Haplotype-resolved genome assembly of a male Dioscorea alata cultivar reveals the structure and evolution of sex chromosomes

Komivi Dossa

komivi.dossa@cirad.fr

CIRAD, UMR AGAP Institut

Xiaojun Su

Wuhan Benagen Technology Company Limited

Yedomon Ange Bovys Zoclanclounon

Rothamsted Research

Yao Chen

Wuhan Benagen Technology Company Limited

Muhammad Amjad Nawaz

National Research Tomsk State University

Peng Ni

Central South University

Ting Qiu

Wuhan Benagen Technology Company Limited

Gang Chen

Wuhan Benagen Technology Company Limited

Erick Malédon

CIRAD, UMR AGAP Institut

Marie-Claire Gravillon

CIRAD, UMR AGAP Institut

Christophe Perrot

CIRAD, UMR AGAP Institut

Elie Nudol

CIRAD, UMR AGAP Institut

Lévy Laurent

CIRAD, UMR AGAP Institut

Moussa Diouf

Université Cheikh Anta Diop

Petre I. Dobrev

Czech Academy of Sciences
Benjamin Karikari
University for Development Studies
Bertrand Fouks
UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro
Gemma Arnau
UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro
Pierre Mournet
UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro
Jianxin Wang
Central South University
Sylvain Santoni
UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro
Hana Chair
UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro

Research Article

Keywords: dioecy, root and tuber crops, structural variants, genetic gain, genome

Posted Date: September 9th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5627324/v2

Additional Declarations: The authors declare no competing interests.

Haplotype-resolved genome assembly of a male *Dioscorea alata* cultivar reveals the structure and evolution of sex chromosomes

Komivi Dossa[1,2,*], Xiaojun Su[3], Yedomon Ange Bovys Zoclanclounon[4], Yao Chen[3], Muhammad Amjad Nawaz[5,6], Peng Ni[7,8], Ting Qiu[3], Gang Chen[3], Erick Malédon[1,2], Marie-Claire Gravillon[1,2], Christophe Perrot[1,2], Elie Nudol[1,2], Lévy Laurent[1,2], Moussa Diouf[9], Petre I. Dobrev[10], Benjamin Karikari[11], Bertrand Fouks[2,12], Gemma Arnau[2,12], Pierre Mournet[2,12], Jianxin Wang[7,8], Sylvain Santoni[2], Hanâ Chair[2,12,*]

[1]CIRAD, UMR AGAP Institut, 97170 Petit Bourg, Guadeloupe, France
[2]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France
[3]Wuhan Benagen Technology Company Limited, 430070 Wuhan, Hubei, China
[4]Plant Sciences and the Bioeconomy, Rothamsted Research, Harpenden, AL5 2JQ, UK
[5]Advanced Engineering School (Agrobiotek), National Research Tomsk State University, Lenin Ave, 36, 634050, Tomsk, Tomsk Oblast, Russia
[6]Laboratory for Research and Application of Supercritical Fluid Technologies in Agro-Food Biotechnology, National Research Tomsk State University, Lenin Ave, 36, 634050, Tomsk, Tomsk Oblast, Russia
[7]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[8]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, China
[9]Département de Mathématiques et Informatique, Faculté des Sciences et Techniques, Université Cheikh Anta Diop, BP 5005 Dakar-Fann, 10700, Dakar, Senegal
[10]Institute of Experimental Botany, Czech Academy of Sciences, Laboratory of hormonal regulations in plants, Rozvojova 263, 165 02 Prague 6, Czech Republic
[11]Department of Agricultural Biotechnology, University for Development Studies, Nyankpala campus, Tamale, Ghana
[12]CIRAD, UMR AGAP Institut, F-34398 Montpellier, France
*Corresponding authors: Komivi Dossa, komivi.dossa@cirad.fr, Tel: +590590386161; Hanâ Chair, hana.chair@cirad.fr, Tel: +33(4)32722298.

Abstract:

Dioscorea alata L., the most widely cultivated yam species, exhibits dioecy (XY system) and a strong male-biased sex ratio. These two major constraints limit parental combinations and hinder breeding progress. To get insight into the sex chromosome structure and evolution in *D. alata*, we present a haplotype-resolved, near-complete genome assembly of the male cultivar 'Kabusa', spanning 958 Mb across 40 chromosomes. DalaChr6A is identified as the Y chromosome which exhibits early signs of heteromorphism, including a slight reduction in size (~400 Kb difference) and a potential centromere shift relative to the X chromosome. Our findings also reveal a sex chromosome turnover between *D. alata* and *D. rotundata*. The sex chromosomes of *D. alata* are evolutionarily young (~4.32 million year ago) and emerged after the divergence from the *D. rotundata* lineage. The sex-determining region (SDR) is refined to ~7.6 Mb, representing ~44% of the Y chromosome. It contains several inversions and a divergence gradient was observed across these inversions. INV4 identified as the oldest pericentric inversion likely marks the early step in *D. alata* SDR evolution. Despite structural divergence, both X and Y chromosomes remain transcriptionally active. Among the 231 genes

annotated in the SDR, 97 are sex-biased and enriched in functions associated with floral organ formation and hormonal signaling pathways. This study enhances our understanding of sex chromosome evolution and sex determination in dioecious plants. It provides a gold-standard reference genome for the *Dioscorea* genus and lays the foundation for accelerated breeding and genetic improvement in yam.

Key words: dioecy, root and tuber crops, structural variants, genetic gain, genome.

Main

Dioecy, the condition of separate male and female individuals within a species is present in only 5–6% of flowering plants species[1]. Dioecy poses specific challenges for plant breeding, particularly when sex expression is stable and strictly unisexual, such as the need for controlled pollination and limited access to one sex within breeding populations. However, it also offers unique opportunities, including the development of hybrid breeding strategies, exploitation of heterosis, and exploration of sex-determining mechanisms[2]. Recent advances have provided insights into the genetic and molecular mechanisms underlying sex determination across a variety of plant species. In crops such as papaya[3], kiwifruit[4,5], and date palm[6], a two-gene model of sex determination has been reported. This model involves two genetically linked genes located within a non-recombining sex-determining region (SDR), where one gene promotes the development of one sexual function and the other suppresses the opposite function[4]. In kiwifruit, the gene *FRIENDLY BOY* (FrBy) has been identified as a crucial male-promoting factor, while another gene, *SHY GIRL* (SyGI), acts as a suppressor of female organ development[4,7]. The interplay between these two genes is essential for determining the sexual phenotype of the plant. Their presence within the non-recombining SDR prevents recombination, thus preserving their functional relationship across generations[4,6]. However, in persimmon[8] and poplar[9], a one-gene model appears to control sex differentiation, pointing to diverse evolutionary pathways. In many species, dioecy is leaky or labile, allowing for a degree of reproductive flexibility that can sometimes be leveraged in breeding programs[2].

To date, more than 55 dioecious species genomes have been sequenced, with sex chromosomes identified in over 35 species[10,11]. However, in most cases, only one sex type has been sequenced, typically the homogametic sex, resulting in incomplete coverage of sex chromosomes in genomic data[11,12]. A near-complete or telomere-to-telomere (T2T) genome assembly is essential for addressing key biological questions covering genomics-assisted

breeding, the analysis of repetitive elements, genome organization and structure, evolutionary studies, gene discovery, and sex determination, among other areas of research[13,14,15,16]. For instance, in maize, the T2T genome assembly has provided insights into the structural variations (SV) within centromeric regions that were overlooked in previous assemblies[17]. This has led to a better understanding of how these variations influence traits like kernel size and yield, showcasing the potential for improving crop performance through informed breeding strategies[17]. Moreover, genome assemblies of diploid organisms have collapsed the two homologous chromosomes into a single mosaic reference genome. While this approach simplifies the assembly process, it inherently obscures important genetic differences between the homologous chromosomes, which can be particularly significant in plant species with high heterozygosity[18] and/or sexual chromosomes[19]. This loss of allelic information limits the understanding of genetic variation and can affect downstream analyses, such as identifying SV, allele-specific gene expression, sex related genes in plant exhibiting dioecy.

Yams (*Dioscorea* spp.) are economically important tuber crops in tropical regions, with approximately 90% of global production concentrated in West Africa[20]. *Dioscorea alata*, commonly known as water yam or greater yam, is one of the most widely cultivated yam species due to its high productivity, resilience to environmental stresses, and rich nutritional profile[21]. *Dioscorea alata* is reported to be strictly dioecious and mixed-cytotype, with chromosome numbers of 2n = 40, 60, or 80[22]. Its haploid genome size is estimated to be ~454 Mb[22]. Breeding efforts to enhance the yield, disease resistance, and overall tuber quality of *D. alata* have been significantly hampered by its dioecious nature and sex ratio distortion towards male cultivars. This limits the availability of parental combinations for hybridization, hinders the development of improved cultivars and slows genetic gains[23]. To overcome these challenges, developing methods to control and manipulate sex phenotype in *D. alata* could facilitate breeding, thereby allowing for more effective selection and improvement of desirable traits. Recent genetic studies have identified Chromosome 6 as the sex chromosome in *D. alata*, supporting an XY (male heterogametic) sex determination system[24,25,26]. Despite these advances, the development of a reliable sex-specific marker, as well as the elucidation of the molecular mechanisms and identification of key genes governing sex determination in *D. alata*, remain elusive.

The available reference genomes for greater yam is a collapsed assembly of two female (XX) individuals[27,28]. The lack of a high-quality Y chromosome sequence restricts comprehensive sex chromosome analysis in this species. To get insight into the sex chromosome structure and evolution in greater yam, we employed a combination of long-reads and chromosome conformation sequencing technologies to assemble a haplotype-resolved,

near-complete genome of a male (XY) individual. We also conducted in-depth investigation of the SDR and analyzed the expression profiles of sex-linked genes underpinning sex phenotype expression in *D. alata*. The genome revealed a high degree of similarity between X and Y chromosomes with a 7.6-Mb SDR representing a major pericentric inversion.

Results

Genome assembly for the greater yam male cultivar 'Kabusa'

To investigate the genetic mechanisms underlying sex determination in *Dioscorea alata*, we constructed a chromosome-level, haplotype-resolved, near-complete assembly of the male cultivar 'Kabusa', allowing us to generate sequences of both the X and Y sex chromosomes. 'Kabusa' is the most popular cultivar in West Indies with a round-shaped tuber, moderate resistance to anthracnose disease and good tuber quality[29] (Fig 1a). The haploid genome size of 'Kabusa', a diploid cultivar (2n = 40; Supplementary Fig. 1), was estimated to be ~494.13 Mb based on 19-mers analysis, with an unsurprising high heterozygosity level (1.46%) (Supplementary Tables 1,2; Supplementary Fig. 2). We generated 126.5× sequence coverage of raw ultra-long Oxford Nanopore Technology (ONT) data, 55.1× coverage of PacBio HiFi data and 123.8× coverage of chromatin conformation capture (Hi-C) data for assembling the 'Kabusa' genome (Supplementary Tables 3,4,5). Sequencing depths were calculated based on the estimated haploid genome size (1C).

Four assembly strategies including Hifiasm (HiFi)[30], Hifiasm (HiFi+ONT), Hifiasm (HiFi+Hi-C), and Nextdenovo (ONT)[31] were evaluated to identify the most effective preliminary backbone assembly to prioritize phasing accuracy. The combination of HiFi and Hi-C reads assembled using Hifiasm produced the most robust foundation for the genome assembly with contig N50 of 12.23 Mb and 17.46 Mb for HaplotypeA and HaplotypeB, respectively (Supplementary Table 6). It is worth noting that, although the Hifiasm (HiFi+ONT) assembly demonstrated good contiguity metrics, the relatively low base-level accuracy of ONT reads may compromise phasing precision. Incorporating Hi-C data in the assembly process theoretically enhances the accuracy of haplotype phasing results and enables chromosome scaffolding. After manual adjustment, a total of 958.71 Mb of contig sequences (contig N50 = 23.83 Mb for HaplotypeA and contig N50 = 24 Mb for HaplotypeB) were mapped to 40 chromosomes in the hybrid assembly version (Supplementary Table 7; Figs 1b, c). The assembly was iteratively polished by HiFi and Illumina data. HiFi reads were aligned to the genome assembly for telomere sequence extension and repair. The telomere sequence

identification results, based on the repetitive units CCCTAAA for the 5' end and TTTAGGG for the 3' end, are shown in Supplementary Table 8 and Supplementary Fig. 4. Gap filling was performed using ONT and HiFi contigs as well as raw reads. After gap filling, we assessed the reliability of the filled regions by aligning both preliminary alternative assemblies and ONT reads to the gap-filled genome. A gap was considered reliably filled if either data type provided continuous coverage spanning both flanking regions (Supplementary Fig. 5). In total, out of the 54 gaps detected in the assembly, 51 were successfully filled (Supplementary Table 9). Only DalaChr4A, DalaChr4B and DalaChr13B contained single gaps and 37 out of the 40 chromosomes were covered by single contigs (Supplementary Table 10). A total of 99.88% of the contig sequences was anchored on the chromosomes. Altogether, statistics shown in Table 1 confirmed that we generated a highly accurate and contiguous chromosome-level haplotype-resolved assembly for 'Kabusa'. The genome assembly is available via Yam Genome Hub (https://yam-genome-hub.southgreen.fr/).

Figure 1. Chromosome-level haplotype-resolved near-complete assembly of the male *D. alata* cultivar 'Kabusa'. (**a**) Pictures of 'Kabusa' cultivar displaying the aerial biomass in the field, the tuber shape and flesh color, and the male flowers. (**b**) Genome-wide Hi-C contact heatmap, visualized with Juicebox[32]. (**c**) Circle diagram of basic genome information. Concentric circles from outside to inside represent chromosomes, gene density, repeat density, and GC density calculated in a 50 Kb window. Lines in the interior indicate collinearity. The scale bar = 10 cm. (**d**) Syntenic mapping between the two haplotypes of the 'Kabusa' genome using SyRI[33]. Collinear regions between HaplotypeA and HaplotypeB are indicated with gray lines, while inversion, translocation, and duplication regions are indicated with orange, green, and blue lines, respectively.

Annotation of the 'Kabusa' genome and analysis of sequence variation between haplotypes

The annotation of the two haplotype genome assemblies revealed remarkably similar features. About 63% of each haplotype genome sequences was annotated as repetitive elements (Supplementary Table 11). LTR retrotransposons were identified as the most abundant sequences (43% and 44%), of which Gypsy (20% and 21%) was the top superfamily in HaplotypeA and HaplotypeB, respectively. Moreover, 8% of each haplotype genome sequence was made of tandem repeat sequences (Supplementary Table 11).

ONT-based full-length transcriptome data generated from eight tissues collected at developmental stages were used for genome annotation (Supplementary Table 12). We employed an automated annotation pipeline supported by ONT-based full-length transcriptome data, *ab initio* gene prediction, and protein homology, successfully annotating 27,767 and

27,609 protein-coding genes in HaplotypeA and HaplotypeB, respectively (Table 1; Supplementary Table 13). In total, 21,652 allelic gene pairs were identified between the two haplotypes. Functional annotations were assigned to more than 94% of genes in each haplotype using multiple reference databases.

BUSCO assessment of gene annotation completeness demonstrated 97.5% and 97.8% completeness in the two haplotypes, confirming the high quality of the gene annotations. (Supplementary Table 14). In addition, 139 miRNAs, 376 tRNAs, 2642 rRNAs, and 8609 snRNAs (HaplotypeA) and 140 miRNAs, 374 tRNAs, 3521 rRNAs, and 8738 snRNAs (HaplotypeB) were annotated (Supplementary Table 13). We also performed putative centromere detection by resolving concentration of transposable elements (TE) and genes along all the 40 chromosomes, with size ranging from 122,139 to 5,871,010 bp (Supplementary Table 15; Supplementary Figs. 4,6). For example, the homologous copies of the chromosomes DalaChr6 and DalaChr20 exhibit significant differences in putative centromere position. To enhance the gene annotation of the Dala_V2 female genome assembly[27], we projected gene models from the high-quality haplotype-resolved assembly of 'Kabusa' (HaplotypeA) onto the Dala_V2 assembly using Liftoff v1.6.3[34] with minimap2 v2.24-r1122[35]. A total of 27,240 genes were successfully mapped to Dala_V2, with only 527 genes failing to align, likely representing Kabusa-specific sequences. Among the mapped genes, 99.9% were placed on chromosomal sequences. The updated annotation (Dala_V3, https://zenodo.org/records/15700131) represents an increase of 2051 genes compared to the previous version (Dala_V2, 25,189 genes), highlighting the improved completeness and accuracy of gene discovery using the 'Kabusa' HaplotypeA as a reference (Supplementary Table 16).

We compared the two haplotype assemblies to highlight their sequence variation. A total of 4.2 M single nucleotide polymorphisms (SNPs) including 1.17 M genic SNPs and 741,026 small insertions and deletions (indels), of which half located in genic regions, were identified between the two haplotypes (Supplementary Table 17). In addition, 49,694 structural variants (SVs) were found, including 5075 interchromosomal translocations, 202 inversions, 24,563 duplications, 9824 large insertions and 10,030 large deletions (Fig 1d). Nearly half of these SVs were located in genic regions (Supplementary Fig. 7A). Functional enrichment analysis of SV-influenced genes revealed a wide range of biological pathways, including nucleic acid binding, DNA integration, defense responses, post-embryonic development, regulation of cell shape, and chloroplast organization (Supplementary Fig. 7B)

Table 1. Summary statistics of genome assembly and gene annotation of 'Kabusa'

Assembly	'Kabusa' HaplotypeA	'Kabusa' HaplotypeB	Combined
Number of chromosomes	20	20	40
Number of contigs	21	22	
Contig N50 (bp)	23,828,568	24,039,800	
Number of gaps	1	2	
Assembly length (bp)	479,656,039	480,287,338	958,714,780
Illumina read-mapping rate (%)	99.73	99.72	
Unanchored scalffods	5	5	10
HiFi read-mapping rate (%)	100	100	
ONT read-mapping rate (%)	99.65	99.62	
Genome assembly BUSCO (%)	98.1	97.9	
Genome assembly QV	53.1006	52.3324	
Annotated genes	27,767	27,609	
Genome annotation BUSCO (%)	97.5	97.8	
Mean mRNA_length (bp)	5,029.52	5,077.03	
Mean cds_length of per gene (bp)	1,117.66	1,117.43	
Mean exon_number of per gene	5.27	5.28	
Mean exon_length (bp)	302.35	300.83	
Mean intron_length (bp)	801.46	811.85	

Detecting 'Kabusa' Y chromosome and sex chromosome evolution in yams

The haplotype-resolved genome of 'Kabusa' offers the opportunity to determine the Y sequence that is absent in the previous female genomes[27,28]. *Dioscorea alata* Chromosome 6 has been identified as the sex chromosome via multiple studies[24,25,26]. Although discrepancies exist in the precise location of the sex-determination region (SDR), evidence points to the presence of large SVs on the sex chromosome. The assembled DalaChr6A was slightly smaller than DalaChr6B (~400 Kb difference) and a large pericentric inversion (from 1,619,245 bp to 9,202,430 bp) was found between them (Fig 1d). A total of 1087 and 1102 genes were annotated in DalaChr6A and DalaChr6B, respectively, with 840 shared genes. Aside from the single pericentric inversion, the two chromosomes exhibit strong collinearity in their distal regions, which likely correspond to the pseudoautosomal regions (PARs).

We compared the sequences of DalaChr6A and DalaChr6B to Chromosome 6 of the female genome Dala_V2 (DalaChr6F)[27] and found that DalaChr6A was the least collinear to DalaChr6F (Supplementary Fig. 8). Remarkably, the region from 1,619,245 bp to 9,202,430 bp on DalaChr6A displayed high variation with DalaChr6F. Moreover, intergenomic comparisons revealed a pericentric inversion between DalaChr6A and DalaChr6F, which was absent between DalaChr6B and DalaChr6F (Supplementary Fig. 9). We confirmed the existence of

this inversion with HiFi long reads spanning the two breakpoints of the inverted regions (Supplementary Fig. 10). We therefore concluded that HaplotypeA genome in 'Kabusa' contains the Y chromosome (DalaChr6A) and the pericentric inversion represents the SDR.

We aligned the 'Kabusa' assemblies to the sequenced genomes of related *Dioscorea* species, focusing on chromosomes previously reported to be associated with sex determination (Supplementary Fig. 11). *Dioscorea alata* and *D. rotundata*, both from section Enantiophyllum[23], diverged approximately 17.4 million years ago (Mya). We observed strong collinearity between DalaChr6 and DrotChr6. Notably, the sex chromosome of *D. rotundata* has been identified as DrotChr11[36], suggesting a sex chromosome turnover between these two species, which also differ in sex determination systems (XY vs. ZW).

In more distantly related species of section Stenophora (2n = 20), such as *D. tokoro* and *D. zingiberensis*, the SDRs have not yet been fully characterized. However, previous work suggests that DtokChr3 harbors sex-determining loci in *D. tokoro*[37]. In our analysis, part of DtokChr3 aligns with DalaChr6 (Supplementary Fig. 11), and a similar pattern is observed between DzinChr2 and DalaChr6. While this may reflect conservation of ancestral chromosomal segments, we cannot conclude that the SDR itself is shared. Moreover, differences in ploidy could influence genome architecture and synteny. With an increasing number of yam genome sequences becoming available[38,39], further studies are needed to identify the SDRs in these species and to determine whether the observed synteny reflects functional conservation or independent evolutionary trajectories.

Structure of the sex-determination region in *D. alata*

To gain insights into the structure of the SDR, we generated Hi-C contact map for the X and Y chromosomes and further conducted a detailed comparison of the X and Y haplotypes of *D. alata* 'Kabusa' using SyRI[33] (Fig. 2a,b). We could observe four different inversions: INV1- INV4 (Table 2). INV1 and INV4, located at the 5′ and 3′ boundaries respectively, were also detected in comparisons between Y chromosome and two independently assembled X chromosomes from female genomes (Dala_V2[27] and Ziyushenshu[28]). INV2 was also detected in the Y vs. Ziyushenshu comparison, but not in Dala_V2, likely due to the lower resolution of Dala_V2 genome. These observations support a model in which inversions have structured the SDR during the evolution of sex chromosomes in *D. alata*. Between the four inversions, blocks enriched in repeated sequences were observed. These regions are non-collinear with the X chromosome assemblies and may represent putative male-specific region of the Y chromosome (MSY), contributing to the structural differentiation of the Y chromosome. Reciprocally, the

corresponding regions on the X chromosome, which lack homology with the Y, may define X-specific regions (XSR). Interestingly, we observed that the SDR is approximately 400 Kb shorter than the homologous X region (HXR). While such size differences can occur in phased assemblies, the reduction here corresponds to specific deletions and rearrangements detected in the Y chromosome, supporting the hypothesis of a deletion event during sex chromosome differentiation (Table 2).

To further delineate the SDR in the *D. alata* genome, we employed five complementary analytical approaches. First, we constructed female and male read pools using resequencing data from 37 *D. alata* accessions (17 females and 20 males) from Mota et al.[40]. Reads were aligned to the HaplotypeA genome of 'Kabusa' to extract variants. This yielded 15.5 million SNPs, from which we calculated the fixation index (*Fst*) between the two sex groups. The highest *Fst* values were observed between 2 and 8 Mb on DalaChr6A (Fig. 2c). Second, analysis of male-specific SNPs, defined as variants homozygous in the female pool and heterozygous in the male pool, identified 4,948 such SNPs. Approximately one-quarter of these variants were located between 2.6 and 9.1 Mb on DalaChr6A (Fig. 2c). Third, we conducted genome-wide association studies (GWAS) using a high-density variant set (>2 M SNPs and 658,000 InDels) from an independent population comprising 68 males and 40 females (Figs. 2d,e). Sex phenotypes were scored over two years at two sites in Guadeloupe (West Indies). GWAS confirmed DalaChr6 as the sole sex chromosome, with 2152 significant SNPs (1,589,777–9,195,813 bp) and 383 significant InDels (1,638,603–9,195,052 bp) associated with sex ($p <$ 1e-6; Supplementary Table 18). Notably, most sex-associated SNPs were located within the four inversions identified between Y and X (Figs. 2f,g). Previously reported sex-linked markers[25,25,26] also localized within INV1 and INV4. The colocalization of INVs with sex-associated markers suggests that these structural rearrangements have played a central role in the evolution and differentiation of the SDR in *D. alata*. Fourth, we examined differential read coverage (10 Kb window) between male and female genotypes along the Y chromosome (Fig. 2h). Several regions within the SDR displayed markedly higher coverage in males, corresponding to the boundaries of the putative MSY. These regions exhibited low variant density and contained few significant GWAS markers, suggesting suppressed recombination and high sequence divergence. Finally, we analyzed male-specific 27-mers (> 8.2 M) along the Y chromosome, which showed a distribution pattern that precisely mirrored the male-biased coverage (Fig. 2i), further supporting the presence of Y-specific sequences.

The SDR contains 231 annotated genes, while the HXR includes 248 genes (Supplementary Table 19). Upstream, the SDR is delimited by *DalaChr6AG081680*, encoding a 3-isopropylmalate dehydrogenase, and downstream by *DalaChr6AG083980*, encoding a transmembrane protein 50. Only 108 genes had homologous counterparts in the HXR and were considered putative gametologues. An additional five genes within the SDR showed homology with PAR-linked genes in the X chromosome. Conversely, 118 genes within the SDR lacked identifiable X-linked homologues and are considered Y-specific while 140 genes are considered X-specific. We identified 43 and 41 genes on the Y and X chromosomes, respectively, that have functional homologues on autosomal chromosomes in the *D. alata* 'Kabusa' genome, including DalaChr18, DalaChr19, DalaChr2, DalaChr7, and DalaChr16A (Supplementary Table 20). Notably, only four autosomal genes (from DalaChr19) have homologues located within INV1 of the SDR and HXR, while another four genes (from DalaChr16A) have homologues in the XSR. The majority of these syntenic gene pairs are found within the PARs of both chromosomes, suggesting that these relationships likely predate sex chromosome differentiation.

Together, these lines of evidence point to a region of pronounced structural and sequence divergence between the X and Y chromosomes, consistent with an evolving non-recombining SDR (Fig. 2j).

Table 2: Structure and gene content of the SDR and HXR in the 'Kabusa' genome

Structure	**SDR (DalaChr6A)**		**HXR (DalaChr6B)**	
	Location/ size (Mb)	Gene content	Location: size (Mb)	Gene content
INV1	1.6-3.8/2.2	144	7.5-9.6/2.1	141
INV2	6.15-6.3/0.15	15	4.6-4.75/0.15	11
INV3	6.65-7.1/0.45	3	6.55-7/0.45	9

INV4	8.86-9.2/0.34	15	1.6-1.79: 0.19	9
MSY/XSR	-/4.46	54	-/5.11	78
Total	1.6-9.2/7.6	231	1.6-9.6/8	248

Figure 2. Structure of the sex determination region (SDR) in *D. alata*. (**a**) Hi-C contact heatmap between DalaChr6A (Y) and DalaChr6B (X), visualized with Juicebox [31]. (**b**) Syntenic mapping between 'Kabusa' Y chromosome and X chromosomes from 'Kabusa', Dala_V2[27] and Ziyushenshu[28] using SyRI[33]. Four inversions (INV1-INV4) were detected. (**c**) The genomic landscape of genetic variants between male and female pools. The outer track represents the 20 chromosomes (chr) of HaplotypeA in the 'Kabusa' genome. The inner tracks from outside to inside represent *Fst* between females and males and male-specific SNP density, respectively (window size 100 Kb). (**d**) Manhattan plot for SNP-sex phenotype genome-wide association study (GWAS), with the peaks indicating significant GWAS signals, and the black and red dotted horizontal lines indicating the suggestive and genome-wide significance thresholds, respectively. (**e**) Manhattan plot for InDel-sex phenotype GWAS, with the peaks indicating significant GWAS signals, and the black and red dotted horizontal lines indicating the suggestive and genome-wide significance thresholds, respectively. Regional Manhattan plots on DalaChr6A showing the dispersion of the significant SNPs (**f**) and Indels (**g**). (**h**) Differential read coverage (10 Kb window) between male and female groups along the Y chromosome. (**i**) Exact matches of male specific 27-mers along the Y chromosome. (**j**) Representation of the X (DalaChr6B) and Y (DalaChr6A) chromosomes along with the gene density (heatmap), centromere and telomere locations, pseudoautosomal regions (PAR), and the SDR containing the 4 inversions (INV1-INV4) detected.

Evolutionary divergence of the sex chromosomes in *D. alata*

We computed synonymous site divergence (*Ks*) between X- and Y-linked allelic gene pairs. In total, we analyzed 87 gametologues within the SDR and 272 one-to-one homologues genes in the PARs. After removing outliers, *Ks* values varied across the SDR–HXR segment of the sex chromosomes (range: 0.00486–0.137; mean = 0.0345; s.d. = 0.0259) and were significantly higher than those in the PARs (range: 0.00205–0.127; mean = 0.0292; s.d. = 0.0245; Kruskal–Wallis $p = 0.0019$) (Fig. 3a), consistent with ongoing divergence between the SDR and the HXR. Notably, the gene *DalaChr6AG083010*, located within INV1, showed the highest *Ks* value and encodes a PLATZ transcription factor previously reported as essential for female flower morphology determination in grapevine[41].

To gain further insight into the evolution of the SDR, we examined *Ks* distributions within the three chromosomal inversions (INV1, INV2, and INV4) that distinguish the Kabusa Y chromosome from the three assembled X chromosomes (Fig. 3b). We analyzed 74, 7, and 6 gametologues within INV1, INV2, and INV4, respectively. As expected, INV4 (range: 0.0202–0.0724; mean = 0.0438; s.d. = 0.0234) and INV1 (range: 0.00535–0.137; mean = 0.0367; s.d. = 0.0275) exhibited significantly higher *Ks* values than INV2 (range: 0.00486–0.0274; mean = 0.0145; s.d. = 0.00743; Kruskal–Wallis $p = 0.0072$). These results suggest that INV1 and INV4 represent ancestral chromosomal rearrangements associated with the origin of the SDR.

Using a synonymous substitution rate of $r = 4 \times 10^{-9}$ substitutions per site per year previously estimated for *Dioscorea*[39], we estimated that the divergence between the SDR and HXR in *D. alata* occurred approximately 4.32 Mya. The estimated divergence times of the three inversions suggest that INV4 (~5.48 Mya), followed by INV1 (~4.58 Mya), predate the SDR. These two inversions may therefore have played a role in the initial expansion of the SDR and the evolution of the sex chromosomes, similarly as reported in *Populus euphratica*[42]. In contrast, INV2 exhibits a considerably more recent divergence (~1.81 Mya). Altogether, these findings suggest that the sex chromosomes of *D. alata* are evolutionarily young, with their emergence occurring well after the split from the *D. rotundata* lineage (~17.4 Mya). The data reveal a gradient of divergence across the inverted regions, with INV4 emerging as the oldest pericentric inversion likely involved in the early steps of SDR evolution.

TEs can influence the dynamics of the SDR by promoting recombination suppression under selective pressure, thereby shaping the organization and structural evolution of sex chromosomes[43,44]. We further analyzed TE abundance and distribution pattern along chromosomes in HaplotypeA and HaplotypeB with an emphasis on LTR/Gypsy and LTR/Copia as the most abundant TEs (Supplementary Fig. 12). Detailed analysis of LTR/Gypsy and LTR/Copia densities across X and Y in ‘Kabusa’ revealed region-specific accumulation patterns. Significant differences in both LTR/Gypsy and LTR/Copia densities were observed when comparing the SDR and HXR to their respective flanking PARs (Figs. 3c, d), indicating a localized enrichment of LTR/Gypsy in sex-linked regions. The comparison between the SDR and HXR did not yield a statistically significant difference in TE density (Fig. 3e), suggesting that the overall levels of TE accumulation in the X and Y are broadly similar. However, smoothed density plots along X and Y highlighted marked peaks of LTR/Gypsy accumulation in regions that correspond to non-colinear segments between haplotypes (Fig. 3f). These regions coincide with the putative MSY, as well as XSR. This finding suggests that both the X and Y

chromosomes harbor structurally TE-rich segments as reservoir of sex-specific divergence. Such patterns align with observations in other dioecious plant species such as papaya[45] or *Silene latifolia*[46] where sex chromosome differentiation is associated with the accumulation/specific localization of LTR elements and chromosomal rearrangements. Although global TE density may not differ significantly between the SDR and HXR in 'Kabusa', the presence of localized LTR/Gypsy-rich regions likely contributes to the disruption of collinearity and may play a structural role in X–Y divergence. In conclusion, our findings support a model where localized accumulation of LTR/Gypsy elements contributes to sex chromosome differentiation in *D. alata*.

Figure 3: Sex chromosome divergence in *D. alata*. (**a**) Comparison of the synonymous substitution rates (*Ks*) of genes between SDR and PAR on the Y chromosome. (**b**) *Ks* comparison between Y chromosome components. ** = significant difference at $p < 0.001$. (**c**) Comparison of the LTR/Copia and LTR/Gypsy densities between PAR and SDR in the Y chromosome. (**d**) Comparison of the LTR/Copia and LTR/Gypsy densities between PAR and HXR in the X chromosome. (**e**) Comparison of the LTR/Copia and LTR/Gypsy densities between SDR and HXR. (**f**) Smoothed density plots along X and Y chromosomes. Densities are calculated per 1 Mb overlapping windows.

Sex-biased gene expression along sex chromosomes and SDR in *D. alata*

In *D. alata*, ontogenetic differences between female and male flowers become morphologically evident between Day 8 and Day 12 after blooming (Fig. 4a), indicating that key sex-determination processes likely occur during this window. To capture relevant transcriptional activity during the differentiation of sexual organs, we collected flower buds at Day 5, when a clear developmental divergence is not yet observed. Transcriptome sequencing was conducted on six female and six male individuals (Supplementary Table 21). Libraries of poor quality were excluded from further analyses (Supplementary Table 22). More than 918 million 150-bp paired-end clean reads were mapped to the haploid consensus genome (HapA + HXR) of Kabusa. While this approach enables consistent gene expression analysis, we acknowledge that some degree of cross-mapping between homologous X and Y regions may occur, potentially introducing minor biases in quantification. These limitations are inherent to mapping in sex chromosome systems with closely related haplotypes. Transcript quantification and differential expression analysis were performed to detect the sex-biased genes (SBG). Male and female groups were clearly separated at the transcriptional level with globally higher gene expression levels within the male group (Fig 4b; Supplementary Fig. 13).

Globally, male-biased genes (4905) were twice as abundant as female-biased genes (2012) (Fig. 4c), a trend frequently reported in dioecious plant species[47]. Gene ontology enrichment analysis of the SBGs revealed several biological pathways including photosynthesis, cellular morphogenesis, pollen wall assembly, oxidoreductase activity, and carbohydrate metabolism. While not all enriched categories are directly involved in sex determination, they may reflect sex-specific physiological processes or developmental demands during floral organ formation in *D. alata* (Supplementary Fig. 14). The most strongly-induced SBGs were located within or in close proximity to the SDR, further supporting the central role of this region in sex differentiation (Fig. 4c). A total of 751 genes on the Y chromosome (69% of total annotated Y genes) and 762 on the X chromosome (69% of total annotated X genes) were transcriptionally active in the sampled tissue types, indicating that both sex chromosomes retain substantial expression capacity and that Y chromosome of *D. alata* has not yet undergone widespread loss of gene expression typically associated with advanced stages of sex chromosome differentiation. A detailed examination of gene expression patterns along the sex chromosomes revealed a clear concentration of male-biased expression along the Y chromosome (Fig. 4d). In total, 230 SBGs were identified on the Y chromosome, compared to 154 SBGs on the X chromosome. As illustrated in Fig. 4e, SBG located in the Y chromosome exhibited significantly higher $\log_2$ fold changes in expression between male and female groups, compared to their counterparts on the X chromosome (p = 4.994e–08). Along the sex chromosomes, the magnitude of differential expression was more pronounced within the SDR and HXR, highlighting a potential regulatory role of these regions in sex-specific transcriptional activity (Fig. 4d). Within the SDR, 97 SBGs were detected (Supplementary Table 23), with all but one—*DalaChr6AG081990*, a *GERMIN-LIKE PROTEIN SUBFAMILY 3*, showing higher expression in male flowers. This supports the idea that the SDR is a transcriptional hub for male-specific functions. Notably, genes located in the MSY displayed markedly stronger expression in males, as would be expected for loci involved in male reproductive development and function (Fig. 4f). In contrast, the HXR contained 45 SBGs, but only a single gene *DalaChr6BG081880*, annotated as *Cytochrome P450* was upregulated in male samples (Supplementary Table 23). Interestingly, these X- and Y- linked SBGs identified in floral tissues also exhibited differential expression in mature tubers (three males *D. alata* accessions compared to three female *D. alata* accessions), suggesting that sex-associated transcriptional divergence may extend beyond reproductive tissues in *D. alata*[42,48] (Fig. 4g).

Notably, we identified 10 genes within the SDR that were only expressed in males (Supplementary Table 23). These genes may contribute to sex differentiation by promoting

male organ development and/or repressing female organ development. For instance, *DalaChr6AG082390*, annotated as a style cell-cycle inhibitor 1-A (SCI1) homolog, showed strong expression in all male flower samples but was undetectable in female flowers. In *Nicotiana tabacum* and *Arabidopsis thaliana*, SCI1 has been shown to control cell proliferation in pistil development[49,50]. While its function in male reproductive tissues has not been reported in these model species, the male-specific expression pattern in *D. alata* raises the possibility of a lineage-specific functional divergence or a novel regulatory role in male floral development. However, functional validation is required to confirm its role in sex differentiation in yam. Several other SBGs within the SDR have known functions in flower sexual organ development, including *DalaChr6AG081680* (3-isopropylmalate dehydrogenase) essential for pollen development and proper embryo sac development[51]; *DalaChr6AG082570* (SWR1) involved in female germline development[52]; *DalaChr6AG083470* (NYFC) known for its role in flowering and organ morphogenesis[53]; and *DalaChr6AG082810* (ARR) involved in cytokinin (CK) hormone signaling[54]. Since INV4 is the most ancestral inversion associated with sex chromosome evolution in *D. alata*, we hypothesize that SBGs located within this region may represent strong candidates for primary sex-determining factors. Among them, four have homologs in model species with functions related to reproductive development. *DalaChr6AG083890* encodes a rhomboid-like protein, homologous to *Arabidopsis thaliana* genes involved in pollen development and pollen wall formation[55]. *DalaChr6AG083930* encodes a PHD-finger protein with strong similarity to genes essential for male fertility in *A. thaliana*, soybean, and rice[56–58]. *DalaChr6AG083880* is a homolog of CORONATINE INSENSITIVE 1 (COI1), an F-box protein known to play a central role in jasmonic acid signaling and reproductive organ development across multiple plant species[59–60]. Finally, *DalaChr6AG083840* is similar to the MATERNAL EFFECT EMBRYO ARREST 9 gene in *A. thaliana*, which is involved in female gametophyte development and function[61]. While functional validation remains to be conducted in *D. alata*, the presence of these homologs highlights candidate genes potentially involved in sex determination and reproductive processes in this species.

To explore whether epigenetic regulation may contribute to the observed biased gene expression, we analyzed DNA methylation levels of 5-methylcytosine (5mC) at CpG sites. Globally, quite similar patterns of methylation levels and distribution across homologous chromosomes of HaplotypeA and HaplotypeB could be observed (Supplementary Fig. 15, Fig. 4h). However, a finer-scale comparison of X and Y chromosomes revealed a significantly higher methylation level within the SDR and HXR compared to the PARs (Figs. 4i,j). No

significant difference was observed between the SDR and HXR, suggesting that the epigenetic landscape of this region is largely conserved despite major structural rearrangements (Fig. 4k). It is important to note that our analysis focused solely on CpG methylation and did not account for other types of DNA methylation (e.g., CHG or CHH contexts), which may also play a role in gene regulation[62].

Figure 4: Sex-biased gene expression in *D. alata*. (**a**) Development of *D. alata* flower at Day5, Day8 and Day12 after blooming in male 'Kabusa' and female '74F'. Ontogenetic differences between female and male flowers become morphologically evident between Day8 and Day12 after blooming. Scale = 1 cm. (**b**) Principal component analysis based on the gene expression levels for flower samples (Day5) expressed as fragments per kilobase of transcript per million fragments mapped in the male and female groups. The first and the second principal components explaining 55% and 20% of the total variation, respectively. (**c**) Volcano plot displaying the differential expression analysis in flowers between male and female groups. Dots in green, red, and gray represent genes downregulated, upregulated and not differentially expressed in male flowers compared to female flowers, respectively. (**d**) Gene expression patterns in flower buds (log_2fold change between male and female groups) along the sex chromosomes. Each dot represents a sex-biased gene (SBG). Statistical comparison of SBG located within the sex chromosomes (**e**) and within the Y chromosome components (**f**). (**g**) Statistical comparison of SGB in mature tuber samples located within the sex chromosomes. Samples represent mature tuber from three female *D. alata* accessions ('Dou', 'Kinabayo', and '74F') and three male accessions ('Divin', 'Plimbite', and 'Toufi-Tetea'). * = significant difference at $p < 0.05$. (**h**) Comparison of the methylation levels (5-methylcytosine at CpG sites) across sex chromosomes and statistical comparison between sex chromosomes components: SDR vs PAR (**i**), HXR vs PAR (**j**) and SDR vs HXR (**k**). Methylation levels are calculated per 100 Kb overlapping windows.

Development of a reliable sex-prediction molecular marker

A sex-prediction PCR allele competitive extension (PACE) marker for *D. alata* was previously developed[26]. However, it does not achieve 100% accuracy in distinguishing male and female individuals. This suggests that the marker is not tightly linked to the sex determination loci on DalaChr6. In this study, we developed a new PACE marker, *sda6*, based on allelic variation within the coding sequence of one candidate sex-controlling gene *DalaChr6AG083080* (psbP domain-containing protein 6) and its homolog *DalaChr6BG082060*. The *sda6* marker is located within the INV1. From the transcriptome data, all female genotypes were homozygous T/T (DalaChr6B: 7,643,820 bp) while all male genotypes were heterozygous G/T (DalaChr6A: 3,714,267 bp) (Supplementary Table 24). Genotyping 146 *D. alata* accessions using the *sda6* marker revealed two distinct clusters corresponding to male and female groups (Supplementary Fig. 16). Among the 74 accessions with known sex phenotypes, *sda6* accurately distinguished male and female plants with 100%

accuracy (Supplementary Table 25). This result underscores *sda6* as a reliable molecular marker for predicting the sex phenotype in *D. alata*, addressing an essential requirement for advancing breeding efforts[24].

Discussion

The haplotype-resolved genome assembly of the 'Kabusa' cultivar surpasses the quality of existing *D. alata* assemblies[27,28], establishing a new reference standard for the *Dioscorea* genus. It features a significantly higher contig N50, improved genome annotation, and greater accuracy. The assembly is nearly gap-free and haplotype-resolved, containing a large repertoire of novel genes. Importantly, it includes the previously missing full Y chromosome sequence. As demonstrated in other plant species, such as pear[63] and walnut[64], haplotype-resolved genome assemblies offer significant advantages over mosaic genomes. These include enhanced resolution for studying complex traits and evolutionary dynamics, especially in species with high heterozygosity or sex-specific chromosomes, such as *D. alata*. The haplotype assemblies of the 'Kabusa' genome provide valuable resources for a wide range of applications, including variant calling, gene expression analysis, genome-wide association studies, genomic prediction, genome editing, and comparative genomics. Additionally, they enable precise evaluation of the specific contributions of each haplotype to phenotypic traits, supporting both functional and applied research efforts[14,65].

The identification of telomeric and centromeric sequences is crucial for ensuring a complete, high-quality genome assembly. Such completeness enhances the precision of trait mapping (e.g., GWAS, QTL), enabling more accurate association between genetic loci and agronomic traits, which in turn facilitates targeted breeding strategies[13]. In the 'Kabusa' genome, we identified the telomeres and putative centromeric positions for all 40 chromosomes, revealing notable putative shifts in centromere positions between some homologous chromosomes. These shifts can profoundly affect chromosome behavior during meiosis and mitosis[66]. The intricate relationship between centromere dynamics, genomic rearrangements, and their influence on trait variation has been well-documented in various species[67,68,69]. In the 'Kabusa' genome, we also detected several para- and pericentric SVs. Consistent with our findings, a paracentric inversion on linkage group 16 (corresponding to DalaChr16) in 'Kabusa' has been previously identified[29]. Their study further demonstrated that this SV reduces fertility and recombination rates in 'Kabusa'. The prevalence and impacts of such SVs on *D. alata* breeding programs warrant further investigation. Overall, we anticipate that the high-quality

564 'Kabusa' genome will play a pivotal role in accelerating greater yam improvement by providing
565 a robust foundation for functional genomics and targeted breeding strategies.

566 The sex chromosomes of *D. alata* are young and at the early stage of differentiation,
567 features observed in other plant species such as willows, poplars, strawberry and silene (*Silene*
568 *otites* and *Silene pseudotites*)[70,71,72,73]. In addition to the pericentric inversion, the Y
569 chromosome exhibits a slight size reduction and a shift in putative centromere position
570 compared to the X chromosome. Cormier et al.[25] demonstrated that *D. alata* sex chromosomes
571 still undergo recombination, albeit at a significantly reduced rate. This reduced recombination
572 can impact the detection of quantitative trait loci (QTL) on these chromosomes. Genes located
573 on sex chromosomes, particularly those linked to important agronomic traits, present both
574 challenges and opportunities for crop breeding[74]. To date, only a single minor QTL associated
575 with tuber oxidative ratio has been identified on DalaChr6[29,75]. It is likely that the reduced
576 recombination rate on DalaChr6 obscures the detection of additional QTL present on this
577 chromosome.

578 The collinearity of sex chromosomes in related species offers valuable insights into their
579 evolutionary dynamics[76]. Our analysis suggests that sex chromosome formation in *D. alata*
580 occurred after its divergence from the *D. rotundata* lineage. In *D. rotundata* (ZW system), a
581 relatively small sex-determining region (SDR) of ~161 Kb containing only three genes has been
582 identified[36]. In contrast, the SDR in *D. alata* spans 7.6 Mb and harbors 231 genes, reflecting a
583 markedly different genomic architecture associated with sex determination. These differences
584 in sex determination systems, as well as in SDR size and structure, indicate a clear case of sex
585 chromosome turnover between the two species. Notably, *D. rotundata* frequently displays sex
586 lability and includes monoecious cultivars[77,78], whereas such reproductive flexibility has not
587 been reported in *D. alata* to date. Given that yams display diverse sex determination systems,
588 including XY, ZW, and XO (*D. reticulata* and *D. sinuata*)[23], the *Dioscorea* genus represents an
589 excellent model for studying sex chromosome evolution and turnovers. However, high-quality
590 genome resources of heterogametic sex, like the 'Kabusa' genome, are essential to enable
591 accurate detection of sex chromosomes, SDR, and sex-determining genes.

592 We specifically investigated the SDR of *D. alata*, which corresponds to a large
593 pericentric inversion on the Y chromosome and is composed of multiple inversion-derived
594 evolutionary strata. Our results perfectly match reports in a wide range of animals and plants,
595 where suppression of recombination between sex chromosomes is thought to occur
596 progressively, often mediated by chromosomal inversions. This leads to the formation of
597 evolutionary strata, marked by distinct levels of sequence divergence between the sex

chromosomes[79,80,81,82,83]. The *D. alata* SDR harbors sex-limited, sex-specific and sex-biased genes critical for sex phenotype expression. Many of these SDR-associated genes are predicted to have functions related to flower organ development. This aligns with previous findings that male/female sterility and sex determination are closely tied to processes such as growth and development, pollen viability and mobility, ovule and embryo sac formation, and other gametogenic processes[84], as well as hormone signaling pathways[85]. The convergent model of dioecy in plants involves a cascade of regulatory pathways, with sex-regulatory genes playing central roles in determining sex[10]. Overall, the candidate sex-determining genes identified in *D. alata* provide a valuable foundation for advancing functional genomics and offer promising targets for future efforts in sex manipulation and breeding in this species.

Methods

Plant material

Dioscorea alata cultivars 'Kabusa' and '74F' were obtained from our germplasm collection of the Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Guadeloupe. Tubers of 'Kabusa' and '74F' were planted at the CIRAD research station located at Roujol (16°10′ 56″ N, 61° 35′ 24″ W, 10 meters above sea level), Guadeloupe in May 2022. We also planted ten offspring ('A36', 'A94', 'A113', 'A74', 'A96', 'A117', 'A88', 'A111', 'A64', 'A89') from the cross 'Kabusa' x '74F'[25]. Genotyping data from 37 and 108 *D. alata* accessions were retrieved from the population studied by Mota et al.[40].

Library construction and sequencing

During vegetative stage, young leaves of 'Kabusa' were collected for high-quality genomic DNA extraction using the cetyl trimethyl ammonium bromide method[86]. DNA quality and concentration were tested by 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). After the DNA quality and integrity were tested, it was randomly sheared by Covaris ultrasonic disruptor. DNBSEQ sequencing pair-end libraries were prepared using VAHTS® Universal Plus DNA Library Prep Kit for MGI V2 (Vazyme, Nanjing, China). For Oxford Nanopore sequencing, the libraries were prepared using the SQK-LSK110 ligation kit according to the standard protocol. The purified library was loaded onto primed R9.4.1 Spot-On Flow Cells and sequenced using a PromethION sequencer

(Oxford Nanopore Technologies (ONT), Oxford, UK) with 72-hour runs. For PacBio sequencing, the libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, Menlo Park, CA, USA) according to the standard protocol. The purified library was loaded and sequenced using REVIO Polymerase Kit (Pacific Biosciences, Menlo Park, CA, USA) and REVIO sequencing plate. For Hi-C sequencing, nuclear DNA collected from young leaves was first cross-linked to capture DNA fragment interaction and then digested with a restriction enzyme, Dpn II. Next, sequencing libraries were constructed based on the biotinylated and ligated DNA fragments and subsequently sequenced on Illumina NovaSeq 6000 (San Diego, CA, USA). Mixed tissues (flower, young leaf, old leaf, vine, tuber flesh, tuber skin, root and petiole) of 'Kabusa' were collected at different developmental stages and used for ONT full-length Iso-seq sequencing. The preparation of long-read cDNA libraries was initiated using the ONT PCR-amplified cDNA kit (SQK-PCS109). Specifically, full-length cDNA libraries were constructed from poly(A)+ mRNA using the cDNA-PCR Sequencing kit (SQK-PCS109). Subsequently, the cDNA was PCR-amplified for 13–14 cycles with barcoded adapters from the Oxford Nanopore PCR Barcoding kit (SQKPBK004). Finally, the 1D sequencing adapter was ligated to the cDNA prior to loading onto an R9.4.1 FLO PRO002 flow cell for sequencing on a PromethION platform. Flower buds were chosen at Day5 because they are RNA-rich and at a stage when a clear developmental divergence between male and female is not yet observed (Fig. 4a). Samples from ten offspring and their parents ('Kabusa' x '74F') were collected in triplicate and used for RNA-seq. Total RNA was isolated from young flower buds using the E.Z.N.A.® Plant RNA Kit (Omega Biotek, Norcross, GA, USA) according to the manufacturer's instructions. Quality, purity and quantity of the isolated RNA was checked on Agarose electrophoresis, Nanodrop (NanoDrop Technologies, Wilmington, DE, USA) and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), respectively. The libraries were prepared using the NEBNext® UltraTM Kit of Illumina® according to the manufacturer's instructions. RNA sequencing was conducted on Illumina Novaseq 6000 (San Diego, CA, USA).

Genome assembly

The k-mer depth distribution was calculated with Jellyfish (v2.2.10)[87] based on the DNBSEQ short reads, and GCE (v1.0.0)[88] was used to estimate the genome size and heterozygosity. Four preliminary genome assemblies were generated:

1- with nextDenovo (v2.5.0, parameters: read_cutoff = 1k, blocksize = 1g, nextgraph_options = -a 1)[31], we performed the preliminary assembly of ONT

sequencing data

2- HiFi data was assembled with hifiasm, v0.16.1-r375[30]

3- A mixed assembly of HiFi and ONT ultra-long data using hifiasm, v0.18.2-r467

4- A mixed assembly of HiFi and Hi-C data using hifiasm, v0.19.5-r587

The quality of the four assemblies were evaluated and the 'HiFi + Hi-C' preliminary assembly version was selected as the foundation for the subsequent near-complete genome assembly.

Contig sequences from the assembly were clustered into 40 chromosomes by the agglomerative hierarchical clustering method and then oriented using the ALLHiC (v0.9.8)[89]. We convert the pairwise interactions between contigs into specified binary files (.hic files) using 3D-DNA (v180419)[90] and juicer (v1.6)[32]. We further refined the ordered and oriented contigs (.assembly) using Juicebox (v1.11.08)[32] to manually adjust into a reviewed assembly (.review.assembly). Gap regions were represented with 100 Ns to obtain chromosome-level genome sequences. The Hi-C heatmaps were generated based on contig interaction strength and genomic positions using HiCExplorer (v3.6)[91].

To proceed to the extension and repair of telomere sequences, we employed Winnowmap (v2.03, parameters: k=15, -MD)[92]. All ONT reads were aligned to the genome and reads aligning once within 50 bp from the chromosome ends were collected. Occurrence of telomere repeat motifs (CCCTAAA/TTTAGGG) was calculated in each read, and we defined the read with the highest occurrence as the reference and others as query. With medaka_consensus (v1.2.1, parameters: -m r941_min_high_g360), we reassembled the reference telomere reads and query telomere reads to obtain the consensus sequence. Thereafter, the consensus telomere sequences were aligned to each chromosome with nucmer (v3.1)[93], and we replaced the telomere sequence at the chromosome ends with the best alignment result.

The next step consists of gap-filling. Winnowmap (v2.03, parameters: k=15, -MD)[92] was used to align gap-filling data (excluding Ns) to the genome gap regions (including Ns). The priority for gap-filling data was as follows: preliminary assembly genome versions > ONT data (consensus.fa) > HiFi data (ccs.fasta). When the alignment spans exactly across both ends of the gap, the best alignment region with the longest length were chosen to replace the gap-containing sequence on the genome using gap-filling data. After the gap-filling is completed, the replaced regions in the gap-filled genome were aligned with other assembly versions/ONT reads/HiFi reads to verify the reliability of the gap-filling results.

Genome evaluation

Genome assembly continuity was assessed by locating the gaps on the genome. For genome assembly consistency, the clean short reads were aligned to the genome using bwa (v0.7.17-r1188) while ONT and HiFi reads were mapped to the genome using minimap2 (v 2.17-r941)[35]. Alignment and coverage rates were calculated. Genome completeness was assessed using BUSCO protein mode with embryophyta_odb10 data set (v 5.3.0; parameters: --evalue 1e-05 --augustus)[94]. Likewise, genome quality and accuracy were evaluated using Meryl (v1.3)[95]. We employed MUMmer (v4.0.0)[96] to align the 'Kabusa' haplotype genome assemblies to the female genome Dala_V2[27] to determine the chromosome numbering and orientation. Then, we used SyRI (v1.6)[33] to identify collinear regions and variations between genomes.

Repeat sequence annotation

The same pipeline was applied for HaplotypeA and HaplotypeB. First, the repeat sequences were identified and annotated. Model sequences were predicted based on the genome sequence using RepeatModeler (v2.0.4, parameters: -database mydb -threads 16, https://github.com/Dfam-consortium/RepeatModeler). Additionally, LTR sequences were predicted using LTR_FINDER (version: Official release of LTR_FINDER_parallel, parameters: -threads 16 -harvest_out -size 1000000 -time 300)[97] and LTRharvest (v1.62, parameters: -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes)[98]. We used LTR_retriever (v2.9.0, parameters: -threads 16 -noanno)[99] to remove redundancy from the sequences predicted by the two methods to obtain non-redundant LTR sequences and merged these two sets of sequences to create a *de novo* repeat sequence library. TEclass (v2.1.3, default parameters)[100] was employed to redefine the content labeled as "Unknown". Then, we merged the RepBase library (v20181026) with the *de novo* library. We also used the RepeatMasker subprogram RepeatProteinMask (v4.1.5, parameters: -noLowSimple -pvalue 0.0001) to predict TE proteins. Finally, all repetitive prediction results were merged, and redundancy was removed to obtain the final collection of repetitive sequences in the 'Kabusa' haplotype assemblies. Tandem repeat sequences were predicted using TRF software (v4.09, parameters: 2 7 7 80 10 50 2000 -d -h)[101] and MISA software (v2.1, default parameters)[102].

Gene annotation

Protein-coding gene models were predicted by integrating three strategies: *ab initio* prediction, homology-based search, and gene expression-based prediction.

For gene expression-based prediction, (a) NGS RNA-seq clean reads were aligned to the genome using hisat2 (v2.2.1, parameters: –dta -p 16), and the resulting BAM files were used by stringtie (v2.2.1, parameters: -p 16)[103] to reconstruct the transcripts; (b) ONT Iso-seq data were filtered using NanoFilt (v2.8.0, parameters: -l 50 -q 7) and full-length sequence identification performed with pychopper (v2.7.5, parameters: -t 16). The obtained full-length sequences were aligned to the 'Kabusa' haplotype genomes using minimap2 (v2.26-r1175, parameters: minimap2 -t 16 -ax splice -uf --secondary=no)[35]. The alignment results in BAM files were used by stringtie (v2.2.1, parameters: -p 16 -R -L)[103] to reconstruct the transcripts. (c) The transcripts predicted by the two methods were merged using tama (v1.0, parameters: -f filelist.txt -p merge, https://github.com/GenomeRIK/tama). The merged transcript results were then used with TransDecoder (v5.7.0, parameters: TransDecoder.Predict –retain_long_orfs_length 150, TransDecoder.LongOrfs -m 50, https://github.com/TransDecoder/TransDecoder) to predict open reading frames and finally obtain predicted coding genes.

For homology-based search, homologous protein sequences from *D. alata* female genome[27], *D. rotundata*[36], *D. dumetorum*[104], *D. zingiberensis*[105] and *T. zeylanicus*[106] were aligned to the 'Kabusa' haplotype genomes with tblastn (v2.13.0, parameters: default)[107] to obtain candidate regions. Then, Exonerate (v2.4.0, parameters: –model protein2genome –showtargetgff 1)[108] was used to predict transcripts and coding regions based on the alignment results.

The *ab initio* prediction was conducted using Augustus (v3.5.0, parameters: –uniqueGeneId=true –noInFrameStop=true –gff3=on –strand=both)[109] and glimmerhmm (v3.0.4，parameters: -f -g)[110].

All the predicted gene sets were integrated using MAKER2 (v3.01.03, default parameters)[111.] The completeness of the predicted annotation sets was assessed using BUSCO protein mode with the Embryophyta_odb10 data set[94]. Functional annotation of protein-coding genes was performed based on existing databases such as KEGG, GO and Uniprot. Concerning non-coding RNAs prediction, tRNAscan-SE (v2.0.12, parameters: -E -j tRNA.gff -o tRNA.result -f tRNA.struct –thread 16)[112] was used to find tRNA sequences in the genome, RNAmmer (v1.2, parameters: -S euk -m tsu,lsu,ssu)[113] was used to predict rRNA, and INFERNAL (v1.1.4)[114] was employed to search for ncRNA in the genome based on Rfam database (v14)[115].

Centromere detection analyses

We used two complementary approaches to predict centromere regions.

1- Tandem repeat density and gene coverage approach: This method leveraged tandem repeat (TR) density and low gene density, as described in previous studies[116,117]. Using BEDTools (v2.30.0)[118], we calculated TR and gene coverage in 50 Kb windows across the 'Kabusa' genome, visualizing continuous regions of high TR and low gene density as potential centromeric regions.

2- TR cluster (TRC)-based approach: Following the methods used in T2T kiwifruit and *Rhodomyrtus tomentosa* genomes[119,120], TR sequences were identified using TRF (v4.09.1, parameters: 2 7 7 80 10 50 500 -d -h)[101] and clustered into TRCs using cd-hit (v4.8.1, parameters: -c 0.9 -n 5 -d 0 -g 1 -aS 0.9, https://github.com/weizhongli/cdhit/releases). The most abundant TRCs were plotted to aid centromere identification.

3- Final prediction and visualization: We used the longest TRC as the representative sequence and minimized overlaps using RepeatMasker. BEDTools (v2.30.0)[118] was then used to merge and define centromere regions, which were visualized with RIdeogram[121] and LINKVIEW2 (v1.0.5, https://github.com/YangJianshun/LINKVIEW2), alongside features like rDNA and telomeres.

Sequence variation between haplotype genomes of 'Kabusa' and evolutionary analysis with other *Dioscorea* species

Haplotype genomes of 'Kabusa' were aligned using MUMmer (v4.0.0rc1; default parameters)[96], with HaplotypeA as the reference. Then, variants were detected using SyRI (v1.5; default parameters)[33]. For synteny analysis between haplotype genomes of 'Kabusa' and the female *D. alata* genome[27,28], gene sequences were aligned using LAST (v1170)[122] to identify similar gene pairs. These gene pairs were then analyzed using JCVI (v0.9.13)[123] to determine if they are adjacent on the chromosomes based on annotation files (gff3), thereby identifying all genes in syntenic blocks. The synteny plots were generated using SyRI (v1.5; default parameters)[33].

To identify shared and unique genes between the two haplotype genomes (HaplotypeA and HaplotypeB), the genomes were clustered into gene families using Orthofinder (v2.3.12)[124]. Protein sequences from HaplotypeA were extracted and compared to the HaplotypeB genome using tblastn (v2.13.0)[107]. Structural integrity of matching genes was verified using gffread (v0.12.7). Protein sequences of structurally complete genes were further compared to HaplotypeA's proteins using blastp (v2.6.0)[107]. Genes with alignment scores >150 were deemed

homologous and removed from HaplotypeA's unique gene list. The remaining genes were considered unique to HaplotypeA. From the unique gene list of HaplotypeA, genes located on Chr6 were identified as DalaChr6A-specific unique genes. The same process was applied to identify unique genes in HaplotypeB and DalaChr6B-specific unique genes.

For phylogenetic tree construction between *Dioscorea* species, a multiple sequence alignment of protein sequences for each single-copy gene family was performed using MUSCLE (v3.8.31)[125]. The alignment results were then filtered using trimAl (v1.2rev59; parameters: -gt 0.2)[126]. The filtered alignments were concatenated into the supergene. A maximum likelihood phylogenetic tree was constructed based on the supergene using RAxML (v8.2.10)[127] with the PROTGAMMAWAG model. Divergence times for the selected species (*D. alata*, *D. rotundata*, *D. tokoro* and *D. zingiberensis*) were estimated using the mcmctree program from the PAML package (v4.9; parameters: nsample=3000000; burnin=8000000; seqtype=0; model=4)[128].

Transcriptome analysis

Newly generated raw reads from flower buds, along with short reads from mature tubers of three female *D. alata* accessions ('Dou', 'Kinabayo', and '74F') and three male accessions ('Divin', 'Plimbite', and 'Toufi-Tetea'), previously generated by Mota et al.[40] (PRJNA918625), were used to perform a comparative gene expression analysis between male and female groups. The package fastp (v0.21.0)[129] was used to filter raw reads (150 bp paired-end) and discard low-quality reads. Several libraries were found to be of poor quality or contaminated after sequencing; therefore, we excluded them from further analyses. The cleaned reads from each sample were mapped to the haploid consensus genome of the 'Kabusa' genome (specifically the HapA assembly, which includes the Y chromosome + the HXR) using STAR (v2.7.9, parameters:default)[130]. This choice was made to enable a focused analysis of male-specific and shared gene expression. The RSEM package[131] was used to calculate gene expression level for each sample expressed as fragments per kilobase of transcript per million fragments mapped. The edgeR (v3.32.1) software[132] was used for differential expression analysis. Selection thresholds of FDR < 0.05 and $|\log_2 FC| > 1$. Functional enrichment analysis was performed using the R package clusterProfiler v4.0[133] based on the functional annotations of all genes in this species, including Gene Ontology[134] and Kyoto Encyclopedia of Genes and Genomes[135].

Refining sex-determination region

Whole genome resequencing raw data of 20 male and 17 female accessions were downloaded from NCBI (PRJNA918625). We constructed a male-pool by merging male data and a female-pool by merging female data. The two pools were mapped to the HaplotypeA of the 'Kabusa' genome. GATK (v4.6.1.0)[136] was used for SNP calling and *Fst* was estimated between the two pools with VCFtools (0.1.17)[137]. Next, SNPs that are homozygous in the female pool but heterozygous in the two pools were filtered. The variant density in a 100 Kb window was calculated and plotted with the circlize package[138] in R (v4.4.2).

Genome wide association studies

A total of 200 *D. alata* accessions were planted at Roujol and Godet research stations in Guadeloupe during seasons 2023 and 2024. Sex phenotypes were recorded during flowering period (September to December) each year. Flowering was recorded in108 accessions including 68 males and 40 females. Genotyping data from the 108 *D. alata* accessions were retrieved from the population studied by Mota et al.[40] (PRJNA918625). SNPs and InDels calls were performed as described above using GATK (v4.6.1.0)[136]. GWAS was conducted independently for SNPs and InDels genotype matrices using PLINK v1.90b6.16[139] with the logistic regression model that included the the top 3 principal components from SNPs and InDels genotype matrices. The Manhattan plots were generated in R (v4.4.2). Markers having a significant association with traits were determined by the adjusted *p* value.

K-mer analysis

To identify sex-specific sequences based on k-mer composition, we performed a k-mer counting analysis using KMC3 (v3.2.1)[140]. Cleaned Illumina paired-end reads from 20 male and 20 female individuals were used for this analysis. For each sample, canonical 27-mers (k = 27) were counted independently using the following parameters: -k27 -ci5 -fq -m64 -t4. The minimum count threshold (-ci5) was applied to filter out low-frequency k-mers likely resulting from sequencing errors. Reads were processed in paired-end mode (-fq) using compressed FASTQ files (*_clean_R1.fq.gz and *_clean_R2.fq.gz). The k-mer databases generated for each individual were stored separately to allow downstream comparison between male and female groups. (**1**) Canonical k-mer databases previously generated using KMC3 were compared using kmc_tools. Specifically, k-mers unique to males were identified by subtracting male using kmers_subtract producing male-specific k-mers. (**2**) The unique k-mers were dumped into text files and filtered to exclude sequences containing ambiguous nucleotides

('N'). Cleaned lists were then converted into FASTA format using awk, with each k-mer assigned a unique identifier. (**3**) Male-specific k-mers were mapped to the *D. alata* HaplotypeA using bwa mem, with relaxed parameters (-k 21 -T 21 -a) to accommodate short reads and allow multi-mapping. The resulting SAM file was converted to sorted BAM format using SAMtools, (v1.18)[141] followed by indexing for downstream visualization and coverage analysis.

Read coverage analysis

To quantify sequencing read coverage, we used Mosdepth (v0.3.3)[142]. A 10 Kb sliding window BED file was generated to cover DalaChr6A. BAM files were grouped by sex based on metadata and processed independently for the 68 male and 40 female samples. Coverage depth was calculated across each window using Mosdepth with the options -n -x to exclude duplicate and secondary alignments. For each window, the mean depth was calculated across all individuals of each sex using custom AWK scripts. The final dataset consisted of a merged table reporting average male and female coverage per window across the DalaChr6A.

Sex chromosome divergence analysis

We estimated nonsynonymous (*Ka*) and synonymous (*Ks*) substitution rates for gene pairs located on the X and Y chromosomes. First, we extracted gene coordinates from specific regions of interest SDR, INV1–4 and XSR, PAR, MSY—using bedtools intersect and GFF annotations from the 'Kabusa' haplotype-resolved genome assembly. The total number of genes in each region was quantified based on gene annotations from chromosomes DalaChr6A (Y) and DalaChr6B (X). Next, their corresponding coding DNA sequences and protein sequences were retrievd using seqtk-1.5 (r133; https://github.com/lh3/seqtk). Orthologous gene pairs between Haplotypes A and B were determined using a anchor file from whole-genome synteny analysis. For each orthologous gene pair, protein sequences were aligned using MAFFT (v7.453)[143]. Codon alignments were then generated using PAL2NAL (v14)[144]. doi: 10.1093/nar/gkl315.). The resulting codon alignments were cleaned, concatenated, and used as input to KaKs_Calculator (v2.0)[145] with the Yang-Nielsen model to estimate pairwise Ka, Ks, and Ka/Ks ratios.

PACE marker development and experimental procedure

SNP genotyping was performed as described by von Maydell[146]. DNA was extracted from leaf tissue 146 *D. alata* accessions using mixed alkyl trimethylammonium bromide buffer

and NucleoMag Plant Kit (Macherey-Nagel, Düren, Germany). We used the PACE™ master mix (3CR Bioscience, UK) (standard protocol), and runs were performed on LightCycler480 (Roche, Basel, Switzerland). Five negative controls (water) were included in the experiment. From the raw fluorescence data, allele calling was performed by defining fluorescence clusters. Plotting was conducted with the R package snpclust (v1.1; https://github.com/jframi/snpclust).

Methylation analysis

HiFi CCS reads were used for estimating the methylation frequencies of the CpGs according to the ccsmethphase pipeline developed by Ni et al.[147]. To phase the reads, we used Meryl (v1.4.1)[95] to find unique k-mers in the reference genome of each haplotype as described by Nie et al.[148] and classify the reads using the unique k-mers. The methylation frequencies were estimated for binned regions (100 Kb) in each haplotype assembly.

Statistical analysis

All statistical analyses were performed in R (v4.4.2). Depending on the data distribution and experimental design, both parametric tests (t-tests and one-way ANOVA) and non-parametric tests (Wilcoxon rank-sum test and Kruskal–Wallis test) were applied. A significance threshold of $p < 0.05$ was used throughout.

Supplementary information

Supplementary Table 1: Statistics of the Illumina short reads data
Supplementary Table 2: Statistics of k-mer analysis based on Illumina short reads data
Supplementary Table 3: Oxford Nanopore Technology ultra-long data statistics
Supplementary Table 4: Pacbio HiFi ultra-long data statistics
Supplementary Table 5: Statistics of the Hi-C data
Supplementary Table 6: Four assembly strategies evaluated on the quality of the contig-level genomes
Supplementary Table 7: Detailed statistical results of each chromosome and its length in the hybrid assembly version (HiFi + Hi-C)
Supplementary Table 8: Position of telomeric repeat sequences along the chromosomes in the 'Kabusa' genome
Supplementary Table 9: Gap filling in the genome assembly
Supplementary Table 10: Detailed statistical results of each chromosome and its length in the final near-complete 'Kabusa' genome assembly
Supplementary Table 11: Statistics of repetitive element annotation in the genome of 'Kabusa'
Supplementary Table 12: Statistics of ONT based full-length transcriptome data for genome annotation
Supplementary Table 13: Statistics of prediction, functional annotation of coding genes and Non-coding RNA annotation in the HaplotypeA and HaplotypeB genome assemblies of 'Kabusa'

Supplementary Table 14: Statistics of gene set BUSCO evaluation (busco_lineage:embryophyta_odb10) in the HaplotypeA and HaplotypeB genome assemblies of 'Kabusa'

Supplementary Table 15: Position of centromeric repeat sequences along the chromosomes in the 'Kabusa' genome

Supplementary Table 16: Report of the Liftoff annotation of the female genome *D. alata*_v2 based on HaplotypeA of the 'Kabusa' genome

Supplementary Table 17: Statistics of variants detected between the two haplotype genome assemblies of 'Kabusa' and their impact on gene function

Supplementary Table 18: PLINK logistic GWAS results based on SNPs and Indels markers. 102 *D. alata* accessions were evaluated for sex phenotype in the field over 2 years at 2 sites in Guadeloupe.

Supplementary Table 19: Genes located within the SDR and HXR and their functional annotation. INV1-INV4 =Inversion 1-Inversion 4. XSR = X-specific region; MSY = Male-Specific region of the Y chromosome

Supplementary Table 20: Sex-linked genes that have functional homologues on autosomal chromosomes in the *D. alata* 'Kabusa' genome

Supplementary Table 21: Samples used for RNA-seq transcriptome

Supplementary Table 22: Statistics of RNA-seq transcriptome data from male and female flower buds. -1, -2, -3 represent the three biological replicates. If a genotype does not have 3 replicates, it means a replicate was contaminated and was excluded.

Supplementary Table 23: Sex-biased genes located within the SDR and HXR in *D. alata*. Up/down expression in male compared to female flower buds

Supplementary Table 24: Development of a PACE sda6 marker linked to sex determination in Dioscorea alata. A sex-linked locus was detected via GWAS in the gene DalaChr6AG083080 at the position 91 bp. All female offspring are homozygous TT (XX) while male are heterozygous GT (XY).

Supplementary Table 25: Details on the 146 *D. alata* germplasm genotyped with the *sda6* PACE marker

Supplementary Table 26: Details on the *D. alata* germplasm used for genetic analyses

Supplementary Fig. 1: Flow cytometry-based determination of the ploidy level in 'Kabusa'. The ploidy level was assessed by comparing the position of the main peak with the references indicated by the arrows. According to the results, 'Kabusa' is identified as a diploid cultivar (2x).

Supplementary Fig. 2: k-mer analysis for estimation of the 'Kabusa' genome size. The genome size was estimated to approximately 494.13 Mb.

Supplementary Fig. 3: Distribution of the length of gene and gene components in the 'Kabusa' genome assembly compared to related species. A. Gene length distribution. B. Coding sequence (CDS) length distribution. C. Exon length distribution. D. Intron length distribution.

Supplementary Fig. 4: Centromere and telomere detection map in the two haplotype genome assemblies of 'Kabusa'. Orange squares and green triangles represent centromere and telomeres within the assembled genome. Blue indicates low gene density; red indicates high gene density. Deep blue and light blue circles represent the distribution along the chromosomes of 45S_rDNA and 5S_rDNA, respectively.

Supplementary Fig. 5: Illustrations of alignments of preliminary assemblies (A and B) or ONT reads (C) that clearly spanned gap regions providing strong evidence of the accuracy of the gap filling stage.

Supplementary Fig. 6: Centromere detection in the two haplotype genome assemblies of 'Kabusa'. The first layer representing LTR/Copia density (blue color), the second layer representing LTR/Gypsy density (orange color), the third layer representing tandem repeat density (ruby red color) and the fourth layer representing the gene density (turquoise color) along the chromosomes.
Supplementary Fig. 7: Structural variation (SV) between the two haplotype assemblies of the 'Kabusa' genome. A. Quantity of SVs mapped to different genomic regions. B. Gene ontology enrichment analysis of SV-influenced genes.
Supplementary Fig. 8: Collinearity analysis between the HaplotypeA, HaplotypeB of the 'Kabusa' genome and the female genome assembly from Bredeson et al. [27].
Supplementary Fig. 9: Intergenomic comparison of the HaplotypeA, HaplotypeB of the 'Kabusa' genome and the female genome assembly from Bredeson et al. [27]. Red circle shows the location of the pericentric inversion in DalaChr6A vs DalaChr6F but absent in DalaChr6B vs DalaChr6F.
Supplementary Fig. 10: IGV snapshots showing the alignment of HiFi reads spanning the breakpoints of the inverted region bewteen the Y chromosome (DalaChr6A) and X chromosome (DalaChr6B) of the 'Kabusa' genome
Supplementary Fig. 11: Sex chromosome evolution in yams. (A) Synteny analysis between genomes of four yam species with emphasis on sex chromosomes reported to date. Dala_V2 = female genome of *Dioscorea alata*, DalA and DalB = Two haplotype assemblies of the 'Kabusa' genome (*D. alata*), Drot = *D. rotundata*, Dtok = *D. tokoro* and Dzin = *D. zingiberensis*. *Dioscorea alata* and *D. rotundata* belong to the section Enantiophyllum while *D. tokoro* and *D. zingiberensis* belong to the section Stenophora. XY and ZW denote the sex determination systems reported in the yam species. Numbers before and within brackets represent divergence time million years ago
Supplementary Fig. 12: Distribution of transposable elements in the two haplotype assemblies of the 'Kabusa' genome.
Supplementary Fig. 13: Gene expression profiling and heatmap clustering of male and female groups.
Supplementary Fig. 14: Gene ontology enrichment analysis of the 6917 sex-biased expressed genes detected between male and female groups in *Dioscorea alata*.
Supplementary Fig. 15: Patterns of methylation level (DNA methylation 5mC at CpG sites) along homologous chromosomes in the two haplotype assemblies of the 'Kabusa' genome. Plot was generated using a window of 100 Kb.
Supplementary Fig. 16: Genotyping of *D. alata* diversity panel (146 accessions from different countries) with the *sda6* PACE marker. Fluorescence signals are plotted by accession and observed sex. In x-axis, the "T" fluorescence allele (X), and in y-axis, the "G" fluorescence allele (Y).

Code

No new code has been developed in this project.

Data availability

All raw high-throughput sequencing data comprising Illumina, Oxford Nanopore, PacBio, and Hi-C reads are deposited in NCBI under project numbers PRJNA1175864 and PRJNA880983. Whole genome assemblies are available under the Umbrella BioProject number PRJNA1182344 with accession numbers PRJNA1181019 (HaplotypeA) and PRJNA1181020

(HaplotypeB). The genome assemblies and annotation files are also available at https://yam-genome-hub.southgreen.fr.

Funding

This work was supported by the grants from CGIAR Research Program on Roots, Tubers and Bananas (CRP-RTB), Bill and Melinda Gates Foundation (BMGF) through the RTB breeding project (grant opportunity INV-003446, formerly OPP1112307), the French National Research Agency (ANR) project EVOSEXYAM (ANR-24-CE20-6201-01), the ERDF program Johannes Amos Comenius project TowArds Next GENeration Crops (CZ.02.01.01/00/22_008/0004581) and European Union and Guadeloupe Region (Programme Opérationnel FEDER – Guadeloupe - Conseil Régional 2021-2027) project: TAESA.

Authors' contributions

Conceptualization: Komivi Dossa and Hanâ Chair; Methodology: Komivi Dossa, Ting Qiu, Yao Chen, Peng Ni, Xiaojun Su, Gang Chen, Petre I. Dobrev, Bertrand Fouks, Gemma Arnau, Pierre Mournet, Jianxin Wang, Sylvain Santoni, Hanâ Chair; Software: Komivi Dossa, Ting Qiu, Yao Chen, Peng Ni, Xiaojun Su, Gang Chen, Jianxin Wang; Formal Analysis: Yedomon Ange Bovys Zoclanclounon, Muhammad Amjad Nawaz, Erick Malédon, Moussa Diouf, Petre I. Dobrev, Benjamin Karikari, Bertrand Fouks, Pierre Mournet; Investigation: Erick Malédon, Marie-Claire Gravillon, Christophe Perrot, Elie Nudol, Lévy Laurent, Pierre Mournet, Sylvain Santoni; Resources: Komivi Dossa and Hanâ Chair; Data Curation: Komivi Dossa, Ting Qiu, Yedomon Ange Bovys Zoclanclounon, Pierre Mournet, Yao Chen, Moussa Diouf, Peng Ni, Xiaojun Su, Gang Chen, Petre I. Dobrev, Bertrand Fouks, Jianxin Wang; Writing—Original Draft Preparation: Komivi Dossa, Ting Qiu, Yedomon Ange Bovys Zoclanclounon, Muhammad Amjad Nawaz; Writing—Review and Editing: Hanâ Chair, Gemma Arnau; Visualization, Komivi Dossa, Ting Qiu, Gang Chen, Muhammad Amjad Nawaz; Supervision: Hanâ Chair, Benjamin Karikari; Project Administration: Komivi Dossa and Hanâ Chair; Funding Acquisition: Komivi Dossa and Hanâ Chair. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

We thank Dr Song Chi for providing access to his computing platform.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Renner, S.S. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am. J. Bot*. **101**, 1588-96 (2014).
2. Barrett, S.C. The evolution of plant reproductive systems: how often are transitions irreversible? *Proc. Biol. Sci*. **280**, 20130913 (2013).
3. Liu, Z. et al. A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* **427**, 348-52 (2004).
4. Akagi, T. et al. Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nat. Plants* **5**, 801-809 (2019).
5. Akagi, T. et al. Recurrent neo-sex chromosome evolution in kiwifruit. *Nat. Plants* **9**, 393-402 (2023).
6. Torres, M.F. et al. Genus-wide sequencing supports a two-locus model for sex-determination in Phoenix. *Nat. Commun.* **9**, 3969 (2018).
7. Akagi, T. et al. A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. *Plant Cell* **30**, 780-795 (2018).
8. Akagi, T. et al. The persimmon genome reveals clues to the evolution of a lineage-specific sex determination system in plants. *PLoS Genet.* **16**, e1008566 (2020).
9. Müller, N.A. et al. A single gene underlies the dynamic evolution of poplar sex determination. *Nat. Plants* **6**, 630-637 (2020).
10. Zhang, X. et al. A convergent mechanism of sex determination in dioecious plants: Distinct sex-determining genes display converged regulation on floral B-class genes. *Front. Plant Sci*. **13**, 953445 (2022).
11. Garcia, S. et al. Sex-chrom v. 2.0: a database of green plant species with sex chromosomes. *Chromosoma* **132**, 55-58 (2023).
12. Baránková, S. et al. Sex-chrom, a database on plant sex chromosomes. *New Phytol*. **227**, 1594-1604 (2020).
13. Garg, V. et al. Unlocking plant genetics with telomere-to-telomere genome assemblies. *Nat. Genet*. **56**, 1788-1799 (2024).
14. Liao, Z. et al. A telomere-to-telomere reference genome of ficus (*Ficus hispida*) provides new insights into sex determination. *Hortic. Res.* **11**, uhad257 (2024).
15. Liu, A.W. et al. H3S28P Antibody Staining of Okinawan Oikopleura dioica Suggests the Presence of Three Chromosomes. *F1000Res*. **9**, 780 (2020).
16. Massonnet, M. et al. The genetic basis of sex determination in grapes. *Nat. Commun*. **11**, 2902 (2020).
17. Chen, J. et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* **55**, 1221-1231 (2023).
18. Zhang, X. et al. Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.* 18, 66-72 (2020).
19. Li, H. and Durbin R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet*. **25**, 658-670 (2024).
20. Owusu Danquah, E. et al. Sustainable Intensification and Climate-Smart Yam Production for Improved Food Security in West Africa*:* A Review. *Front. Agron*. **4** (2022).
21. Neina, D. Ecological and Edaphic Drivers of Yam Production in West Africa. *Appl. Environ. Soil. Sci.* **2021**, 5019481 (2021).
22. Arnau, G. et al. Revision of ploidy status of *Dioscorea alata* L. (*Dioscoreaceae*) by cytogenetic and microsatellite segregation analysis. *Theor. Appl. Genet.* **118**, 1239-49 (2009).
23. Sugihara, Y. et al. Population Genomics of Yams: Evolution and Domestication of *Dioscorea* Species, in Population Genomics: Crop Plants, O. Rajora, Editor. Springer International Publishing: Cham. 837-864 (2024).
24. Mondo, J.M. et al. Genome-Wide Association Studies for Sex Determination and Cross-Compatibility in Water Yam (*Dioscorea alata L.*). *Plants* **10** (2021).

25. Cormier, F. et al. A reference high-density genetic map of greater yam *(Dioscorea alata L.). Theor. Appl. Genet*. **132**, 1733-1744. (2019)
26. Cormier, F. et al. Genetic control of flowering in greater yam (*Dioscorea alata L.*). *BMC Plant Biol.* **21**, 163 (2021).
27. Bredeson, J.V. et al. Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nat. Commun.* **13**, 2001 (2022).
28. Zhang YM, et al. A telomere-to-telomere genome assembly for greater yam (*Dioscorea alata*). *Plant Commun*. **2**, 101326 (2025).
29. Ehounou, A.E. et al. Identification and validation of QTLs for tuber quality related traits in greater yam (*Dioscorea alata* L.). *Sci. Rep*. **12**, 8423 (2022).
30. Cheng, H. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Method* **18**, 170-175 (2021).
31. Hu, J. et al. Next*Denovo:* an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol*. **25**, 107 (2024).
32. Durand, N.C. et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*. **3**, 99-101 (2016).
33. Goel, M. et al. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. **20**, 277 (2019).
34. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **19**, 1639-1643 (2021).
35. Li, H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* **32**, 2103-10 (2016).
36. Tamiru, M. et al. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC Biol*. **15**, 86 (2017).
37. Natsume, S. et al. Whole genome sequencing of a wild yam species *Dioscorea tokoro* reveals a genomic region associated with sex. *BioRxiv* 2022.06.11.495741 (2022).
38. Ried, T. et al. Genome Sequencing of Chinese yam (*Dioscorea polystachya*, analysis of PEBP gene family diversity and identification of a potential tuber inducing factor. *BioRxiv* 2025.03.10.641988 (2025).
39. Hu K, Feng Y, Li P, Chen M, Shen ZJ, Sun XQ, Lu RS. Haplotype-resolved genome and population genomics provide insights into dioscin biosynthesis and evolutionary history of the medicinal species *Dioscorea nipponica*. *Plant J.* **121**, e17237 (2025).
40. Mota, A.Z. et al. Whole-genome sequencing and comparative genomics reveal candidate genes associated with quality traits in *Dioscorea alata. BMC Genomics* **25**, 248 (2024).
41. Iocco-Corena, et al. VviPLATZ1 is a major factor that controls female flower morphology determination in grapevine. *Nat. Commun.* **12**, 6995 (2021).
42. Zhang, S. et al. Chromosome-scale assemblies of the male and female *Populus euphratica* genomes reveal the molecular basis of sex determination and sexual dimorphism. *Commun. Biol.* **5**, 1186 (2022).
43. Srikulnath, K., Ahmad, S.F., Singchat, W. & Panthum, T. Do Ty3/Gypsy Transposable Elements Play Preferential Roles in Sex Chromosome Differentiation? *Life* **12**, 522 (2022).
44. Hobza, R. et al. Impact of Repetitive Elements on the Y Chromosome Formation in Plants. *Genes* **8**, (2017).
45. Wang, J. et al. Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc. Natl. Acad. Sci. U.S.A*. **109**, 13710-13715 (2012).
46. Moraga, C. et al. The *Silene latifolia* genome and its giant Y chromosome. *Science* **7**, 630-636 (2025).
47. Prentout, D. et al. An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of *Cannabis sativa. Genome Res*. **30**, 164-172 (2020).
48. Cossard G.G., Toups M.A. & Pannell J.R. Sexual dimorphism and rapid turnover in gene expression in pre-reproductive seedlings of a dioecious herb. *Ann. Bot.* **8**, 1119-1131. (2019).
49. DePaoli, H.C. et al. Stigma/style cell cycle inhibitor 1 (SCI1), a tissue-specific cell cycle regulator that controls upper pistil development. *New Phytol.* **190**, 882-895 (2011).
50. DePaoli, H.C., Dornelas, M.C. & Goldman, M.H.S. SCI1 is a component of the auxin-dependent control of cell proliferation in Arabidopsis upper pistil. *Plant Sci*. **229**, 122-130 (2014).
51. He, Y. et al. Functional characterization of *Arabidopsis thaliana* isopropylmalate dehydrogenases reveals their important roles in gametophyte development. *New Phytol*. **189**, 160-75 (2011).
52. Huang, Y. et al. Epigenetic regulation of female germline development through ERECTA signaling pathway. *New Phytol*. **240**, 1015-1033 (2023).
53. Siriwardana, C.L. et al. NUCLEAR FACTOR Y, Subunit A (NF-YA) Proteins Positively Regulate Flowering and Act Through *FLOWERING LOCUS T. PLoS Genet.* **12**, e1006496 (2016).
54. Luo, Y. et al. Developmental basis for flower sex determination and effects of cytokinin on sex determination in *Plukenetia volubilis* (*Euphorbiaceae*). *Plant Reprod*. **33**, 21-34 (2020).

55. Kanaoka, M.M. et al. *KOMPEITO,* an Atypical *Arabidopsis* Rhomboid-Related Gene, Is Required for Callose Accumulation and Pollen Wall Development. *Int. J. Mol. Sci.* **23**, (2022).
56. Thu, S.W. et al. Mutation in a PHD-finger protein MS4 causes male sterility in soybean. *BMC Plant Biol.* **19**, 378 (2019).
57. Yang X. Makaroff C. A. & Ma H. The Arabidopsis MALE MEIOCYTE DEATH1 gene encodes a PHD-finger protein that is required for male meiosis. *Plant Cell* **15**, 1281–1295 (2003).
58. Yang Z. F. et al. TDR INTERACTING PROTEIN 3, encoding a PHD-finger transcription factor, regulates Ubisch bodies and pollen wall formation in rice. *Plant J.* **99**, 844–861 (2019).
59. Wasternack, C. Determination of sex by jasmonate. *J. Integr. Plant Biol.* **62**, 162-164 (2020).
60. Xie, D.X. et al. *COI1:* an *Arabidopsis* gene required for jasmonate-regulated defense and fertility. *Science* **1280**, 1091-4 (1998).
61. Pagnussat, G.C. et al. Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. *Development* **132**, 603–614 (2005).
62. Fernandes, J.B. et al. Structural variation and DNA methylation shape the centromere-proximal meiotic crossover landscape in Arabidopsis. *Genome Biol.* **22**, 30 (2024).
63. Li, Q. et al. Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. *Plant Commun*. **5**, 101000 (2024).
64. Han, L. et al. A haplotype-resolved genome provides insight into allele-specific expression in wild walnut (*Juglans regia L.*). *Sci. Data* **11**, 278 (2024).
65. Raiyemo, D.A. et al. A phased chromosome-level genome assembly provides insights into the evolution of sex chromosomes in *Amaranthus tuberculatus. BioRxiv* 2024.05.30.596720 (2024).
66. Lu, M. & He, X. Centromere repositioning causes inversion of meiosis and generates a reproductive barrier. *Proc. Natl. Acad. Sci. U S A* **116**, 21580-21591 (2019).
67. Han, Y. et al. Centromere repositioning in cucurbit species: implication of the genomic impact from centromere activation and inactivation. *Proc. Natl. Acad. Sci. U S A* **106**, 14937-41 (2009).
68. Gent, J.I. et al. Stable Patterns of CENH3 Occupancy Through Maize Lineages Containing Genetically Similar Centromeres. *Genetics* **200**, 1105-16 (2015).
69. Taagen, E., Bogdanove, A.J. & Sorrells, M.E. Counting on Crossovers: Controlled Recombination for Plant Breeding. *Trends Plant Sci*. **25**, 455-465 (2020).
70. Hou, J. et al. Different autosomes evolved into sex chromosomes in the sister genera of Salix and Populus. *Sci. Rep.* **5**, 9076 (2015).
71. Cauret, C.M.S. et al. Chromosome-scale assembly with a phased sex-determining region resolves features of early Z and W chromosome differentiation in a wild octoploid strawberry. *G3* **12** (2022).
72. Martin, H. et al. Evolution of Young Sex Chromosomes in Two Dioecious Sister Plant Species with Distinct Sex Determination Systems. *Genome Biol. Evol.* 11, 350-361 (2019).
73. Pucholt, P. et al. Recent Sex Chromosome Divergence despite Ancient Dioecy in the Willow *Salix viminalis. Mol. Biol. Evol*. **34**, 1991-2001 (2017).
74. Charlesworth, D. and A. Harkess, Why should we study plant sex chromosomes? Plant Cell, 2024. 36(5, 1242-1256.
75. Dossa, K. et al. Genome-wide association studies reveal novel loci controlling tuber flesh color and oxidative browning in *Dioscorea alata. J. Sci. Food Agric*. **104**, 4895-4906 (2024).
76. Zhou, Q. et al. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**, 1246338.
77. Denadi, N. Déterminisme du sexe chez les ignames cultivées *Dioscorea rotundata* (Poir.) du Bénin. PhD Thesis, UCLouvain, Belgium. pp : 1-364. available at http://hdl.handle.net/2078.1/270468. (2022).
78. Denadi, N. et al. Plant Sex Prediction Using Genetic Markers in Cultivated Yams *(Dioscorea rotundata* Poir.) in Benin. *Agronomy* **10** (2020).
79. Charlesworth, D. et al. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128 (2005).
80. Hoffmann, A.A. & Rieseberg L.H. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu. Rev. Ecol. Evol. Syst*. **39**, 21-42 (2008).
81. She, H. et al. Evolution of the spinach sex-linked region within a rarely recombining pericentromeric region. *Plant Physiol*. **193**, 1263-1280 (2023).
82. Wang, Y. et al. Evolution of Sex-linked Genes and the Role of Pericentromeric Regions in Sex Chromosomes: Insights from Diploid Willows. *Mol. Biol. Evol*. **41** (2024).
83. Olito, C. and J.K. Abbott, The evolution of suppressed recombination between sex chromosomes and the lengths of evolutionary strata. *Evolution* **77**, 1077-1090 (2023).
84. Renner, S.S. & Müller N.A. Plant sex chromosomes defy evolutionary models of expanding

recombination suppression and genetic degeneration. *Nat. Plants* **7**, 392-402 (2021).
85. Golenberg, E.M. & West N.W. Hormonal interactions and gene regulation can link monoecy and environmental plasticity to the evolution of dioecy in plants. *Am. J. Bot.* **100**, 1022-37 (2013).
86. Porebski, S. L.G. Bailey, & Baum, B.R. *Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol. Biol. Rep*. **15**, 8-15 (1997).
87. Marçais, G. and C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of *k-mers. Bioinformatics* **27**, 764-70 (2011).
88. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* arXiv:1308.2012 (2013).
89. Zhang, X. et al. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833-845 (2019).
90. Dudchenko, O. et al. *De novo a*ssembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).
91. Wolff, J. et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res*. **48**, W177-W184 (2020).
92. Jain, C. et al. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111-i118 (2020).
93. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol*. **5**, R12 (2004).
94. Manni, M. et al. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc*. 1, e323 (2021).
95. Rhie, A. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. **21**, 245 (2020).
96. Marçais, G. et al. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
97. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. **35**, W265-8 (2007).
98. Ellinghaus, D. S. Kurtz, & Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
99. Ou, S. & Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**, 1410-1422 (2018).
100. Abrusán, G. et al. TEclass--a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329-30 (2009).
101. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. **27**, 573-80 (1999).
102. Beier, S. et al. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583-2585 (2017).
103. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. **20**, 278 (2019).
104. Siadjeu, C. E. Mayland-Quellhorst, & Albach D.C. Genetic diversity and population structure of trifoliate yam (*Dioscorea dumetorum* Kunth) in Cameroon revealed by genotyping-by-sequencing (GBS). *BMC Plant Biol*. **18**, 359 (2018).
105. Li, Y. et al. The genome of *Dioscorea zingiberensis* sheds light on the biosynthesis, origin and evolution of the medicinally important diosgenin saponins. *Hortic. Res*. **9**, uhac165 (2022).
106. Vadakkemukadiyil Chellappan, B. et al. High Quality Draft Genome of Arogyapacha (Trichopus zeylanicus), an Important Medicinal Plant Endemic to Western Ghats of India. *G3* **9**, 2395-2404 (2019).
107. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
108. Slater, G.S. & Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
109. Stanke, M. et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-44 (2008).
110. Delcher, A.L. et al. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673-9 (2007).
111. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
112. Chan, et al. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*. 49, 9077-9096 (2021).
113. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Re*s. **35**, 3100-8 (2007).
114. Nawrocki, E. and S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-5 (2013).
115. Kalvari, I. et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic*

Acids Res. **49**, D192-D200 (2021).
116. Deng, Y. et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plan*t **15**, 1268-1284 (2022).
117. Song, J.M. et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* **14**, 1757-1767 (2021).
118. Quinlan, A.R. & Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-2 (2010).
119. Han, X. et al. Two haplotype-resolved, gap-free genome assemblies for Actinidia latifolia and Actinidia chinensis shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Mol. Plant* **16**, 452-470 (2023).
120. Li, F. et al. Gap-free genome assembly and comparative analysis reveal the evolution and anthocyanin accumulation mechanism of *Rhodomyrtus tomentosa. Hortic. Res.* **10**, uhad005 (2023).
121. Hao, Z. et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci*. **6**, e251 (2020).
122. Frith, M.C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80 (2010).
123. Tang, H. et al. JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211 (2024).
124. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. **20**, 238 (2019).
125. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**, 1792-7 (2004).
126. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-3 (2009).
127. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **20**, 1312-3 (2014).
128. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-91 (2007).
129. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fast *Imeta* **2**, e107 (2023).
130. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
131. Li, B. & Dewey C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
132. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. *edgeR*: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).
133. Wu, T. et al. *clusterProfiler* 4.0: A universal enrichment tool for interpreting omics data. *Innovation* (Camb) **2**, 100141 (2021).
134. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. **25**, 25-9 (2000).
135. Kanehisa, M. & Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28**, 27-30 (2000).
136. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297-303 (2010).
137. Danecek, et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).
138. Gu, Z. et al. *circlize* Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2 (2014).
139. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81** (2007).
140. Kokot M., Długosz, M., Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
141. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
142. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
143. Rozewicki, J. et al. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research* **47**, W5-W10 (2019).
144. Suyama, M., Torrents, D. & Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. **1**, W609-12 (2006).
145. Wang, D. et al. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77-80 (2010).
146. von Maydell, D. PCR Allele Competitive Extension (PACE). *Methods Mol. Biol.* **2638**, 263-271 (2023).
147. Ni, P. et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nat. Commun.* **14**, 4054 (2023).

148. Nie, F. et al. De novo diploid genome assembly using long noisy reads. *Nat. Commun*. **15**, 2964 (2024).

Figures

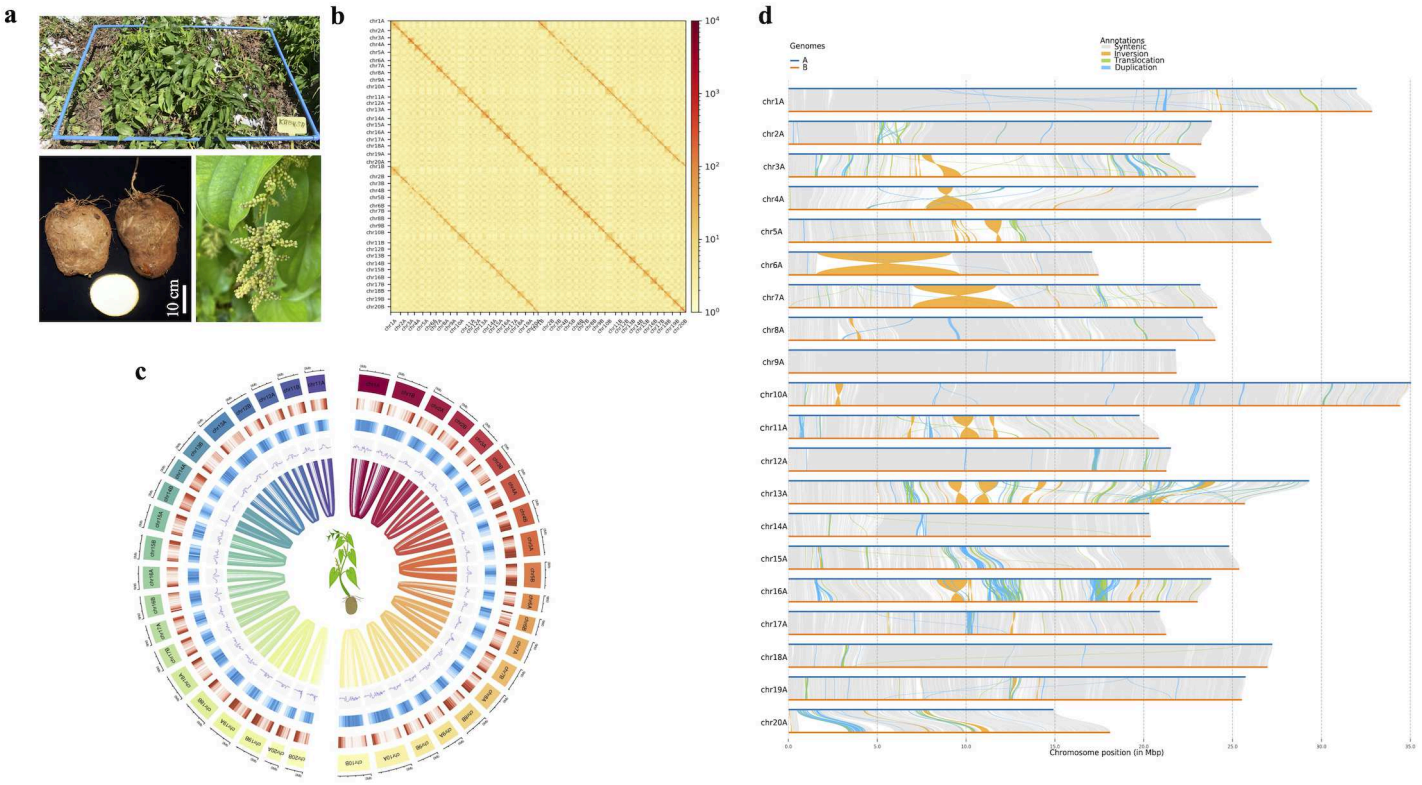

Figure 1

Chromosome-level haplotype-resolved near-complete assembly of the male *D. alata* cultivar 'Kabusa'. (**a**) Pictures of 'Kabusa' cultivar displaying the aerial biomass in the field, the tuber shape and flesh color, and the male flowers. (**b**) Genome-wide Hi-C contact heatmap, visualized with Juicebox[32]. (**c**) Circle diagram of basic genome information. Concentric circles from outside to inside represent chromosomes, gene density, repeat density, and GC density calculated in a 50 Kb window. Lines in the interior indicate collinearity. The scale bar = 10 cm. (**d**) Syntenic mapping between the two haplotypes of the 'Kabusa' genome using SyRI[33]. Collinear regions between HaplotypeA and HaplotypeB are indicated with gray lines, while inversion, translocation, and duplication regions are indicated with orange, green, and blue lines, respectively.

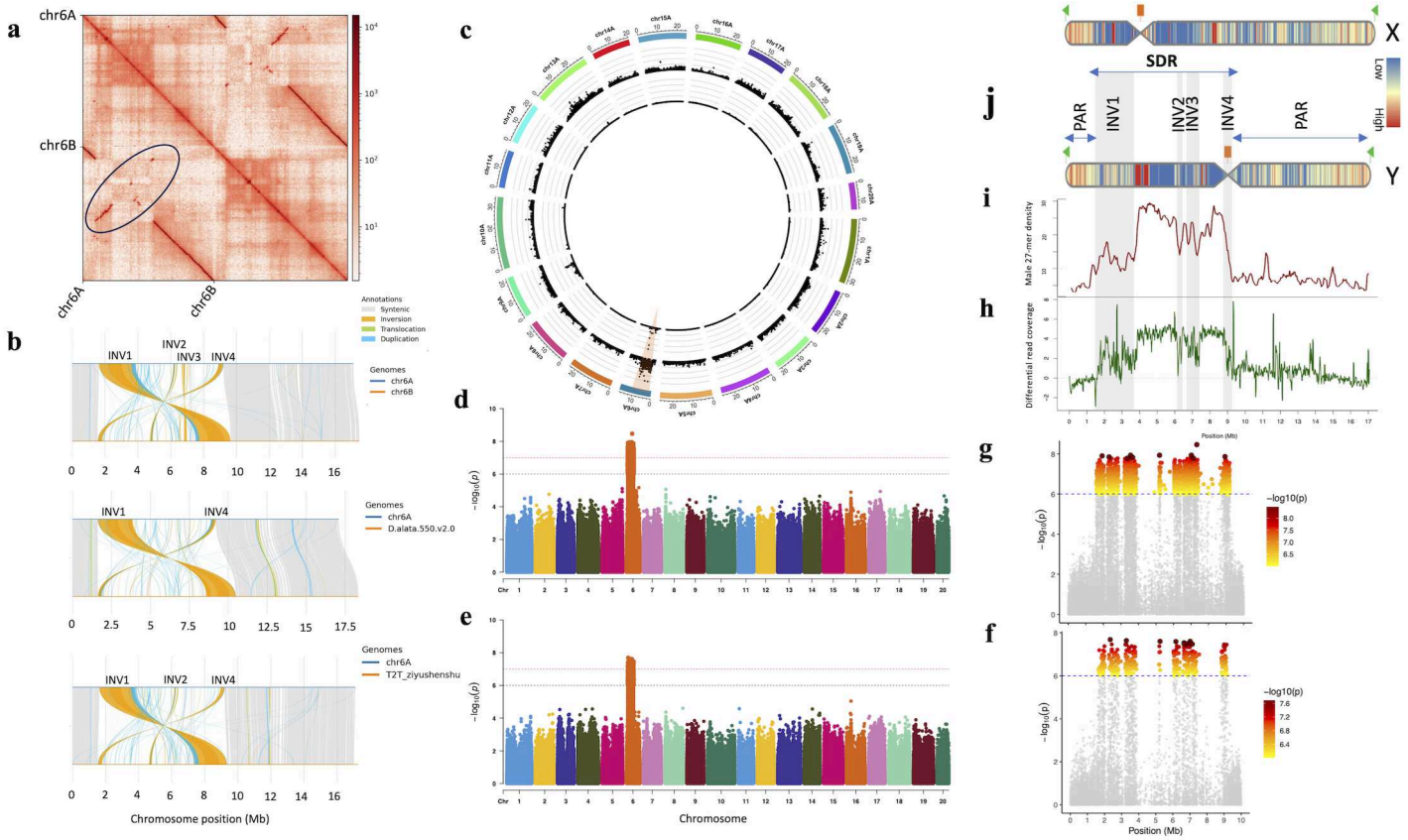

Figure 2

Structure of the sex determination region (SDR) in *D. alata*. (**a**) Hi-C contact heatmap between DalaChr6A (Y) and DalaChr6B (X), visualized with Juicebox [31]. (**b**) Syntenic mapping between 'Kabusa' Y chromosome and X chromosomes from 'Kabusa', Dala_V2[27] and Ziyushenshu[28] using SyRI[33]. Four inversions (INV1-INV4) were detected. (**c**) The genomic landscape of genetic variants between male and female pools. The outer track represents the 20 chromosomes (chr) of HaplotypeA in the 'Kabusa' genome. The inner tracks from outside to inside represent *Fst* between females and males and male-specific SNP density, respectively (window size 100 Kb). (**d**) Manhattan plot for SNP-sex phenotype genome-wide association study (GWAS), with the peaks indicating significant GWAS signals, and the black and red dotted horizontal lines indicating the suggestive and genome-wide significance thresholds, respectively. (**e**) Manhattan plot for InDel-sex phenotype GWAS, with the peaks indicating significant GWAS signals, and the black and red dotted horizontal lines indicating the suggestive and genome-wide significance thresholds, respectively. Regional Manhattan plots on DalaChr6A showing the dispersion of the significant SNPs (**f**) and Indels (**g**). (**h**) Differential read coverage (10 Kb window) between male and female groups along the Y chromosome. (**i**) Exact matches of male specific 27-mers along the Y chromosome. (**j**) Representation of the X (DalaChr6B) and Y (DalaChr6A) chromosomes along with the gene density (heatmap), centromere and telomere locations, pseudoautosomal regions (PAR), and the SDR containing the 4 inversions (INV1-INV4) detected.

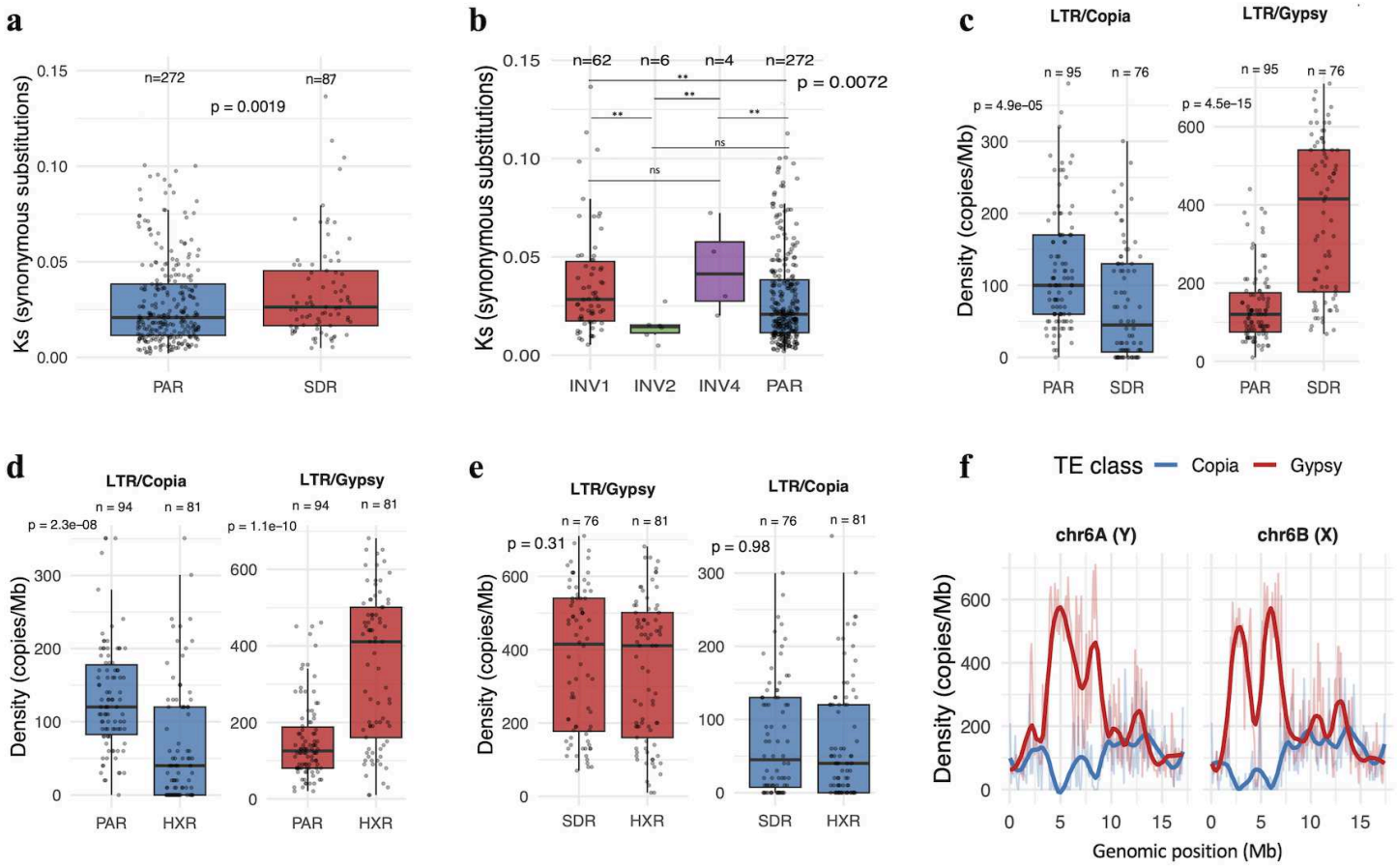

Figure 3

Sex chromosome divergence in *D. alata*. (**a**) Comparison of the synonymous substitution rates (*Ks*) of genes between SDR and PAR on the Y chromosome. (**b**) *Ks* comparison between Y chromosome components. ** = significant difference at $p < 0.001$. (**c**) Comparison of the LTR/Copia and LTR/Gypsy densities between PAR and SDR in the Y chromosome. (**d**) Comparison of the LTR/Copia and LTR/Gypsy densities between PAR and HXR in the X chromosome. (**e**) Comparison of the LTR/Copia and LTR/Gypsy densities between SDR and HXR. (**f**) Smoothed density plots along X and Y chromosomes. Densities are calculated per 1 Mb overlapping windows.

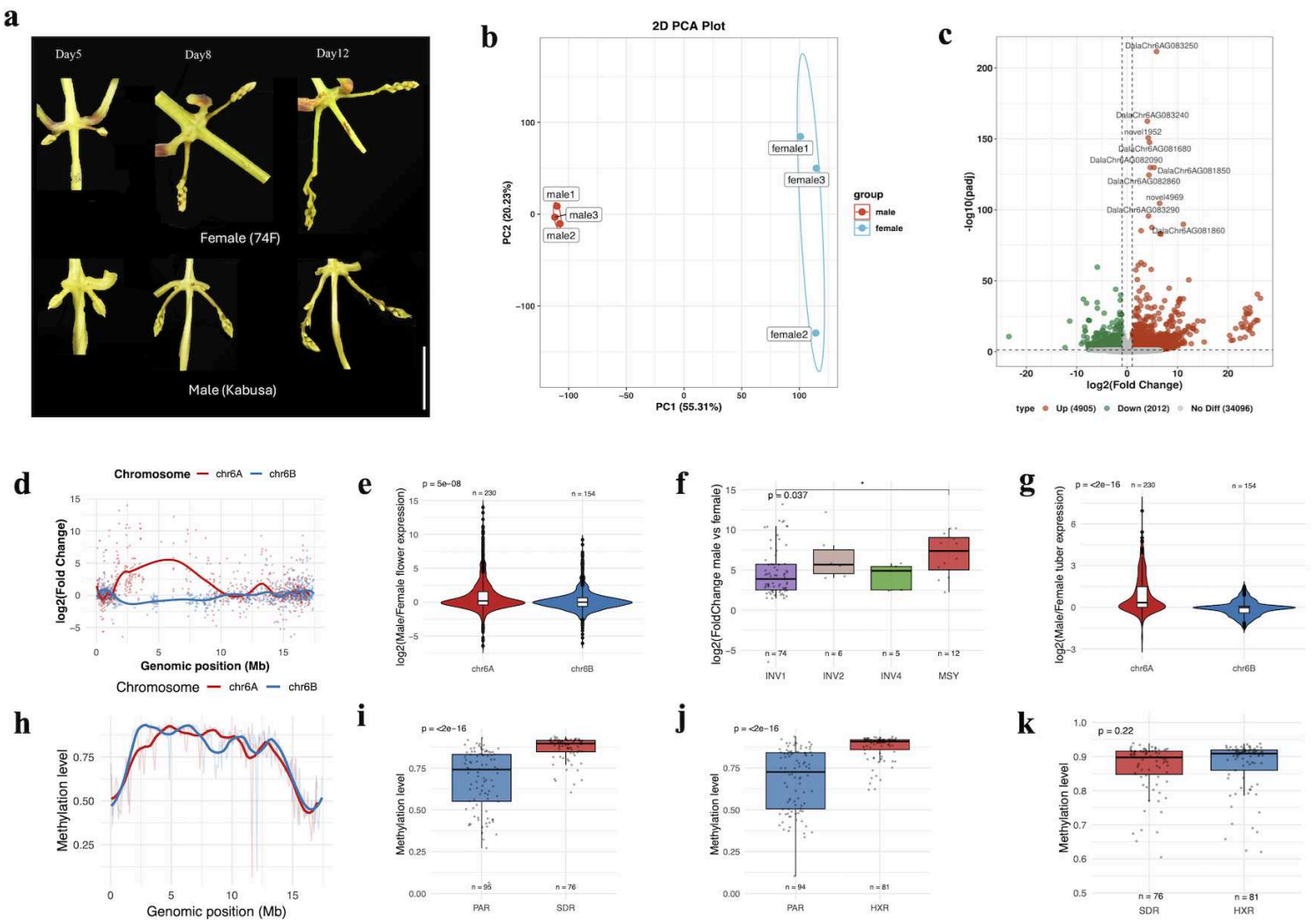

Figure 4

Sex-biased gene expression in *D. alata*. (**a**) Development of *D. alata* flower at Day5, Day8 and Day12 after blooming in male 'Kabusa' and female '74F'. Ontogenetic differences between female and male flowers become morphologically evident between Day8 and Day12 after blooming. Scale = 1 cm. (**b**) Principal component analysis based on the gene expression levels for flower samples (Day5) expressed as fragments per kilobase of transcript per million fragments mapped in the male and female groups. The first and the second principal components explaining 55% and 20% of the total variation, respectively. (**c**) Volcano plot displaying the differential expression analysis in flowers between male and female groups. Dots in green, red, and gray represent genes downregulated, upregulated and not differentially expressed in male flowers compared to female flowers, respectively. (**d**) Gene expression patterns in flower buds (log_2fold change between male and female groups) along the sex chromosomes. Each dot represents a sex-biased gene (SBG). Statistical comparison of SBG located within the sex chromosomes (**e**) and within the Y chromosome components (**f**). (**g**) Statistical comparison of SGB in mature tuber samples located within the sex chromosomes. Samples represent mature tuber from three female *D. alata* accessions ('Dou', 'Kinabayo', and '74F') and three male accessions ('Divin', 'Plimbite', and 'Toufi-Tetea'). * = significant difference at $p < 0.05$. (**h**) Comparison of the methylation levels (5-methylcytosine at CpG

sites) across sex chromosomes and statistical comparison between sex chromosomes components: SDR vs PAR (**i**), HXR vs PAR (**j**) and SDR vs HXR (**k**). Methylation levels are calculated per 100 Kb overlapping windows.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryFigures.pdf