


Genome architecture and genetic diversity of allopolyploid okra (*Abelmoschus esculentus*)

Ronald Nieuwenhuis¹, Tamara Hesselink¹, Hetty C. van den Broeck¹, Jan Cordewener¹, Elio Schijlen¹, Linda Bakker¹, Sara Diaz Trivino¹, Darush Struss², Simon-Jan de Hoop², Hans de Jong³ and Sander A. Peters^{1,*} 

¹Business Unit of Bioscience, Cluster Applied Bioinformatics, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands,

²East-West International B.V., Heiligeweg 18, 1601 PN Enkhuizen, The Netherlands, and

³Laboratory of Genetics, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

Received 21 June 2023; revised 17 October 2023; accepted 6 December 2023.

*For correspondence (e-mail sander.peters@wur.nl).

SUMMARY

The allopolyploid okra (*Abelmoschus esculentus*) unveiled telomeric repeats flanking distal gene-rich regions and short interstitial TTTAGGG telomeric repeats, possibly representing hallmarks of chromosomal speciation. Ribosomal RNA (rRNA) genes organize into 5S clusters, distinct from the 18S–5.8S–28S units, indicating an S-type rRNA gene arrangement. The assembly, in line with cytogenetic and cytometry observations, identifies 65 chromosomes and a 1.45 Gb genome size estimate in a haploid sibling. The lack of aberrant meiotic configurations implies limited to no recombination among sub-genomes. k-mer distribution analysis reveals 75% has a diploid nature and 15% heterozygosity. The configurations of Benchmarking Universal Single-Copy Ortholog (BUSCO), k-mer, and repeat clustering point to the presence of at least two sub-genomes one with 30 and the other with 35 chromosomes, indicating the allopolyploid nature of the okra genome. Over 130 000 putative genes, derived from mapped IsoSeq data and transcriptome data from public okra accessions, exhibit a low genetic diversity of one single nucleotide polymorphisms per 2.1 kbp. The genes are predominantly located at the distal chromosome ends, declining toward central scaffold domains. Long terminal repeat retrotransposons prevail in central domains, consistent with the observed pericentromeric heterochromatin and distal euchromatin. Disparities in paralogous gene counts suggest potential sub-genome differentiation implying possible sub-genome dominance. Amino acid query sequences of putative genes facilitated phenol biosynthesis pathway annotation. Comparison with manually curated reference KEGG pathways from related Malvaceae species reveals the genetic basis for putative enzyme coding genes that likely enable metabolic reactions involved in the biosynthesis of dietary and therapeutic compounds in okra.

Keywords: okra, allopolyploid, genome, telomere, flavonoid biosynthesis pathway, sub-genome fractionation.

INTRODUCTION

The well-known vegetable, okra (*Abelmoschus esculentus*), belongs to the family Malvaceae, comprising more than 244 genera and over 4200 species. The Malvaceae are divided into nine subfamilies of which okra belongs to the subfamily Malvoideae. *Abelmoschus* is closely related to *Hibiscus* species like *Hibiscus rosa-sinensis* or Chinese rose and *Hibiscus cannabinus* or Kenaf, which is exemplified by the beautiful characteristic Hibiscus-like flowers that both genera display. Based on genetic differences, *Abelmoschus* has been placed in a separate genus from *Hibiscus* though Okra demonstrates continuous flowering

and is self-compatible; however, research indicates cross-pollination ranging from 4 to 19% (Choudhury & Choomsai, 1970; Purewal & Randhawa, 1947), with instances of it reaching as high as 42% (Mitidieri & Vencovsky, 1974). Its characteristic hermaphroditic flowers usually have white or yellow perianths, consisting of five petals and five sepals, whereas calyx, corolla, and stamens are fused at the base. Other well-known species in the Malvaceae are cocoa (*Theobroma cacao*), cotton (*Gossypium hirsutum*), and *Tilia* species like lime tree, the mangrove *Heritiera* species, and durian (*Durio zibethinus*). Currently, according to Plants of the World online database (<https://powo.science.kew.org>), the genus *Abelmoschus* currently contains 12

accepted species, four of which have economic value (Li et al., 2020). Okra or 'lady's finger' is a low-calorie vegetable, mainly cultivated for its fruits that are harvested while still unripe, containing a large variety of nutrients and elements essential for daily human consumption, such as vitamins, flavonoids, minerals, and other health components such as folate and fibers (Muimba-Kankolonga, 2018; Wu et al., 2020). For example, total polyphenol extracts from okra fruits, containing flavonoids such as myricetin and quercetin, have been demonstrated for their antidiabetic activity in obese rats suffering from type 2 diabetes mellitus (Peter et al., 2021). These health compounds and additional nutritional qualities make okra an appreciated vegetable in many parts of the tropics and subtropics of Asia, Africa, and America, gaining rapidly in popularity. Global production has increased yearly since 1994, reaching 10M tonnes in 2019 and covering some 2.5M ha (<https://www.fao.org/faostat/en/#data/QCL>), with Asia having the largest production share of almost 70%, of which India alone is currently annually producing more than 4M tonnes. However, its production is challenged by a range of pathogens and insect pests, such as powdery mildew and blackmold (*Cerospora abelmoschii*), bacterial blight disease (*Xanthomonas campestris* p.v. *malvacearum*), mycoplasmas, nematodes, worms and insects such as whitefly (*Bemisia tabaci*), thrips (*Thrips palmi*), cotton leafhopper (*Amarasca biguttula*), and aphids (*Aphis gossypii*). Besides feeding damage, these vectors can transmit viruses such as Yellow Vein Mosaic Virus (YVMV), a geminivirus, causing crop losses of up to 80–90% without pest control (Benchari, 2012; Dankhar & Koundinya, 2020; Lata et al., 2021; Muimba-Kankolonga, 2018). Typical symptoms of YVMV-infected okra plants are stunted growth, with veins and veinlets turning yellow in color, producing seed pods that are small, distorted, and chlorotic. Crop loss may be reduced to 20–30%, by controlling insect pests with rather harmful and toxic chemicals and insecticides (Ali et al., 2005), causing considerable collateral damage to the ecosystem. Moreover, increased insect tolerance to pesticides has led to over-use and mis-use of chemicals, leaving unhealthy high levels of pesticide residues (Benchari, 2012). Although there are some YVMV-tolerant okra genotypes, such as Nun1144 and Nun1145 (Venkataravanna et al., 2013), the genetic basis for this tolerance has not been identified. Besides a need for disease resistance, other breeding challenges and demands include maximizing production, unraveling the genetic basis for abiotic stress tolerance, and the need to develop double haploid lines enabling the study of recessive gene traits (Dankhar & Koundinya, 2020).

To meet current demands and challenges, accelerated breeding is urgently needed. Presently, several methods of breeding for improvement in okra are being used, such as pure line selection, pedigree breeding, as well as mutation

and heterosis breeding (Dankhar & Koundinya, 2020). These methods are very time-consuming though, and often involve laborious analyses over multiple generations. Despite wide genetic variation available among wild relatives of okra, significant crop improvement by introgression breeding, has not been achieved due to hybridization barriers. Advanced breeding is further hampered due to the lack of sufficient molecular markers, although several studies have reported on the use of RAPD (Random Amplified Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), and SSR (Simple Sequence Repeat) markers to study genetic diversity in okra (Lata et al., 2021), linkage maps and reference genome, and this in turn has impeded genome and transcriptome studies. Molecular studies have further been complicated due to the presence of large amounts of mucilaginous and polyphenolic compounds in different tissues, interfering with the preparation of genetic materials (Lata et al., 2021; Takakura & Nishio, 2012). Furthermore, correct *de novo* assembly is presumed to be complex because of the expected large genome and transcriptome size and the highly polyploid nature of the genome. Salameh (2014) reported flow cytometric estimates of nuclear DNA size estimations with 2C values ranging from 3.98 to 17.67 pg, equaling to genome sizes between 3.8 and 17.3 Gbp. In addition, chromosome counts demonstrated a huge variation, ranging from $2n=62$ to $2n=144$, with $2n=130$ as the most frequently observed chromosome number (Benchari, 2012; Merita et al., 2012). These findings have led to further assumptions on the geographical origin of cultivated *A. esculentus*, speculating that a $2n=58$ species *A. tuberculatus* native from Northern India and a $2n=72$ species *A. ficulneus* from East Africa might have hybridized followed by a chromosome doubling, giving rise to an allopolyploid *Abelmoschus* hybrid with $2n=130$ (Benchari, 2012; Joshi & Hardas, 1956; Merita et al., 2012; Siemonsma, 1982). However, genomic, genetic, and cytological information is scanty, limiting the possibility to further understand the hereditary constituent of the crop. In this study, we benefited from naturally occurring okra haploids, circumventing heterozygosity in the reconstruction of composite genome sequences, and supporting faithful genome reconstruction (Langley et al., 2011). In this study, we provide a detailed insight into the complex genome and transcriptome architecture of an okra cultivar and its haploid descendant, using cytogenetic characterization of its mitotic cell complements and meiosis, and advanced sequencing and assembly technologies of the haploid genome, providing basic scientific knowledge for further evolutionary studies and representing a necessary resource for future molecular-based okra breeding. Furthermore, we provide a structural and functional genome annotation that is of paramount importance to understanding plant metabolism (Weißborn & Walther, 2017) and the genetic basis for the

enzyme coding genes, enabling metabolic reactions involved in the biosynthesis of dietary and therapeutic compounds in okra.

RESULTS AND DISCUSSION

The cytogenetic characterization of the okra crop

As okra is known to contain large numbers of chromosomes that differ between cultivars, we first established chromosome counts and morphology in the cultivar used in this study. Actively growing root tips were fixed and prepared for cell spread preparations following a standard pectolytic enzyme digestion and air-drying protocol, and 4',6-diamidino-2-phenylindole (DAPI) fluorescence microscopy (Kantama et al., 2017). In these diploid plants with a red petiole phenotype, we observed a count of 130 chromosomes in late prophase and metaphase cell complements (Figure 1a). Chromosomes measured 1–2 μm , often show telomere to telomere interconnections (Figure 1b) and were clearly monocentric (Figure 1a). In addition, a few chromosomes displayed a less condensed distal region at one of the chromosome arms and satellites (Figure 1c,d), which we interpreted as the nucleolar organizer region of the satellite chromosome. Interphase nuclei showed well-differentiated heterochromatic domains or chromocenters, most of them with more than 130 spots, although a small number of nuclei decondensed most of its heterochromatin, leaving only a striking pattern of about 10 chromocenters (Figure 1e). We used flow cytometry on DAPI-stained nuclei, utilizing young leaf material from five normally growing red petiole phenotype plantlets. Since we expected a considerable DNA content for okra nuclei, we decided to use a reference sample from Agave (*Agave americana*), which has a known DNA content of 15.90 pg. Surprisingly, in comparison to the Agave reference flow cytometric profile, the 2C DNA amount for the normal okra plant was estimated at $2.99 \text{ pg} \pm 0.01$ (Table S1). This amount is equivalent to a genome size of approximately 2.92 Mbp. In contrast to the 130 chromosomes observed in diploid okra, the haploid okra exhibited a chromosome number of 65 (Figure 1f). The genome size for this haploid okra was estimated at 1.45 Mbp. This plant was feeble, lagging in growth and unfortunately died precociously, but encouraged us to seek for haploid offspring in later samples of reared young okra plants. Such natural haploids, of which a dwarf form of cotton (*Gossypium*) was discovered in 1920 as the first haploid angiosperm with half the normal chromosome complement (Dunwell, 2010), are assumed to result from asexual egg cell (gynogenetic) reproduction (Naumova, 2008). We took advantage of the fact that the diploid hybrid cultivar has a green recessive petiole female parent and a red dominant petiole male (Figure S1) (Portemer et al., 2015). Offspring with the green petiole trait lacks the dominant paternal

allele and hence can be used as a diagnostic marker for identifying haploid offspring. Accordingly, we selected additional haploid offspring of which one plant was used for sequencing.

For the analysis of homologous pairing, chiasma formation, and chromosome segregation in diploid okra plants, we studied male meiosis in pollen mother cells from young anthers (Kantama et al., 2017). Pollen mother cells at meiotic stages are filled with fluorescing granular particles in the cytoplasm, which makes it notoriously difficult to see fine details in chromosome morphology. By long enzymatic digestion and acetic acid maturation, we still could make the following details visible: pachytene was strikingly diploid-like with clear bivalents showing denser pericentromere regions and weaker fluorescing euchromatin distal parts (Figure 1f). We did not observe clear inversion loops indicative of inversion heterozygosity or pairing partner switches that demonstrate homoeologous multivalents or heterozygous translocation complexes, however, the occurrence of such chromosome structure variants could not be excluded. Cell complements at diakinesis displayed that most (if not all) chromosome configurations were bivalents, supporting a diploid-like meiosis (Figure 1g). We did not see univalents or laggards at later stages, and pollen was strikingly uniform and well-stained.

In a study conducted by Hu et al. (2019) regarding the origin and evolution of the phylogenetically related cotton genomes, a synteny analysis between the homoeologous chromosomes of the diploid progenitors *Gossypium arboreum* (A sub-genome) and *G. raiimondii* (D sub-genome) and the allotetraploid cotton genomes from *G. hirsutum* and *G. barbadense* revealed multiple large translocations between the homoeologous chromosomes. To prevent irregular pairing switches and complex multivalent configurations, chromosome pairing leading to a strict diploid-like meiosis is crucial in avoiding disruptions caused by structural and numerical variants events in the tetraploid cotton genome, ensuring stable pairing and crossover patterns. Indeed, a strict diploid-like cytological behavior was described for tetraploid cotton by Endrizzi (1962).

Previously, Joshi and Hardas (1956) suggested the allopolyploid nature of *A. esculentus* and proposed that the 65 chromosome pair cultivars comprise one genome with 29 chromosomes and one with 36 chromosomes. The allopolyploid origin of *A. esculentus* was also considered by Siemonsma (1982) and Hamon and van Sloten (1995), who assumed that one parental species is close to *A. tuberculatus* and the other unknown but related to *A. ficulneus*. We currently do not know which progenitor genomes have contributed to the allopolyploid okra genome or the timepoint of the polyploidization event. It remains uncertain whether the okra sub-genomes harbor structural variations similar to those observed in the *Gossypium* sub-genomes.

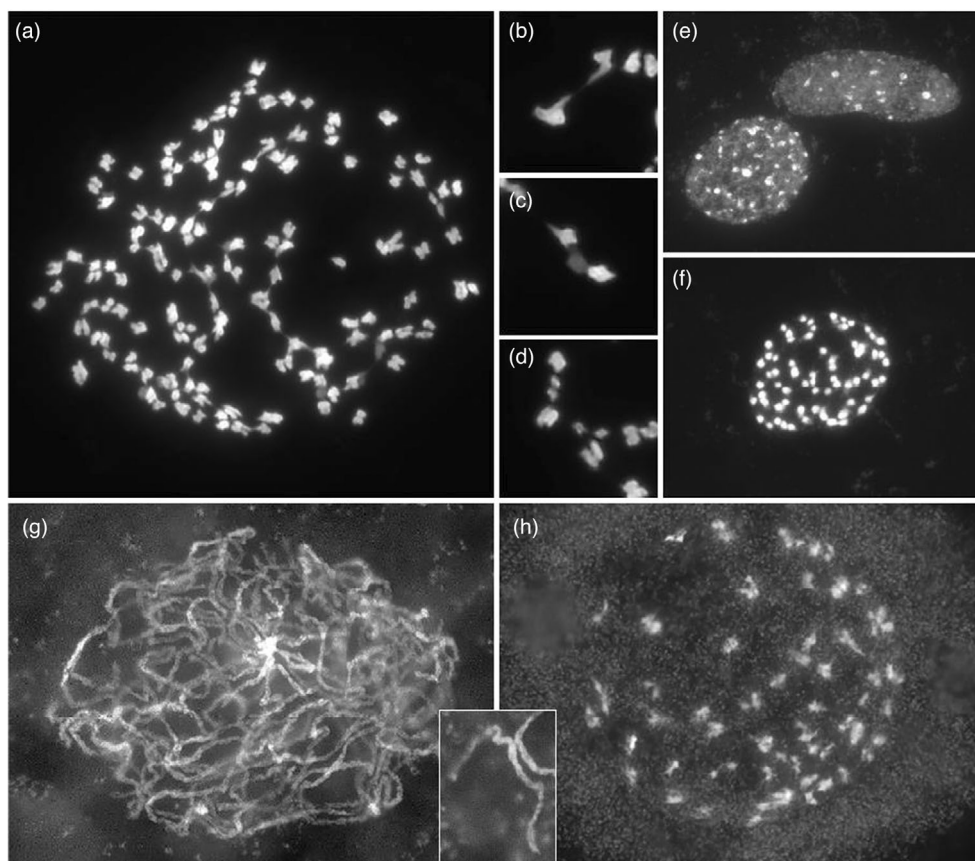


Figure 1. Mitotic cells in root tip meristems.

- (a) Example of a well-spread metaphase complement of a diploid okra plant with $2n = 130$.
 (b) Magnification of two chromosomes that are joined by telomere connectives.
 (c) Chromosome pair with a less fluorescing region, likely representing a decondensed Nucleolar Organizer Region (NOR).
 (d) Chromosomes with small satellites.
 (e) Two interphase nuclei display a striking difference in number of highly condensed chromocenters, regions of the pericentromere heterochromatin, and NORs. The top nucleus has about 10 chromocenters; some other nuclei can have more than 100 of such condensed regions.
 (f) Metaphase complement of a haploid okra plant ($2n = 65$).
 (g, h) Meiotic chromosomes in pollen mother cells of a diploid okra plant.
 (g) Cell at pachytene stage. Most of the chromosomes are fully and regularly paired without clear indications for multiple synapsis, pairing loops, or pairing partner switches. The brightly fluorescing regions are the pericentromeres, see also the inset between the figure (g, h).
 (h) Cell at diakinesis. A greater part of the chromosomes clearly forms bivalents.

However, based on the cytogenetic characterization, we speculate that okra has evolved a strict diploid-like meiosis similar to the phylogenetically close *Gossypium* species to potentially mitigate any effects of such structural variations.

Okra haploid genome reconstruction

Based on public reports (Benchasri, 2012; Salameh, 2014) and cytological analysis presented above, we applied several technologies for genome reconstruction of the okra haploid individual. 10x Genomics linked read information was used to obtain sequence information in the 100 to 150 kbp range. This microfluidics-based technology combines barcoded short-read Illumina sequencing, allowing a set of 150 bp paired-end reads to be assigned to large

insert molecules. We produced 800 Gbp of linked read sequencing data from three libraries with an average GC content of 34% (Table S2). Furthermore, we applied PacBio Circular Consensus Sequencing (CCS), generating 1400 Gbp of polymerase reads of up to 150 kb from circularized insert molecules from three sequence libraries with average fragment insert sizes of 10, 14, and 18 kbp, respectively. Polymerase reads were subsequently processed into consensus or so-called highly accurate long reads (high fidelity [HiFi] reads) with an average read length of approximately 12.2 kbp and a sequence error rate of less than 1% (Table S2). Over 93% of 1000 randomly sampled CCS reads had the best BlastN hit to species from the Malvaceae family with *Gossypium* ranking first in number of hits, indicating the consistent taxonomic origin, in contrast to the

Abelmoschus species that are apparently less represented in the NCBI sequence database (Table S3). Furthermore, the organellar DNA content was sufficiently low (Table S4), illustrating the efficiency of our nuclear DNA sample preparations. Upon assembling the HiFi reads with the Hifiasm assembler (Cheng et al., 2021), we obtained 3051 high-quality primary contigs (Table S5). The incremental sequence assembly displayed in the A50 plot (Figure S2) shows a plateau genome size of approximately 1.35 Gbp, which agrees with the nuclear 2C DNA content. The assembly also resulted in 972 alternative contigs, although their total size of 31 Mbp was small, indicating a highly consistent primary assembly. We nevertheless assessed the origin of alternative contigs, using a taxon-annotated GC (TAGC) screen (Kumar et al., 2013), providing a means to discriminate between on-target and off-target genomic sequence based on the combined GC content and read coverage and corresponding best matching sequence in annotated databases. The distribution and specific classes of Blast hits indicated that approximately 30% of the alternative contigs could be mapped against annotated sequences (Figure S3), while two-thirds were of unknown origin. Alternative contigs had a GC content of 47.6%, which was proportionally higher compared to primary contigs. Furthermore, BlastN hits pointed to a fungal and, to a lesser extent, a bacterial origin. Thus, the smaller-sized alternative contigs represented yet a minor contamination in the gDNA sample. In addition, we assembled the okra chloroplast and mitochondrial genome to further check for possible contaminations in the nuclear genome assembly (Table S5). However, we did not encounter any.

We next physically mapped the genome with BioNano Genomics technology to determine the genome structural organization. We produced 4.88 Tbp of unfiltered genome map data with an N50 molecule size of 90.18 kbp (Table S2) and a label density of 15.9 per 100 kbp (Table S6) from nuclei preparations of leaf samples. Size filtering for molecules larger than 100 kbp left approximately 1.2 Tbp of genome mapping data with an N50 molecule size of 206.8 kbp (Table S6). Next, molecules, having matching label position and distance, were *de novo* assembled into 216 genome maps with an N50 length of 12.98 Mbp and a total length of 1248.8 Mbp, representing an effective coverage of 375 \times (Table S6). The genome map size was consistent with the genome sequence assembly size of 1.2 Gbp and thus provided high-quality ultra-long-range information for further scaffolding. For that, genome maps were aligned with the *in silico* DLE restriction maps from primary sequence contigs and assembled into higher-order scaffolds. The alignment required the cutting of one optical map and four sequence assembly contigs to resolve conflicts between Bionano maps and sequence contigs, respectively, indicating a consistent orientation and order between both. The resulting hybrid assembly

was substantially less fragmented, yielding 80 scaffolds with an N50 scaffold size of 18.93 Mbp and a total length of 1.19 Gbp, of which the largest scaffold sized more than 29 Mbp (Table S5). Additional scaffolding with 10 \times Genomics linked reads finally yielded 65 scaffolds with an N50 length of 19.2 Mbp (Table S5). Approximately 57% of the individual Bionano molecules and 97.5% of the 10 \times Genomics linked reads could be mapped back to the final hybrid assembly (Table S6). To further assess the genome assembly quality, evaluate completeness, and identify possible copy-number errors in our non-reference-based *de novo* assembly, we generated additional assembly assessment metrics using MERQURY (Rhie et al., 2020). The k-mer profiling of the assembly compared to that of high-quality reads indicates a completeness of 99.58% and a consensus quality (QV) Phred-scaled score of 69.95. Regions flagged for possible misassembly were visually inspected and found to appear as deletions in short homopolymer tracts. While MERQURY does not explicitly validate the structural accuracy of the okra assembly, the k-mer profiling metrics along with the ultra-long-range mapping statistics for Bionano and 10 \times Genomics linked reads, suggests a structurally accurate, highly confident genome scaffold and complete genome, without obvious copy number errors.

BUSCO analysis and topology of orthologs

To assess the completeness of the genome assembly we screened for the presence or absence of BUSCO (Benchmarking Universal Single-Copy Ortholog) genes (Li et al., 2015) using a dataset 2326 reference orthologs from the eudicots_odb10 dataset (https://busco-archive.ezlab.org/frame_plants.html). Based on the best tBlastN hit, 2270 (98%) core genes in 65 scaffolds were detected as 'Complete' orthologs (Table 1). Of these, 284 (12.2%) genes were detected as a single-copy ortholog. A very small amount (0.3%) was classified as 'Fragmented', whereas 32 core genes (1.3%) could not be detected, classifying them as 'Missing'. These missing BUSCO genes were confirmed to be missing in the alternative contigs as well. We further grouped 2004 (86.2%) multiplied ortholog genes according to their copy number. A majority of 1150 (49.4%) and 843 (36.2%) orthologs were detected as duplicated and triplicated genes, respectively. Interestingly, we found seven and three core genes that were quadruplicated (0.7%) and quintuplicated (0.1%), respectively, and detected one septuplicate core gene, pointing to a complex polyploid nature of the okra genome (Table 2). To get more insight into the sub-genome organization, the genomic position and topology of ortholog gene copies were assessed. This revealed duplicated BUSCO genes predominantly occurring on two scaffolds, whereas only a single duplicated ortholog was detected on one scaffold (Table S7). Both tandem copies were spaced within 1 kbp, thus likely representing paralogous genes. Out of 800 triplicate BUSCOs, 794 (99%)

Table 1 Detection of ortholog core genes

Class	BUSCO statistics				
	1	2	3	4	4
Assembly version	1	2	3	4	4
Coverage	20x	84x	95x	95x	95x
Ctgs/scfcs	All	All	All	Primary	Alternative
Complete	2288 (98.3%)	2271 (97.7%)	2269 (97.6%)	2288 (98.4%)	66 (2.8%)
Single copy	266 (11.4%)	311 (13.4%)	313 (13.5%)	284 (12.2%)	66 (2.8%)
Multiplied	2022 (86.9%)	1960 (84.3%)	1956 (84.1%)	2004 (86.2%)	0 (0%)
Duplicated	n.d.	n.d.	n.d.	1150 (49.4%)	0 (0%)
Triplicated	n.d.	n.d.	n.d.	843 (36.2%)	0 (0%)
Quadruplicated	n.d.	n.d.	n.d.	7 (0.3%)	0 (0%)
Quintuplicated	n.d.	n.d.	n.d.	3 (0.1%)	0 (0%)
Sextuplicated	n.d.	n.d.	n.d.	0 (0%)	0 (0%)
Septuplicated	n.d.	n.d.	n.d.	1 (0.04%)	0 (0%)
Fragmented	5 (0.2%)	5 (0.2%)	6 (0.3%)	6 (0.3%)	9 (0.4%)
Missing	33 (1.5%)	50 (2.1%)	51 (2.1%)	32 (1.3%)	2251 (96.8%)
Total	2326	2326	2326	2326	2326

Genome assemblies at different coverage levels were analyzed to assess the assembly completeness. BUSCO classes are shown as single copy or multiplied ortholog. Multiplied orthologs are subdivided into additional copy classes as indicated. For each assembly coverage level, BUSCO counts in primary, alternative, and all contigs are shown in absolute numbers and percentages of total expected orthologs (between brackets), or n.d. (not determined).

occurred on three scaffolds, representing three alleles of the same core gene, whereas only six sets (1%) of triplicate core genes were positioned on two scaffolds. Also, quadruplicate, quintuplicate, and septuplicate BUSCOs mainly occurred on three contigs. The copies of these groups manifested in a tandem configuration, probably representing paralogs. Tandemly arranged copies on the same scaffolds always showed less sequence distance than between copies on different scaffolds. Moreover, the copies of the septuplicate core gene were dispersed over three scaffolds. Of these, one contig displayed a triplet, whereas the two other contigs each contained gene copies in a doublet configuration. The triplet consisted of two closely related and one more distantly related paralog. The observed distribution of BUSCO paralogs thus pointed to at least two sub-genomes. However, at this point, we could not rule out a higher number of sub-genomes, which might not be discriminated because of a low allelic diversity. Given the sub-genomic organization for okra, we presume that 284 'single copy' BUSCO genes are either truly unique or they are maintained as gene copies with indistinguishable alleles.

Several examples of BUSCO duplication levels in different plant species, including homozygous and heterozygous diploids as well as in auto and allopolyploids, highlight the variation in gene duplication. For instance, in allotetraploid ($2n=4x=38$) *Brassica napus*, a close relative of *B. campestris* or Chinese cabbage, 90% of BUSCOs are duplicated, whereas only 14.7% in its diploid relative *B. campestris* ($2n=2x=18$) or Chinese cabbage show duplication (Table S8). Similar trends have been observed in the allotetraploid white clover ($2n=4x=32$) (*Trifolium repens*). White clover showing

57% of duplication BUSCOs is thought to be evolved from two related ancestral diploid species *T. occidentale* ($2n=2x=16$) and *T. pallescens* ($2n=2x=16$), which show 10% and 11% of BUSCO duplicates, respectively (Griffiths et al., 2019). Duplicated BUSCOs in hexaploid bamboo *B. amplixicaulis* ($2n=6x=72$) have increased to 57% compared to 35% in its diploid bamboo relative *Olyra latifolia* ($2n=2x=22$) (Guo et al., 2019). Significant differences were also observed in BUSCO scores between heterozygous and homozygous diploid Solanaceae. For instance, the heterozygous *S. tuberosum* RH potato ($2n=2x=24$) showed 74.1% of its BUSCOs duplicated, which was significantly more than 4.3% of duplicated BUSCOs detected in diploid inbred *S. chacoense* M6 potato ($2n=2x=24$), and 9.5% detected in autotetraploid inbred *S. tuberosum* potato (Kyriakidou et al., 2020). Considering these trends, the BUSCO gene copy numbers in the okra genome suggest an allopolyploid nature and support the previous findings, as reported by Joshi and Hardas (1956).

k-mer counts and smudgeplot analysis

To estimate heterozygosity level, repetitiveness, genome size, and ploidy levels, we determined k-mer counts from raw Illumina and HiFi reads. We compared the 21-mers counts for okra to two related allotetraploid cotton species (*G. barbadense* and *G. hirsutum*), each having a genome of approximately 2.3 Gbp, and subsequently visualized the readout with SMUDGE PLOT (Ranallo-Benavidez et al., 2020) (Figure 2). The k-mer-based genome size estimation for the haploid okra amounted to 1.2 Gbp, approximating the NGS assembly size. Approximately 75% of the okra k-mers was assigned to an 'AB' type (Figure 2). Thus, a major part of

Table 2 Structural annotation for the 65 largest okra scaffolds

Class	Count	Av. size	Total length
Total Scfds	4023	328 851	1 322 968 356 (100%)
Large Scfds	65	14 932 447	1 194 595 770 (90.30%)
Gene	130 324	2537	330 639 435 (24.99%)
CDS	150 032	2497	374 629 904 (28.32%)
mRNA	150 032	2497	374 629 904 (28.32%)
Start	150 004	3	-
Stop	150 009	3	-
Intron	676 681	307	207 741 067 (15.70%)
sRNA	2308	797	1 839 960 (0.139%)
Total repeats	1 351 943	-	677 354 628 (51.20%)
Unclassified	834 282	321	268 086 453 (20.26%)
[TTTAGGG]n	123	1810	222 579 (0.017%)
Retroelements	442 480	858	379 556 626 (28.69%)
LTRs	389 339	924	359 876 901 (27.20%)
Gypsy	146 673	1376	201 862 257 (15.26%)
Ty1/Copia	122 317	975	119 289 622 (9.02%)
LINES	26 571	650	17 281 993 (1.31%)
SINES	312	507	158 078 (0.01%)
DNA transposon	46 849	478	15 456 661 (1.17%)
Hobo-Ac	16 781	354	5 941 881 (0.45%)
Tc1-Pogo	645	212	136 523 (0.01%)
5S rDNA	9644	90	871 856 (0.066%)
5S rDNA partial	129	31	4035 (<0.001%)
5.8S rDNA	201	622	125 073 (0.009%)
5.8S rDNA partial	23	231	5311 (<0.001%)
18S rDNA	183	1750	320 241 (0.024%)
18S rDNA partial	170	372	63 169 (0.005%)
28S rDNA	167	3368	562 376 (0.043%)
28S rDNA partial	266	628	167 176 (0.013%)

Features are classified into genic and repeat elements as indicated. Statistics are in nucleotide length and in fractions of total scaffold length.

the okra genome apparently behaved as a diploid, which is consistent with our cytological observations of a diploid-like meiosis, and also coincides with the high number of duplicated BUSCO scores. Approximately 15% of all k-mer pairs showed a triploid behavior ('AAB type'). Furthermore, the 'AAAB' k-mer type seemed more prominent than the 'AABB' k-mer type. Previously, published k-mer readouts for the allopolyploids *G. barbadense* and *G. hirsutum* (Ranallo-Benavidez et al., 2020) showed that at least 50% of the cotton genomes behaved like a diploid, almost a quarter displayed a triploid behavior and 14% of the k-mers showed tetraploid characteristics. Furthermore, cotton k-mer distributions showed the 'AAAB' type more frequently occurring than the 'AABB' k-mer type. This relative proportion of k-mer types, which was suggested to be a characteristic for allopolyploids and in particular for the two allotetraploid cotton species (Ranallo-Benavidez et al., 2020), was also observed for okra (Figure 2). Thus the GENOMESCOPE and SMUDGE PLOT readouts for okra point to an allopolyploid nature of the genome, though less complex and smaller sized than anticipated.

Transcriptome profiling and structural annotation

In addition to sequencing the nuclear genome, we generated approximately 1.2 Tbp of IsoSeq data from multiple tissues including leaf, flower buds, and immature fruits to profile the okra transcriptome. The polymerase mean read lengths of up to 86 kbp benefitted the processing into high-quality CCS reads with a mean length of 4.7 kbp ($\sigma = 378$ bp), indicating the efficient full-length transcript sequencing (Table S9). The CCS reads were used as transcript evidence for okra gene modeling with the AUGUSTUS and GENEMARK algorithms from BRAKER2. We subsequently annotated the 65 largest okra scaffolds with 130 324 genes. Predicted genes had an average length of 2537 nucleotides (nts), whereas the average per gene intron and coding sequence lengths amounted to 307 and 2497 nts, respectively (Table 2). Coding regions showed low sequence diversity, as only 1109 and 8127 single nucleotide polymorphisms (SNPs) could be called from full-length transcripts of the okra haploid and an unrelated diploid individual, respectively (Data S1). Strikingly, the genes we discovered in the okra genome, appeared to be predominantly located at the distal ends of scaffolds, gradually decreasing in abundance toward more centrally positioned scaffold domains. In contrast, LTR-retrotransposons were more abundant in centrally located scaffold domains, while less frequently represented in the distal ends (Figure 3). A comparable distribution between gene and LTR-retrotransposon regions has been observed for other species such as tomato. The gene and LTR-retrotransposon predict a heterochromatin organization of pericentromere heterochromatin and distal euchromatin as shown in Figure 1g and inset. This pattern is common in species with small or moderate chromosome size like Arabidopsis, rice, and tomato. Apparently, okra also has relatively small-sized chromosomes as is substantiated by our cytological observations. The gene-rich regions predominantly occur in euchromatin-rich distal chromosome ends and gradually decrease toward the repeat-rich more condensed pericentromeric heterochromatin, whereas LTR-retrotransposons were more frequently distributed in pericentromeric heterochromatin (Aflitos et al., 2014; Peters et al., 2009; The Tomato Sequencing Consortium, 2012). Our observations thus suggest a similar chromatin architecture to okra chromosomes. Approximately 51% of the assembled genome was found in the repetitive fraction with 20.26% of repeats unclassified (Table 2). A substantial part (28.69%) consisted of retroelements, of which 24.8 and 1.17% was identified as retrotransposon and DNA transposon, respectively. Gypsy and Ty1/Copia retroelements, spanning 15.26 and 9.02% of the assembled genome, respectively, appeared to be most abundant (Table 2). The annotated transposons were also utilized to assess the assembly quality further, employing the LTR

8 Ronald Nieuwenhuis et al.

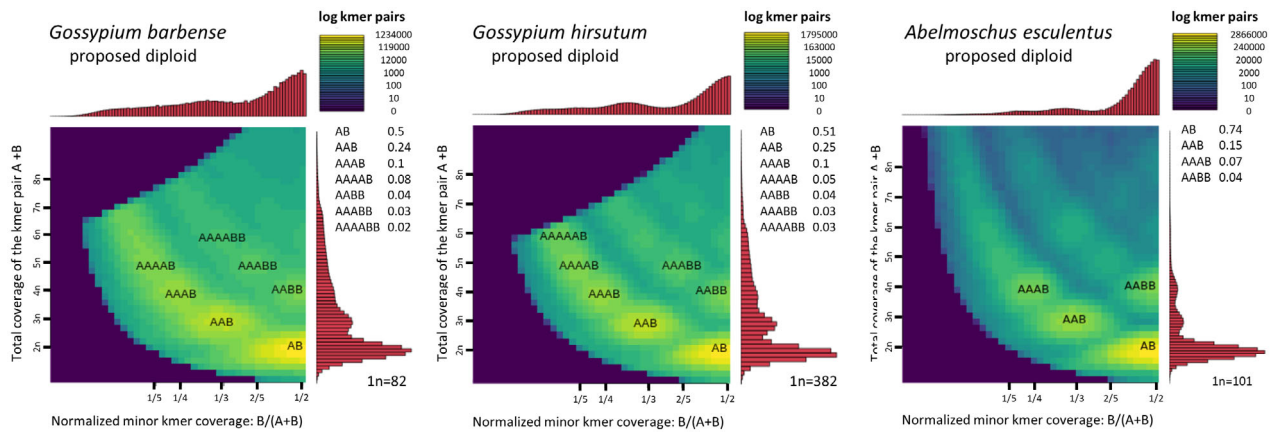


Figure 2. Smudge plot for the allotetraploids. Smudgeplots for *Gossypium barbense* (left panel), *Gossypium hirsutum* (middle panel), and *Abelmoschus esculentus* (right panel) are shown using a log scale. The color scale and digits at the top right of each panel refer to the proportion and absolute k-mer counts per bin, respectively. Below that, the proportion of k-mer heterozygosity forms for diploid, triploid, and tetraploid cases is stated.

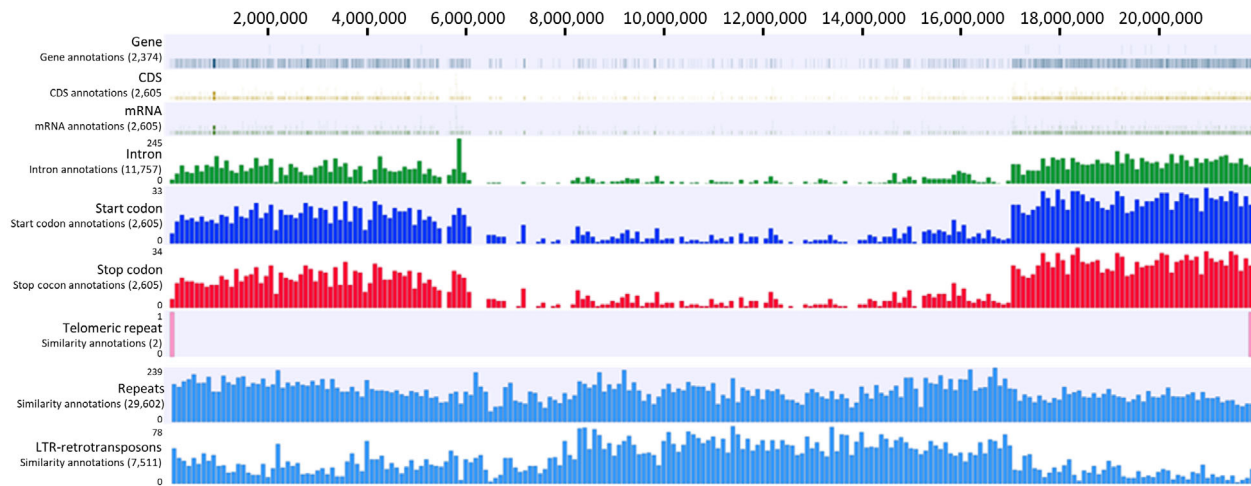


Figure 3. Structural annotation. Annotation feature classes for a 25 Mbp okra chromosome scaffold B30 are indicated at the left side of each row and from top to bottom as gene, cds (coding sequence), mRNA, intron, start codon, stop codon, telomeric repeat, repeats, and LTR-retrotransposon, respectively. The absolute number of features per class is given in digits in between brackets at the left of each row. Colored bar heights in each row correspond to the relative frequency of a feature class per scaffold segment of approximately 115 kb. The top horizontal ruler represents the scaffold coordinate positions in base pairs (bp).

Assembly Index (LAI) metric. With an LAI index score of 13, we achieved an assembly quality score comparable to that obtained for the Arabidopsis and grape reference genomes (Ou et al., 2018).

The repeat screening also revealed stretches of the Arabidopsis telomere TTTAGGG motif, flanking gene-rich regions at distal scaffold ends (Figure 3). In plants such repeats usually occur in high copy numbers at the distal ends of chromosomes, constituting telomeres that protect the terminal chromosomal DNA regions from progressive degradation and prevent the cellular DNA repair mechanism from mistaking the ends of chromosomes for a double-stranded break. Indeed, we found blocks of TTTAGGG units in high copy numbers positioned at both ends for 55 scaffolds, whereas 9 scaffolds had a telomere

repeat block at one end, and 1 scaffold had no telomere ends, in total 119 telomeres at the end of 130 chromosome arms (Data S2). This repeat distribution suggested full-length chromosome scaffolds and capturing the majority of 65 chromosome ends of the haploid okra genome and again confirms the relatively small-sized okra chromosomes. Besides long distal blocks of TTTAGGG repeats, we also detected short interstitial TTTAGGG blocks. These interspersed non-telomeric short TTTAGGG repeats possibly reflect footprints of internalized telomeres that may have arisen from end-to-end fusion of chromosomes (Baird, 2017), possibly representing hallmarks of chromosomal speciation upon allopolyploidization of okra.

Another substantial fraction of repeats originated from ribosomal RNA (rRNA) genes. Using BlastN analysis with

G. hirsutum ribosomal gene query sequences, we clearly observed an arrangement of 18S–5.8S–28S rRNA gene clusters in okra. Two clusters of these genes are located at scaffold ends, though they do not coincide with, or flanking telomere blocks. An additional two clusters are positioned toward the scaffold center, while three scaffolds are almost entirely composed of 18S–5.8S–28S gene clusters. Notably, these scaffolds do not contain 5S rRNA clusters; instead, the 5S rRNA genes are organized in separate clusters, distinctly indicating an S-type rRNA gene arrangement (Goffová & Fajkus, 2021). We identified four 5S rRNA clusters on four different scaffolds, with the largest cluster consisting of almost 8700 copies tandemly arranged on a single scaffold (Table S10). However, we did not observe any clear signatures of underlying chromosome evolution involving telomere fusion at rRNA gene clusters, as interstitial telomere repeats were not found within rRNA clusters.

Genetic diversity in okra

To assess the genetic diversity in public okra germplasm, we used the okra reference genome to call SNPs from several publicly available Illumina RNA-seq datasets that we retrieved from the short-read archive (<https://www.ncbi.nlm.nih.gov/sra>). Most of these samples represent accessions originating from the Indian and Chinese parts of the Asian continent. The average mapping rate for a panel of 11 samples, indicating the high proportion of reads that successfully mapped to the reference genome, was $93.2 \pm 5.6\%$ (Figure 4; Data S1). The combined samples cover 20–25% of the reference genome, which is in line with the 26.6% genic portion of the genome and suggests a faithful structural annotation of the reference genome. The unusually high coverage of the reference genome for the ‘Xianzhi’ dataset (~65%) possibly was due to a deviating library preparation or DNA contamination, while the lower coverage for the ‘IIHR-299’, ‘Mahyco Arka Abhay’, and ‘Commercial’ samples (10–12%) was likely due to their smaller data size (Data S1). The total panel size comprised 412 185 loci, for which 690 145 SNPs were detected from coding regions of okra genes. The ‘Commercial’ sample yielded only 1741 unique SNPs with reads mapping to 12.5% coverage of the reference genome. This SNP rate is substantially lower compared to ‘Mahyco Arka Abhay’ and ‘IIHR-299’, while these three samples cover approximately equal portions of the genome, suggesting that the ‘Commercial’ breeding line apparently shares a large part of its ancestry with the reference cultivar. However, differences in SNP rate may be due to differences in tissue types, growth conditions, data generation, and processing workflows complicating direct sample comparison. Nevertheless, most of the samples attain 20 000 unique SNPs, and only ‘Arka Anamika’ and ‘Danzhi’ exceed this level with 110 368 and 44 461 unique SNPs, respectively (Figure 4;

Data S1). Although we could not yet assess the genetic diversity in the non-genic portion, it appears that the okra accessions in the panel represent low genetic diversity.

Allopolyploid composition of okra

We attempted to divide okra sub-genomes by searching for patterns based on hierarchical clustering of repeat k-mer counts without a supposition of ancestral species. We compared the clustering results for okra with two k-mer test sets, of which the first comprised an artificially constructed hybrid genome, consisting of merged tomato (*Solanum lycopersicum* cv. Heinz 1706) and a diploid potato (*S. tuberosum* cv. Solyntus) genomes. The second set was generated from the allotetraploid cotton genome (*G. hirsutum*). As expected the hierarchical clustering analysis of the artificial hybrid dataset produced a cluster map that separated into two distinct subclusters, each representing repetitive k-mers from 12 tomato and potato chromosomes in the merged genome (Figure S4). Similarly, the analysis of the cotton genome also resulted in two distinct groups, effectively separating the repetitive k-mers from the sub-genome A and D chromosomes (Hu et al., 2019). Applying the same method to the okra reference genome, of which the five smallest scaffolds were removed, yielded two distinct clusters of 30 and 35 scaffolds with a length of 636 and 557 Mbp, respectively (Figure S5). Scaffolds from cluster 1 are overall larger in length than those from cluster 2 (Figure S6). Apparently, the repeat k-mer pattern for okra points to two distinct sub-genomes A and B, consisting of 30 and 35 chromosomes, respectively, and, together with the apparent absence of multivalent pairing of metaphase chromosomes, suggests an allotetraploid nature of *A. esculentus*. In addition, we could not find clear evidence of erosion, as the clusters had comparable BUSCO completeness scores of 90.2 and 89.3%, with duplicate rates of 21.8 and 19.4%, respectively. This suggests a relatively recent hybridization of ancestral species, yet without clear evidence indicating the dominance of one sub-genome over the other sub-genome.

We subsequently aligned the clustered chromosome scaffolds to further investigate the orthology between the two sub-genomes. Although we found only partial alignments and several inversions, there is substantial homology between scaffolds (Figure S7), confirming in general there are no homoeologs within a single cluster. Specifically, the network displayed the portion of BUSCO genes shared between chromosome scaffolds (Figure S8), indicating the strong homology between chromosome scaffolds of clusters 1 and 2 and corroborating the alignment dot-plot.

Further, repeat annotation reveals only 784 out of 22 100 repeat classes are present in >90% of chromosome scaffolds in both clusters. A small number of 14 repeat

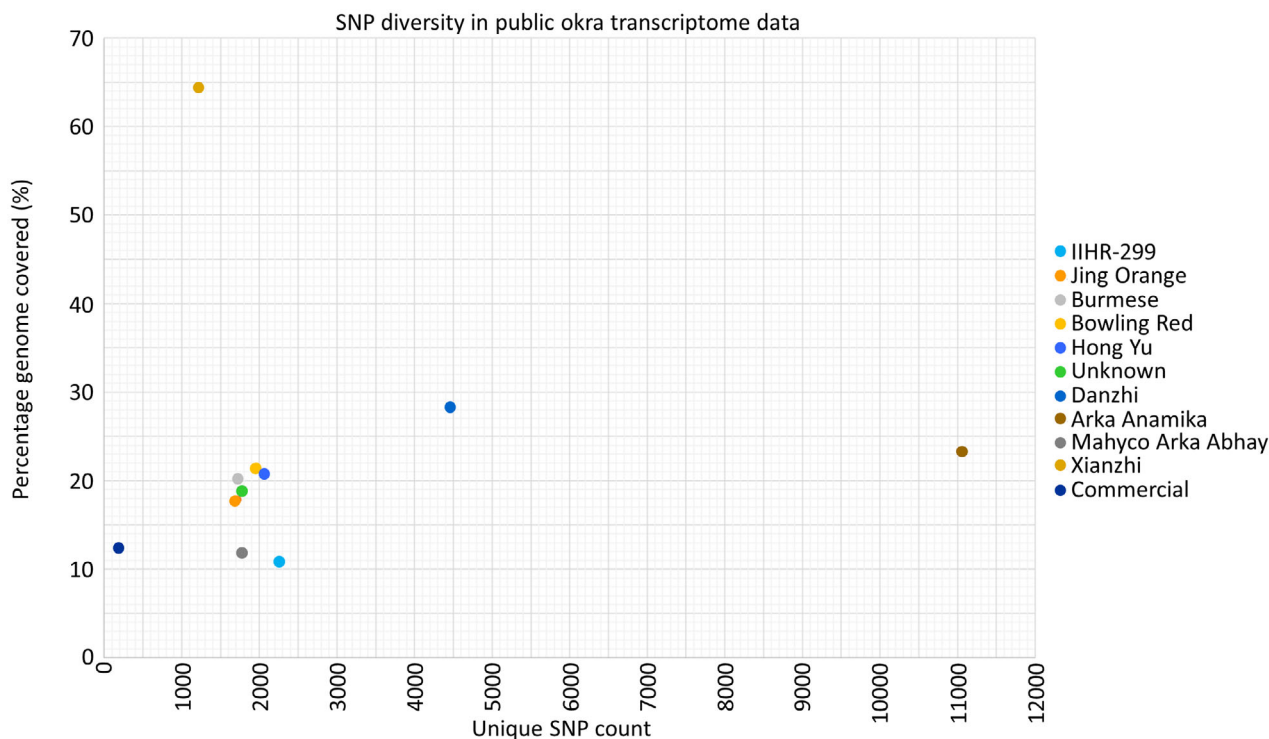


Figure 4. Genetic diversity in okra germplasm. The unique single nucleotide polymorphism (SNP) count, derived from transcriptome data mapping for 11 okra accessions, is shown along the x-axis (in millions). The percentage of genome coverage per transcriptome dataset for each accession is shown along the y-axis. Accession identities are according to the legend inset at the right. The average count of 1 SNP per 2.1 kb in the genic portion confirms a low genetic diversity in the panel of accessions.

classes are present in more than 90% of the scaffolds in either cluster, while occurring in less than 10% of scaffolds in the other. These include cluster-specific repeat classes, occurring in high copy numbers in cluster 1 but not in cluster 2 and vice versa (Figure S9). The distribution of specific repeats over two distinct clusters again points to two distinct sub-genomes A and B. Although such sub-genome-specific repeats do not show similarity to known repeats and we can only speculate about their origin, they are clearly related to different ancestral progenitor species, further pointing to an allotetraploid nature of okra. At this point, we assigned the 30 chromosome scaffolds from cluster 1 to sub-genome A and 35 chromosome scaffolds from cluster 2 to sub-genome B.

Candidate genes assigned to phenylpropanoid, flavonoid, and flavone and flavonol biosynthesis pathways

Polyphenols represent one of the most ubiquitous class of secondary metabolites in okra fruits. An important subclass of polyphenols are flavonoids, including myricetin, quercetin, isoquercitrin, and quercetin-3-*O*-gentiobioside derivatives that have been implicated in antidiabetic activity (Lei et al., 2018; Liu et al., 2005; Peter et al., 2021; Wu

et al., 2020). Myricetin was previously detected in *A. moschatus* (Liu et al., 2005). Recently, the bioactive phytochemicals isoquercitrin and quercetin-3-*O*-gentiobioside, and to a lesser extent also rutin and catechin, were detected as the major phenolic compounds in okra fruits (Wu et al., 2020). Their biosynthesis in the flavonoid and, flavone, and flavonol biosynthesis pathways (KEGG reference pathways 00941 and 00944) is thought to start with *p*-coumaroyl-CoA and cinnamoyl-CoA precursors that are synthesized in the phenylpropanoid pathway (KEGG reference pathway 00940).

To find putative enzyme coding okra genes that may function in phenylpropanoid, flavonoid, flavone, and flavonol biosynthesis, 142571 extracted amino acid query sequences from predicted okra genes and putative splice variants were mapped against the manually curated KEGG GENES database, using the KEGG Automatic Annotation Server (KAAS) [KAAS – KEGG Automatic Annotation Server (genome.jp) (Moriya et al., 2007)]. We identified 33641 amino acid sequences that could be assigned to 395 KEGG metabolic pathway maps based on a best bidirectional hit (BBH). Currently, in total there are $N=1302$ manually annotated genes from Malvaceae species

T. cacao (cacao), *G. arboreum*, *G. hirsutum* (cotton), *G. raimondii*, and *Durio zibethinus* (durian) of which $K=99$ enzyme coding reference genes are known for the phenylpropanoid ($K_1=36$), flavonoid ($K_2=30$), or flavone and flavonol ($K_3=33$) biosynthesis pathway in KEGG. Out of the 33 641 okra amino acid queries, $n=47$ putative okra orthologs were assigned to the KEGG phenylpropanoid biosynthesis ($n_1=16$), flavonoid biosynthesis ($n_2=17$), and flavone and flavonol biosynthesis ($n_3=14$) pathways, respectively, adding up to $n=41$ distinct putative okra enzyme orthologs. We subsequently assessed the mapping probability of okra orthologs to the reference pathways based on the known Malvaceae enzymes and the okra BBH. Mapping confidence values $P1=1.43e^{-12}$, $P2=1.15e^{-12}$, and $P3=0.0$ pointed to a confident assignment of okra orthologs to phenylpropanoid biosynthesis, flavonoid biosynthesis, and flavone and flavonol biosynthesis pathways, respectively. Copy numbers for putative genes possibly involved in the conversion of *p*-coumaroyl-CoA and cinnamoyl-CoA precursors varied extensively. Only a single putative gene orthologous to a 5-*O*-(4-coumaroyl)-*D*-quinic acid 3'-monooxygenase (EC:1.14.14.96) from *Durio zibethinus* (XP_022742205) with an amino acid identity of 93.9% was found on sub-genome B, whereas 14 putative orthologs on sub-genome A to shikimate *O*-hydroxycinnamoyltransferase (EC2.3.1.133) were detected, with a highest amino acid identity (94.9%) to the ortholog from *G. arboreum* (XP_017607223). The coverage of okra orthologs mapped to these biosynthesis pathways is shown in Figure 5 and Figure S10a,b. The alternative metabolic routes, leading to the biosynthesis of quercetin, myricetin, isoquercitrin, and quercetin-3-*O*-gentiobioside derivatives in these pathways, involve three critical flavonol synthases CYP74A (flavonoid 3',5'-hydroxylase, EC:1.14.14.81), CYP75B1 flavonoid 3'-monooxygenase (EC:1.14.14.82) and (FLS) dihydroflavonol,2-oxoglutarate: oxygen oxidoreductase (EC:1.14.20.6). Apparently, putative orthologs for CYP74A and CYP75B1 are encoded by a single copy gene in okra, whereas the FLS oxidoreductase catalytic activity, that is thought to catalyze the conversion of several dihydroflavonol intermediates into quercetin and myricetin, appears represented by five putative okra homologs, suggesting that the conversion into quercetin, isoquercitrin, rutin, and catechin mainly runs via a dihydrokaempferol intermediate.

Conclusions

We successfully applied multiple DNA sequencing and scaffolding techniques to reconstruct the 65 chromosomes of a haploid okra sibling. The final result consists of 55 chromosomes, each with telomeres at both ends, illustrating chromosome completeness. Additionally, there are nine chromosome scaffolds representing nearly resolved chromosomes, although the later have a telomere at only

one end. This assembly contiguity for okra surpasses that of other crop genome assemblies. The use of PacBio HiFi reads and availability of a haploid sample were key factors contributing to the high quality of the reconstructed okra genome. The more pronounced diploid-like behavior of okra, as exemplified by the SMUDGE PLOT spectrum showing 75% AB dominance when compared to the k-mer spectrum for allotetraploid cotton with 50% AB dominance, illustrates a decreased complexity that benefitted the genome reconstruction. This diploid-like nature aligns with our cytogenetic observations on pollen mother cells at pachytene, showing chromosomes strikingly diploid-like with clear bivalents. The okra genome is characterized by a high chromosome count and relatively small chromosomes, each not extending beyond 30 Mbp in length. The total haploid assembly size of approximately 1.2 Gbp apparently is consistent with the slightly larger k-mer-based haploid genome size estimation. Mapping of telomeric sequences and gene-dense sections toward both scaffold ends, together with repeat-dense sections mapping toward more centrally located scaffold sections, is in line with the observed chromatin landscape at pachytene for several okra chromosomes that display pericentromeric heterochromatin and distal euchromatin. These regions have been shown to be repeat-rich and gene-rich, respectively, in many species' genomes. While the overall repeat content (57%) is lower than described for *G. hirsutum* (67.2%) (Li et al., 2015), and *G. raimondii* (70.7%), it is higher than for *T. cacao* (29.4%) (Novák et al., 2020), despite the smaller size of these genomes. However, our repeat classification remains incomplete due to a lack of diversity in annotated repeat libraries. Moreover, with over 130 000 putative genes, the gene prediction count appears inflated compared to other species genomes like *Arabidopsis* and rice (*Oryza sativa*), which contain around 38 000 and 35 000 coding, non-coding, and pseudogenes, respectively. Some increase in okra gene count may be attributed to allopolyploidization. Conversely, decreased selective pressure may cause many genes to accumulate *de novo* mutations and convert into pseudogenes (Bird et al., 2018). Nevertheless, over 88% of the predicted exons were supported by high-quality long-read data and predicted proteins mostly returned partial matches to several different databases, substantiating our gene prediction. In this respect, the structural and functional annotation revealed putative enzyme-coding genes that we could map to phenylpropanoid, flavonoid, flavone, and flavonol metabolic pathways, likely underlying the biosynthesis of an array of secondary metabolites that have been implicated in dietary and therapeutic bioactivity.

Additionally, the identification of sub-genomes from distinct k-mer repeat profiles point to an allotetraploid composition of the genome. The two sub-genomes apparently have unequal chromosome counts and differ slightly

certain genes are preferentially retained, possibly even accumulated, or preferentially lost (Mandáková & Lysak, 2018; Renny-Byfield et al., 2013). For instance, we noted a notable difference in copy number for BUSCO genes and specific okra genes likely involved in phenylpropanoid biosynthesis, flavonoid biosynthesis, and flavone and flavonol biosynthesis pathways. These disparities in paralogous gene counts may be attributed to a phenomenon known as biased gene fractionation (Mandáková & Lysak, 2018), which in turn could have led to sub-genome fractionation. Indeed, we identified a small difference in the overall lengths of the two sub-genomes, which suggests potential sub-genome differentiation and implies the possibility of sub-genome dominance (Mandáková & Lysak, 2018). Finally, we emphasize that the annotated high-quality genome provides a solid basis for advancing okra breeding, encompassing diversity and compatibility screening, and marker development.

MATERIALS AND METHODS

Chromosome analysis

Plants of the Green Star F1 hybrid of okra (*A. esculentus*) were grown in small for collecting actively growing rootlets that appeared at the outside of the pot soil. The root tips were pretreated with 8-hydroxyquinolin and then fixed in freshly prepared glacial acetic acid: ethanol 96% (1:3) and 1 day later transferred to ethanol 70% for longer storage at 4°C. Young flower buds were collected from nurse fields in Kamphaeng Saen, Thailand, and directly fixed in acetic acid ethanol without pretreatment. Microscopic preparations of root tip mitoses and pollen mother cells at meiotic stages were prepared following pectolytic enzyme digestion of cell walls and acetic acid maceration and cell spreading following the protocol of Kantama et al. (2017). Air-dried slides were stained in 300 nm DAPI in Vectashield (Vector Laboratories, Newark, CA, USA) and studied under a Zeiss fluorescence microscope equipped with 1.4 N.A. objectives and appropriate epifluorescence filters for DAPI. The captured images were optimized for best contrast and brightness in Adobe Photoshop, and slightly sharpened with the Focus Magic (www.focusmagic.com) 2D deconvolution sharpening to remove excessive blurring of the DAPI fluorescence (Kantama et al., 2017).

Bionano optical maps

Sequence-specific labeling of approximately 700 ng genomic DNA from okra cv. Green Star and subsequent backbone staining and DNA quantification for BioNano mapping was done using a Direct Label Enzyme (DLE-1, CTTAAG) according to the manufacturer protocol 30206F BioNano Prep Direct Label and Stain Protocol (<https://bionanogenomics.com/wp-content/uploads/2018/04/30206-Bionano-Prep-Direct-Label-and-Stain-DLS-Protocol.pdf>). Chip loading and real-time analysis was carried out on a BioNano Genomics Saphyr[®] analyzer, using the green color channel on three flow cells, according to the manufacturer system guide protocol 30143C (<https://bionanogenomics.com/wp-content/uploads/2017/10/30143-Saphyr-System-User-Guide.pdf>). Using the DLE-1 enzyme, 1.18 Tbp of filtered DNA molecules with an average length of 215 kbp was produced, with a label density of 15.9/100 kb and a molecule N50 of 207 kbp. Subsequently, a *de*

novo assembly was constructed using Bionano Access[™] (v.3.2.1) and the non-haplotype aware assembly program without extending and splitting but with the cutting of the complex multi-path regions. Per the default settings, molecules <150 kbp were removed before assembly. Next, a hybrid scaffolding of assembled sequence contigs was performed with Bionano Genomics Solve (v.3.2.1) with a 375x-fold coverage for the DLE-1 molecules. Molecule quality hybrid scaffold reports were carried out using the BioNano Solve[™] analysis pipeline (<https://bionanogenomics.com/support-page/data-analysis-documentation/>).

Pacbio HiFi, linked-read sequencing, and de novo assembly

We produced three Pacbio HiFi libraries using gDNA isolated from okra leaf tissue according to the manufacturer's protocol (<https://www.pacb.com>). HiFi reads of 15–20 kbp were generated by Circular Consensus Sequencing, using six SMRT cells, in total yielding 1400 Gbp of sequence data. Subsequent consensus calling was done using the pbccs v5.0.0 command line utility. HiFi reads were defined as CCS reads having a minimum number of three passes and a mean read quality score of Q20. Reads from different libraries were then combined into a single dataset for further analysis. Assembly of HiFi Reads was done using hifiasm v0.12-r304 for coverages of ~20x, ~84x, and ~95x (Cheng et al., 2021). Primary contigs of the ~95x coverage assembly were scaffolded using BIONANO GENOMICS SOLVE v3.6_09252020 and an optical *de novo* assembly. Solve scaffolded output was further scaffolded, in contrast to the unscaffolded output, using Arcs v1.2.2 (<https://github.com/bcgsc/arcs>) and Links v1.8.7 (<https://github.com/bcgsc/links>) based on the 10x genomics data that was mapped using LONGRANGER v2.2.2. Scaffolds resulting from the final step were renamed to fit the naming scheme from the Bionano scaffolding.

The 10x Genomics libraries were constructed with the Chromium[™] Genome Reagent Kits v2 (10x Genomics[®]) according to the Chromium[™] Genome v2 Protocol (CG00043) as described by the manufacturer (<https://www.10xgenomics.com>). 10x Genomics libraries were sequenced on two separate runs using the Illumina Novaseq6000 platform and S2 flow cells. Base calling and initial quality filtering of raw sequencing data were done using BCL2FASTQ v2.20.0.422 using default settings. The Long Ranger pipeline from 10x Genomics was used to process the 800 Gbp sequencing output and align the reads to the okra draft genome. After detecting the conflict region with the BIONANO GENOMICS ACCESS SUITE (v.1.3.0), we manually inspected the conflict regions using 10X linked reads mapped to the superscaffolds. Mapping and visualization of scaffolds were done with LONGRANGER WGA v.2.2.2 and LOUPE v.2.1.1, respectively.

Assembly QC

Statistics on NX lengths, GC-percentage, mean-, median-, maximum-, and minimum lengths of contigs or scaffolds for each assembly step were collected from software output and when not available generated with custom python scripts. HiFi reads were mapped back to the assembly using MINIMAP2 v2.17-r941 to assess purging correctness with PURGE_HAPLOTIGS (Roach et al., 2018). MINIMAP2 alignment was also used for BLOOTOOLS v1.1.1 analysis to check the taxonomic origin, coverage, and GC-percentage of unscaffolded output (Laetsch et al., 2017). Base-level accuracy and assembly completeness were evaluated with MERQURY (Rhie et al., 2020). The completeness of the assembly was further benchmarked with the LTR Index Assembly (LAI) metric (Ou et al., 2018) and by using BUSCO with the eudicots_odb10 (Eukaryota, 2020-09-10) lineage set to scan for single-copy orthologs

(Kriventseva et al., 2019; Simão et al., 2015). AUGUSTUS v3.2.2 was subsequently used for gene prediction. BUSCO output was used for topology analysis of duplicated genes (Stanke, Keller, et al., 2006; Stanke et al., 2006). To assess repeat content and synteny within the scaffolds, the assembly was self-aligned using a combination of MINIMAP2 and DGENIES, and NUCMER v4.0.0beta2 together with MUMMERPLOT v3.5 (Kurtz et al., 2004).

Iso-Seq sequencing and data analysis

Total RNA was isolated from leaf (10 µg), flower buds (21 µg), and young fruits (31 µg) from okra cv. 'Green Star'. RNA quality was checked on a Bioanalyzer platform (<https://www.agilent.com>) by comparing it to standard samples of 25S and 18S ribosomal RNA. Transcript samples were subsequently used for the construction of three sequence libraries and sequenced with PacBio SMRT technology (<https://www.pacb.com/smrt-science/smrt-sequencing>) using four SMRT cells. Consensus reads were produced with the ccs v5.0.0 command-line utility of PacBio. HiFi reads were classified as such using the same specifications as the genomic reads. Primer sequences from reads were removed and demultiplexed with LIMA v2.0.0. Poly-A tails were trimmed and concatemer was removed with ISOSEQ3 v3.4.0 refine to generate full-length non-concatemer reads and subsequently clustered with ISOSEQ3 CLUSTER. Since distributions of mean read quality showed over 90% of data to have a mean quality score in range of 90–93, no final polishing was applied. High-quality full-length transcripts obtained from the SMRT analysis pipeline were then mapped to the hybrid assembly using GMAP (Wu & Watanabe, 2005).

Read analysis

Quality control of reads was performed using SMRTLINK v9 and FASTQC v0.11.9. A random sample of 1000 reads per SMRT cell was taken using SEQTK v1.3-r106 seq with a random seed from the BASH v4.2.46 internal pseudorandom generator. Samples were screened for chloroplast content, plastid content, and taxonomic contamination by applying BLAST v2.10.1+ with parameter settings *-evalue 0.001* and *-max_target_seqs 1* (Altschul et al., 1990, 1997; Camacho et al., 2009). The databases used for each screening were NCBI nt, plastid, and mitochondrion publicly available FTP downloads dated 2020-11-15 (Agarwala et al., 2016).

K-mers were counted for both 10X linked reads and HiFi reads, using KMC 3.1.1 (Kokot et al., 2017) with parameter settings *-m64 -ci1 -cs10000* for *k = 16, 21, 28, 37, 48, and 61* to determine the k-mer size for best-model fit. For 10X linked reads 23 bp of R1, containing 16 bp barcode plus the 7 bp long spacer sequence, were trimmed off before counting. The kmc histogram of counts was subsequently used for the estimation of genome parameters and visualization of the k-mer spectrum using GENOMESCOPE2 (Ranallo-Benavidez et al., 2020). The polyploid nature of okra was further examined by applying a locally developed, publicly available fork of SMUDGE PLOT labeled v0.2.3dev_rn that is true to the original algorithm, allowing for parallelization. In the original algorithm, k-mer pairs with a hamming distance of 1 are found by a recursive method that iterates over all positions within the k-mer. The redesigned algorithm parallelizes the search by each thread looking at a given position within the k-mer. To reduce the number of false negatives, results are then filtered using a bloom filter with an *error rate* set at 0.0001.

Annotation

Repeats annotation was carried out with REPEATMODELER, REPEAT-CLASSIFIER, and REPEATMASKER tools and the combined REPBASE

(2014) and DFAM (2020) databases for the classification of repeats (Bao et al., 2015; Hubley et al., 2016; Smith et al., 2013). Full-length non-concatemer IsoSeq reads were mapped against the genome assembly using MINIMAP2 with parameter settings *-ax splice -uf --secondary=no -C5*. A repeat masked genome and transcriptome read mapping was then input to the BRAKER2 v2.1.5 pipeline to generate a structural annotation of genes using *ab initio* prediction (Borodovsky & Lomsadze, 2011; Hoff et al., 2019). Using the general feature format (GFF) file output, open reading frame translations of the predicted genes were made with the default eukaryotic translation table. Produced polypeptide sequences were then annotated with INTERPROSCAN v5.39-77 10 that collected annotations from the default set of databases including PANTHER, GEN3D, CDD, etc. (Hunter et al., 2009; Jones et al., 2014).

Variant calling

To investigate the diversity among okra accessions, multiple publicly available transcriptome datasets were mapped to the assembled okra reference as presented in this study. Public datasets include SRR620228 (<https://www.ncbi.nlm.nih.gov/sra/?term=SRR620228>), PRJNA393599 (<https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA393599>), and PRJNA430490 (<https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA430490>). In case multiple tissues or runs were available, data were merged. Mapping was done using STAR v2.7.8a (Dobin et al., 2013; Leinonen et al., 2011). Duplicates were marked by applying GATK v4.2.0 MarkDuplicatesSpark mode (McKenna et al., 2010). Then variants were called using GATK HAPLOTYPECALLER with default filters. Obtained variants were then filtered for SNPs and subsequently quality filtered by GATK SELECTVARIANTS with parameter filter settings *QD > 2.0 && MQ > 40.0 && FS < 60.0 && SOR < 3.0 && QUAL > 30.0 && MQRankSum > -12.5 && ReadPosRankSum > -8.0*. IsoSeq reads were mapped as described above.

Sub-genome separation

To separate the sub-genomes in the okra genome assembly, a clustering approach on the repetitive part of the okra genome was applied to generate clusters of scaffolds with similar repeat patterns. For each scaffold, forward and reverse complement non-canonical k-mers were counted setting a length *k = 13*. The counts were normalized by scaffold length and then filtered to have a minimal count of 100. Sets of k-mer counts for both forward and reverse strands of each scaffold were then merged. All counts were subsequently increased by one and \log_{10} was transformed to generate a scale that was appropriate for visualization. Euclidean distances between the scaffolds were stored and scaffolds were subsequently clustered using Ward's (minimum variance) method. To visualize the clustering along the k-mer patterns, we generated a cluster map from a heatmap for the k-mer counts combined with a hierarchical clustering of the scaffolds. In addition, the clustering of k-mers generated from an in-house generated artificial potato tomato hybrid assembly, and the allopolyploid cotton genome were used as a test set. Clusters were then compared based on BUSCO score completeness and scanned for homoeology in a network visualization using an adjacency matrix. Scores in the adjacency matrix were taken from the edge counts between scaffolds that have identical BUSCO genes in common, implying that scaffold x and scaffold y have an edge count of z when sharing a number of z BUSCO genes. Based on a threshold of 50% BUSCO gene presence we subsequently constructed a BUSCO connectivity graph.

Statistical analysis of metabolic pathway assignment for okra orthologs

The mapping probability for okra orthologs to KEGG reference pathways was based on a hypergeometric test (one-sided Fisher's exact test) to measure the statistical significance of pathway assignment of a putative okra ortholog set. Pathway P -values were calculated according to Equation (1), where K equals the unique enzymes known for a pathway p , k for the number of searched enzymes uniquely mapping on pathway p , N as the number of unique enzymes of all reference species known for all pathways, and n the number of searched enzymes uniquely mapping on all pathways.

$$P(X = k) = f(k, N, K, n) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \quad (1)$$

ACKNOWLEDGEMENTS

We wish to thank Hortigenetics Research of East-West Seed (S.E. Asia) Ltd., ENZA Zaden Research and Development B.V., Genetwister Technologies B.V., Nunhems Netherlands B.V., Syngenta Seeds B.V., Takii & Company Ltd., HM. Clause, SA., UPL Ltd., Namdhari Seeds Pvt. Ltd., Maharashtra Hybrid Seeds Co. Pvt. Ltd., and Acsen HyVeg Pvt. Ltd. for providing material and support to the okra genome project.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Datatypes for this study are available at <https://www.ncbi.nlm.nih.gov/> under BioProject number PRNJA985739.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Data S1. Transcriptome mapping of public okra accessions to the okra reference genome and SNP calls.

Data S2. Telomere configuration for chromosome scaffolds. Chromosome scaffolds have been divided over sub-genome A and B. Each chromosome scaffold has a telomere count for its right and left scaffold end.

Figure S1. Phenotypes of diploid and haploid okra plants. The magnifying glass in the image is placed over the position of the green and the red petiole for the haploid (left) and diploid (right) Okra plant, respectively.

Figure S2. The incremental genome assembly size for okra. The A50 plot for contigs larger than 100 bp shows the assembly size in Gbp on the y -axis is plotted against the incremental contig count at 20 \times , 84 \times , and 95 \times sequence coverage indicated by the light blue, red, and orange curves, respectively. The dashed line indicates the anticipated genome size.

Figure S3. Taxon annotated GC coverage plot. In the low left panel, the proportion of GC bases (x -axis) and read coverage (y -axis) for 527 alternative contigs are shown. Each colored dot in the graph corresponds to a contig. Colors correspond to species classes for which a best BlastN match was found in annotated databases. In the top left graph and the low right graph, the relative proportion for each class is depicted with respect to the coverage and GC content, for which color codes match species classes as indicated in the legend at the top right.

Figure S4. Cluster maps of repetitive 13-mer counts. Hierarchical cluster maps for *Gossypium hirsutum* and an artificially constructed hybrid diploid genome from tomato (*S. lycopersicum* cv. Heinz 1706) and a diploid potato (*S. tuberosum* cv. Solentus) is shown in the left and right panel, respectively. The index at the top left of each panel refers to the log₁₀ count of 13-mers occurring with at least 100-fold coverage. k -mer counts have been normalized for chromosome length.

Figure S5. Cluster maps of repetitive 13-mer counts for the okra reference genome. Hierarchical cluster maps for the okra (*Abelmoschus esculentus*) reference genome. The index at the top left of the panel refers to the log₁₀ count of 13-mers occurring with at least 100-fold coverage. k -mer counts have been normalized for chromosome length.

Figure S6. Repeat content analysis in chromosome scaffold. Chromosome scaffolds assigned to sub-genome A and B are represented by yellow and green dots, respectively, and have been separated by scaffold length (x -axis) and repeat percentage (y -axis).

Figure S7. Dot plot alignment of chromosome scaffolds from sub-genomes. Superscaffolds are assigned to cluster A or B according to their k -mer clustering profile. The top right graph shows two homoeologous chromosome scaffolds A16 and B04 having 72% of BUSCO genes in common. The bottom right alignment detail of the aforementioned chromosome scaffolds is partially syntenic, sharing a large inversion.

Figure S8. BUSCO connectivity graph. Yellow and green color-coded nodes correspond to chromosome scaffolds from sub-genomes A and B, respectively. Edges between the nodes indicate the percentage of shared BUSCO genes between each scaffold pair. Note that a single node can have multiple edges. Pairs of scaffolds point to links of homoeology between chromosomes from sub-genome A and B.

Figure S9. Repeat count of cluster-specific and shared repeats between sub-genomes. The occurrence of three distinct repeat families is shown as counts per chromosome scaffold (y -axis) that are divided over sub-genome A and B (x -axis). Counts per chromosome scaffold are represented by gray dots. The repeat family identifier is indicated above each plot. The left panel shows the occurrence of an unclassified repeat family in sub-genome A-specific scaffolds while absent in sub-genome B scaffolds. The right panel shows an unclassified sub-genome B-specific repeat family. The unclassified repeat family in the middle graph is sub-genome unpecific.

Figure S10. (a) The phenylpropanoid KEGG bio-synthesis pathway in *Abelmoschus esculentus* (<http://www.kegg.jp/kegg/kegg1.html>). Putative okra enzyme coding genes for which a bi-directional best hit was found to enzyme pathways are shown with colored EC identifiers. (b) The flavone and flavonol KEGG bio-synthesis pathway in *Abelmoschus esculentus* (<http://www.kegg.jp/kegg/kegg1.html>). Putative okra enzyme coding genes for which a bi-directional best hit was found to enzyme pathways are shown with colored EC identifiers.

Table S1. DNA amount of nuclei samples from okra root tip cells. DNA amount of okra replicate samples in picogram quantities was compared to a reference sample from *Agave Americana*. In the right column flow histograms of Okra samples and the Agave reference sample are shown. The count of observed nuclei in each histogram is depicted on the y -axis and is proportional to the fluorescent intensity of each peak. The position of the peak along the x -axis is proportional to the relative DNA amount in each nuclei.

Table S2. Genome sequencing and genome map data statistics. Linked read sequencing for three 10 \times Genomics libraries was

performed using Illumina paired-end (PE) sequencing. Circular consensus sequencing (CCS) was performed for three Pacbio HiFi sequence libraries. Genome map data was produced for one BioNano DLE labeled library.

Table S3. BlastN screening statistics for Pacbio HiFi reads. Screening was performed against the NCBI nucleotide database. Readouts are indicated in counts of Malvaceae species-specific and non-Malvaceae hits for a subset of 1000 HiFi reads.

Table S4. Pacbio sequence library contamination statistics for 1000 HiFi. Organelle content was determined using a BlastN screening against mitochondrial and chloroplast databases.

Table S5. NGS assembly and hybrid scaffolding statistics. Sequences were assembled using the Hifiasm assembler and scaffolded with Bionano Genomics genome maps. The number of chromosome scaffolds was obtained with 10x Genomics linked reads.

Table S6. *De novo* genome map assembly statistics. Assembled molecules were mapped back to genome maps to estimate the effective coverage and average confidence of the *de novo* assembly.

Table S7. BUSCO distribution and topology. BUSCO genes are classified according to their copy number in the genome. Distribution counts for gene copies have been indicated according to their position either on one, two, or three contigs. Configuration of gene copies is depicted by a horizontal line representing a contig and superimposed small gray-colored boxes representing a gene copy.

Table S8. BUSCO scores in homozygous and heterozygous diploids, and autopolyploids, and allopolyploids. White clover *Trifolium repens* has evolved as an allotetraploid from fused genomes of two related diploid species *occidentale* and *palescens*. The heterozygous RH potato is highly heterozygous, whereas the diploid tomato or the M6 potato are highly inbred.

Table S9. Transcriptome sequencing statistics. Pacbio IsoSeq libraries were constructed for three different tissues as indicated.

Table S10. Ribosomal gene clusters in the okra genome. Ribosomal gene clusters are characterized by unit configuration and number of tandemly arranged unit copies per superscaffold. Total lengths of clustered units that can be derived from the start and end position.

REFERENCES

- Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R. *et al.* (2014) Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing. *The Plant Journal*, **80**, 136–148.
- Agarwala, R., Barrett, T., Beck, J., Benson, D.A., Bollin, C., Bolton, E. *et al.* (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **44**(Database issue), D7–D19.
- Ali, S., Khan, M.A., Rasheed, H.S. & Iftikhar, Y. (2005) Management of yellow vein mosaic disease of okra through pesticide/bio-pesticide and suitable cultivars. *International Journal of Agriculture and Biology*, **7**, 145–147.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.
- Baird, D.M. (2017) Telomeres and genomic evolution. *Philosophical Transactions of the Royal Society B*, **373**, 20160473.
- Bao, W., Kojima, K.K. & Kohany, O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 1–6.
- Benchasri, S. (2012) Okra (*Abelmoschus esculentus* (L.) Moench) as a valuable vegetable of the world. *Ratarstvo i povrtarstvo*, **49**, 105–112.
- Bird, K.A., VanBuren, R., Puzey, J.R. & Edger, P.P. (2018) The causes and consequences of sub-genome dominance in hybrids and recent polyploids. *New Phytologist*, **220**, 87–93.
- Borodovsky, M. & Lomsadze, A. (2011) Eukaryotic gene prediction using GeneMark.Hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics*, **4**, 610.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.
- Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods*, **18**, 170–175.
- Choudhury, B. & Choomsai, M.L.A. (1970) Natural cross-pollination in some vegetable crops. *Indian Journal of Agricultural Sciences*, **40**, 805–812.
- Dankhar, S.K. & Koundinya, A.V.V. (2020) Accelerated breeding in Okra. In: Gosal, S.S. & Wani, S.H. (Eds.) *Accelerated plant breeding. Volume 2, vegetable crops*. Switzerland: Springer Nature Switzerland AG, pp. 337–354.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
- Dunwell, J.M. (2010) Haploids in flowering plants: origins and exploitation. *Plant Biotechnology Journal*, **8**, 377–424.
- Endrizzi, J.E. (1962) The diploid-like cytological behaviour of tetraploid cotton. *Evolution*, **16**, 325–329.
- Goffová, I. & Fajkus, J. (2021) The rDNA loci-intersections of replication, transcription, and repair pathways. *International Journal of Molecular Sciences*, **22**, 1302.
- Griffiths, A.G., Moraga, R., Tausen, M., Gupta, V., Bilton, T.P., Campbell, M.A. *et al.* (2019) Breaking free: the genomics of allopolyploidy-facilitated niche expansion in white clover. *Plant Cell*, **31**, 1466–1487.
- Guo, Z.-H., Ma, P.-F., Yang, G.-Q., Hu, J.-Y., Liu, Y.-L., Xia, E.H. *et al.* (2019) Genome sequence provides insights into the reticulate origin and unique traits of woody bamboos. *Molecular Plant*, **12**, 1353–1365.
- Hamon, S. & Van Sloten, D.H. (1995) Okra: *Abelmoschus esculentus*, *A. caillei*, *A. manihot*, *A. moschatus* (Malvaceae). In: Smartt, J. & Simmonds, N.W. (Eds.) *Evolution in crop plants*. Harlow: Longman, pp. 350–357.
- Hoff, K.J., Lomsadze, A., Borodovsky, M. & Stanke, M. (2019) Whole-genome annotation with BRAKER. In: Kollmar, M. (Ed.) *Gene Prediction. Methods in Molecular Biology*, Vol. 1962. New York, NY: Humana.
- Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y. *et al.* (2019) *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nature Genetics*, **51**, 739–748.
- Hubble, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W. *et al.* (2016) The Dfam database of repetitive DNA families. *Nucleic Acids Research*, **44**, D81.
- Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research*, **37**, D211–D215.
- Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
- Joshi, A.B. & Hardas, M.W. (1956) Allopolyploid nature of okra, *Abelmoschus esculentus* (L.) Moench. *Nature*, **178**, 1190.
- Kantama, L., Wijnker, E. & de Jong, H. (2017) Optimization of cell spreading and image quality for the study of chromosomes in plant tissues. *Methods in Molecular Biology*, **1669**, 141–158.
- Kokot, M., Dlugosz, M. & Deorowicz, S. (2017) KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, **33**, 2759–2761.
- Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A. *et al.* (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, **47**(D1), D807–D811.
- Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. (2013) Biology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics*, **4**, 237.
- Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biology*, **5**, 1–9.
- Kyriakidou, M., Anglin, N.L., Ellis, D., Tai, H.H. & Strömvik, M.V. (2020) Genome assembly of six polyploid potato genomes. *Scientific Data*, **7**, 88.
- Laetsch, D.R., Blaxter, M.L., Eren, A.M. & Leggett, R.M. (2017) BlobTools: interrogation of genome assemblies. *F1000Research*, **6**, 1287.

- Langley, C.H., Crepeau, M., Cardeno, M., Corbett-Detig, R. & Stevens, K. (2011) Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics*, **188**, 239–246.
- Lata, S., Yadav, R.K. & Tomar, B.S. (2021) Genomic tools to accelerate improvement in okra (*Abelmoschus esculentus*). In: Elkelish, A. (Ed.) *Landraces - Traditional Variety and Natural Breed*. London: IntechOpen, pp. 1–240. <https://doi.org/10.5772/intechopen.97005>
- Lei, Z., Zhou, C., Ji, X., Wei, G., Huang, Y., Yu, W. *et al.* (2018) Transcriptome analysis reveals genes involved in flavonoid biosynthesis and accumulation in *Dendrobium catenatum* from different locations. *Scientific Reports*, **8**, 6373.
- Leinonen, R., Sugawara, H. & Shumway, M. (2011) International nucleotide sequence database collaboration. The sequence read archive. *Nucleic Acids Research*, **39**, D19–D21.
- Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J. *et al.* (2015) Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnology*, **33**, 524–530.
- Li, J., Ye, G.-Y., Liu, H.-L. & Wang, Z.-H. (2020) Complete chloroplast genomes of three important species, *Abelmoschus moschatus*, *A. manihot* and *A. sagittifolius*: genome structures, mutational hotspots, comparative and phylogenetic analysis in *Malvaceae*. *PLoS One*, **15**, e0242591.
- Liu, I.M., Liou, S.S., Lan, T.W., Hsu, F.L. & Cheng, J.T. (2005) Myricetin as the active principle of *Abelmoschus moschatus* to lower plasma glucose in streptozotocin-induced diabetic rats. *Planta Medica*, **71**, 617–621.
- Mandáková, T. & Lysak, M.A. (2018) Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, **42**, 55–65.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **9**, 1297–1303.
- Merita, K., Kattakunnel, J.J., Yadav, S.R., Bhat, K.V. & Rao, S.R. (2012) Chromosome counts in wild and cultivated species of *Abelmoschus medikus* from the Indian sub-continent. *The Journal of Horticultural Science and Biotechnology*, **87**, 593–599.
- Mitidieri, J. & Vencovsky, R. (1974) *Rivista de Agricultura. Brazil*, **49**, 3–6.
- Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. (2007) KAA3: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, **35**, 182–185.
- Muimba-Kankolonga, A. (2018) Vegetable production. In: Demetre, C. (Ed.) *Food crop production by smallholder farmers in Southern Africa*. London: Academic Press, pp. 205–273.
- Naumova, T.N. (2008) Apomixis and amphimixis in flowering plants. *Cytology and Genetics*, **42**, 53–65.
- Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblížková, A. *et al.* (2020) Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature Plants*, **6**, 1325–1329.
- Ou, S., Chen, J. & Jiang, N. (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research*, **46**, e126.
- Peter, E.L., Nagendrappa, P.B., Ajayi, C.O. & Sesaazi, C.D. (2021) Total polyphenols and antihyperglycemic activity of aqueous fruits extract of *Abelmoschus esculentus*: modeling and optimization of extraction conditions. *PLoS One*, **16**, e0250405.
- Peters, S.A., Datema, E., Szinay, D., van Staveren, M.J., Schijlen, E.G.W.M., van Haarst, J.C. *et al.* (2009) *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *The Plant Journal*, **58**, 867–869.
- Portemer, V., Renne, C., Guillebaux, A. & Mercier, R. (2015) Large genetic screens for gynogenesis and androgenesis haploid inducers in *Arabidopsis thaliana* failed to identify mutants. *Frontiers in Plant Science*, **6**, 581–586.
- Purewal, S.S. & Randhawa, G.S. (1947) Studies in *Hibiscus esculentus* (Lady's finger). Chromosome and pollination studies. *The Indian Journal of Agricultural Sciences*, **17**, 129–136.
- Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. (2020) Genomescope 2.0 and smudgeplot for reference free profiling of polyploid genomes. *Nature Communications*, **11**, 1432.
- Renny-Byfield, S., Kovarik, A., Kelly, L.J., Macas, J., Novak, P., Chase, M.W. *et al.* (2013) Diploidization and genome size changes in allopolyploids is associated with differential dynamics of low and high-copy sequences. *The Plant Journal*, **74**, 829–839.
- Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, **21**, 245.
- Roach, M.J., Schmidt, S.A. & Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, **19**, 460.
- Salameh, N. (2014) Flow cytometric analysis of nuclear DNA of okra landraces (*Abelmoschus esculentus* L.). *American Journal of Agricultural and Biological Sciences*, **9**, 245–250.
- Siemonsma, J.S. (1982) West African okra - morphological and cytogenetical indications for the existence of a natural amphidiploid of *Abelmoschus esculentus* (L.) Moench and *A. manihot* (L.) Medikus. *Euphytica*, **31**, 241–252.
- Simaão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single copy-orthologs. *Bioinformatics*, **31**, 3210–3212.
- Smith, A., Hubley, R. & Green, P. (2013) *RepeatMasker Open-4.0* [2013–2015]. Available from: <http://www.repeatmasker.org> [Accessed 21st November 2018].
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. & Morgenstern, B. (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, **34**, W435–W439.
- Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.
- Takakura, K.-I. & Nishio, T. (2012) Safer DNA extraction from plant tissues using sucrose buffer and glass fiber filter. *Journal of Plant Research*, **125**, 805–807.
- The Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit tomato. *Nature*, **485**, 635–641.
- Venkataravanappa, V., Lakshminarayana Reddy, C.N. & Krishna Reddy, M. (2013) Begomovirus characterization, and development of phenotypic and DNA-based diagnostics for screening of okra genotype resistance against Bendi yellow vein mosaic virus. *3 Biotech*, **3**, 461–470.
- Weißborn, S. & Walther, D. (2017) Metabolic pathway assignment of plant genes based on phylogenetic profiling – a feasibility study. *Frontiers in Plant Science*, **8**, 1831.
- Wu, D.-T., Nie, X.-R., Li, H.-Y. *et al.* (2020) Phenolic compounds, antioxidant activities, and inhibitory effects on digestive enzymes of different cultivars of okra (*Abelmoschus esculentus*). *Molecules*, **25**, 1276.
- Wu, T.D. & Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.