–104.

9. Gramazio P, Blanca J, Ziarsolo P, et al. Transcriptome analysis and molecular marker discovery in *Solanum incanum* and *S. aethiopicum*, two close relatives of the common eggplant (*Solanum melongena*) with interest for breeding. *BMC Genomics* 2016;**17**(1):300.
10. Mennella G, Rotino GL, Fibiani M, et al. Characterization of health-related compounds in eggplant (*Solanum melongena* L.) lines derived from introgression of allied species. *J Agric Food Chem* 2010;**58**(13):7597–603.
11. Cappelli C, Stravato V, Rotino G, et al. Sources of resistance among *Solanum* spp. to an Italian isolate of *Fusarium oxysporum* f. sp. *melongenae*. In: Andrasfalvi A, Moor A, Zatyko L. Proceeding of the 9th Meeting of EUCARPIA, Genetics and Breeding of Capsicum and Eggplant, EUCARPIA. 1995.
12. Fock I. Source of resistance against *Ralstonia solanacearum* in fertile somatic hybrids of eggplant (*Solanum melongena* L.) with *Solanum aethiopicum* L. *Plant Sci* 2001;**160**(2):301–13.
13. Gisbert C, Prohens J, Raigón MD, et al. Eggplant relatives as sources of variation for developing new rootstocks: effects of grafting on eggplant yield and fruit apparent quality and composition. *Sci Hortic* 2011;**128**(1):14–22.
14. Rizza F, Mennella G, Collonnier C, et al. Androgenic dihaploids from somatic hybrids between *Solanum melongena* and *S. aethiopicum* group gilo as a source of resistance to *Fusarium oxysporum* f. sp. *melongenae*. *Plant Cell Rep* 2002;**20**(11):1022–32.
15. Kim S, Park J, Yeom SI, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol* 2017;**18**(1):210.
16. Sierró N, Battey JN, Ouadi S, et al. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol* 2013;**14**(6):R60.
17. Qin C, Yu C, Shen Y, et al. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc Natl Acad Sci U S A* 2014;**111**(14):5135–40.
18. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* 2011;**475**(7355):189–95.
19. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 2012;**485**(7400):635.
20. Bombarely A, Moser M, Amrad A, et al. Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nat Plants* 2016;**2**(6):16074.
21. Moscone EA, Baranyi M, Ebert I, et al. Analysis of nuclear DNA content in capsicum (Solanaceae) by flow cytometry and Feulgen densitometry. *Ann Bot* 2003;**92**(1):21.
22. Kajitani R, Toshimoto K, Noguchi H, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* 2014;**24**(8):1384–95.
23. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;**408**(6814):796–815.
24. Kim S, Park M, Yeom SI, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat Genet* 2014;**46**(3):270–8.
25. Waterhouse RM, Seppely M, Simão FA, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 2017;**35**:3.
26. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**(9):2178–89.
27. Wang Y, Tang H, DeBarry JD, et al. MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**(7):e49.
28. Jaillon O, Aury J-M, Noel B, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007;**449**(7161):463.
29. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 2008;**9**(1):18.
30. Steinbiss S, Willhoeft U, Gremme G, et al. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res* 2009;**37**(21):7002–13.
31. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* 2004;**101**(34):12404–10.
32. McHale L, Tan X, Koehl P, et al. Plant NBS-LRR proteins: adaptable guards. *Genome Biol* 2006;**7**(4):212.
33. Yue JX, Meyers BC, Chen JQ, et al. Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. *New Phytol* 2012;**193**(4):1049–63.
34. Xia R, Xu J, Arikiti S, et al. Extensive families of miRNAs and PHAS loci in Norway spruce demonstrate the origins of complex phasiRNA networks in seed plants. *Mol Biol Evol* 2015;**32**(11):2905–18.
35. Ratnaparkhe MB, Wang X, Li J, et al. Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny. *New Phytol* 2011;**192**(1):164–78.
36. Lester R, Niakan L. Origin and domestication of the scarlet eggplant, *Solanum aethiopicum*, from *S. anguivi* in Africa. In: D'Arcy WG, ed. *Solanaceae: Biology and Systematics*. New York, NY: Columbia University Press; 1986:433–56.
37. Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2004;**21**(2):263–5.
38. Pascual L, Desplat N, Huang BE, et al. Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol J* 2015;**13**(4):565–77.
39. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009;**19**:1655–64.
40. Li H, Richard D. Inference of human population history from whole genome sequence of a single individual. *Nature* 2012;**475**(7357):493–6.
41. Manning K, Timpson A. The demographic response to Holocene climate change in the Sahara. *Quat Sci Rev* 2014;**101**:28–35.
42. Wang Y, Wu Y, Tang D. The autophagy gene, ATG18a, plays a negative role in powdery mildew resistance and mildew-induced cell death in *Arabidopsis*. *Plant Signal Behav* 2011;**6**(9):1408–10.
43. Suttangkakul A, Li F, Chung T, et al. The ATG1/ATG13 protein kinase complex is both a regulator and a target of autophagic recycling in *Arabidopsis*. *Plant Cell* 2011;**23**(10):3761–79.
44. Larsen PB, Cancel J, Rounds M, et al. Arabidopsis ALS1 encodes a root tip and stele localized half type ABC transporter required for root growth in an aluminum toxic environment. *Planta* 2007;**225**(6):1447.



45. Kim D-Y, Bovet L, Kushnir S, et al. AtATM3 is involved in heavy metal resistance in *Arabidopsis*. *Plant Physiol* 2006;**140**(3):922–32.
46. Giritch A, Herbig A, Balzer HJ, et al. A root-specific iron-regulated gene of tomato encodes a lysyl-tRNA-synthetase-like protein. *Eur J Biochem* 1997;**244**(2):310–7.
47. Perera IY, Hung C-Y, Moore CD, et al. Transgenic *Arabidopsis* plants expressing the type 1 inositol 5-phosphatase exhibit increased drought tolerance and altered abscisic acid signaling. *Plant Cell* 2008;**20**(10):2876–93.
48. Cheng S, Gutmann B, Zhong X, et al. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *Plant J* 2016;**85**(4):532–47.
49. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012;**1**(1):18.
50. Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23):3150–2.
51. Galindo-González L, Mhiri C, Deyholos MK, et al. LTR-retrotransposons in plants: engines of evolution. *Gene* 2017;**626**:14–25.
52. Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* 2003;**33**(1):102.
53. Kashkush K, Khasdan V. Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. *Genetics* 2007;**177**(4):1975–85.
54. Hollister JD, Smith LM, Guo Y-L, et al. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A* 2011;**108**(6):2322–7.
55. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* 2009;**19**:1419–28.
56. Ma L-J, Van Der Does HC, Borkovich KA, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* 2010;**464**(7287):367.
57. Song B, Morse D, Song Y, et al. Comparative genomics reveals two major bouts of gene retroposition coinciding with crucial periods of *Symbiodinium* evolution. *Genome Biol Evol* 2017;**9**(8):2037–47.
58. Song B, Chen S, Chen W. Dinoflagellates, a unique lineage for retrogene research. *Front Microbiol* 2018;**9**:1556.
59. Stoffel K, van Leeuwen H, Kozik A, et al. Development and application of a 6.5 million feature Affymetrix Genechip® for massively parallel discovery of single position polymorphisms in lettuce (*Lactuca spp.*). *BMC Genomics* 2012;**13**(1):185.
60. Liu B, Shi Y, Yuan J, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* 2013:13082012.
61. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 2013;**30**(1):31–7.
62. Huang J, Liang X, Xuan Y, et al. (2018): BGISEQ-500 WGS library construction. [protocols.io](http://dx.doi.org/10.17504/protocols.io.ps5dng6). <http://dx.doi.org/10.17504/protocols.io.ps5dng6>.
63. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**(2):573–80.
64. RepeatMasker Open-4.0. <http://www.repeatmasker.org>. 2015. Accessed on 5 November 2017.
65. Smit AF, Hubley R. RepeatModeler Open-1.0. <http://www.repeatmasker.org>. 2008. Accessed on 5 November 2017.
66. GtRNAdb. <http://gtrnadb.ucsc.edu/>. Accessed on 19 August 2019.
67. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 2015;**6**(1):11.
68. Stanke M, Keller O, Gunduz I, et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 2006;**34**(suppl.2):W435–W9.
69. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. *J Mol Biol* 1990;**215**(3):403–10.
70. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res* 2004;**14**(5):988–95.
71. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**(4):357.
72. Perteza M, Perteza GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**(3):290.
73. Elsik CG, Mackey AJ, Reese JT, et al. Creating a honey bee consensus gene set. *Genome Biol* 2007;**8**(1):R13.
74. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;**25**(5):955.
75. Kalvari I, Argasinska J, Quinones-Olvera N, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* 2017;**46**(D1):D335–D42.
76. Viales D, D'Ambrosio U, Gálvez F, et al. Third release of the plant rDNA database with updated content and information on telomere composition and sequenced plant genomes. *Plant Syst Evol* 2017;**303**:1115–21.
77. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;**46**(5):2699.
78. Mitchell AL, Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* 2019, **47**, D1, D351–60.
79. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25.
80. Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 2016;**45**(D1):D331–8.
81. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**(1):41.
82. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;**28**(1):45–8.
83. Hirakawa H, Shirasawa K, Miyatake K, et al. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the Old World. *DNA Res* 2014;**21**(6):649–60.
84. Nakamura T, Yamada KD, Tomii K, et al. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 2018;**1**:3.
85. Guindon S, Dufayard J-F, Lefort V, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;**59**(3):307–21.
86. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* 2005;**23**(1):212–26.



87. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* 2000;**16**(6):276–7.
88. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
89. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**(1):139–40.
90. Kolde R, Kolde MR. Package "pheatmap". <http://www.rdocumentation.org>. 2018. Accessed on 20 August 2018.
91. Seo E, Kim S, Yeom S-I, et al. Genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among Solanaceae plants. *Front Plant Sci* 2016;**7**:1205.
92. HMMER. <http://hmmmer.org/>. Accessed 19 August 2019.
93. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**(3):e9490.
94. GATK. <https://software.broadinstitute.org/gatk/>. Accessed on 19 August 2019.
95. Picard. <https://broadinstitute.github.io/picard/>. Accessed on 19 August 2019.
96. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
97. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**(1):7.
98. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 2012;**6**(2):80–92.
99. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007;**81**(5):1084–97.
100. He W, Zhao S, Liu X, et al. ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis. *Geneti Mol Res* 2013;**12**(4):6275–83.
101. Yang J, Lee SH, Goddard ME, et al. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. In: Condro C, Van der Werf J, Hayes BJ. *Genome-Wide Association Studies and Genomic Prediction*. Springer; 2013:215–36.
102. Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;**21**:263–5.
103. Song B, Song Y, Fu Y, et al. Supporting data for “Draft genome sequence of *Solanum aethiopicum* provides insights into disease resistance, drought tolerance, and the evolution of the genome.” *GigaScience Database* 2019. <http://dx.doi.org/10.5524/100642>.