A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons

Susanne S. Renner[a,1,2], Shan Wu[b,2], Oscar A. Pérez-Escobar[c], Martina V. Silber[a], Zhangjun Fei[b,d,3], and Guillaume Chomicki[e,1,3]

[a]Institute of Systematic Botany and Mycology, University of Munich, 80638 Munich, Germany; [b]Boyce Thompson Institute, Ithaca, NY 14853; [c]Royal Botanic Gardens, Kew, Richmond TW9 3AE, United Kingdom; [d]Robert W. Holley Center for Agriculture and Health, Agricultural Research Service, US Department of Agriculture, Ithaca, NY 14853; and [e]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Edited by Peter H. Raven, Missouri Botanical Garden, St. Louis, MO, and approved March 31, 2021 (received for review January 24, 2021)

Wild relatives or progenitors of crops are important resources for breeding and for understanding domestication. Identifying them, however, is difficult because of extinction, hybridization, and the challenge of distinguishing them from feral forms. Here, we use collection-based systematics, iconography, and resequenced accessions of *Citrullus lanatus* and other species of *Citrullus* to search for the potential progenitor of the domesticated watermelon. A Sudanese form with nonbitter whitish pulp, known as the Kordofan melon (*C. lanatus* subsp. *cordophanus*), appears to be the closest relative of domesticated watermelons and a possible progenitor, consistent with newly interpreted Egyptian tomb paintings that suggest that the watermelon may have been consumed in the Nile Valley as a dessert by 4360 BP. To gain insights into the genetic changes that occurred from the progenitor to the domesticated watermelon, we assembled and annotated the genome of a Kordofan melon at the chromosome level, using a combination of Pacific Biosciences and Illumina sequencing as well as Hi-C mapping technologies. The genetic signature of bitterness loss is present in the Kordofan melon genome, but the red fruit flesh color only became fixed in the domesticated watermelon. We detected 15,824 genome structural variants (SVs) between the Kordofan melon and a typical modern cultivar, "97103," and mapping the SVs in over 400 *Citrullus* accessions revealed shifts in allelic frequencies, suggesting that fruit sweetness has gradually increased over the course of watermelon domestication. That a likely progenitor of the watermelon still exists in Sudan has implications for targeted modern breeding efforts.

watermelon | domestication | chromosome-level genome assembly | phylogenetics | iconography of Egyptian tomb paintings

Wild relatives or progenitors of crops are important resources for breeding and for understanding domestication (1–3). Identifying them, however, is difficult because of extinction (4), hybridization (5), and the challenge of distinguishing them from feral forms (6). Domesticated watermelon (*Citrullus lanatus* subsp. *vulgaris*; for all taxonomic authorities, refer to *SI Appendix*, Table S1) is among the 10 most important crops in Central Asia (7), and knowing its geographic origin and potential progenitor would help targeted breeding efforts (8–10). The geographic region of watermelon domestication has long remained unclear with competing hypotheses favoring southern Africa, West Africa, and Northeast Africa, especially the Kordofan region (9, 11–19), a former province of Sudan bordering North and South Darfur, and part of the western Sahel savannas. This uncertainty resulted from unclear species circumscriptions, and hence, unclear geographic ranges of wild species combined with a lack of sampling of the watermelon's closest relatives.

The hypothesis that watermelon might descend from the South African citron melon (*Citrullus amarus*), which predominated between the 1930s and 2013, derived from a taxonomic mistake involving the oldest collection from South Africa, which in the 1930s was wrongly synonymized with the cultivated watermelon (20). Molecular phylogenies, including sequences from the 1773 Cape Town type collection, showed that no South African material is closely related to the domesticated watermelon (21). Genomic data, albeit with limited geographic sampling, also clarified the proximity of a West African watermelon (*Citrullus mucosospermus*) to domesticated watermelons (9, 21, 22). Since 2015, however, attention has turned back to the possibility that the watermelon may have been domesticated in Northeast Africa (16, 18), perhaps in Sudan where sweet, white-fleshed forms have been collected (11–14).

Besides *C. lanatus*, the genus *Citrullus* contains six other species, of which four (*C. amarus*, *Citrullus ecirrhosus*, *Citrullus naudinianus*, and *Citrullus rehmii*) are native in the Namib–Kalahari region (although the range of *C. amarus* extends further east), one (*C. mucosospermus*) in West Africa (Benin, Ghana, and Nigeria), and one (*Citrullus colocynthis*) in northern Africa to West India; *C. amarus* is also naturalized in Australia (21). All wild species have white pulp that cannot be eaten raw due to the presence of bitter terpene compounds called cucurbitacins. Only fruits of *C. mucosospermus* are sometimes not bitter but instead bland tasting (23); the large, soft seeds of this species are used in West African "egusi" stews (23).

Collection-Based Phylogenomics and Large-Scale Resequencing Both Imply That the Kordofan Melon Could Be the Closest Relative to the Domesticated Watermelon

For phylogenetic analyses of *Citrullus*, we assembled entire plastid genomes (121 genes and 33 spacers) and 6,183 single-copy nuclear

Significance

Wild progenitors of crops are important resources for breeding and for understanding domestication, but identifying them is difficult. Using an integrative approach, we discovered that a Sudanese form of melon with nonbitter whitish pulp, known as the Kordofan melon, is the closest relative of domesticated watermelons and a possible progenitor. To gain insights into the genetic changes that occurred from the progenitor to the domesticated watermelon, we assembled and annotated the genome of a Kordofan melon at the chromosome level. Our analyses imply that early farmers brought into cultivation already nonbitter watermelons, different from other domesticated Cucurbitaceae crops such as cucumber. The Kordofan melon genome is a significant new resource for watermelon breeding.

Author contributions: S.S.R. and G.C. designed research; S.S.R., S.W., O.A.P.-E., M.V.S., Z.F., and G.C. performed research; S.S.R., M.V.S., Z.F., and G.C. contributed new reagents/analytic tools; S.W., O.A.P.-E., Z.F., and G.C. analyzed data; and S.S.R., S.W., Z.F., and G.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: renner@lmu.de or g.chomicki@sheffield.ac.uk.

[2]S.S.R. and S.W. contributed equally to this work.

[3]Z.F. and G.C. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2101486118/-/DCSupplemental.

Published May 21, 2021.

genes for nine accessions from a taxonomic collection (Dataset S1 and *Data Availability*). Maximum likelihood (ML) analysis of the plastid alignment as well as coalescence-based analysis of the nuclear data yielded the same topology (with a minor difference in the concatenated ML analysis of the nuclear alignment, consistent with gene flow; Fig. 1*A* and *SI Appendix*, Fig. S1) and a backbone similar to earlier phylogenies that, however, lacked multiple Northeast African samples (18, 21). It now appears that Sudanese watermelons (collected in 1958, 1982, and 2017 in Darfur; *SI Appendix*, Table S1 and Dataset S1) are closest to the domesticated watermelon (Fig. 1*A*), with West African *C. mucosospermus* the next-closest relative. Similar to the West African egusi melon (*C. mucosospermus*) (23), which is cultivated in its local range, the Kordofan melon is grown by local farmers in Darfur where the very short rainy season requires locally adapted plants.

To complement our collection-based approach, we took advantage of publicly available genome resequencing data of various wild and cultivated watermelon accessions (8). We aligned the *cordophanus* genomic reads to the genome of *C. lanatus* cultivar "97103," a typical East Asian cultivar with sweet, red, and crispy flesh, to identify single nucleotide polymorphisms (SNPs), which were combined with SNPs identified in 415 accessions reported in Guo et al. (8) The phylogenetic relationship between *cordophanus* and other *Citrullus* accessions were inferred using 114,193 SNPs at fourfold degenerate sites. Our *C. lanatus* subsp. *cordophanus* accession was located in the deepest branches of the *C. lanatus* clade and most closely related to PI 254622 and PI 481871 from Sudan (Fig. 2 *A* and *B*). Principal component analysis (PCA) showed that *cordophanus* located between *C. mucosospermus* and *C. lanatus* (Fig. 2*C*). Consistently, *cordophanus* was inferred to share ancestry with both *C. mucosospermus* and *C. lanatus* based on STRUCTURE (24) analyses (Fig. 2*A*). To test for possible gene flow between the Kordofan melon and other accessions, we performed ABBA-BABA tests (25). The results suggest significant gene flows between *C. mucosospermus* and *C. lanatus* subsp. *cordophanus*, *C. lanatus* landraces and *C. lanatus* subsp. *cordophanus*, and *C. mucosospermus* and *C. lanatus* landraces (*SI Appendix*, Fig. S2). These results are consistent with Guo et al. (8) and indicate admixture that could have occurred because of larger population sizes during the African Humid period (14,800 to 5,500 y ago) (26) or cultivation close to wild populations. Altogether, these results support that the Sudanese Kordofan melon could be a direct progenitor of the domesticated watermelon.

An alternative to the hypothesis that Kordofan melons are the watermelon's progenitor is that they are instead a feral form (implying that the watermelon would have been domesticated in West Africa and subsequently brought to Sudan). If this were the case, the nucleotide diversity of Kordofan melons should be similar to that of watermelon landraces. We therefore calculated the nucleotide diversity (π) using the three *cordophanus* accessions and obtained a value of 0.68×10^{-3}. The nucleotide diversity of our 87 accessions of landraces (Dataset S1) is 0.56×10^{-3}. We then randomly selected three accessions from the 87 landraces, calculated their nucleotide diversity, and repeated this 10 times. This yielded a mean nucleotide diversity of 0.44×10^{-3} and an SE of 0.02×10^{-3}. The nucleotide diversity in *cordophanus* is thus higher than that in the 87 sampled landraces. This finding is consistent with *cordophanus* being a progenitor rather than a feral form.

Iconography of Egyptian Tomb Paintings Supports That Sweet Watermelon Was Consumed in Ancient Egypt by 4300 BP

Prior to this study, two Ancient Egyptian illustrations of plausible watermelons were known. One comes from the tomb of Chnumhotep near Saqqara, dated to 4360 to 4350 BP (27,

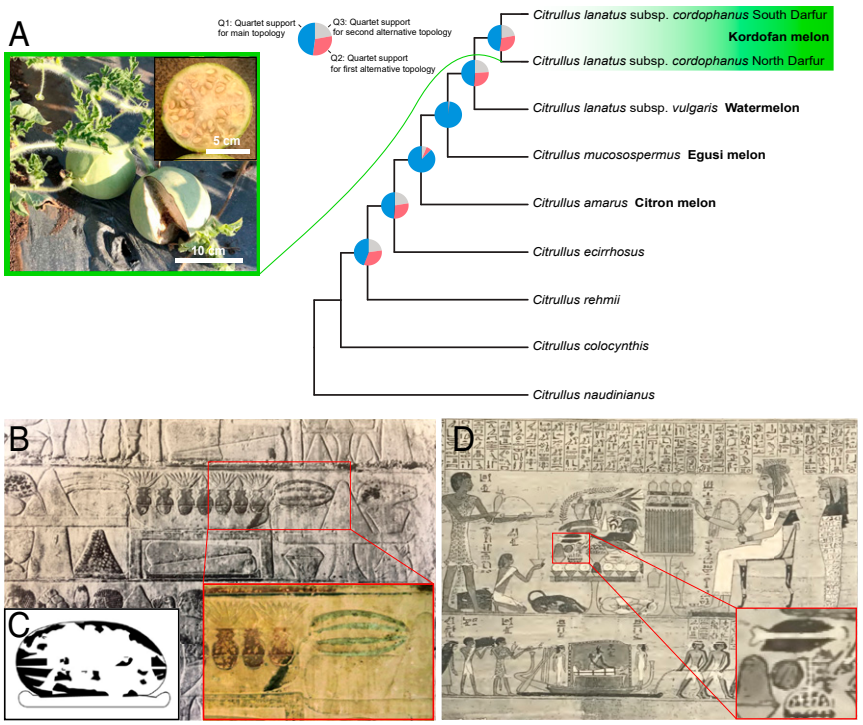

Fig. 1. Collection-based phylogenomics point to the Kordofan melon as the closest relative of the cultivated watermelon, matching Egyption tomb paintings. (*A*) Coalescence-based phylogeny of all *Citrullus* species based on 6,183 nuclear genes rooted on the relevant outgroups (18, 21) (see *SI Appendix*, Fig. S1 for statistical support). (*Left*) A Kordofan melon from North Darfur (*C. lanatus* subsp. *cordophanus*). (*B*) Wall illustrations from the tomb of Chnumhotep, Saqqara, ca. 4450 BP (27). The color photo of Moussa and Altenmüller (27) is a courtesy of L. Manniche, 22 May 2018. (*C*) A drawing from a tomb from Meir (28, 29), Northwest of Asyut, dated to 4350 to 4200 BP. (*D*) Papyrus de Kamara (30), illustrating a *Citrullus* fruit (*inset*), interpreted as a wild watermelon by Keimer (31). The globose striped fruit is reminiscent of the morphology of the Kordofan melon.

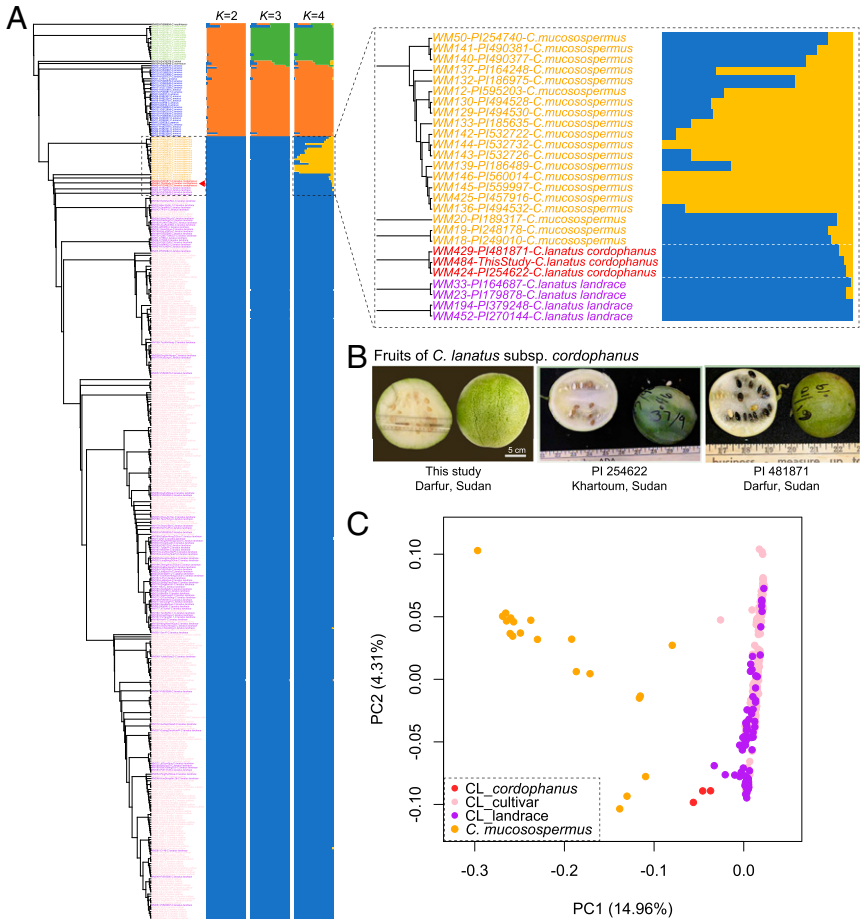

Fig. 2. Phylogeny and population structure of *C. lanatus* subsp. *cordophanus* and other *Citrullus* accessions. (*A*) ML phylogenetic tree and model-based clustering with *K* from 2 to 4. (*B*) Fruits of *cordophanus* at 35 DAP (*Left*) and those from two other Sudanese accessions (*Middle* and *Right*). Fruit pictures of PI 254622 and PI 481871 were obtained from the US National Plant Germplasm System. (*C*) PCA of *C. mucosospermus* and *C. lanatus* (CL) accessions.

Fig. 1*B*). It shows an oblong fruit with dark green stripes on a flat surface. To the left of this fruit in this illustration are seven lotus flowers, and to the left of these are two snake melons (*Cucumis melo* var. *flexuosus*) on a tray. One can also see grapes, suggesting that this was a table laden with sweet foods. The other illustration comes from a tomb at Meir, northwest of Asyut, and shows a large oblong watermelon with dark longitudinal stripes served on a tray (28, 29) (Fig. 1*C*). Its precise age has not been resolved, but it probably dates to 4350 to 4200 BP (R. Schiestl, Alte Geschichte und Altertumskunde, University of Munich, February 2018). A third relevant illustration shows a perfectly globose fruit with longitudinal stripes on a stalk with two leaves slightly longer than the fruit (Fig. 1*D*). It comes from a papyrus of the 21st dynasty, 1069 to 945 BC (30), reproduced by Keimer (31), who thought it showed a wild form of *C. lanatus*. In its small shape (relative to the leaves) and stripes on the fruit, it matches Kordofan melons in photos of Ter-Avanesyan, a Russian breeder who in the early 1960s obtained seeds from the Kordofan region, which were propagated at a research station of the Vavilov Center in Tashkent (22). These watermelons have fruits of 23.5 × 21 cm in size with white, nonbitter pulp that has an aromatic taste. Both Ter-Avanesyan (32) and Fursa (33, 34) considered these watermelons to represent the progenitor of cultivated watermelon, describing them formally as subspecies or varietas *cordophanus* (*C. lanatus* subsp. *cordophanus* Ter-Avan., *C. lanatus* subsp. *vulgaris* var. *cordophanus* [Ter-Avan.] Fursa). Because they imply early consumption of raw (hence nonbitter) watermelon in the Nile Valley and because one depiction morphologically matches the Kordofan melon (Fig. 1*D*), these archaeological records are consistent with the Kordofan melon being a direct progenitor of the cultivated watermelon.

Generating a Chromosome-Level Kordofan Melon Genome

Our identification of Kordofan melons from collection-based phylogenomics and large-scale resequencing as potential progenitors of domesticated watermelon prompted us to generate a high-quality genome assembly to compare its genome structure and content to the domesticated form, assessing structural variants (SVs) in genes linked to the domestication. We generated ∼165 Gb Pacific Biosciences (PacBio) sequences for *C. lanatus* subsp. *cordophanus*, covering ∼388.8× of the genome, which were de novo assembled into contigs, followed by polishing with both PacBio and Illumina reads. The resulting assembly contained 86 contigs with a total size of 367.9 Mb and an N50 length of 9.34 Mb, which had higher contiguity than the watermelon "97103" PacBio genome assembly (8) (*SI Appendix*, Table S2). Combining the Hi-C contact maps and collinearity with the "97103" genome, 98.94% of the *cordophanus* contigs were clustered into 11 pseudomolecules (Fig. 3 and *SI Appendix*, Table S3 and Figs. S3 and S4). The completeness of the *cordophanus* genome assembly was assessed with BUSCO (35). About 97.2% of the core conserved plant genes were found complete in the assembly (*SI Appendix*, Table S4). We then aligned the RNA sequencing (RNA-seq) data generated from various tissues of *cordophanus* to the assembly, which showed mapping rates of up to 97.9%

(*SI Appendix*, Table S5). These results indicate that the *cordophanus* genome assembly is of high quality.

About 57.7% of the *cordophanus* assembly were repetitive (*SI Appendix*, Table S6). A total of 23,043 protein-coding genes were predicted in the *cordophanus* genome, of which 20,554 (89.2%) were assigned with a putative function.

Comparative Genomics of the Bitterness Regulator *ClBt* Across *Citrullus* Suggests That Fruit Bitterness Could Have Been Lost Prior to Watermelon Domestication

Bitterness in cucurbits depends on terpene compounds called cucurbitacins. The *Bi* gene, which encodes an oxidosqualene cyclase (OSC) that catalyzes the first committed step in cucurbitacin C biosynthesis, is critical for determining bitterness (36–38). Two bHLH transcription factors (Bl and Bt) regulate cucurbitacin C biosynthesis by up-regulating *Bi* expression in the leaves (Bl) and fruits (Bt) directly via binding to the E-box elements of the *Bi* promoter (37, 38). In addition to up-regulating *Bi*, the watermelon ClBl and ClBt transcription factors up-regulate most other cucurbitacin metabolism genes, including those encoding the two cytochrome P450 enzymes that convert cucurbitadienol in cucurbitacin precursor (Cl890A and Cl890B) and the acyltransferase that converts the cucurbitacin precursor in cucurbitacin E (38). In cucumber (*Cucumis sativus*), honey melon (*Cucumis melo*), and watermelon,

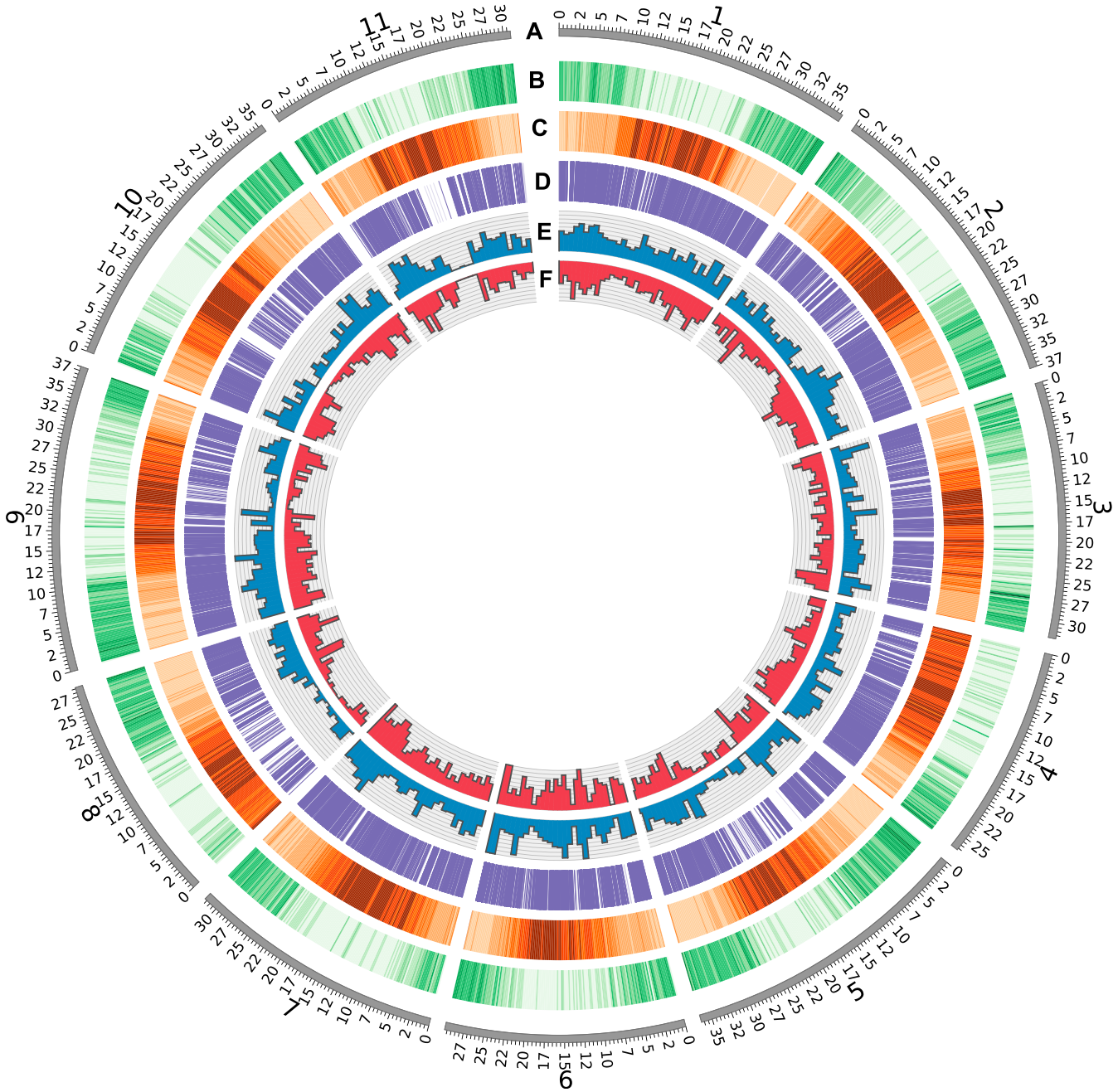

Fig. 3. Genomic landscape of the Kordofan melon (*C. lanatus* subsp. *cordophanus*). (*A*) Ideograms of the 11 chromosomes in mega base pairs. (*B* and *C*) Gene (*B*) and transposable element (TE) (*C*) density represented by percentage of genomic regions covered by genes and TEs, respectively, in 200-kb windows (white to green or orange, low to high). (*D*) Genomic positions of SVs between *cordophanus* and watermelon cultivar "97103." (*E* and *F*) Numbers of insertions (*E*) and deletions (*F*) in 1-Mb windows (maximum = 50) in the *cordophanus* genome compared to the "97103" genome.

the examination of different lines with varying bitterness has revealed that the domestication of nonbitter fruits occurred via a convergent nucleotide substitution, leading to a premature stop codon in the *Bt* gene and resulting in a truncated, nonfunctional protein (36–38). Building on this knowledge, we compared genes in the cucurbitacin biosynthetic pathway and its regulation across *Citrullus* species. All cucurbitacin metabolic genes were conserved across species.

Analysis of the *Bt* gene across all *Citrullus* species revealed that the Kordofan melon *Bt* gene shares the substitution leading to a premature stop codon with the modern watermelon as well as *C. mucosospermus* (*SI Appendix*, Fig. S5*A* and https://figshare.com/s/7dc6e938e304d7854920), implying nonbitter fruits. The Kordofan melon has the homozygous nonbitter genotype at Chr01:3216322 (same as "97103") (*SI Appendix*, Table S7), while *C. mucosospermus* is variable for bitterness, with about 20% of individuals having acid, plain, or sweet pulp, while the rest are bitter (table 5 in ref. 23). These results suggest a scenario wherein loss of pulp bitterness occurred in the most recent common ancestor of *C. lanatus* subsp. *cordophanus* and *C. lanatus* subsp. *vulgaris* following a transition phase in *C. mucosospermus* in which the trait was variable. Loss of pulp bitterness therefore appears to be a preadaptation for the domestication of watermelon, implying that early farmers probably took into cultivation nonbitter plants from the wild. However, our data on Kordofan bitterness come from few accessions, and we therefore cannot ascertain if the absence of fruit bitterness is fixed in the population. Nevertheless, the presence of nonbitter forms suggests that those were the ones brought into cultivation.

Red-Fleshed Watermelons Are Likely the Result of Selection by Early Farmers

The red pulp color in modern watermelon is due to lycopene accumulation, likely by blocking the conversion of lycopene into β-carotene, a step mediated by the enzyme LCYB (lycopene β-cyclase) (39–41). LCYB is encoded by the *lcyb* gene, and watermelon accessions with red flesh are characterized by having a polymorphism in the *lcyb* gene on chromosome 4, resulting in a valine instead of a phenylalanine (V226F) (39, 41). These distinct *LYCB* alleles affect flesh color by modulating the abundance of the LYCB protein (41). Comparison of lycopene metabolic genes across all *Citrullus* species revealed a high conservation of these metabolic genes, with the same copy number for all genes in the lycopene pathway across all species, all of which were functional (lacked a premature stop codon; https://figshare.com/s/7dc6e938e304d7854920). With the exception of the domesticated, red-fleshed sweet watermelon, all *Citrullus* species including the Kordofan melons lack the V226F mutation of LYCB and have white to greenish pulp (*SI Appendix*, Fig. S5*B*). Indeed, the Kordofan melon has a homozygous non-red genotype at Chr04:15442987 (*SI Appendix*, Table S7). These results suggest that the red-fleshed color of watermelon could have arisen during the early stage of the domestication process.

Structural Variants (SVs) between the *cordophanus* and "97103" Genomes and Allele Distribution in Different *Citrullus* Accessions

To characterize genome-wide differences and look for differences in key domestication-related genes between the Kordofan melon and modern domesticated watermelon, we investigated SVs. The *cordophanus* and "97103" genome assemblies were compared with each other to identify SVs, including medium-sized [10 to 49 base pairs (bp)] and large SVs (≥50 bp). In addition, PacBio genomic reads of *cordophanus* and "97103" were mapped to the opposite genomes for SV detection. We found 13,884 medium-sized and 1,940 large SVs, hence a total of 15,824 SVs (Fig. 3 *D*–*F* and *SI Appendix*, Fig. S6). Based on the gene annotations of *cordophanus* and "97103," 242 SVs led to changes in the coding sequences. Another 3,110 SVs in gene bodies were found only in introns or untranslated regions (UTRs), and an additional 3,363 SVs were located within 3 kb upstream of the translation start site (Dataset S2). Approximately 47.6% of the SV sequences were *gypsy*-like retrotransposons (*SI Appendix*, Fig. S7), higher than in the entire genome (37.8%), while the contents of other types of transposable elements were similar between the indel regions and the whole genome (*SI Appendix*, Fig. S7), suggesting that SVs are overrepresented in genome regions occupied by *gypsy*-like retrotransposons, similar to the pattern found in tomato (42).

The 15,824 identified SVs were then genotyped in 408 watermelon accessions using the *Citrullus* resequencing data reported in Guo et al. (8) (Dataset S1) After removing accessions with insufficient genotyping data, 393 accessions were used to investigate the SV dynamics in different populations, including *C. amarus*, *C. colocynthis*, *C. mucosospermus*, *C. lanatus* landraces, and *C. lanatus* cultivars. To evaluate the SV genotyping accuracy using genome resequencing data, we performed SV genotyping in the two reference accessions ("97103" and *cordophanus*) using Illumina short reads, which showed an accuracy rate of 97.8% for "97103" and 90.3% for *cordophanus*. The lower SV genotyping accuracy in *cordophanus* than in "97103" was mainly due to the higher heterozygosity of the *cordophanus* genome, which sometimes caused one allele to be assembled in the genome and another to be captured in the short reads.

The *cordophanus* alleles of a total of 1,414 (8.9%) SVs were not found in cultivated watermelons, of which 1,026 were also absent in watermelon landraces, indicating that the *cordophanus* genome provides breeders with new genetic elements that have been lost during watermelon domestication and improvement. In *C. mucosospermus*, SVs with homozygous *cordophanus* alleles comprised 54% of the SVs in each accession. In *C. lanatus* landraces, the *cordophanus* alleles become significantly (ANOVA; $P < 0.0001$) less prevalent (40%) and in *C. lanatus* cultivars, even less (31%) (Fig. 4*A*). Pairwise comparison of the *cordophanus* allele frequencies with all other species and populations identified 7,302 SVs with significantly changed frequencies ($P < 0.01$) in at least one comparison (Fig. 4*B*, *SI Appendix*, Fig. S8*A*, and Dataset S2). At the loci of these differential SVs, the frequency of *cordophanus* alleles was lower in *C. lanatus* (cultivars and landraces) than in *C. mucosospermus* and also in cultivars compared to landraces (*SI Appendix*, Fig. S8*A*). Out of the 3,075 SVs with significantly changed frequencies during domestication (*C. mucosospermus* to *C. lanatus* landraces), 312 (10.15%; compared to 4.97% of the total SVs) were found to be located in the previously identified domestication sweeps (8), and 319 (7.18%; compared to 3.70%) out of the 4,440 SVs with significantly changed frequencies from *C. lanatus* landrace to *C. lanatus* cultivar were found in the improvement sweeps. Our data showed that SVs with changed frequencies were significantly (Fisher's exact tests; $P < 0.0001$) enriched within the regions under selection during domestication and improvement.

These changes of *cordophanus* allele frequencies are the result of evolution, domestication, and modern breeding. Fruit flesh sweetness was selected during domestication and continues to be an important breeding target. We identified a 12-bp indel (sv20309) in the promoter (~1.7 kb upstream of the translation start site) of a sucrose synthase gene, *Clc10G10370*/*Cla97C10G194010*, located in a previously identified genomic region that is significantly associated with flesh sugar content (8) (Dataset S2). The "97103" allele (12-bp deletion) at sv20309 is the predominant genotype in *C. lanatus* landraces and cultivars ("97103" allele frequency > 0.97 in both), and the *cordophanus* allele is not found in the East Asian cultivars but is present in some American cultivars (Fig. 4*C*). We also identified five SVs whose *cordophanus* allele frequencies were significantly reduced during domestication (*C. mucosospermus* to *C. lanatus* landraces) and/or improvement (*C. lanatus* landraces to *C. lanatus* cultivars) (Fig. 4*C* and Dataset S2) in the promoter or intron regions of sugar metabolism/transport genes, including one sucrose phosphate synthase gene (*Clc06G08580*/*Cla97C06G117750*; sv11422), one arabinose 5-phosphate isomerase

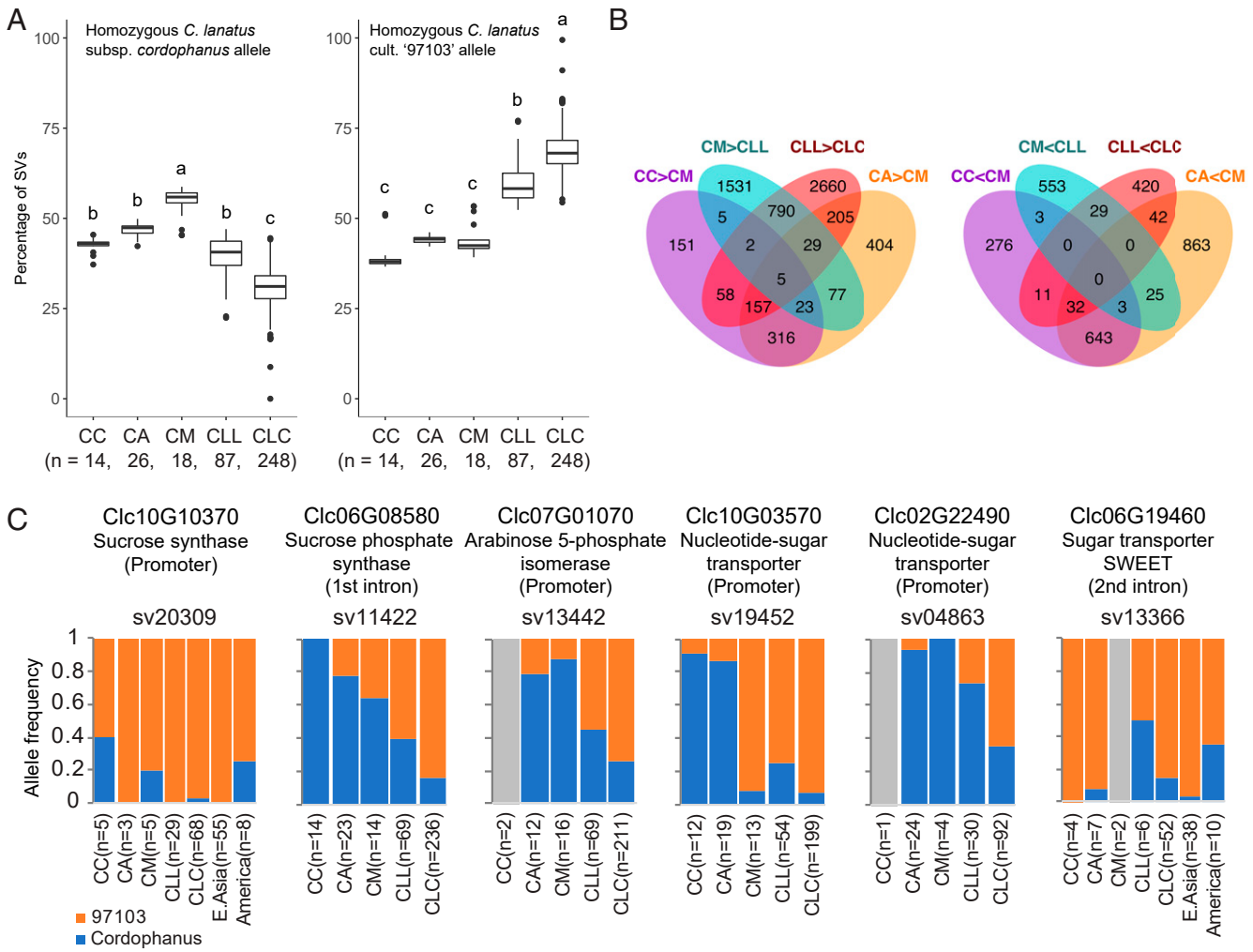

Fig. 4. Population dynamics of *cordophanus* alleles and SVs between *cordophanus*, *lanatus* cultivar "97103," and other species of *Citrullus*. (*A*) Percentages of SVs with homozygous *cordophanus* and cultivar "97103" genotypes in each accession of *C. colocynthis* (CC), *C. amarus* (CA), *C. mucosospermus* (CM), *C. lanatus* landraces (CLL), and *C. lanatus* cultivars (CLC). The lower and upper bounds of each box indicate the first and third quartiles, respectively, and the center line indicates the median. Significant different means are labeled with "a," "b," and "c." (*B*) Numbers of SVs with significantly changed allele frequencies ($P < 0.01$) between populations and species (CC-CM, CA-CM, CM-CLL, or CLL-CLC). (*C*) *Cordophanus* and "97103" allele frequencies of SVs in the promoter, and intron regions of sugar metabolism or transport genes in populations and species. Numbers of accessions with determined genotypes are shown under the allele frequency bar graphs. The allele frequencies were calculated with ≥3 accessions. Missing data are shown in gray.

gene (*Clc07G01070*/*Cla97C07G129400*; sv13442), two nucleotide sugar transporter genes (*Clc10G03570*/*Cla97C10G187870*; sv19452 and *Clc02G22490*/*Cla97C02G047270*; sv04863), and one sugar transporter gene (*Clc06G19460*/*Cla97C06G127910*; sv13366). These results point to a role of these variants in determining pulp sugar content during watermelon domestication and improvement.

In addition to the SVs in genes potentially contributing to pulp sweetness, a 12-bp indel (sv18841) was detected in the 5′ UTR of an expansin gene, *Clc09G19820*/*Cla97C09G179520*, that may play a role in fruit development through modifying cell walls; its allele frequency increased from 0.70 in *C. lanatus* landraces to 0.99 in *C. lanatus* cultivars (*SI Appendix*, Fig. S8*B* and Dataset S2).

We also detected three SVs in the promoters or introns of disease resistance genes (CC-NBS-LRR and TIR-NBS-LRR) that had significantly reduced *cordophanus* allele frequency after domestication and/or improvement (*SI Appendix*, Fig. S8*C*). Functional genomic work, aided by CRISPR-cas9 genome editing—which is now established in watermelon (43)—will enable researchers to test whether the successive changes in the frequency of SVs in disease resistance genes during watermelon domestication mirrors a genetic erosion of the plants' defense genetic toolkit as has been postulated (9). This in turn could prove useful to engineer disease-resistant watermelons.

Conclusion

This study identifies Kordofan melons from Sudan as closest relatives, and perhaps progenitors, of modern watermelons and implies that progenitor populations still exist in the eastern Sahel savannas in the Darfur region, possibly as relicts of a formerly more widespread range into parts of the Sahara during the Holocene African humid period. The sweet, red watermelon may have been domesticated there by Nubian or other Nilo-Saharan ethnicities (44), and the crop could then have spread northward, where it appears to have been consumed as a dessert by 4360 BP (Fig. 1*B*). Alternatively, but less likely, watermelon could have been domesticated in West Africa where the soft-seeded *C. mucosospermus* is endemic (Fig. 2), the natural range of which might have extended north into Libya, where ancient seeds of *Citrullus* have been collected at Uan Muhugiagg (15) close to the Takarkori rock shelter, which covers four millennia of human occupation in the central Sahara (45). Human population levels reached a peak around 7500 BP, with an expansion of populations

into new Saharan territories (46), and early farmers could have brought seeds with them as well as encountering new local variants to select on. While it is still unsure which people domesticated the watermelon, our high-resolution, chromosome-level genome for the Kordofan melon provides a substantial resource for watermelon breeding. For example, Kordofan melons possess significantly different disease resistance alleles compared to domesticated watermelons.

Materials and Methods

Plant Taxon Sampling for the Collection-Based Phylogenomic Approach. The nine accessions from a taxonomic collection sequenced for this study are listed in *SI Appendix*, Table S1, which also reports herbarium vouchers and GenBank accession numbers for the genome resequencing reads. *Citrullus* taxonomy follows Chomicki and Renner (21) and Renner et al. (18) For de novo genome assembly, seeds of *C. lanatus* subsp. *cordophanus* were obtained from two locations in the Darfur region of Sudan (for coordinates, see *SI Appendix*, Table S1) and cultivated in the greenhouses of the Munich Botanical Garden, and two other accessions of *C. lanatus* subsp. *cordophanus* were received by the US Department of Agriculture and collected from Darfur in 1958 (PI 254622) and in 1982 (PI 481871) (see Dataset S1 for links).

Genome resequencing data of the 415 *Citrullus* accessions sampled by Guo et al. (8) were retrieved from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession number SRP188834). All accessions are listed in Dataset S1, which also provides links to photos of the fruits and seeds and information on geographic origins. We checked all taxonomic assignments based on morphology, genetic, geographic, and historical data (i.e., date of seed collection). Modern watermelon breeding in China began in the late 1970s (47), and the Chinese inbred lines in our dataset are mainly from the breeding programs of the Germplasm Bank of National Engineering Research Centre for Vegetables of China and the National Midterm GenBank for Watermelon and Melon of China (for their addresses, see Dataset S1). Our use of the term "landrace" follows these sources.

Genome Resequencing of all *Citrullus* Species. We extracted genomic DNA from the nine *Citrullus* samples using the QIAGEN DNeasy Plant Kit following the manufacturer's protocol. Genome library construction and sequencing were performed by Genewiz. Paired-end DNA libraries were sequenced on an Illumina HiSeq platform, and 350 GB of sequence data were produced. An average of 100 million reads were produced for each sample, corresponding to over 30× coverage of the genome.

Plastid Data Processing, Assembly, and Annotation. The Illumina raw reads were quality filtered using Cutadapt (48), discarding sequences with an averaged phred33 score below 20. Pre- and post-trimming read quality was assessed using FASTQC v.0.1 (49). Whole plastid genomes were assembled by blasting quality-filtered reads against a fully annotated reference plastid genome of elite watermelon line "97103" (GenBank accession NC032008) using the tool *blastn* and a maximum number of target sequences of one. We obtained the best-scoring blast hits from the blastn output by filtering all hits with bitscore and query coverage values less than 100 and 80%, respectively. The best-scoring reads were mapped against the reference plastid genome using the Geneious mapping tool (50) with the following parameters: five iterations, minimum mapping quality of 30, maximum gap size of 100, and maximum mismatches per read of 40%. Consensus plastid genome sequences were assembled following a modified statistical base-calling approach of Li et al. (51), that is, minimum depth coverage of 10 and bases matching at least 50% of the reference sequence. Plastid genomes were fully annotated by transferring intron, exon, and spacer annotations from the reference plastid genome to the de novo assembled plastid genomes.

Plastid and Nuclear Phylogenomics of *Citrullus*. De novo plastid genomes were aligned with Mauve (52) using a progressive algorithm and assuming collinearity. The resulting ∼150,000-bp alignment was subjected to ML tree inference in RAxML v8.0 (53), using the general time reversible substitution model, 25 gamma categories, and 1,000 bootstrap replicates. To generate a nuclear phylogeny from the deep genome resequencing data, we first used the existing annotation of the reference genome to extract sequences of all nuclear coding sequences. These sequences were used as a reference to produce de novo assembled sequences for each gene and sample using the python package HybPiper 1.3.1 (54). Next, we filtered out the genes giving paralog warning based on reciprocal Blast and only retained genes that were present in at least 90% of the samples, with SD in length sequence between samples lower than 10 and sequence length between 1 and 8 kb. This yielded an alignment of 6,183 genes, provided as a dataset at https://figshare.com/s/0534e1ad32732184d803. ML tree inference using the concatenated nuclear matrix relied on the same settings as employed for plastid tree inference.

In addition to ML tree inference, we also performed coalescence-based analyses. We generated ML trees for every nuclear gene alignment as well as the concatenated nuclear matrix using the same settings as employed for plastid tree inference in RAxML. Species tree analyses relied on the coalescence-based method ASTRAL III (55), using as input the annotated consensus gene trees derived from RAxML but collapsing branches with likelihood bootstrap support < 10. Gene tree conflict was visualized by plotting the local posterior probabilities derived from ASTRAL as pie diagrams at nodes (Fig. 1*A*).

PacBio Sequencing of the Kordofan Melon (*C. lanatus* subsp. *cordophanus*). Fresh leaf material from four 3-4-leaf seedlings of *cordophanus* from North Darfur was flash frozen in liquid nitrogen and sent on dry ice to Dovetail Genomics (https://dovetailgenomics.com), which performed high molecular weight (HMW) DNA extraction, PacBio SMRT library construction, and PacBio long read data generation. Specifically, the DNA sample was quantified using Qubit 2.0 Fluorometer (Life Technologies). The PacBio SMRTbell library with an insert size of ∼20 kb was constructed using SMRTbell Template Prep Kit 1.0 (PacBio) following the manufacturer's instructions and sequenced on four PacBio Sequel SMRT cells.

Hi-C and Chicago Library Construction. The Chicago and Dovetail Hi-C libraries were constructed by Dovetail Genomics using the same material as for the PacBio sequencing described above.

A Chicago library was prepared as previously described (56). Briefly, ∼500 ng of HMW DNA (mean fragment length of 60 kb) was reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5′ overhangs filled in with biotinylated nucleotides, and then the free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to fragments with sizes of ∼350 bp, and the sequencing library was generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment. The library was sequenced on an Illumina HiSeqX sequencer.

A Dovetail Hi-C library was prepared in a similar manner as described previously (57). Briefly, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5′ overhangs filled in with biotinylated nucleotides, and then the free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to fragments with sizes of ∼350 bp, and the sequencing library was generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment. The library was sequenced on an Illumina HiSeqX sequencer.

Construction and Sequencing of Illumina DNA and RNA-Seq Libraries. *Citrullus lanatus* subsp. *cordophanus* plants from North Darfur (seeds from Munich) were grown in the greenhouse at Boyce Thompson Institute in Ithaca, New York, under a 16-h light period and day/night temperature of 25/20 °C. For DNA extraction, young leaves pooled from seedlings at 2 wk after germination were collected, flash frozen in liquid nitrogen, ground into powder, and stored at −80 °C. DNA was extracted from the leaf sample and used to construct a DNA library using the Illumina Genomic DNA Sample Preparation Kit following the manufacturer's instructions (Illumina). Total RNA was extracted from root, stem, leaf, male flower at anthesis, hermaphrodite flower at anthesis, tendril, and young fruit at 5 d after pollination (DAP) and fruit rind and fruit flesh at 25 and 35 DAP (two biological replicates) using the QIAGEN RNeasy Plant Mini Kit (QIAGEN). Strand-specific RNA-seq libraries were constructed using the protocol described in Zhong et al. (58) The DNA and RNA-seq libraries were sequenced on an Illumina NextSeq 500 platform with the paired-end mode and the read length of 150 bp.

De Novo Assembly of the C. *lanatus* subsp. *cordophanus* Genome. The PacBio reads of *C. lanatus* subsp. *cordophanus* were assembled into contigs using the FALCON 1.8.8 pipeline (https://github.com/PacificBiosciences/pb-assembly).

The assembled contigs were polished using the raw PacBio reads with the Arrow algorithm implemented in SMRT Link 5.0.1 (PacBio).

A total of ~84.0 Gb (covering ~200× of the *cordophanus* genome) Illumina short read sequences were generated to further correct base errors in the *cordophanus* PacBio assembly using Pilon (v1.22) (59). The cleaned Illumina reads were aligned to the *cordophanus* assembly using bwa-aln (v0.7.13) (60). Alignments with mapping quality >20 were used, and four rounds of Pilon error correction were performed. The corrected contigs were compared to the NCBI nucleotide database, and those with more than 50% of their length aligned to nonplant sequences, including those from microorganisms and insects, were identified as contamination and removed. Furthermore, contigs with >95% identity and >90% of their length covered by other longer contigs were identified as redundant sequences and removed.

The Chicago and Hi-C reads were aligned to the *cordophanus* assembled contigs using a modified SNAP read mapper (http://snap.cs.berkeley.edu). Based on the mapping, the contigs were clustered into scaffolds using the HiRise pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (56).

Repeat Annotation and Gene Prediction. Libraries of miniature inverted-repeat transposable elements (MITEs) and long terminal repeats (LTRs) were constructed by scanning the *cordophanus* genome using MITE-Hunter (61) and LTRharvest (v1.5.9) (62) and were used to mask the *cordophanus* genome with RepeatMasker (v4.0.7; http://www.repeatmasker.org/) followed by de novo repeat library construction using RepeatModeler (v1.0.11; http://www.repeatmasker.org/RepeatModeler/). Repeat sequences in the MITE, LTR, and de novo repeat libraries were combined, classified, and used to mask the *cordophanus* genome.

Gene prediction was performed with MAKER (v2.31.10) (63), which integrated ab initio predictions and evidence from transcript and protein homology. RNA-seq reads were assembled into transcriptomes using both de novo and genome-guided approaches with Trinity (v2.6.6) (64), and the assembled transcriptomes served as transcript evidence. Protein homology evidence was obtained by aligning protein sequences from watermelon "97103," cucumber, melon, *Arabidopsis*, and Swiss-Prot to the *cordophanus* genome with Spaln (v2.1.4) (65). Protein-coding genes were also predicted with BRAKER (v2.1.2) (66), using the repeat-masked *cordophanus* assembly, and refined with PASA (v2.4.1) (67). The BRAKER gene models were used to replace the MAKER ones when the predictions were different. Functional annotation for the predicted genes was performed by comparing their protein sequences against the *Arabidopsis* proteome, Pfam, and Swiss-Prot databases.

Identification of SVs and SNPs between *cordophanus* and "97103." To identify SVs between the genomes of *cordophanus* and "97103" (v2), the two assemblies were aligned using minimap2 (68), and the resulting alignments were analyzed using Assemblytics (69) for SV calling. In addition, PacBio reads were mapped to the opposite genomes, and SVs were detected with pbsv (https://github.com/PacificBiosciences/pbsv). SVs spanning gap regions were removed. SVs with flanking sequences mapped to the correct positions on the opposite genomes were kept. Overlapping SVs identified by different methods were combined, and the integrated set of SVs was annotated based on the gene predictions of the two genomes. SNPs between the two genomes were identified by aligning the *cordophanus* genomic reads to the "97103" v2 reference genome using bwa-mem (v0.7.13) (60), and SNP calling was performed using GATK (v4.0.7.0) (70).

Phylogeny and Population Structure Analyses. The SNPs identified between *cordophanus* and "97103" were combined with those of the accessions reported in Guo et al. (8) We reclassified several of these accessions based on morphology, genetic, geographic, and historical data (Dataset S1) and five accessions (PI 195927, PI 482271, PI 500301, PI 482378, and PI 482255) with unclear taxonomy were excluded from the analysis. Phylogenetic relationships among the 411 *Citrullus* accessions (33 *C. amarus*, 16 *C. colocynthis*, 3 *cordophanus*, 1 *C. ecirrhosus*, 20 *C. mucosospermus*, 87 *C. lanatus* landraces, 249 *C. lanatus* cultivars, 1 *C. naudinianus*, and 1 *C. rehmii*) were inferred using 114,193 SNPs at the fourfold degenerate sites. The ML tree was constructed using iqtree (v1.6.12) (71) with 1,000 bootstrap replicates and a *C. naudinianus* accession, PI 596694, as the outgroup. Population structure was investigated using FastSTRUCTURE (72) with the same set of SNPs, and PCA was then performed using PLINK (v1.9) (73). Only *C. mucosospermus* and *C. lanatus* (a total of 359 accessions: 20 *C. mucosospermus*, 3 *cordophanus*, 87 landraces, and 249 cultivars) were used in the PCA analysis. The nucleotide diversity (π) was calculated using VCFtools (74) with all identified SNPs.

Testing for Gene Flow between the Kordofan Melons and Close Relatives. Potential gene flows between different groups were identified using the ABBA-BABA test (25) (also called the D test) with biallelic SNPs. For each group, the D value (sum(ABBA) − sum(BABA))/(sum(ABBA) + sum(BABA)) was calculated, where the ancestral ("A") and derived ("B") states of the sites were determined based on the allele frequency in the outgroup. SEs and the significance of the weighted D values were calculated based on *Z*-scores obtained using the jackknife method (75).

SV Genotyping in Watermelon. Illumina genome resequencing data from Guo et al. (8) with the same sampling as above except for one fewer *cordophanus* accession were used to genotype the identified SVs between subsp. *cordophanus* and cultivated "97103." Raw reads were processed to remove duplicated read pairs defined by identical bases in the first 90 bp of both left and right reads. Adaptor and low-quality sequences were trimmed with Trimmomatic (v0.36) (76). The cleaned reads from each accession were aligned to both *cordophanus* and "97103" genomes using bwa-mem (v0.7.13) (60). For the detection of a deletion in an accession, at least three split reads at the identified breakpoint or <50% coverage in the deleted regions were required. In each accession, SVs were genotyped as *cordophanus* genotype, "97103" genotype, heterozygous, or undetermined (insufficient split read and read coverage evidence). A total of 12 accessions with >80% of SVs having undetermined genotypes as well as the *cordophanus* accessions (population size not large enough) were excluded, and 393 accessions (26 *C. amarus*, 14 *C. colocynthis*, 87 *C. lanatus* landrace, 248 *C. lanatus* cultivar accessions, and 18 *C. mucosospermus*) were used in the downstream population analysis.

Data Availability. The genome sequence of *Citrullus lanatus* subsp. *cordophanus* has been deposited at DDBJ/ENA/GenBank under the accession JADPLL000000000. The version described in this paper is JADPLL010000000. Raw genome and transcriptome sequencing reads of *cordophanus* and the nine genome resequencing reads have been deposited into the NCBI BioProject database under accessions PRJNA676157, PRJNA676179, and PRJNA708291, respectively. The concatenated alignments and the individual gene alignments of the phylogenomic dataset shown in Fig. 1 are in the dataset available at https://figshare.com/s/0534e1ad32732184d803. The 73 alignments of the genes related to cucurbitacin and lycopene biosynthesis pathways and their regulations are in the dataset available at https://figshare.com/s/7dc6e938e304d7854920.

ACKNOWLEDGMENTS. This work was funded by DFG 603/27-1, the US Department of Agriculture National Institute of Food and Agriculture Specialty Crop Research Initiative (Grant 2020-51181-32139), and generous support from the Elfriede and Franz Jakob Foundation. G.C. is funded by a Natural Environment Research Council Independent Research Fellowship (Grant NE/S014470/1). We thank two anonymous reviewers for their excellent suggestions and Richard Parkinson, Egyptologist at the University of Oxford, for discussion.

1. J. R. Harlan, Genetic resources in wild relatives of crops. *Crop Sci.* **16**, 329–333 (1976).
2. R. Hajjar, T. Hodgkin, The use of wild relatives in crop improvement: A survey of developments over the last 20 years. *Euphytica* **156**, 1–13 (2007).
3. M. A. Steinwand, P. C. Ronald, Crop biotechnology and the future of food. *Nat. Food* **1**, 273–283 (2020).
4. V. Caracuta *et al*., 14,000-year-old seeds indicate the Levantine origin of the lost progenitor of faba bean. *Sci. Rep.* **6**, 37399 (2016).
5. K. A. Hodgins *et al*., Genomics of Compositae crops: Reference transcriptome assemblies and evidence of hybridization with wild relatives. *Mol. Ecol. Resour.* **14**, 166–177 (2014).
6. L. E. Newstrom, Evidence for the origin of chayote, *Sechium edule* (Cucurbitaceae). *Econ. Bot.* **45**, 410–428 (1991).
7. FAO, Statistical Year Book 2014. www.fao.org/3/a-i3621e.pdf. Accessed 11 November 2020.
8. S. Guo *et al*., Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.* **51**, 1616–1623 (2019).
9. S. Guo *et al*., The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51–58 (2013).
10. S. Wu *et al*., Genome of 'Charleston Gray', the principal American watermelon cultivar, and genetic characterization of 1,365 accessions in the U.S. National Plant Germplasm System watermelon collection. *Plant Biotechnol. J.* **17**, 2246–2258 (2019).
11. G. Schweinfurth, Sur l'origine de quelques plantes cultivées en l'Égypte. *Bull. Inst. Egypte* **12**, 200–206 (1873).
12. G. Schweinfurth, La flore de l'ancienne Egypte. *Rev. Sci.* **32**, 72–77 (1883a).

13. G. Schweinfurth, The flora of ancient Egypt. *Nature* **28**, 109–114 (1883b).
14. G. Schweinfurth, Ueber Pflanzenreste aus altaegyptischen Gräbern. *Ber. Dtsch. Bot. Ges.* **2**, 351–371 (1884).
15. K. Wasylikowa, M. van der Veen, An archaeobotanical contribution to the history of watermelon, *Citrullus lanatus* (Thunb.) Mats. & Nakai (syn. *C. vulgaris* Schrad.). *Veg. Hist. Archaeobot.* **13**, 213–217 (2004).
16. H. S. Paris, Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus*. *Ann. Bot.* **116**, 133–148 (2015).
17. H. S. Paris, Overview of the origins and history of the five major cucurbit crops: Issues for ancient DNA analysis or archaeological specimens. *Veg. Hist. Archaeobot.* **25**, 405–414 (2016).
18. S. S. Renner, A. Sousa, G. Chomicki, Chromosome numbers, Sudanese wild forms, and classification of the watermelon genus *Citrullus*, with 50 names allocated to seven biological species. *Taxon* **66**, 1393–1405 (2017).
19. G. Chomicki, H. Schaefer, S. S. Renner, Origin and domestication of Cucurbitaceae crops: Insights from phylogenies, genomics and archaeology. *New Phytol.* **226**, 1240–1255 (2020).
20. L. H. Bailey, Three discussions in Cucurbitaceae. *Gentes Herb.* **2**, 175–186 (1930).
21. G. Chomicki, S. S. Renner, Watermelon origin solved with molecular phylogenetics including Linnaean material: Another example of museomics. *New Phytol.* **205**, 526–532 (2015).
22. S. S. Renner, How Russian breeders discovered *Citrullus mucosospermus* and *Citrullus lanatus* var. *cordophanus*, the likely closest relatives of domesticated watermelon. *Rep. Cucurbit Genet. Coop.* **42**, 10–12 (2020).
23. E. G. Achigan-Dako *et al*., Phenetic characterization of *Citrullus* spp. (Cucurbitaceae) and differentiation of egusi-type (*C. mucosospermus*). *Genet. Resour. Crop Evol.* **62**, 1159–1179 (2015).
24. D. Falush, M. Stephens, J. K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
25. S. H. Martin, J. W. Davey, C. D. Jiggins, Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
26. T. M. Shanahan *et al*., The time-transgressive termination of the African Humid period. *Nat. Geosci.* **8**, 140–144 (2015).
27. A. M. Moussa, H. Altenmüller, *Das Grab des Nianchchnum und Chnumhotep* (Archäologische Veröffentlichungen 21, Mainz am Rhein, Germany, 1977), p. 89.
28. L. Manniche, *An Ancient Egyptian Herbal* (British Museum, London, 1989).
29. J. Janick, H. S. Paris, D. C. Parrish, The cucurbits of mediterranean antiquity: Identification of taxa from ancient images and descriptions. *Ann. Bot.* **100**, 1441–1457 (2007).
30. E. Naville, *Le papyrus hiéroglyphique de Kamara, le papyrus hiératique de Nesikhonsou au Musée de Caire*, (Papyrus funéraires de la XXI dynastie., Ernest Leroux, Paris, 1912), vol. 1.
31. L. Keimer, *Die Gartenpflanzen im alten Ägypten* (Hoffmann und Campe Verlag, Hamburg, Berlin, 1924), vol. I.
32. D. V. Ter-Avanesyan, Arbuz Kordofanskyi *Citrullus lanatus* Mansf. ssp. *cordophanus* Ter-Avan. *Bot. Z.* **51**, 423–426 (1966).
33. T. B. Fursa, K sistematike roda *Citrullus* Schrad. *Bot. Zhurn.* **57**, 31–41 (1972).
34. T. B. Fursa, I. P. Gavrilyuk, Phylogenetic relations of the genus *Citrullus* Schrad. based on the immunochemical analysis of seed proteins. *Sborn. Nauchn. Trudov Prikl. Bot. Genet. Selekts.* **133**, 19–26 (1990).
35. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
36. J. Qi *et al*., A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**, 1510–1515 (2013).
37. Y. Shang *et al*., Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084–1088 (2014).
38. Y. Zhou *et al*., Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat. Plants* **2**, 16183 (2016).
39. H. Bang, S. Kim, D. Leskovar, S. King, Development of a codominant CAPS marker for allelic selection between canary yellow and red watermelon based on SNP in lycopene β-cyclase (LCYB) gene. *Mol. Breed.* **20**, 63–72 (2007).
40. S. Grassi *et al*., Comparative genomics reveals candidate carotenoid pathway regulators of ripening watermelon fruit. *BMC Genomics* **14**, 781 (2013).
41. J. Zhang *et al*., Decreased protein abundance of Lycopene β-Cyclase contributes to red flesh in domesticated watermelon. *Plant Physiol.* **183**, 1171–1183 (2020).
42. X. Wang *et al*., Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat. Commun.* **11**, 5817 (2020).
43. S. Tian *et al*., Efficient CRISPR/Cas9-based gene knockout in watermelon. *Plant Cell Rep.* **36**, 399–406 (2017).
44. R. Blench, *Archaeology, Language, and the African Past* (Altamira Press, Lanham, Maryland, 2006).
45. A. M. Mercuri, R. Fornaciari, M. Gallinaro, S. Vanin, S. di Lernia, Plant behaviour from human imprints and the cultivation of wild cereals in Holocene Sahara. *Nat. Plants* **4**, 71–81 (2018).
46. K. Manning, A. Timpson, The demographic response to Holocene climate change in the Sahara. *Quat. Sci. Rev.* **101**, 28–35 (2014).
47. M. Wang, P. Hou, Origin, history, taxonomy and breeding achievement of watermelon [in Chinese]. *Modern Vegetable* **3**, 18–19 (2006).
48. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
49. S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data (2010), https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
50. M. Kearse *et al*., Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
51. H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
52. A. C. Darling, B. Mau, F. R. Blattner, N. T. Perna, Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
53. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
54. M. G. Johnson *et al*., HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**, 1600016 (2016).
55. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
56. N. H. Putnam *et al*., Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
57. E. Lieberman-Aiden *et al*., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
58. S. Zhong *et al*., High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* **2011**, 940–949 (2011).
59. B. J. Walker *et al*., Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
60. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
61. Y. Han, S. R. Wessler, MITE-hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
62. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
63. B. L. Cantarel *et al*., MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
64. M. G. Grabherr *et al*., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
65. O. Gotoh, A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.* **36**, 2630–2638 (2008).
66. K. J. Hoff, A. Lomsadze, M. Borodovsky, M. Stanke, Whole-genome annotation with BRAKER. *Methods Mol. Biol.* **1962**, 65–95 (2019).
67. B. J. Haas *et al*., Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
68. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
69. M. Nattestad, M. C. Schatz, Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
70. GATK pipeline. https://github.com/broadinstitute/gatk/issues/6254.
71. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
72. A. Raj, M. Stephens, J. K. Pritchard, fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
73. S. Purcell *et al*., PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
74. P. Danecek *et al*.; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
75. B. Efron, C. Stein, The jackknife estimate of variance. *Ann. Stat.* **9**, 586–596 (1981).
76. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).