scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A haplotype-phased male genome sequence of the stinging nettle, *Urtica dioica* ssp. *dioica*

Kaede Hirabayashi[1], Diana Percy[2,3], Eric González-Segovia[2,3], Michael Deyholos[4], Quentin Cronk[2,3,5] & Marco Todesco[1,2,3,4] ✉

Stinging nettle (*Urtica dioica* L.) is a widespread weed of economic significance with a dioecious mating system. Previously, we generated a high-quality genome assembly of a diploid female plant, which showed extreme levels of structural variation between haplotypes. Here, we present a chromosome-level, haplotype-resolved sequence of a diploid male plant; since the male is believed to be the heterogametic sex in *Urtica dioica*, this assembly represents a first step towards elucidating the control of sex determination in this species. This independently assembled genome confirms three previously reported nettle genome features, including (1) a high degree of structural variation between haplotypes, including large inversions, (2) the likely existence of polycentric centromeres, and (3) the presence of urticaceous "pain peptide" sequences. Chromosome 8 stands out for its multiple large, nested inversions and high levels of repetitive sequences, features that are often associated with sex determining regions (SDRs). This chromosome is therefore a candidate for further investigations to characterize the sex determination in nettle.

Background & Summary

The stinging nettle (*Urtica dioica* L. ssp. *dioica*) is a widespread weed common throughout Europe, temperate Asia, and North America. It owes its name to its stinging hairs, which act as a powerful herbivore deterrent. It is often associated with nutrient-enriched human habitats as it has high phosphate demands, ceasing growth when phosphate is limiting, but responding vigorously to its addition[1]. Its natural habitat is probably rich alluvial river floodplains, where its phosphate demands are met by sediment deposition. More recently, *Urtica dioica* has exploited human disturbance and eutrophication to expand as an aggressive ruderal and weedy species. However, stinging nettle has also a long history of positive associations with humans and has been used as a source of medicine[2], food[3], and fibre[4]. Ecologically, it acts as a "super-host" for a multitude of invertebrates across Europe[5]. It is a crucial food source for numerous specialist and generalist insects, particularly within the orders Lepidoptera, Coleoptera, and Hemiptera[6], and for molluscs. Its attractiveness to invertebrate herbivores may be due to its status as a general mineral accumulator, containing high concentrations of calcium, nitrogen, and phosphorus in its tissues[7].

Urtica dioica ssp. *dioica* is a dioecious taxon (i.e. it has separate sexes on different plants). Evidence for male heterogamety was first obtained by Strasburger, who selfed a female plant that had formed some aberrant individual male flowers, obtaining only female progeny[8,9]. Subsequent studies reported complex inheritance patterns and significant maternal effects on seed sex ratios in dioecious *Urtica*, while monoecious individuals produced varying numbers of female flowers based on environmental conditions, suggesting a wholly genetic basis for sex determination in dioecious types but environmental influences in monoecious types[10,11].

The first published genome sequence of *U. dioica* was the haplotype-resolved assembly of a diploid female individual (clone 11-4)[12]. A haplotype from a tetraploid individual (the more common ploidy level for stinging nettle) was also subsequently generated[13]. Analysis of the diploid female genome assembly revealed several unusual features: (1) Extensive structural variation (SV) between the two haplotypes, particularly on chromosomes 1, 2, 3, and 8; (2) Presence of two types of centromeres, based on the analysis of patterns of repetitive sequences:

[1]Michael Smith Laboratories, University of British Columbia, 2185 East Mall, Vancouver, BC, V6T 1Z4, Canada. [2]Botany Department, University of British Columbia, Vancouver, Canada. [3]Biodiversity Research Centre, University of British Columbia, Vancouver, Canada. [4]Biology Department, University of British Columbia, Kelowna, BC, Canada. [5]Beaty Biodiversity Museum, University of British Columbia, Vancouver, Canada. ✉e-mail: mtodesco@msl.ubc.ca

acrocentric or near telocentric centromeres in five chromosomes (8, 9, 11, 12, 13) and polycentric centromeres in eight others (chr. 1–7, 10). The occurrence of polycentric centromeres had not been previously reported in the Urticaceae family, but is known in the closely related Moraceae; (3) Identification of two copies of a gene putatively encoding the sting peptide urthionin on chromosome 9, suggesting a recent gene duplication. Urthionin is a 42 amino acid-long peptide with cytolytic activity that contributes to nettle's sting[14].

Here, we report a chromosome-level, haplotype-resolved genome assembly for a diploid male individual, the putative heterogametic sex of stinging nettle. In this assembly, we observe many of the same genomic features mentioned above for the female stinging nettle genome assembly[12], confirming them and validating the correctness and accuracy of this new male stinging nettle genome assembly. This male assembly will be foundational to further studies aimed at characterizing the sex-determining region (SDR) in *Urtica* and determining the cause of the male heterogamety. We provide, for each chromosome, information on which haplotype is maternally- or paternally-derived, taking advantage of the fact that the male individual that we sequenced is the offspring of the female individual for which a genome assembly has been recently released[12]. This knowledge will help future efforts aimed at identifying the SDR of stinging nettle. In particular, a highly variable region in chromosome 8 presents several of the hallmarks of an SDR, including complex large inversions and a possible 8 Mbp insertion in the paternally-inherited haplotype, and will be a promising candidate for further studies on sex determination in stinging nettle.

Methods

Plant materials and sequencing. We selected a male plant (U48) from a progeny panel derived from a cross between two wild-collected *U. dioica* diploid cytotypes: 11-4 (female, River Jiu, north of Rovinari, Romania, for which we previously produced a genome assembly)[12] and 7-5 (male, River Struma, north of Boboshevo, Bulgaria)[15]. A voucher specimen has been deposited in the herbarium of the Beaty Biodiversity Museum at the University of British Columbia (UBC). Leaf tissue was frozen in liquid nitrogen and high-molecular weight DNA was isolated using a modified CTAB method[16]. A PacBio HiFi library was constructed and sequenced on a PacBio Revio instrument at Canada's Michael Smith Genome Sciences Centre (GSC) in Vancouver, BC, Canada, generating a total of 3.23 million HiFi reads (~77.5 Gbp; ~129X genome-wide coverage) with a mean quality of Q32 and mean length of 24,025 bp. Quality control of HiFi reads was performed using SMRT Link v13.1 with the command runqc-reports v10.4.6 (https://www.pacb.com/support/software-downloads/). We also prepared a Hi-C library according to the method described in[12]. The Hi-C library was sequenced with PE150 reads on an Illumina NovaSeq X Plus instrument at GSC, producing 108 million paired Hi-C reads (32.2 Gbp; ~54X genome-wide coverage).

***De novo* genome assembly and quality assessment.** Raw HiFi reads were filtered for >Q20 using fastq-filter v0.3.0 (https://github.com/LUMC/fastq-filter), resulting in 88.2% of the HiFi reads (~113X genome coverage) being retained and used for the *de novo* assembly. A preliminary k-mer analysis using Genomescope v1.0[17] with the k-mer histogram of >Q20 HiFi reads (parameter k = 21) revealed high levels of genome-wide heterozygosity (1.85%). Hi-C reads were quality-checked using fastqc v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and were not filtered or trimmed at this stage. To produce the draft haplotype-resolved genome assembly, we used Hifiasm v0.19.8-r603[18] with HiFi and Hi-C reads as input files (command: hifiasm -o Nettle_male_hifiasm.asm -t 48–h1 $HiC_1–h2 $HiC_2 $Hifi_fastq). Tuning parameters such as–hom-cov and -s did not improve the quality of the initial assembly; therefore, we proceeded with the assembly produced using the default settings.

We used Hi-C data to scaffold the two haplotype assemblies (male haplotype 1: MH1; male haplotype 2: MH2) independently, following the YaHS pipeline v1.2.2[19], which combines assisted scaffolding and manual examination of the contigs. In brief, Hi-C reads were mapped to the draft genome assembly using Juicer v1.9.9[20]. The resulting merged_nodups.txt file was then converted to a bed file using a custom awk script (available at https://github.com/ericgonzalezs/ASSEMBLIES/blob/main/Yahs) where each line has (1) contig ID a read was mapped to, (2) start of the alignment, (3) end of the alignment, (4) read ID with /1 and /2 denoting a paired read, and (5) mapping quality. To avoid inaccurate alignment for the initial scaffolding process, we retained and visualized only read alignments with mapping quality above 0. Next, the draft genome was scaffolded with YaHS's default parameters. We then manually examined the scaffolds for possible misassemblies using Juicebox v1.11.08[21], according to the YaHS manual curation pipeline, and assigned chromosomal boundaries. To further correct any switch errors (i.e. contigs that were placed in the wrong haplotype), we mapped Hi-C reads to the combined MH1 + MH2 assembly using Juicer and converted the merged_nodups.txt to Juicebox compatible files using the 3D-DNA pipeline v189022[22] with no mapping quality filter (parameter -q 0). Although allowing mapping quality 0 reads is generally not recommended because it can lead to erroneous alignment, we found that this significantly improved our ability to resolve repetitive regions when aligning Hi-C reads to both haplotypes at the same time. While we identified no obvious switch errors, we observed many small unassigned contigs (~100–200 kbp) that showed a strong association with a particular chromosome's centromeric or telomeric region (Supplementary Fig. 1). We positioned these contigs within the chromosome scaffolds manually, provided that the linear interaction pattern within the chromosome remained continuous after their insertion. Additionally, a large contig, which appeared to be mainly composed of a repetitive region, showed strong interactions only with the centre of chromosome 8 of MH2, and was manually inserted into the existing gap between the contigs within that chromosome. While these manual curation steps caused a noticeable decrease in contig N50 and scaffold N50, they produced a cleaner Hi-C interaction heatmap and haplotype assemblies that are likely to be more complete (Fig. 1).

Lastly, we used purge_dups v1.2.5[23] to remove duplicates and small contigs/unassembled reads that could not be placed within the main scaffolds. For this step, manual cutoff values (MH1: 5 48 78 96 158 360; MH2:

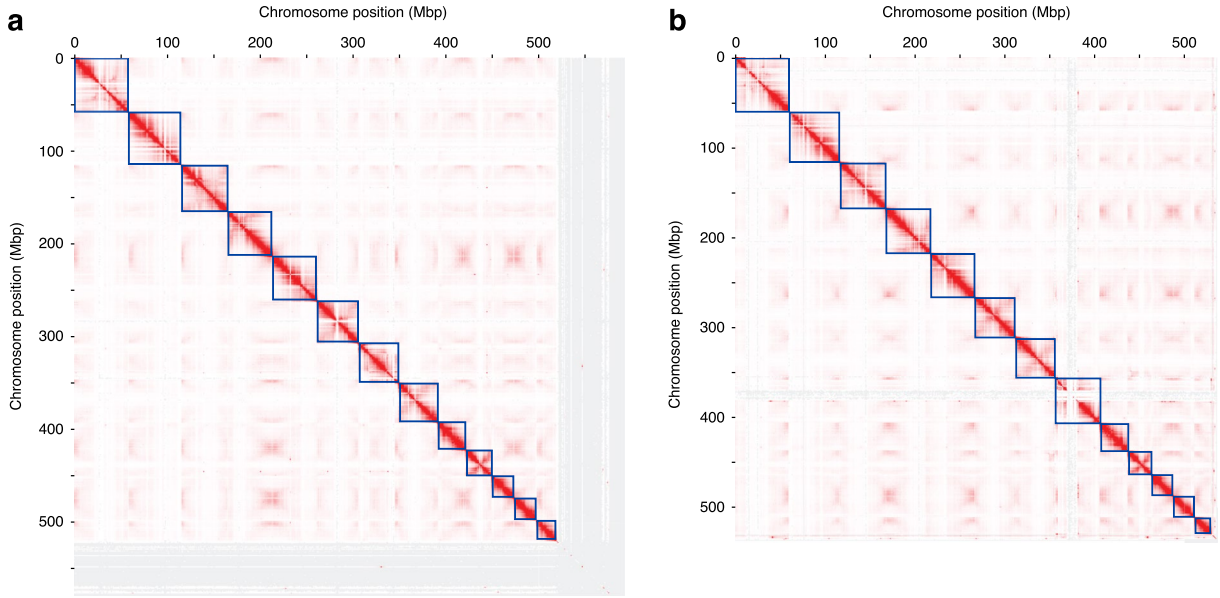

Fig. 1 Hi-C heatmap of the final assemblies for a) haplotype 1 – MH1, and b) haplotype 2 – MH2, of the *Urtica dioica* ssp. *dioica* male diploid genome. Dark blue lines are manually-assigned chromosome boundaries. Stronger Hi-C interactions are inferred by the intensity of the red colour.

Parameters\Haplotype	MH1	MH2	FH1	FH2
Total length (bp)	566,811,450	533,953,390	574,934,600	521,157,583
Number of contigs	770	289	1,598	376
Number of scaffolds	685	66	1,459	248
Conting N50 (Mbp)	21.14	11.17	10.89	13.53
Scaffold N50 (Mbp)	45.34	49.80	43.96	47.99
Number of scaffolds >50 kbp	558	23	236	76
% main genome in scaffolds >50 kbp	99.15	99.80	92.6	99.0
% genome anchored to 13 chromosomes	91.72	99.43	89.52	97.83
BUSCO (C%)	93.1	92.9	92.6	92.2
BUSCO (S%)	90.7	89.9	90.5	90.1
BUSCO (D%)	2.5	3.0	2.1	2.1
Number of protein coding genes annotated	20,237	20,431	20,333	20,140
Protein BUSCO (C%)	90.4	90.5	90.5	90.4
TE coverage (%)	69.51	68.72	69.14	68.59

Table 1. Genome assembly statistics of the male assembly haplotypes (MH1 and MH2) and the female assembly haplotypes (FH1 and FH2) published in[12].

5 59 59 60 155 360) were estimated based on the k-mer histogram produced from the >Q20 PacBio HiFi data, according to the purge_dups guidelines. An additional round of manual curation was performed using Juicer, YaHS, and Juicebox to make sure that both haplotype assemblies did not contain any obvious misassembly. Chromosome numbers and orientations in the finalized assemblies were then modified to match those of the female haplotype-resolved assembly[12]. Unscaffolded contigs contained several non-plant contaminations and were removed; no such contamination was detected in the chromosomal scaffolds. Assembly quality was assessed using BBMap v39.06[24] and BUSCO scores (dataset: eudicot_odb10) with BUSCO v5.1.2[25,26]. Detailed stats for all the rounds of assembly are reported in Supplementary Table 1.

The final scaffolded assemblies for MH1 and MH2 had a total length of 557,148,949 bp and 533,955,990 bp, consisting of 535 and 66 scaffolds (N50 = 45.3 Mbp and 49.8 Mbp), respectively; 93.3% and 99.4% of the total assembly length were scaffolded inside the 13 chromosomes of the stinging nettle genome. These assembly sizes are consistent with previous estimates of genome size for stinging nettle[15], and with the size of the published female stinging nettle assembly[12] (Table 1). BUSCO scores for the final assemblies showed high completeness (C ≈ 93%) and low duplication rates (D ≤ 3%; Table 1). Despite its relatively small size, the genome of stinging nettle contains a high amount of structural variants, especially on chromosomes 1, 2, 3, 6, and 8 (Fig. 2a); with the exception of chromosome 8, these structural variants coincide with regions of low gene density (Fig. 2a,b). To compare the two haplotypes at the nucleotide level, we used SyRI v1.7.0[27] and visualized the output synteny and structural variations, including inversions, duplications, and translocations, using Plotsr[28] (Supplementary Fig. 2).

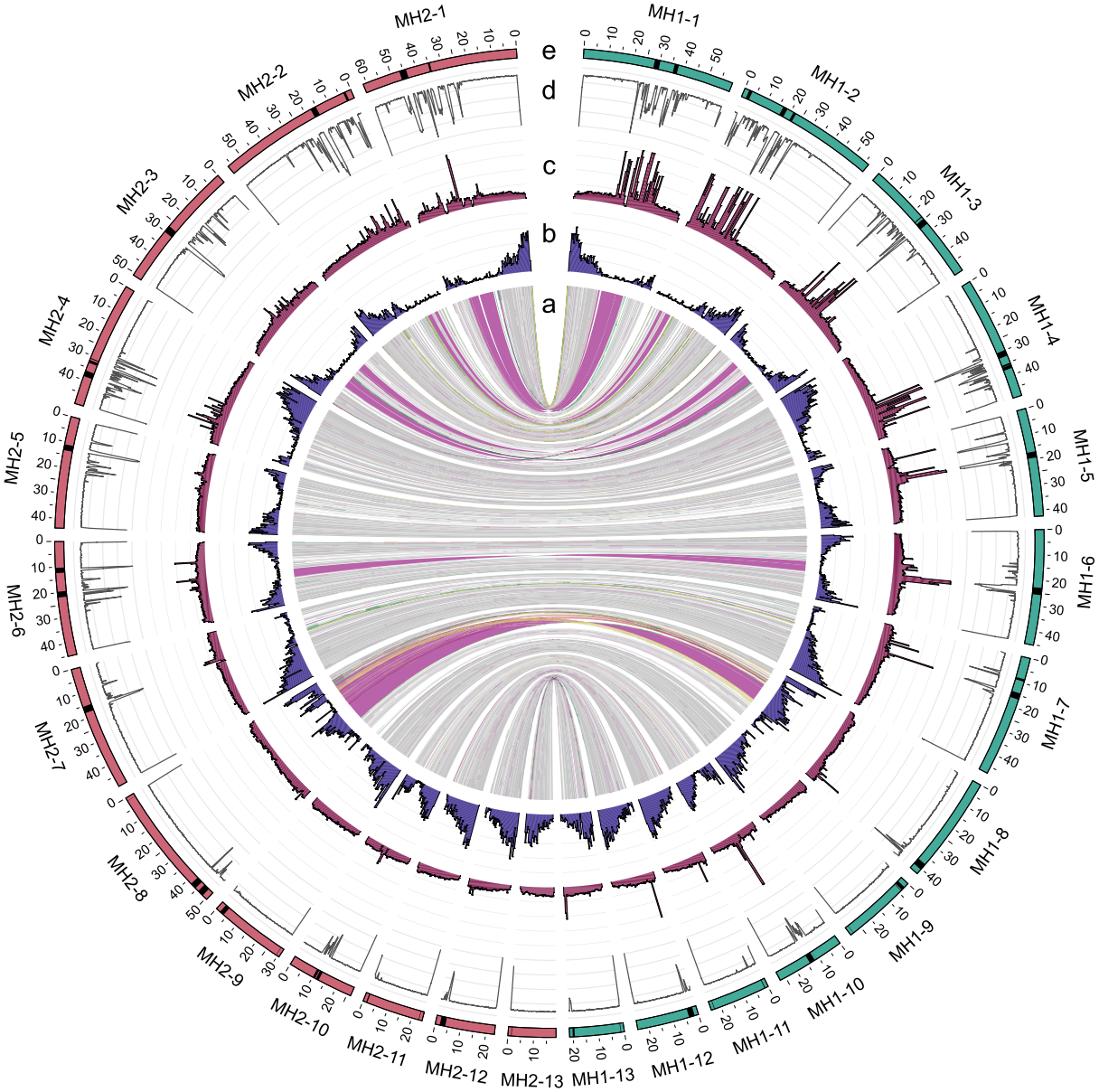

Fig. 2 Circos plot for the male nettle genome assembly. From the inside out, tracks show (**a**) structural variants identified by SyRI: grey = syntenic regions, purple = inversions, green = translocations, yellow = duplications, mint = inverted translocations, pink = inverted duplications; (**b**) gene density in 500 kbp windows; (**c**) transposable elements (TE) density in 500 kbp windows; (**d**) Shannon diversity of repeats mean score per 250 kbp window; (**e**) putative centromeric locations (black bands) based on RepeatOBserver Shannon diversity score. Haplotype 1 is shown in green on the right, haplotype 2 in pink on the left. The maximum y-axis value for the TE counts was set to 4000 to show variation across the genome, resulting in a total of 25 windows that exceed the plot's range. TE counts for these windows with >4000 TEs are reported in Supplementary Table 4.

Gene and repeat annotations. We performed whole genome gene and repeat annotation using BRAKER3 v3.0.8[29,30] and Extensive Denovo TE Annotator v2.2.1[31] pipelines, respectively. To do so, we first soft-masked the genome with the Red command: redmask.py v0.0.2[32]. Publicly available RNAseq reads from *Urtica dioica*[33] were downloaded and filtered with Trimmomatic v0.39 (parameters: ILLUMINACLIP: TruSeq. 3-PE.fa:2:30:10:2:True SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36)[34], then aligned to our genome using Hisat2 v2.2.1[35]. The resulting output was then converted to BAM format using samtools v1.17[36]. Using the AUGUSTUS parameters that we previously generated for the female nettle assembly resulted in a less complete annotation of the male assembly. Therefore, we ran the braker.pl command with the new species flag–species Urtica_dioica_male to train AUGUSTUS, with–prot_seq 'Viridiplantae' dataset downloaded from OrthoDB[37,38],–bam RNAseq alignment file, and–softmasking options. This means that the male MH1 was annotated *de novo* and independently of the female assembly annotation data. To annotate the male MH2 haplotype, we used the same–species Urtica_dioica_male with–useexisting flag to carry over the AUGUSTUS training parameters from MH1.

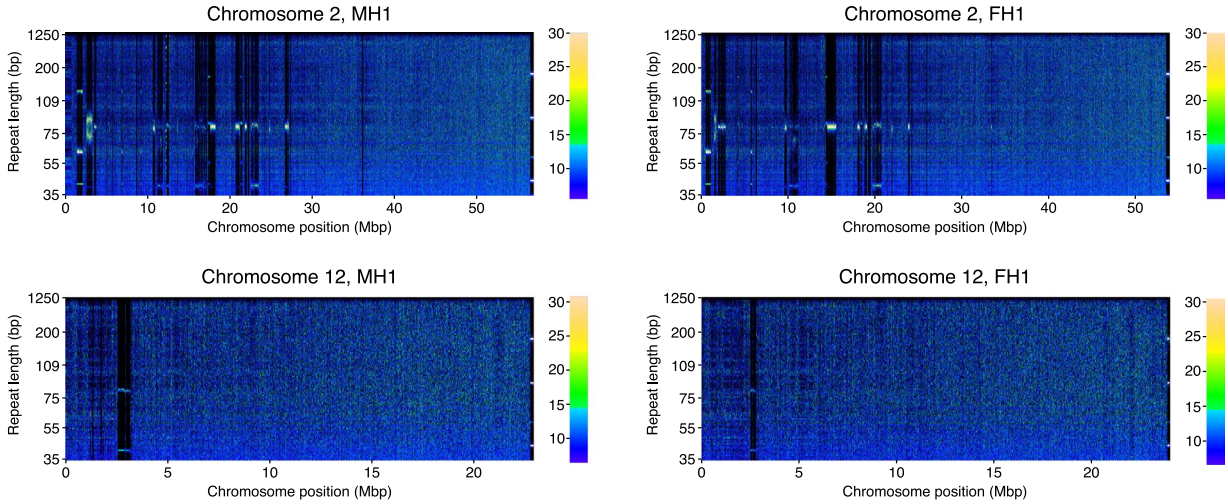

Fig. 3 Chromosome-wide pattern of tandem repeats obtained with RepeatOBserver, showing representative nettle chromosomes displaying different centromeric behaviours: polycentric (chromosome 2) and acrocentric (chromosome 12). Fourier transformed repeat spectra are shown for the H1 haplotype from the male assembly (MH1, current study) on the left and for the H1 haplotype from the female assembly (FH1[12]) on the right. Colour intensity corresponds to the number of times a specific repeat is found in a 5 kbp window. Regions with high proportions of a specific tandem repeat show characteristic banding patterns (bands corresponding to larger repeats sizes are harmonics of the smaller base repeat). Fourier transformed repeat spectra for all chromosomes are reported in Supplementary Fig. 3.

For TE annotation, we ran EDTA with default parameters (command: perl EDTA.pl–genome EDTA_input/Nettle_male_H1_genome.fa–sensitive 1–anno 1–overwrite 1). EDTA is a comprehensive pipeline that has optimal set of various TE annotation programs (LTRharvest, LTR_FINDER, LTR_retriever, GRF, TIR-Learner, HelitronScanner, and RepeatModeler), determined based on the benchmarking test using manually curated, high-quality TE libraries from rice[31]. It can perform the TE annotation fully *de novo* without additional supply of known TEs by running each TE annotation program of different TE category independently, then filtering the resulting TE libraries curated by accuracy and redundancy. We did not supply TEs identified in the nettle female assembly to annotate the male assembly to avoid inflating false discoveries since the TEs in the female genome were also fully *de novo* annotated without further validation. Annotation of putative pain peptides was performed as previously described[12], based on the peptide sequences published in[14,39]. The results of orthologous hits are presented in Supplementary Table 2.

Gene annotation identified 20,237 unique genes and 22,384 transcripts in MH1, with protein BUSCO scores of 90.4% (C), 79.7% (S), and 10.7% (D). Similarly, the MH2 annotation contained 20,431 unique genes and 22,607 transcripts, with BUSCO scores of 90.5% (C), 79.4% (S), 11.2% (D) (Table 1; Supplementary Table 3). We identified 530,940 and 611,955 TEs spanning 393,960,178 bp (69.51%) and 366,876,778 bp (68.72%) of the genome in MH1 and MH2, respectively (Supplementary Table 4). The most abundant type of TEs was Long Terminal Repeat (LTR) retrotransposons [MH1: 49.19% (16.00% Copia, 9.32% Gypsy, 23.87% unknown) and MH2: 47.83% (13.52% Copia, 10.82% Gypsy, 23.49% unknown)], followed by Terminal Inverted Repeats (TIR; MH1: 8.36%, MH2: 15.47%).

We used RepeatOBserver v0.1.0[40] to generate Fourier heatmaps for each chromosome. The resulting tandem repeat patterns were consistent with the presence of polycentric chromosomes in the nettle genome (chromosomes 1-7), alongside metacentric (chromosome 10) and acrocentric or near telomeric centromeres (chromosomes 8, 9, 11, 12, 13; Fig. 3; for all 13 chromosomes, see Supplementary Fig. 3). A circos plot[41] was generated to visualize gene/repeat counts per 500 kbp windows, as well as the mean Shannon diversity scores from RepeatOBserver averaged per 250 kbp windows (Fig. 2d). We also used RepeatOBserver with default cut-off values to identify putative centromeric regions (Fig. 2e, Supplementary Table 5).

Parental assignment of haplotypes. We took advantage of the fact that the female individual sequenced in[12] is the mother of the newly sequenced male individual to identify which of its haplotypes are maternally- or paternally-derived. We used minimap2 v2.24-r1122[42] to perform pairwise alignments of four haplotypes from the *U. dioica* ssp. *dioica* female[12] and male (this study) individuals, denoted as female H1 (FH1), female H2 (FH2), male H1 (MH1), male H2 (MH2), to the two female haplotypes (FH1 and FH2). We expected one of the haplotypes of the male assembly to be identical to one of the haplotypes of the female assembly, or to be a mix of the two female haplotypes in cases where recombination has occurred between them. To verify this, we filtered the alignment file between female and male haplotypes to retain hits with quality >30 and calculated the percent identity of the nucleotide match as (number of matching nucleotides/alignment length) × 100 for each alignment detected. We then plotted the alignment results filtered by ≥95% matching nucleotides to identify the maternally-inherited (and, by exclusion, the paternally-inherited) chromosomes in the male nettle assembly (Fig. 4, Supplementary Fig. 4). In cases where there appear to be no recombination (chromosomes 1 – 3,

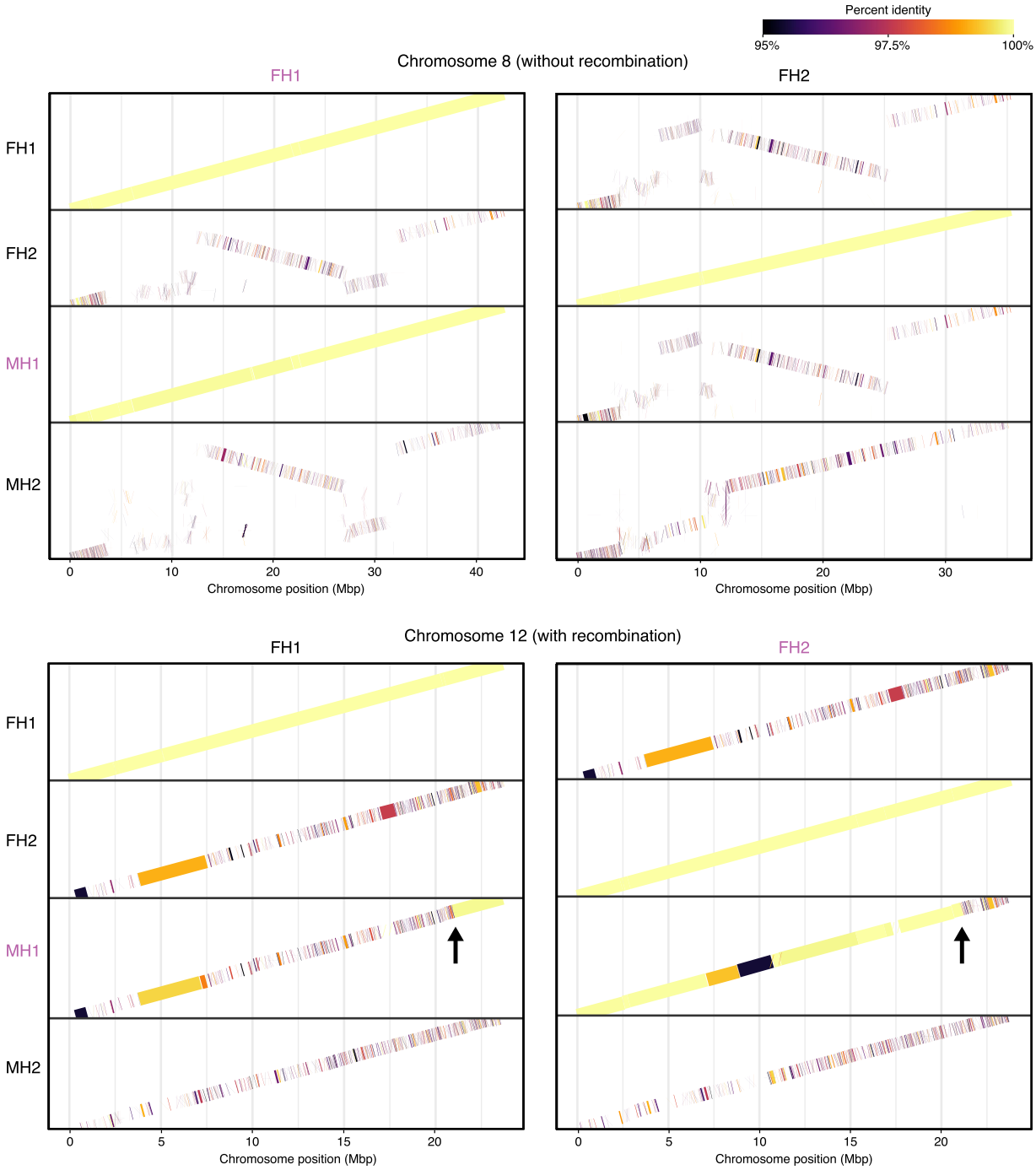

Fig. 4 Comparison of four haplotypes; FH1 = female haplotype 1, FH2 = female haplotype 2, MH1 = male haplotype 1, MH2 = male haplotype 2. The two panel plots show the pairwise alignments at the nucleotide level using FH1 (left) and FH2 (right) as a reference for each chromosome pair. Matching chromosome pairs are highlighted in pink. An example of parental assignment in the absence (top) and presence of visible recombination events (bottom, highlighted by arrows) is shown. The alignments are coloured by nucleotide match in percent identity; the lighter the colour, the higher the nucleotide match over an alignment, which is represented by a solid line connecting the start to end of the alignment (i.e., yellow = 100%, black/dark purple = 95% aligned segment). Since we are only visualizing alignment hits above 95% nucleotide match, the darkest colour represents the 95% match, and lines are absent if there is no region with >95% nucleotide alignment. Pairs of male and female haplotypes that have the highest similarity (coloured in yellow) are inferred to be the maternally-inherited male haplotype and its corresponding haplotype in the female assembly. The full list of chromosome assignments is found in Table 2 and plots for all chromosomes are found in Supplementary Fig. 4.

Chromosome	Proportion of MH1 aligning with		Proportion of MH2 aligning with		Male haplotypes	
	FH1	FH2	FH1	FH2	Maternal haplotype	Paternal haplotype
1	1.2%	98.4%	0.6%	0.2%	MH1	MH2
2	2.6%	97.4%	0.0%	0.2%	MH1	MH2
3	0.0%	0.0%	87.3%	4.8%	MH2	MH1
4	46.4%	23.1%	1.6%	0.3%	MH1	MH2
5	0.0%	0.1%	84.8%	19.2%	MH2	MH1
6	0.0%	0.0%	99.6%	2.1%	MH2	MH1
7	86.2%	1.3%	0.1%	0.0%	MH1	MH2
8	99.5%	0.5%	0.3%	0.3%	MH1	MH2
9	29.9%	63.6%	0.2%	0.9%	MH1	MH2
10	34.1%	13.6%	1.0%	0.0%	MH1	MH2
11	7.5%	93.1%	0.2%	0.2%	MH1	MH2
12	11.6%	72.0%	0.0%	0.1%	MH1	MH2
13	0.8%	47.8%	0.0%	0.0%	MH1	MH2

Table 2. Maternal and paternal chromosome assignment to the male genome assembly based on alignment to the maternal female genome assembly[12]. Maternally-inherited chromosomes were identified based on visual inspection of the pairwise alignment plot coloured by the percent nucleotide identity (Fig. 4, Supplementary Fig. 4), where FH1 and FH2 were used as reference. Higher alignment rates between of a male chromosome haplotype with its female counterpart was considered as evidence of maternal inheritance. To calculate the proportion of a male haplotype (MH1/MH2) in perfect alignment with the female haplotypes (FH1/FH2), we used >99.7% alignment blocks and calculated the (total length of alignment/total chromosome length × 100%) for each chromosome pair.

6 – 8), near-perfect alignment was observed between one of the female haplotypes and one of the male haplotypes, but not in pairwise comparisons with the other male haplotype. For example, in chromosome 8 (Fig. 4), only MH1 has a complete, identical nucleotide match to FH1, while MH2 does not align perfectly to FH2. The MH2 haplotype does not align perfectly to either FH1 or FH2. This indicates that, for chromosome 8, MH1 is maternally-derived and MH2 is paternally-derived. This analysis also identified multiple recombination events, as shown for instance in chromosome 12 (Fig. 4). We observed a clear indication of one recombination event each near the extremities of chromosomes 4, 5, 10, 11, and 12, where high similarity of the male haplotype switches from one to the other female haplotype. A larger recombined region, corresponding to more than one-third of the chromosome, was found in chromosome 9. While most recombination events were easily detectable, a few possible cases were difficult to assign confidently based only on the alignment percentage distribution. For example, a recombination seemed to have occurred at the beginning of chromosome 13, but in the putatively recombined region percent identity is high with either FH1 or FH2 (~97%). We also note that a percent identity was slightly lower around centromeric regions, possibly due to poor alignment caused by the presence of extensive repeats. Contrary to expectations[43] we did not observe recombination in several chromosomes. Intriguingly, recombination events were rarer in chromosomes with predicted polycentric centromeres (two recombination events in seven chromosomes) than in chromosomes with metacentric or acrocentric ones (five recombination events in six chromosomes). While based on very limited data (one generation of recombination in a single individual), this observation is consistent with known patterns of reduced recombination in holocentric chromosomes[44]. The inferred maternal and paternal chromosome assignment is summarised in Table 2. To quantify how much of each chromosome was inherited from the corresponding maternal haplotype in the male offspring, we filtered the alignment file to retain only alignments with nucleotide match >99.7% and divided the sum of the length of these alignments in each chromosome by the total length of that chromosome (= total alignment length per chromosome/chromosome length × 100; Table 2).

Data Records

The raw sequence data (PacBio HiFi, Illumina Hi-C) reported in this paper have been deposited in the NCBI Sequence Read Archive (SRA) under the accession numbers SRR35053641[45] (PacBio HiFi) and SRR35053640[46] (Hi-C) with the BioProject accession PRJNA1308592[47], BioSample accession SAMN50702769[48]. The assembled genomes have been deposited in the NCBI GenBank under the GenBank accessions JBQRAE000000000[49] (MH1; PRJNA1308592[47]), JBQRAF000000000[50] (MH2; PRJNA1308630[51]). We have also deposited the corresponding files, including all the supplementary tables, supplementary figures, genome assemblies, genome annotation, and TE annotation files, at Figshare: https://figshare.com/projects/A_high-quality_phased_genome_assembly_of_stinging_nettle_Urtica_dioica_ssp_dioica/230981.

Technical Validation

The availability of a previous, independently generated diploid assembly for a female stinging nettle individual allows direct validations of many of the features of the male assembly presented here. As mentioned above, we performed pairwise comparisons between the two male and the two female haplotypes (Fig. 4; Supplementary Fig. 4). Additionally, we used GENESPACE v1.3.1[52] to independently compare synteny

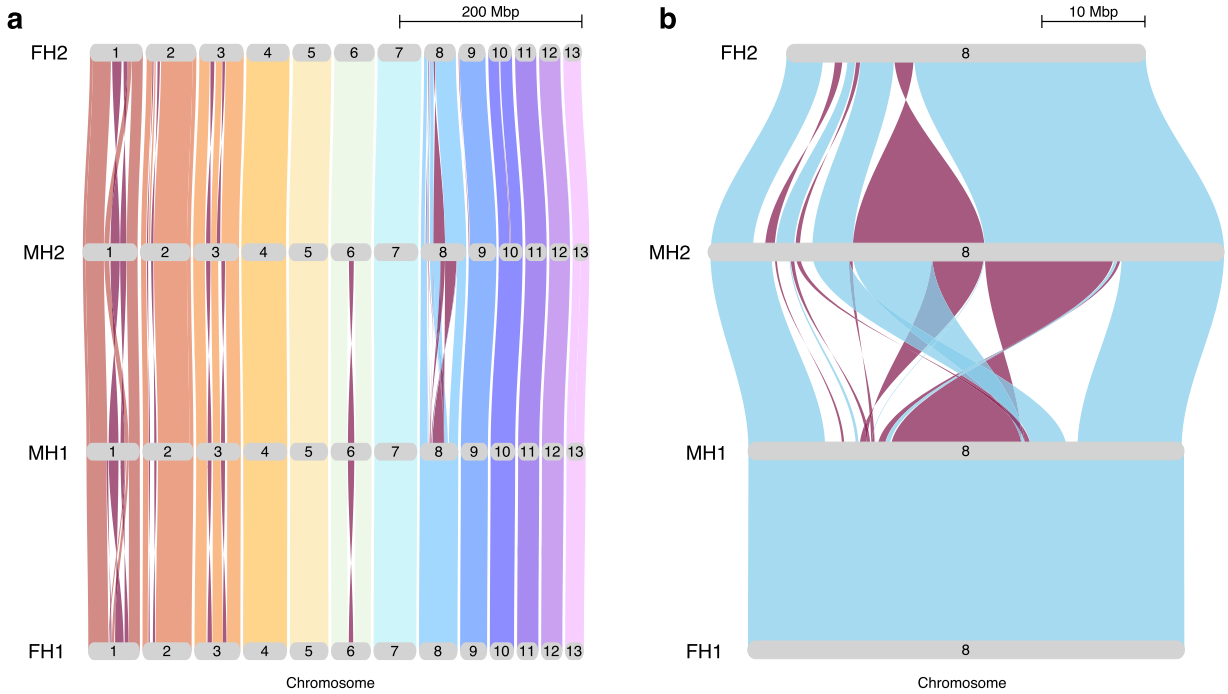

Fig. 5 Riparian plots for the four haplotypes of the male and female nettle assemblies showing (**a**) all 13 chromosomes and (**b**) chromosome 8. The plot was created in GENESPACE by comparing the order of gene blocks across the haplotypes (blkSize = 5, blkRadius = 5). Purple = inverted blocks for both panels. Gene order synteny between the four haplotypes is shown by the coloured ribbons between the chromosomes.

between these haplotypes based on gene order, by using the transcripts annotated on the four haplotypes as input files (Fig. 5a). To run GENESPACE, we used OrthoFinder v2.5.4[53] to find orthogroups and then used MCScanX[54] to find synteny blocks. These analyses confirmed that most of the regions of high structural variability found between haplotypes in the male assembly correspond to regions of high structural variability in comparisons with and between the female haplotypes presented in[12]. This supports the correctness of the male assembly and further highlights the large extent of interspecific structural variation found in stinging nettle. It also identifies a highly variable region of chromosome 8 as a candidate for the SDR in stinging nettle; this region displays multiple multi-Mbp chromosomal inversions and other rearrangements, including a possible 8 Mbp insertion in the paternally-inherited haplotype that is not observed in the maternal haplotypes (Fig. 5b). Additionally, comparison of RepeatOBserver analyses found that the same putative centromere type (polycentric, metacentric, acrocentric) was predicted for the different chromosomes between male and female assemblies of stinging nettle. Finally, two copies of genes putatively encoding the pain peptide urthionin A[14] were found on chromosome 9 in both haplotypes (Supplementary Table 2), as is the case in the female stinging nettle genome assembly.

Data availability

Raw sequencing data and assembled genomes are available at NCBI under BioProject/accession numbers PRJNA1308592/JBQRAE000000000 (haplotype 1; MH1) and PRJNA1308630/JBQRAF000000000 (haplotype 2; MH2). The male plant used in this study has the BioSample ID: SAMN50702769, and is vouchered as a specimen in the herbarium at UBC Beaty Biodiversity Museum. PacBio HiFi reads and Illumina Hi-C reads are provided in.fastq format at the Sequence Read Archive with accession numbers SRR35053641 and SRR35053640, respectively. The genome assemblies and the genome annotations are also deposited in Figshare: https://figshare.com/projects/A_high-quality_phased_genome_assembly_of_stinging_nettle_Urtica_dioica_ssp_dioica/230981.

Code availability

All the code used for this project is an extension from previous work[12] and is found at https://github.com/kaede0e/stinging_nettle_genome_assembly. This includes in particular: the full pipeline for the assembly presented in this paper in 1_genome_assembly_with_PBHiFi_HiC/how_to_run_assembly_workflow_v2.md; the gene annotation pipeline in 2_annotation/how_to_run_BRAKER3_with_conda.sh; the R script for pairwise haplotype alignment visualization in 3_comparative_genomics/PLOT_cont2_allCHR_PER.R; and the coupled calculation for maternal chromosome assignment in 3_comparative_genomics/calculate_chromosome_percent_inheritance.sh.

Received: 5 September 2025; Accepted: 23 February 2026;
Published online: 02 March 2026

References

1. Taylor, K. Biological Flora of the British Isles: *Urtica dioica* L. *Journal of Ecology* **97**, 1436–1458, https://doi.org/10.1111/j.1365-2745.2009.01575.x (2009).
2. Grauso, L., de Falco, B., Lanzotti, V. & Motti, R. Stinging nettle, *Urtica dioica L.*: botanical, phytochemical and pharmacological overview. *Phytochemistry Reviews* **19**, 1341–1377, https://doi.org/10.1007/s11101-020-09680-x (2020).
3. Mohammadian, M., Biregani, Z. M., Hassanloofard, Z. & Salami, M. Nettle (*Urtica dioica* L.) as a functional bioactive food ingredient: Applications in food products and edible films, characterization, and encapsulation systems. *Trends Food Sci Technol* **147**, 104421, https://doi.org/10.1007/s13197-015-1916-y (2024).
4. Xu, X. *et al*. Cell wall composition and transcriptomics in stem tissues of stinging nettle (*Urtica dioica* L.): Spotlight on a neglected fibre crop. *Plant Direct* **3**, e00151, https://doi.org/10.1002/pld3.151 (2019).
5. Wonglersak, R., Cronk, Q. & Percy, D. Salix transect of Europe: structured genetic variation and isolation-by-distance in the nettle psyllid, *Trioza urticae* (Psylloidea, Hemiptera), from Greece to Arctic Norway. *Biodivers Data J* **5**, https://doi.org/10.3897/bdj.5.e10824 (2017).
6. Perrin, R. M. The role of the perennial stinging nettle, *Urtica dioica*, as a reservoir of beneficial natural enemies. *Annals of Applied Biology* **81**, 289–297, https://doi.org/10.1111/j.1744-7348.1975.tb01644.x (1975).
7. Müllerová, V., Hejcman, M., Hejcmanová, P. & Pavlů, V. Effect of fertilizer application on *Urtica dioica* and its element concentrations in a cut grassland. *Acta Oecologica* **59**, 1–6, https://doi.org/10.1016/j.actao.2014.05.004 (2014).
8. Strasburger, E. Sexuelle und apogame Fortpflanzung bei Urticaceen. *Jahrbücher für wissenschaftliche Botanik* **97** (1910).
9. Correns, C. *Bestimmung, vererbung und verteilung des geschlechts bei den höheren pflanzen*. Handbuch der vererbungswissenschaft, Bd. 2, C. (Berlin, Gebrüder Borntraeger, 1928).
10. Shannon, R. K. & Holsinger, K. E. The genetics of sex determination in stinging nettle (*Urtica dioica*). *Sex Plant Reprod* **20**, 35–43, https://doi.org/10.1007/s00497-006-0041-5 (2007).
11. Glawe, G. A. & De Jong, T. J. Inheritance of progeny sex ratio in *Urtica dioica*. *J Evol Biol* **20**, 133–140, https://doi.org/10.1111/j.1420-9101.2006.01215.x (2007).
12. Hirabayashi, K. *et al*. A high-quality phased genome assembly of stinging nettle (*Urtica dioica* ssp. *dioica*). *Plants* **14**, 124, https://doi.org/10.3390/plants14010124 (2025).
13. Christenhusz, M. J. M. & Leitch, I. J. The genome sequence of the stinging nettle, *Urtica dioica* L. (Urticaceae). *Wellcome Open Res* **10**, 31, https://doi.org/10.12688/wellcomeopenres.23399.1 (2025).
14. Xie, J. *et al*. Neurotoxic and cytotoxic peptides underlie the painful stings of the tree nettle *Urtica ferox*. *Journal of Biological Chemistry* **298**, 102218, https://doi.org/10.1016/j.jbc.2022.102218 (2022).
15. Cronk, Q., Hidalgo, O., Pellicer, J., Percy, D. & Leitch, I. Salix transect of Europe: variation in ploidy and genome size in willow-associated common nettle, *Urtica dioica* L. sens. lat., from Greece to arctic Norway. *Biodivers Data J* **4**, https://doi.org/10.3897/bdj.4.e10003 (2016).
16. Stoffel, K. *et al*. Development and application of a 6.5 million feature Affymetrix Genechip® for massively parallel discovery of single position polymorphisms in lettuce (*Lactuca* spp.). *BMC Genomics* **13**, 185, https://doi.org/10.1186/1471-2164-13-185 (2012).
17. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
19. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808, https://doi.org/10.1093/bioinformatics/btac808 (2023).
20. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
21. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
22. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
23. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
24. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. *Lawrence Berkeley National Laboratory*. LBNL Report #: LBNL-7065E. Retrieved from https://escholarship.org/uc/item/1h3515gn (2014).
25. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
26. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
27. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 1–13, https://doi.org/10.1186/s13059-019-1911-0 (2019).
28. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **btac196**. https://doi.org/10.1093/bioinformatics/btac196 (2022).
29. Gabriel, L. *et al*. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *Genome Res* **34**, 769–777, https://doi.org/10.1101/gr.278090.123 (2024).
30. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
31. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 1–18, https://doi.org/10.1186/s13059-019-1905-y (2019).
32. Girgis, H. Z. Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16**, 1–19, https://doi.org/10.1186/s12859-015-0654-5 (2015).
33. Xu, X., Legay, S., Berni, R., Hausman, J. F. & Guerriero, G. Transcriptomic changes in internode explants of stinging nettle during callogenesis. *Int J Mol Sci* **22**, https://doi.org/10.3390/ijms222212319 (2021).
34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **btu170**, https://doi.org/10.1093/bioinformatics/btu170 (2014).
35. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
36. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4, https://doi.org/10.1093/gigascience/giab008 (2021).
37. Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M. & Kriventseva, E. V. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* **41**, D358–D365, https://doi.org/10.1093/nar/gks1116 (2013).
38. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445–D451, https://doi.org/10.1093/nar/gkac998 (2023).

39. Gilding, E. K. *et al*. Neurotoxic peptides from the venom of the giant Australian stinging tree. *Sci Adv* **6**, eabb8828, https://doi.org/10.1126/sciadv.abb8828 (2020).
40. Elphinstone, C., Elphinstone, R., Todesco, M. & Rieseberg, L. H. RepeatOBserver: Tandem repeat visualisation and putative centromere detection. *Mol Ecol Resour* **25**, e14084, https://doi.org/10.1111/1755-0998.14084 (2025).
41. Krzywinski, M. I. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645, https://doi.org/10.1101/gr.092759.109 (2009).
42. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
43. Brazier, T., Stetsenko, R., Roze, D. & Glémin, S. Mating system and the evolution of recombination rates in seed plants. *J Evol Biol* **38**, 920–929, https://doi.org/10.1093/jeb/voaf008 (2025).
44. Zedek, F. *et al*. Chromosome size as a robust predictor of recombination rate: Insights from holocentric and monocentric systems. *Genetics* **232(1)**, iyaf247, https://doi.org/10.1093/genetics/iyaf247 (2026).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35053641 (2025).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35053640 (2025).
47. *NCBI BioProject* https://identifiers.org/ncbi/bioproject:PRJNA1308592 (2025).
48. *NCBI BioSample* https://identifiers.org/ncbi/biosample:SAMN50702769 (2025).
49. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBQRAE000000000 (2025).
50. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBQRAF000000000 (2025).
51. *NCBI BioProject* https://identifiers.org/ncbi/bioproject:PRJNA1308630 (2025).
52. Lovell, J. T. *et al*. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, e78526, https://doi.org/10.7554/elife.78526 (2022).
53. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 1–14, https://doi.org/10.1186/s13059-019-1832-y (2019).
54. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49, https://doi.org/10.1093/nar/gkr1293 (2012).

Acknowledgements

We thank the Digital Research Alliance of Canada (DRAC) for access to their computational resources. We acknowledge the following funding sources: Natural Science and Engineering Research Council (NSERC) Discovery Grants to MT (RGPIN-2023-03344) and QC (RGPIN-2019-04041), and the Natural History Museum, London (UK) for funding to DP for fieldwork in Europe.

Author contributions

Conceptualization: Q.C. and M.D.; funding and resources: Q.C., D.P. and M.T.; data production: K.H.; formal analyses, investigation, and visualization, K.H., E.G. and M.T.; sample collection: D.P. and Q.C.; genotype maintenance and crossing, Q.C.; sample preparation and laboratory work: K.H.; writing, review and editing: all authors. All the authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.T.