

# TITLE

Genomic analyses demonstrate the absence of genetic sex determination in the dioecious conifer  
*Taxus baccata*

# AUTHORS

Daniel Bross<sup>1</sup>, Jannes Mittelbach<sup>1,2</sup>, Martin Pippel<sup>3,4,5</sup>, Malte Mader<sup>1</sup>, Desanka Lazić<sup>1</sup>, Laura Uelze<sup>3,4</sup>,  
Hilke Schroeder<sup>1</sup>, Emilia Pers-Kamczyc<sup>6</sup>, Stefan Kurtz<sup>2</sup>, Sylke Winkler<sup>3,4</sup>, Niels A. Müller<sup>1\*</sup>, Birgit  
Kersten<sup>1\*</sup>

<sup>1</sup> Thünen Institute of Forest Genetics, Großhansdorf, Germany

<sup>2</sup> Faculty of Mathematics, Informatics and Natural Sciences, ZBH - Center for Bioinformatics,  
Universität Hamburg, Hamburg, Germany

<sup>3</sup> Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

<sup>4</sup> Dresden-concept Genome Center, c/o CMCB Center for Molecular and Cellular Bioengineering,  
Technology Platform of the TUD Dresden University of Technology, Dresden, Germany

<sup>5</sup> SciLifeLab, National Bioinformatics Infrastructure Sweden (NBIS), Department of Cell and Molecular  
Biology, Uppsala University, Uppsala, Sweden

<sup>6</sup> Institute of Dendrology, Polish Academy of Sciences, Kórnik, Poland

\* corresponding authors

# ORCID

|                     |                     |
|---------------------|---------------------|
| Daniel Bross        | 0009-0006-3356-7136 |
| Jannes Mittelbach   | 0009-0004-6180-3277 |
| Martin Pippel       | 0000-0002-8134-5929 |
| Malte Mader         | 0000-0002-4808-2717 |
| Desanka Lazić       | 0000-0002-5932-6019 |
| Laura Uelze         | 0000-0002-6020-3841 |
| Hilke Schroeder     | 0000-0002-0908-3397 |
| Emilia Pers-Kamczyc | 0000-0002-5610-2124 |
| Stefan Kurtz        | 0000-0001-5783-0054 |
| Sylke Winkler       | 0000-0002-0915-3316 |
| Niels A. Müller     | 0000-0001-5213-042X |
| Birgit Kersten      | 0000-0001-9900-9133 |

## ABSTRACT

Hundreds of plant lineages have independently evolved dioecy, i.e., separation of female and male flowers on different individuals. In all dioecious plants investigated at the molecular level to date, sex is determined genetically through a sex-determining region (SDR). SDRs have mostly been studied in angiosperms, although dioecy is relatively more common among gymnosperms. Here, we investigate sex determination in the gymnosperm *Taxus baccata*. We assembled four haplotype-resolved chromosome-level genomes for one female and one male tree, with an average size of 10.04 Gb, and generated resequencing data for 100 phenotypically sexed individuals. Strikingly, *k*-mer analyses, genome-wide association studies and differential coverage analyses demonstrate the absence of an SDR in the *T. baccata* genome. This indicates a non-genetic mechanism of sex determination, most likely via a sex-specific epiallele. Given that *T. baccata* is the first species studied among a large group of conifers, our findings suggest that such a mechanism might be widespread.

## MAIN

### > Introduction

Dioecious plants have evolved a variety of sex-determining systems<sup>1,2</sup>. These range from the well-known X/Y or Z/W systems, also found, for example in mammals and birds, respectively, to X/A balance (e.g., *Rumex hastatulus*<sup>3</sup>, *Humulus* and *Cannabis*<sup>4</sup>), or U/V systems (e.g., *Marchantia polymorpha*<sup>5</sup>). Dioecy has evolved many times independently in plants<sup>6-8</sup>, and reversions from a dioecious to a monoecious or hermaphroditic system have also been reported<sup>9,10</sup>. New genomic methods finally allow the identification and characterization of the underlying sex-determining region (SDR), that is, the genomic region in which the sex-determining genes are located<sup>11</sup>. Notably, most studies on SDRs so far were conducted in angiosperms. While dioecy is present only in 6% of angiosperms<sup>12</sup>, it is common among gymnosperms, where about 65% of species are dioecious<sup>6</sup>. Within the gymnosperms, the SDRs have been studied at a molecular level in *Ginkgo biloba*, *Cycas panzhihuaensis* and *Welwitschia mirabilis*, with all three species exhibiting an X/Y sex-determining system<sup>13-15</sup>. Beyond this, several dioecious gymnosperms have been studied by karyotyping (reviewed in Ohri and Rastogi<sup>16</sup>).

According to the WFO Plant List (<http://www.worldfloraonline.org/taxon/wfo-4000037650><sup>17</sup>) the gymnosperm genus *Taxus* consists of 13 accepted species, of which 12 are dioecious and only *T. canadensis* is monoecious<sup>18</sup>. While cosexual individuals were reported in *T. brevifolia*, where DiFazio<sup>19</sup> found female reproductive structures on 29.3% of predominantly male trees (see also Hogg *et al.*<sup>20</sup>), cosexuals appear to be exceedingly rare in most cases. In *T. cuspidata*, for example, Allison *et al.*<sup>21</sup> found no cosexual individual in three studied populations. For *T. baccata*, Iszkuło and Jasińska<sup>22</sup> reported a fraction of 0.13% cosexual individuals in seven Central European populations, in which singular branches of male trees formed female strobili in addition to male strobili. Moreover, populations of *T. baccata* appear to generally have an even sex ratio<sup>23</sup>.

While *Taxus* has received attention from numerous researchers in recent years, mainly due to its capacity to produce the anticancer drug paclitaxel (e.g., <sup>24,25</sup>), the sex-determining systems in this genus remain elusive. Recently, Li *et al.*<sup>26</sup> published a reference genome for *T. wallichiana* and proposed an Z/W system, with a putative sex-determining region located on chromosome 12. However, this assumption is based on indirect evidence such as length differences between the two reference haplotypes. For *T. baccata*, a karyotype analysis by Tomasino *et al.*<sup>27</sup> showed the absence of heteromorphic sex chromosomes and Zarek<sup>28</sup> found no sex-specific marker.

In this study, we aimed to elucidate the sex-determining system of *T. baccata* by taking advantage of technological advancements in genome sequencing. Specifically, we generated high-quality long-read (PacBio HiFi) sequence data and chromatin conformation (Hi-C) data to assemble chromosome-level haplotype-resolved genome assemblies for a female and a male individual of *T. baccata*. Using whole-genome short-read sequencing data from 100 individuals of known sex, we analyzed potential sex-specific sequences. Strikingly, our analyses demonstrate the absence of any sex-linked sequence in the *T. baccata* genome, indicating sex determination by a non-genetic molecular mechanism.

## > Results

### *Taxus baccata* genome assemblies

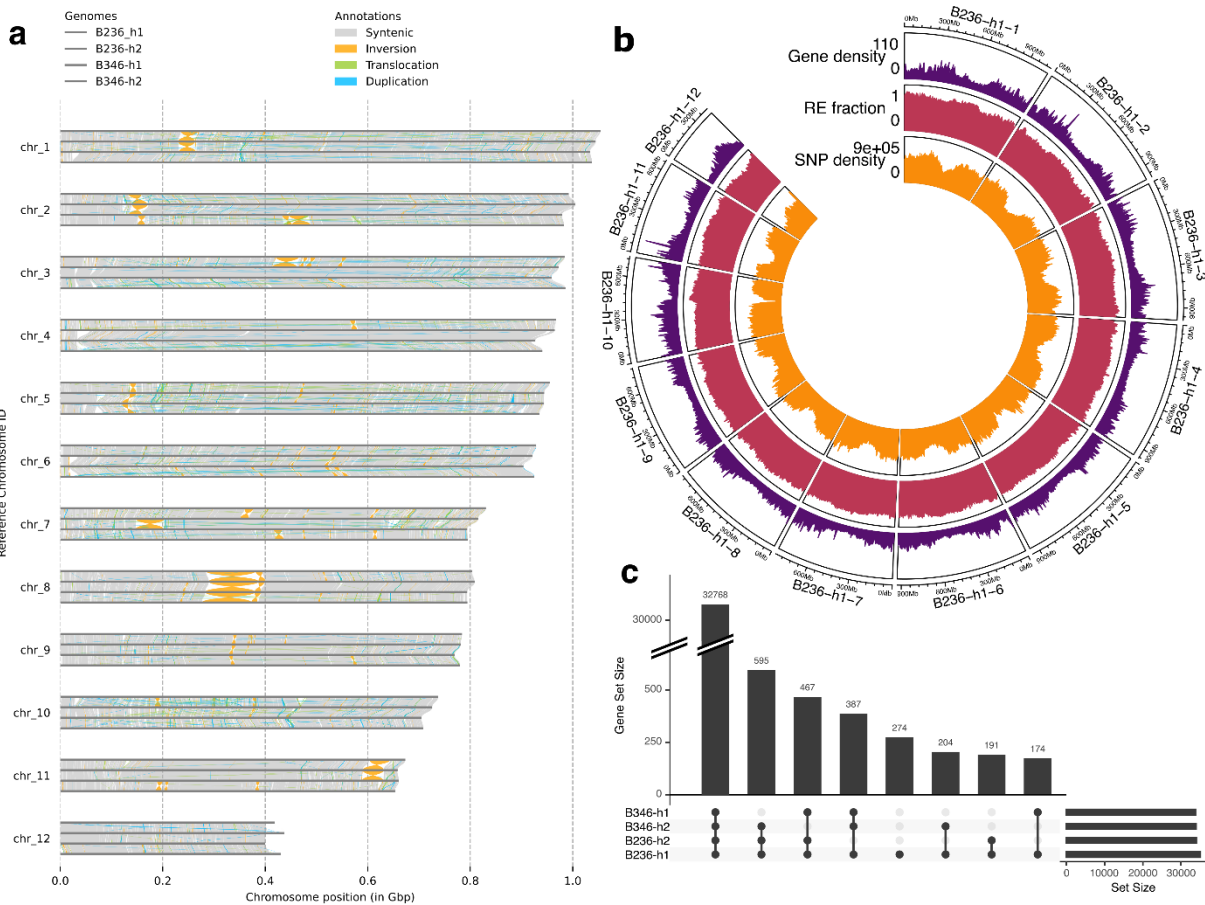
To construct female and male haplotype-resolved chromosome-level genome assemblies of *T. baccata*, we generated PacBio HiFi sequencing data with a coverage of ~36× and ~41× (359.32 and 410.02 Gb), and short reads from Hi-C libraries with a coverage of ~63× and ~50× (632.11 and 499.10 Gb) for a female and a male individual, respectively.

The total sizes of the four final haplotype assemblies range from 9.87 to 10.19 Gb. Merquy<sup>29</sup> *k*-mer completeness values range from 98.6% to 99.3% for the combined female and male haplotypes, respectively, or from 75.7% to 76.6% for the individual assemblies (QVs 65.5 – 66.2) (Table 1, Supplementary Table 1). In each assembly, 12 chromosomes contain 99.5% to 99.9% of the haplotype-assigned sequences, and N50 values range from 903.6 to 928.7 Mb. The four assemblies contained between 89.4% and 90.7% of embryophyte BUSCOs (odb10), and the GC content ranged from 36.58% to 36.70% (Table 1, Supplementary Table 1). Telomeric repeats were identified in the terminal regions of 40 out of the 48 chromosomes across the four assemblies. The Hi-C heatmap of the four haplotypes shows the successful scaffolding into complete chromosomes (Extended Data Fig. 1).

| Table 1. Overview of the four genome assemblies and their annotation. B236 is the female individual and B346 the male (IDs TXBAC_17_1 and TXBAC_7_1 in Supplementary Table 2, respectively). |         |         |         |         |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|---------|---------|---------|
| Assembly                                                                                                                                                                                     | B236-h1 | B236-h2 | B346-h1 | B346-h2 |
| Genome size [Gb]                                                                                                                                                                             | 10.19   | 10.11   | 9.87    | 9.98    |
| GC content [%]                                                                                                                                                                               | 36.70   | 36.66   | 36.58   | 36.59   |
| number of scaffolds                                                                                                                                                                          | 408     | 266     | 100     | 87      |
| number of gaps                                                                                                                                                                               | 1667    | 1593    | 1558    | 1586    |
| N50 [Mb]                                                                                                                                                                                     | 928.74  | 921.31  | 903.58  | 924.79  |
| auN [mio.]                                                                                                                                                                                   | 873.60  | 870.15  | 856.89  | 863.82  |
| anchored [%]                                                                                                                                                                                 | 99.55   | 99.76   | 99.96   | 99.95   |
| Merquy <i>k</i> -mer completeness [%]                                                                                                                                                        | 76.59   | 76.36   | 75.68   | 76.51   |
| Merquy QV                                                                                                                                                                                    | 65.57   | 65.64   | 66.20   | 66.23   |
| combined Merquy <i>k</i> -mer completeness [%]                                                                                                                                               | 99.29   |         | 98.59   |         |
| combined Merquy QV                                                                                                                                                                           | 65.60   |         | 66.21   |         |
| complete (C) BUSCOs [%]                                                                                                                                                                      | 90.7    | 90.2    | 89.4    | 90.1    |
| single-copy (S) [%]                                                                                                                                                                          | 85.1    | 84.8    | 84.0    | 84.5    |
| duplicate (D) [%]                                                                                                                                                                            | 5.6     | 5.4     | 5.5     | 5.6     |
| fragmented (F) [%]                                                                                                                                                                           | 3.8     | 4.3     | 4.5     | 4.1     |
| missing (M) [%]                                                                                                                                                                              | 5.5     | 5.5     | 6.1     | 5.8     |
| complete with internal stop codon (E) [%]                                                                                                                                                    | 4.5     | 5.4     | 4.7     | 4.7     |

|                                    |        |        |        |        |
|------------------------------------|--------|--------|--------|--------|
| total gene count                   | 35,060 | 39,283 | 33,907 | 36,096 |
| total transcript count             | 44,597 | 48,657 | 42,568 | 45,383 |
| mono:multi-exonic transcript ratio | 0.48   | 0.66   | 0.46   | 0.52   |
| functional annotations [%]         | 80.29  | 74.98  | 82.04  | 79.71  |

The four haplotype assemblies exhibit considerable structural variation, with the largest inversion on chromosome 8, which is heterozygous in both reference individuals, spanning almost 100 Mb (Fig. 1a, Extended Data Fig. 2, Supplementary Table 3). Syntenic regions cover between 85.5% and 86.9% of the genome sequences for all pairs of genome assemblies.



**Fig. 1: Four haplotype-resolved genome assemblies of *Taxus baccata* reveal high levels of genomic variation and haplotype-specific genes.** **a**, Pairwise sequence alignments highlight large structural variants between the four haplotype-resolved assemblies. **b**, Circos plot for assembly B236-h1 (female individual). Gene density, repeat element fraction (RE) and SNP density are shown based on sliding windows (size: 5 Mb, step: 2.5 Mb). **c**, Comparisons of the gene sequences annotated in the B236-h1 assembly with the sequences of the other three assemblies, as determined by lift-over, reveal haplotype-specific presence-absence variation of genes. Each bar refers to one combination indicated below the bar (present in the marked assemblies and absent in the unmarked assemblies). Set size refers to the total number of sequences found (or annotated sequences, in case of B236-h1) in the respective assembly.

Using both short-read and long-read RNA sequencing data, we generated annotations for each of the four haplotypes. Across the four assemblies, structural annotations with BRAKER3 yielded between

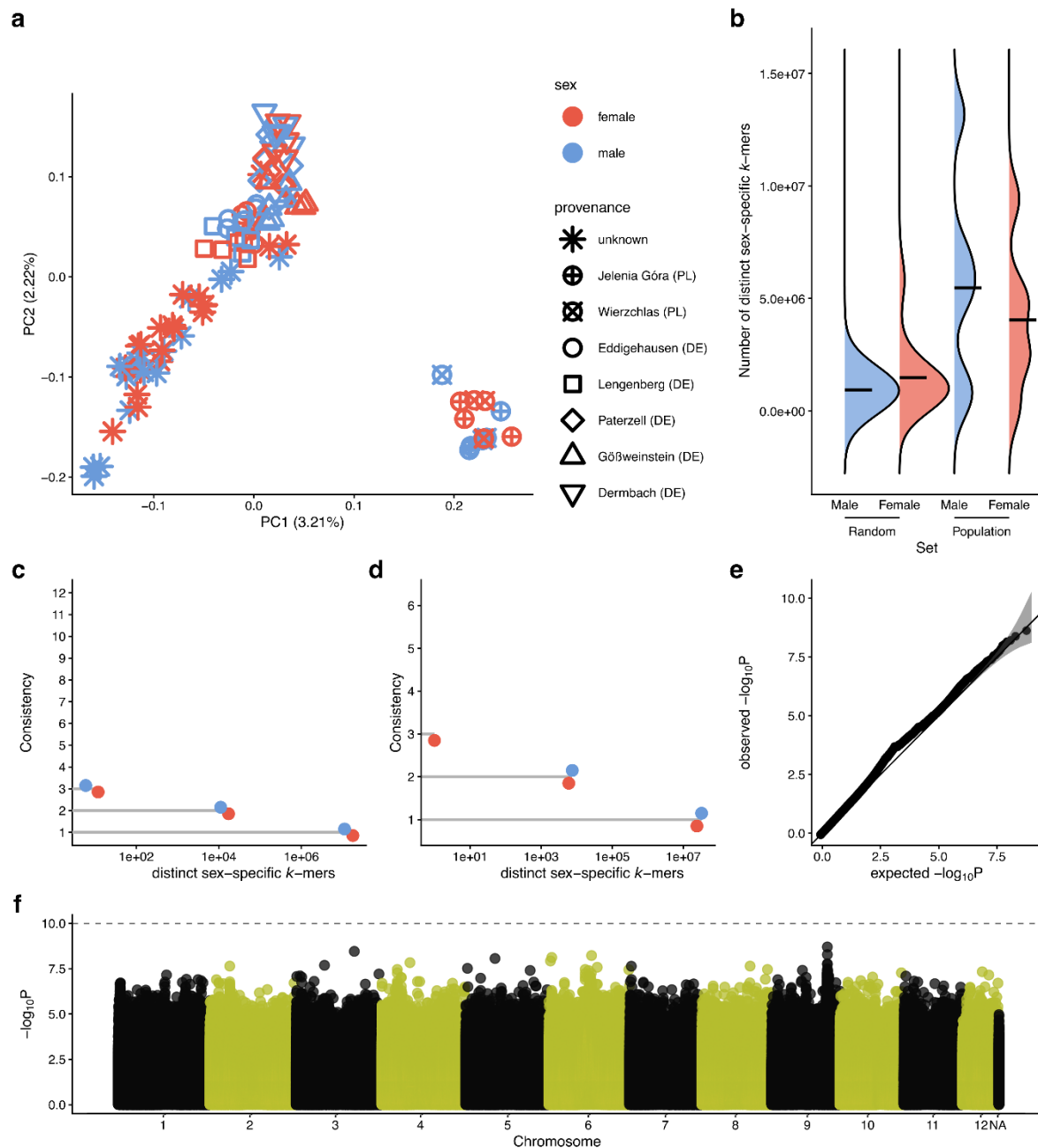
33,907 and 39,283 gene models (Table 1, Supplementary Table 1), excluding transposable elements, which were soft-masked prior to annotation. The ratio of monoexonic to multiexonic gene models varied between 0.456 and 0.664. Functional annotations were achieved for 74.98% to 82.04% of the transcripts in these gene models. While the gene density in the centromeric regions decreases as expected, the density of the repeat elements appears to be relatively evenly distributed across the chromosomes (Fig. 1b, Extended Data Fig. 3).

By transferring the sequences corresponding to the gene annotations of the B236-h1 assembly to the other three assemblies using lift-over, we identified genes shared among the four haplotypes but also genes specific to certain assemblies (Fig. 1c). We applied the same approach to each of the other three haplotypes (Supplementary Fig. 1). As expected, the majority of genes were identified in all four haplotypes. Nevertheless, more than 2,000 genes were missing in at least one haplotype, highlighting the potential relevance of gene presence-absence variation. A list of all haplotype-specific genes is given in Supplementary Table 4.

### *No evidence of a sex-determining region in the Taxus baccata genome*

To identify the sex-determining region, we generated whole-genome sequencing (WGS) data for 103 *T. baccata* samples for which the phenotypic sex was assessed before sample collection. From these 103 sequenced samples (Supplementary Table 2), species identity was confirmed with molecular markers<sup>30</sup> for 101, while two samples were identified as *T. × media* hybrids and excluded from further analysis. One additional sample was excluded due to a monoecious phenotype detected in the year after sampling. This resulted in a final set of 50 female and 50 male *T. baccata* samples, with 62 samples originating from seven putatively autochthonous populations and 38 samples from botanical gardens and various other sources (Extended Data Fig. 4). A microsatellite analysis confirmed that no samples were clonal duplicates prior to sequencing (Supplementary Table 5).

Mappings these WGS reads to the four haplotype assemblies showed an average coverage of 20.8× per sample. Variant calling detected between 969 million and 971 million SNPs before filtering. The population structure of the dataset was explored by selecting a random subset of SNPs and generating a principal component analysis (PCA) (Fig. 2a, Supplementary Fig. 2). In general, the samples from autochthonous populations in Germany formed partially overlapping clusters adjacent to (and overlapping with) the samples from unknown provenances, while samples originating from Poland clustered separately (Fig. 2a). Non-metric multidimensional scaling (NMDS) based on the same dataset showed a similar result, but with a high stress value (Supplementary Fig. 3). In a PCA with all of the original 103 samples, the two *T. × media* samples were placed distant from all other samples as expected (Extended Data Fig. 5).



**Fig. 2: No sex-determining region can be detected in the *Taxus baccata* genome using  $k$ -mer analyses or GWAS.** **a**, Population structure of the WGS dataset comprising 50 female and 50 male *T. baccata* individuals depicted as a PCA based on SNPs from mappings to haplotype B236-h1. Unknown provenances include all samples from non-autochthonous areas, e.g., botanical gardens. **b**, Number of distinct sex-specific  $k$ -mers is not significantly different between sexes in 12 random subsets (6 females versus 6 males in each set) or 6 subsets of samples from autochthonous populations (the two polish populations were considered as one). Black crossbars show means. **c,d**, Consistency of sex-specific  $k$ -mers across the 12 random (c) or 6 population (d) subsets **e**,  $p$ -values from the genome-wide association study (GWAS) show no notable deviation from a uniform distribution in a Q-Q plot. **f**, GWAS in B236-h1 shows no significant association with sex. Dashed horizontal line indicates Bonferroni-corrected significance threshold.

To determine the heterogametic sex in *T. baccata*, we applied a reference-free approach by analyzing sex-specific  $k$ -mers (i.e.,  $k$ -mers that occur in all samples of a given sex and in none of the other) based on the short reads from the 100 sexed individuals described above. First, we tested for

an enrichment of either male-specific *k*-mers (MSKs) or female-specific *k*-mers (FSKs). For this, we generated 12 sets, each containing six randomly chosen female individuals and six randomly chosen male individuals (“random sets”), as well as six population-specific sets, containing four to six female and male individuals from autochthonous populations, to minimize neutral differences in the genetic background (“population sets”). For both random and population sets, we did not detect any significant differences between the total numbers of FSKs and MSKs (Fig. 2b). Furthermore, nearly all MSKs and FSKs occurred only once across replicate sets. In the random sets, 99.9% of FSKs and MSKs were observed in a single replicate, while only 0.1% recurred in two replicates and fewer than 0.001% appeared in three replicates. No *k*-mer was detected in more than three of the replicates (Fig. 2c). In the population sets, 99.9% of FSKs and MSKs were detected in only one replicate, 0.02% in two replicates, and one FSK in three (Fig. 2d). None of the recurring *k*-mers showed statistically significant association with sex after false discovery rate correction in a permutation test.

To further investigate possible genomic differences between female and male individuals and potentially identify genomic regions quantitatively associated with sex, we performed a genome-wide association study (GWAS) based on the generated SNP dataset. Remarkably, the GWAS showed no significant SNP after Bonferroni or Benjamini–Hochberg correction for multiple testing, and the Q–Q plot confirmed the absence of SNPs deviating from the uniform *p*-value distribution expected by chance (Fig. 2e,f). This result was not influenced by the choice of the reference genome used for mapping and variant calling (Extended Data Fig. 6). Furthermore, no sex-specific region with significantly reduced mapping coverage was identified for any of the four haplotype assemblies (Supplementary Fig. 4).

Together, these results demonstrate the absence of a genetic sex-determining system in *T. baccata*. While the terminology to categorize sex-determining systems is not fully consistent across scientific fields, non-genetic sex determination can be defined as environmental sex determination (ESD) in the broad sense. This not only includes cues like temperature, as reported in various gonochoric reptiles (e.g., <sup>31</sup>), but also the cellular environment. ESD in the narrow sense appears unlikely in a tree species, since low predictability of environmental variables could easily lead to distorted sex ratios. Instead, the cellular environment during gamete formation could be controlling sex determination in *T. baccata*. For example, random (stochastic) epigenetic modification of one allele, similar to the sex ratio distorter in persimmon or X chromosome inactivation in placental mammals<sup>10,32</sup>, could lead to epigenetic formation of heterogamety. This situation stands in stark contrast to sexual plasticity, which, in plants, can be controlled by the environment and is thus often classified as ESD (e.g., <sup>8,33–35</sup>). Instead, *T. baccata* appears to feature stable dioecy and ESD.

#### *Sex-specific gene expression reveals potential epiX/Y and epiZ/W candidates*

If sex in *T. baccata* is in fact determined by an epigenetic modification, we would expect differential expression of the sex-determining gene between female and male trees (see different potential epi-systems for sex determination in Supplementary Table 6). Additionally, the (epi)heterogametic sex should exhibit monoallelic expression. To explore this possibility, we sampled flower buds at an early stage of their development from three female and three male trees and performed RNA sequencing. We mapped these RNA sequencing data to our genome assemblies and analyzed differential gene expression (Supplementary Table 7). First, we searched for female- and male-specific expression, as expected for an active X/Y or Z/W system, which highlighted 14 genes (Supplementary Figs. 5 and 6). Second, we searched for genes with expression patterns consistent with an X/A or Z/A balance system, which identified 74 genes (Supplementary Figs. 7 and 8). Of the 88 candidate genes, 74 exhibited heterozygous DNA sequence variants and thus allowed the assessment of allele-specific

expression. Only 9 exhibited monoallelic expression patterns (Supplementary Table 8). Interestingly, two of them also showed sex-specific methylation differences between our two reference individuals for 5mC sites called from the PacBio sequencing data (a carboxylesterase 15 (g4483) and an uncharacterized transcript (g16126), see Supplementary Fig. 9). While these analyses provide an exciting starting point, extensive RNA expression analyses, including non-coding transcripts, and exploration of DNA methylation patterns will be necessary to unambiguously identify the possible epiSDR and provide further insights into the underlying molecular mechanism.

## DISCUSSION

Our haplotype-resolved chromosome-level genome assemblies for a female and a male *T. baccata* individual represent the fourth publicly available *Taxus* reference genome, besides assemblies of *T. wallichiana*<sup>26,36</sup>, *T. yunnanensis*<sup>37</sup> and *T. chinensis*<sup>38</sup>, although *T. yunnanensis* is often considered a synonym of *T. wallichiana*<sup>17</sup>. Our assemblies appear to be of comparable quality to the most recent genome of *T. wallichiana*<sup>26</sup>, with a slightly higher anchoring rate, higher Merqury QV values, but some incomplete telomeres.

Sex-specific *k*-mers have proven powerful for locating SDRs in a range of plants, e.g., the large SDRs of *Ginkgo biloba*<sup>39</sup> or *Amborella trichopoda*<sup>40</sup>. Grayson *et al.*<sup>41</sup> detected the SDR of *Takifugu rubripes*, where a single SNP is considered responsible for sex determination, by analyzing *k*-mers. Hence, a small size of the SDR does not appear to limit the approach in general. Given that we analyzed multiple random sets as well as population-specific sets of females and males in *T. baccata* (Fig. 2b), it seems unlikely that potential phenotyping errors, insufficient local coverage of sex-specific sequences or population structure could have obscured sex-specific signals. However, the binary nature of the applied presence–absence *k*-mer analysis might miss sex-associated loci that are not completely sex-linked. GWAS on the other hand should identify such sequences with quantitative effect. Also, since SDRs often exhibit suppressed recombination<sup>42,43</sup>, SNPs in these regions co-segregate with the sex-determining sequences and can thus be used to detect the presence of the SDR. However, our GWAS results considering four haplotypes of a male and a female reference individual indicate that sex-linked SNPs do not exist in our dataset (Fig. 2e-f, Extended Data Fig. 6). Finally, the absence of genomic regions in which the mapping coverage in samples of one sex was consistently reduced (Supplementary Fig. 4) complements the results of our *k*-mer and GWAS analyses.

The stable sex expression of *T. baccata* and the karyological similarities between the sexes<sup>27</sup> match the expected outcome of an X/Y (or Z/W) sex-determining system with homomorphic sex chromosomes. However, our results show that the heritable sex-determining information does not appear to reside in the genetic sequence. The absence of GSD in turn infers ESD. Knowledge about the molecular mechanism determining sex is rare for dioecious plants (e.g.,<sup>11</sup>), and the combination of ESD and dioecy appears to be a novel observation in a dioecious plant species. The species showing dioecy or ESD listed in Renner<sup>8</sup> do not include examples exhibiting both. Pannell<sup>44</sup> notes that “angiosperms do not seem to have evolved fully environmental sex determination (ESD) that might parallel, for example, temperature-dependent sex determination in some reptiles”.

Given the generally stable sex expression in *T. baccata*, it seems likely that sex is determined only once in an individual’s lifetime. In the absence of a genetic factor, a haplotype-specific epigenetic modification could be responsible, with the downstream pathway promoting either female or male function. Specifically, an allele-specific demethylation of a single gene switch (male factor) at the epi-Y-chromosome and stable hypermethylation of the X-chromosomal alleles could result in a monoallelic paternal expression of a gene sex switch, similar to the situation reported in the fish

*Ictalurus punctatus*<sup>45</sup>. This molecular mechanism would be just as feasible for an epiZ/W system. It is also possible that sex is controlled by an epigenetically regulated X/A (or Z/A) balance system. In a classic genetic X/A model, the sex-determining factor resides solely on the X chromosome and the X/A ratio dictates sex (e.g., <sup>4</sup>). For example, if the same factor were present on both X and Y but silenced via methylation on the Y chromosome, males would show monoallelic expression for the gene while the other sex would retain biallelic expression. It should be noted that the causal epiallele is probably not stable in the case of *T. baccata*, since any fixed (epi-)allele would lead to sex-linked SNPs over evolutionary times, but no such SNPs were detected in our GWAS. Instead, it seems more likely that the underlying epiallele is randomly formed during meiosis of the epi-heterogametic sex, with a mechanism similar to X-chromosome inactivation<sup>46</sup>. An example of such sex-linked random inactivation in plants would be the *HaMster* gene in *Diospyros lotus*<sup>10</sup>.

In conclusion, our data provides exciting evidence for a novel sex-determination mechanism in a dioecious plant species without sex-specific DNA sequence variation. Taken together with past observations of balanced sex ratios and stable expression of dioecy, this highlights the possibility of a dioecious plant with environmental sex determination (*sensu lato*). Given that *T. baccata* is the first conifer studied from a large group of almost 400 dioecious species of the “conifer II” clade<sup>6</sup>, such a mechanism may be widespread and thus relevant for conservation genetics and breeding of many species. Future work in *T. baccata* should focus on the identification of epigenetic marks and transcriptional signatures that differentiate females from males. The genome assemblies of both a female and a male *T. baccata* individual presented here provide a solid foundation for these investigations.

## MATERIAL AND METHODS

### *Sample collection*

*T. baccata* samples used for high molecular weight (HMW) DNA extraction and subsequent PacBio HiFi sequencing or for Hi-C Illumina sequencing were collected in the arboretum of the Thünen Institute of Forest Genetics (Großhansdorf, Germany). Specifically, needles were collected in October 2022 from one male (TXBAC\_7\_1, Supplementary Table 2) and one female reference individual (TXBAC\_17\_1, Supplementary Table 2) whose phenotypic sexes were repeatedly assessed over the course of the two previous years. These samples were frozen directly after collection in liquid nitrogen, and stored at -70 °C.

For whole genome (short-read) sequencing, *T. baccata* needle samples were collected or acquired from various sources, including arboreta, botanical gardens and natural reserves (Supplementary Table 2) and stored at -20 °C. Needle samples from Poland were collected with permission of the Regional Directorate for Environmental Protection, Bydgoszcz, Poland (WOP.6400.37.2022, WOP.6205.72.2022.KLD, WOP.6205.73.2022.KLD, WOP.6400.36.2022.MKW).

Tissue samples of the male and female reference individual destined for RNA extraction and subsequent PacBio Iso-Seq were collected in the arboretum of the Thünen Institute of Forest Genetics in Großhansdorf, Germany. Needle samples were collected in May 2023 and frozen in liquid nitrogen directly after collection. Cambium samples were collected in June 2023 near the base of the trunk, using a 16 mm diameter hollow punch to remove a part of the bark and scraping cambium tissue off the removed bark piece with a sharp knife. Root samples were collected in June 2023 by digging up thin roots near the tree trunk. Roots were cut off, briefly washed in tap water and dried with paper towels. Strobili were collected in November 2022.

*T. baccata* samples destined for RNA extraction and subsequent short-read RNA sequencing were collected from three male and three female individuals, including the two reference individuals, in the arboretum of the Thünen Institute of Forest Genetics in Großhansdorf, Germany. For this, buds in early stages of development were collected in June 2023 by either plucking them off directly or scraping off the buds from shoots with a knife. Brachyblast buds were targeted as far as the bud identity could be asserted.

Species identity of all samples was confirmed using the species markers TA\_ITS, TA\_InDel1, TA\_InDel2, TA\_cox1<sup>30</sup> and trnL-F<sup>47</sup>. In the case of root samples, sample identity was further tested via microsatellite analysis to confirm the association between root sample and the respective reference tree. The phenotypic sex of a sample was assigned based on observed strobili and/or arils. When needed, cone buds were assessed under a binocular to avoid confusion with vegetative buds.

### *Microsatellite analysis*

To ensure genetic distinctiveness among the sampled individuals, all samples destined for WGS were genotyped with the microsatellite markers Tax36, Tax92 (both: Dubreuil *et al.*<sup>48</sup>), TS09 (Huang *et al.*<sup>49</sup>), Ma-14186-166D (Ueno *et al.*<sup>50</sup>), as well as two novel SSR markers named TABAC01 (for: 5'-TATGTGCCTAGGCGTTAGTC-3', rev: 5'-TTGTAGGTTGATAGACAAATGGA-3'; gSSR (TCC)<sub>7</sub> at chr2 of *T. chinensis*<sup>38</sup>) and TABAC02 (for: 5'-TTTGGAAGTAACTAAACACATG-3', rev: 5'-TAATTGTGTTCTCCCTAATAAGG-3'; gSSR (CTT)<sub>11</sub> at chr6 of *T. chinensis*<sup>38</sup> (Supplementary Table 5). Genotyping was performed by PCR and subsequent fragment size analysis. Forward primers were tailed at their 5'-end and dye-labeled DNA fragments with complementary sequences to the tails were added to the PCR mix (Supplementary Table 9, Supplementary Table 10). Forward primers of markers Tax36, TS09, TABAC01 and TABAC02 were tailed with the adapter sequence 5'-CAGGACCAGGCTACCGTG-3', marker Tax92 with 5'-GCCTTGCCAGCCCGC-3' and marker Ma-14186-166D with 5'-CAGGACCAGGCTACCGTG-3'. The PCR reaction setups and PCR programs are given in Tables S6 and S7, respectively. PCR products were analyzed on a CEQ 8000 Genetic Analysis System (Beckmann Coulter) and evaluated using GeneMarker v3.0.1 (SoftGenetics). When fragment length signals showed stutter bands, the highest peak was chosen. When only one peak was visible, the sample was considered to be homozygous for this marker. Potential clones were identified by pairwise comparison of fragment sizes using Excel (Microsoft), as they correspond to the alleles. For heterozygous alleles, the fragments were sorted by size. Each fragment was then compared with the respective fragments of all other samples.

### *PacBio HiFi sequencing and Hi-C of the male and the female Taxus baccata reference tree*

Nuclei preparation from sample tissue, high molecular weight genomic DNA extraction, Pacific Biosciences (PacBio) HiFi library preparation and sequencing for the samples used in the construction of the reference genomes in this study were previously described in Krautwurst *et al.*<sup>51</sup>, corresponding to sample IDs 2,6 and 12 in Table 7 of that publication. In short, at total of 2.3 g (over two extractions) and 1.6 g of snap-frozen *T. baccata* needles of the male and female sample, respectively, were used as described in Krautwurst *et al.*<sup>51</sup> to retrieve a total of 26.45 µg (from male sample) and 21.2 µg (female sample) of HMW gDNA. Yield and quality of the extracted gDNA was assessed via NanoDrop and Qubit (BR assay) measurements (Thermo Fisher Scientific) and the fragment lengths was determined on the Femto Pulse (Agilent). HMW gDNA was then sheared to approximately 20 kbp with the MegaRuptor™ (Diagenode). Using 3 µg of high-molecular weight DNA as input, PacBio HiFi libraries were prepared according to the protocol "Preparing whole genome and

metagenome libraries using SMRTbell® prep kit 3.0” (Pacific Biosciences) and size selected for fragments larger than 8 kbp with the BluePippin Instrument (SAGE) or larger than 5 kbp with AMPure beads (Beckman Coulter) according to the library preparation protocol. Sequencing was performed on a total of 24 SEQUEL II SMRT™ Cells with 30-hour movies (10 for the female sample, 14 for the male sample), using the Sequel II Binding Kits 3.2 and the Sequel II sequencing kit 2.0 (Pacific Biosciences). Raw reads were processed using ccs v6.4.0 (<https://github.com/PacificBiosciences/ccs>), actc v0.6.0 (<https://github.com/PacificBiosciences/actc>) and deepconsensus v1.2.0<sup>52</sup>. Sequencing resulted in 410.02 Gb and 359.32 Gb circular consensus sequence reads for the male and female individual, respectively.

For Hi-C data, nuclei from sample tissue were extracted as described in Krautwurst *et al.*<sup>51</sup>. Hi-C libraries were prepared for each sample using the Arima-HiC Kit (Arima Genomics, Material Nr. A510008) and KAPA HyperPlus Kit (Kapa Biosystems) according to the manufacturer’s protocols. For the female sample, a 0.78X AMPure bead (Beckman Coulter) size selection was performed to eliminate < 200 bp fragments. Both libraries have been sequenced to approximately 80x genome coverage with 200 cycles on an Illumina NovaSeq 6000 platform.

#### *Whole genome sequencing of 103 Taxus samples*

DNA extraction for whole genome sequencing (WGS) were performed as described in Bruegmann *et al.*<sup>53</sup>, or Ziegenhagen *et al.*<sup>54</sup> for difficult samples (see Supplementary Table 2). DNA was checked via NanoDrop and Qubit (BR Assay) measurements (both: Thermo Fisher Scientific). DNA was sheared to 400 bp fragments (on average) using a Covaris LE220 ultrasonicator (Covaris), and 250 ng of sheared DNA was used as input for the KAPA HyperPlus Kit (Kapa Biosystems) to construct paired-end Illumina libraries, using 5 amplification cycles. Sequencing to approximately 20X coverage as 2×150 bp paired-end reads was performed with NovaSeq S4 flowcells and v1.5 reagents on an Illumina NovaSeq 6000 platform (Illumina).

#### *RNA extraction and sequencing*

Frozen sample material was ground using mortar and pestle and RNA was extracted following the Spectrum™ Plant Total RNA-Kit (Sigma-Aldrich) protocol B instructions (including the optional DNase digestion step) with the addition of adding 30 mg Polyclar (SERVA) to the lysis buffer for each sample. RNA quality was assessed via NanoDrop, Qubit (BR assay) measurements (both: Thermo Fisher Scientific) and Bioanalyzer, using a Plant RNA Nano assay (Agilent).

For PacBio Iso-Seq sequencing, Iso-Seq libraries of total RNA of different tissues of the male and female reference sample were generated according to the PacBio protocol ‘Preparing Iso-Seq® libraries using SMRTbell® prep kit 3.0 (PN 102-396-000 REV02 APR2022). In brief, 300 ng total RNA was reversely transcribed with the NEBNext® Single Cell / Low input cDNA synthesis and amplification module, with barcoded cDNA primers in the cDNA amplification step. Barcoded cDNAs were pooled for each individual, with the exception of libraries from cambium samples, which were pooled separately, for a total of 3 pools and PacBio sequencing adapters have been ligated to these pools. The libraries were pooled for each individual, with the exception of libraries from cambium samples, which were pooled separately, for a total of 3 pools. The pools were sequenced on three PacBio Sequel2 SMRT cells for 30 hours with the Sequel II Binding Kits 3.2 and the Sequel II sequencing kit 2.0). After circular consensus calling of the raw sequencing reads, high-quality isoforms have been called and clustered with the default SMRTlink Iso-Seq pipeline (v11 and v13).

For short-read RNA sequencing, libraries with an average insert size of 335 bp were constructed using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (poly-dT enrichment workflow) (New England Biolabs). Sequencing to approximately 60 million reads as 2×150 bp paired-end reads was performed with NovaSeq S4 flowcells and v1.5 reagents on an Illumina NovaSeq 6000 platform (Illumina).

## *Genome assembly*

The assembly procedures of the female and male individuals differed slightly: For the reference assembly of the female sample, assemblies were generated with Hifiasm<sup>55,56</sup> v0.18.9-r527. Remaining haplotype duplications were manually removed, guided by purge\_dups v1.2.6 ([https://github.com/dfguan/purge\\_dups](https://github.com/dfguan/purge_dups)), to avoid over-purging in repetitive regions. Scaffolding of the contigs was performed by YaHS<sup>57</sup> v1.2a.1. Chloroplast and mitochondrial contigs were identified with Tiara<sup>58</sup> v1.0.3 and MitoHifi<sup>59</sup> v3.2.1, respectively, and manually removed. Blobtools<sup>60</sup> v4.4.5 was used to identify and remove contaminations. The male reference assemblies were generated with Hifiasm<sup>55,56</sup> v0.19.8-r603, duplications were removed with purge\_dups v1.2.5 ([https://github.com/dfguan/purge\\_dups](https://github.com/dfguan/purge_dups)) and scaffolded with YaHS<sup>57</sup> v1.2.

All scaffolds were then manually curated in PretextView v1.0.3 (<https://github.com/sanger-tol/PretextView>), GRIT\_Rapid (<https://gitlab.com/wtsi-grit/rapid-curation>, commit 1a3d79a8), HiGlass<sup>61</sup> v0.10.4, and the DAmr pipeline (<https://github.com/MartinPippel/DAmr>; branch: master-v1, commit 1e428cddaa79fee04edbc41bcc41a12a014b11).

## *Assembly structural analysis*

To compare the haplotype assemblies and identify large structural variants, alignments between the assembly sequences were generated with mm2plus<sup>62</sup> v1.1 using the parameters “-x asm5 -k 28 -l 50G -c --eqx”. The mappings were then processed with SyRI<sup>63</sup> v1.7.0 and visualized with Plotsr<sup>64</sup> v1.1.1. A complementary approach using protein orthologs (generated by structural genome annotation, see below) was conducted with GENESPACE<sup>65</sup> v1.3.1.

## *Structural genome annotation*

To generate structural annotations of the female and male reference assemblies (two haplotypes each), the TETools container v1.90 (<https://github.com/Dfam-consortium/TETools>) and BRAKER<sup>66,67</sup> container v3.0.7.6 (<https://github.com/Gaius-Augustus/BRAKER>) were utilized. First, a library of transposable elements was generated *de novo* by RepeatModeler<sup>68</sup> v2.0.6 and contained software, including RepeatScout<sup>69</sup> v1.0.7, Tandem Repeats Finder<sup>70</sup> v4.09, RECON<sup>71</sup> v1.08, LTR\_retriever<sup>72</sup> v2.9.0, NINJA<sup>73</sup> v1.00-cluster\_only, CD-HIT<sup>74</sup> v4.8.1, MAFFT<sup>75</sup> v7.471, HMMER v3.4 (<http://hmmer.org/>), GenomeTools<sup>76</sup> v1.6.4, and a selection of UCSC utilities (see <https://github.com/Dfam-consortium/TETools>). For this step, the LTR Structural Analysis option (“-LTRStruct”) was enabled, apart from this default options were used. Based on this library, the reference assemblies were soft-masked with RepeatMasker v4.1.7-p1 (<http://repeatmasker.org>) using default parameters (except for the soft-mask setting “-xsmall”). Subsequently, the assemblies were soft-masked further with Tandem Repeats Finder v4.09<sup>70</sup>, using the parameters “2 7 7 80 10 50 500 -d -m -h”.

Iso-Seq transcript data for usage in BRAKER was prepared separately to account for large introns in the target species. To estimate the maximum intron size of the *Taxus* genome while avoiding false positives (i.e., wrong alignments that erroneously indicate very large introns), Iso-Seq transcripts were mapped to the female reference assembly haplotype 1 using minimap2<sup>77,78</sup> v2.28 multiple times with the parameters “-uf -a -x splice:hq” and varying “-G” parameter settings. Mappings were then filtered to exclude alignments with the “secondary” or “supplemental” flag, and alignments with a mapping quality < 40. From each mapping, the longest N-operation in the CIGAR string (indicating a skip in the reference compared to the read, i.e., possibly an intron) was determined and plotted against the used -G parameters. The resulting plot showed a region in which the maximum N-operation length stagnated even under increasing “-G” parameter settings while at even higher “-G”, maximum N-operation length rose quickly again (i.e., the plot formed a plateau). The “-G” parameter which corresponded to the middle of this region was then used for the mappings in BRAKER.

The soft-masked reference assemblies were each used as genome data input for two BRAKER<sup>79-82</sup> workflows: (1) BRAKER3 and (2) BRAKER3 with long-read RNA input<sup>83</sup>. For (1), RNA data input consisted of the short-read RNA sequencing data generated for this study, and a public dataset of a *T. baccata* transcriptome by<sup>84</sup> (SRA accessions SRX2999991, SRX2999990). For this workflow, v3.0.7.6 of BRAKER was used. For (2), PacBio Iso-Seq transcripts that were generated for this study were mapped to the respective reference genome with minimap2<sup>78</sup> v2.28 using the parameters “-G 500k -uf -a -x splice:hq”. The resulting mapping dataset was used as RNA input. In both workflows, protein input data consisted of the publicly available OrthoDB (<https://www.orthodb.org/>) partition Viridiplantae\_v12<sup>85</sup> and a UniProt<sup>86</sup> (<https://www.uniprot.org/>) dataset of 10991 peptide sequences (using the filters “(taxonomy\_id:25628) AND (existence:3)”, downloaded on 2024-11-27). For this version of the workflow, the modified BRAKER docker container for Iso-Seq input (docker://teambraaker/braker3:isoseq) was used. In addition to BRAKER itself, the following software was used in the pipeline: AUGUSTUS<sup>66,67</sup> v3.5.0, HISAT2<sup>87</sup> v2.2.1, GeneMark-ETP<sup>88</sup> v1.0.2, DIAMOND<sup>89</sup> v2.0.15, Spaln2<sup>90,91</sup> v2.3.3f, StringTie2<sup>92</sup> v2.2.1, GFF utilities<sup>93</sup> v0.12.7, TESBRA<sup>94</sup> v1.1.2.5, SAMtools<sup>95</sup> v1.13, BamTools<sup>96</sup> v2.5.1 and BEDTools<sup>97</sup> v2.30.0. Merging of the output of the two workflows was performed using TSEBRA<sup>94</sup> v1.1.2.5 with the “--filter\_single\_exon\_genes” option. The two workflows each used new separate species profiles for AUGUSTUS.

Descriptive statistics of the predicted gene sets were generated with a custom awk script (see Code Availability). BUSCO scores of the gene sets were generated via BUSCO<sup>85</sup> v5.8.2 in conjunction with the embryophyta\_odbv10 lineage dataset. Gene predictions were visually inspected in CLC Genomics Workbench v24.0.2 (QIAGEN) by importing the reference genome FASTA file, annotation GTF and the intermediate “hintsfile” GTF (containing extrinsic evidence from RNA and protein data, i.e., intron, CDS, start and stop codon hints) files, and RNA mapping BAM files, converting these files into “tracks” and overlaying these datasets as “track list” in CLC.

To assess annotation consistency across haplotypes and identify unique and shared genes, we used LiftOff<sup>98</sup> v1.6.3. LiftOff transfers annotations by aligning gene sequences from the annotated query genome to the target genome. We applied sequence identity and alignment coverage thresholds of 0.5. The resulting gene-presence sets were visualized using UpSetR<sup>99</sup> v1.4.0.

## Functional genome annotation

Protein-coding sequences were further annotated using the EnTAP<sup>100</sup> container v2.2.0. EnTAP was used in conjunction with the frame selection step with TransDecoder<sup>101</sup> v5.7.1

(<https://github.com/TransDecoder/TransDecoder/>), the similarity searching step with DIAMOND<sup>89</sup> v2.1.8 and the NCBI RefSeq databases (<https://www.ncbi.nlm.nih.gov/refseq/>) for plant, fungi and bacteria, and the ontology analysis step with EggNOG<sup>102</sup>, using the eggno-mapper<sup>103</sup> v2.1.12.

#### *K-mer analysis*

To identify sex-specific sequence regions, a binary presence–absence *k*-mer analysis adapted from Carey *et al.*<sup>40</sup> was applied. Symmetric female/male groups with six individuals per sex were assembled into 12 random replicate pairs, and six additional replicate pairs were constructed, each restricted to one autochthonous population. The samples from Jelenia Góra and Wierzchlas were considered as one autochthonous population in this analysis due to low geographical distance of about 10 km. For each individual, *k*-mers were counted from trimmed paired end reads with KMC3<sup>104</sup> v3.2.4 using canonicalization and a *k*-mer length of 31. The minimum count threshold was set to one (`-ci 1`) to retain low coverage signals, and no upper threshold was applied by setting `-cs` and `-cx` to the maximum 32 bit unsigned value. Sex-specific candidates were defined by two set operations. First, within the target sex, only *k*-mers observed in every target individual were retained using `kmc_tools intersect`. Second, from this intersection, any *k*-mer observed in at least one individual of the comparison sex was removed by iterative `kmc_tools simple subtract` across all comparison individuals. Running the procedure with female as target yielded female-specific *k*-mers (FSKs), and with male as target yielded male-specific *k*-mers (MSKs). Final lists were exported with `kmc_tools transform dump`.

Group level differences were quantified by comparing the sizes of FSK and MSK sets between female and male replicates using a paired Wilcoxon signed rank test. Replicate wise consistency of individual *k*-mers was assessed by counting for each *k*-mer in how many replicate-specific difference sets it recurred, and empirical significance was evaluated by a permutation test with 10<sup>5</sup> randomizations and seed 42, followed by Benjamini–Hochberg correction at  $\alpha = 0.05$ . Only *k*-mers recurring in at least two replicates were tested.

#### *Mapping of WGS data*

WGS paired-end reads were trimmed with Trimmomatic<sup>105</sup> v0.39 using the settings “ILLUMINACLIP:\$adapter:2:30:10:1:true”, “SLIDINGWINDOW:4:15” and “MINLEN:60”. The adapter sequence file (\$adapter) corresponds to the TruSeq3-PE-2.fa file in the trimmomatic github repository (<https://github.com/usadellab/Trimmomatic/blob/main/adapters/TruSeq3-PE-2.fa>). Before and after trimming, the reads were assessed using Fastqc<sup>106</sup> v0.12.1 and Multiqc<sup>107</sup> v1.27.1. Unpaired trimmed reads were excluded from further analysis. The remaining trimmed reads were mapped onto reference genome assemblies (four haplotypes in total) using bwa-mem2<sup>108</sup> v2.2.1. Mapping quality was assessed with Qualimap<sup>109</sup> v2.3. Read mappings corresponding to the same biological sample were then merged using the SAMtools<sup>95</sup> v2.12 function “merge” with the options “-r -t SQ”. Potential duplication artifacts (i.e., reads originating from the same DNA fragment) in the merged files were marked with MarkDuplicates v3.1.0 of the Picard tool suite<sup>110</sup>, using the option “--OPTICAL\_DUPLICATE\_PIXEL\_DISTANCE 2500”.

#### *Variant calling*

Variant calling was performed on the read mappings using the bcftools<sup>95</sup> v1.21 functions “mpileup” with the options “-a AD,DP,INFO/FS,INFO/AD” and “call” with the options “--ploidy 2 -m -v”. From the resulting VCF files, indels were excluded from all further analyses.

### *Population structure analysis*

For PCAs (intended to correct for population structure during GWAS), subsets of SNPs were generated, containing 140853, 101647, 128570 and 134459 SNPs in the sets of B236-h1, B236-h2, B346-h1 and B346-h2, respectively. For these sets, descriptive statistics for the filtered VCF files were generated and visualized using bcftools<sup>95</sup> v1.21 and the ggplot2<sup>111</sup> package v3.5.1 in R<sup>112</sup> v4.4.2. Based on these visualizations and the GATK recommendations for hard-filtering of germline variants (<https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants>), the VCF files were filtered to exclude variant sites as follows: (1) ‘INFO/FS < 0.000001 || INFO/MQ < 40 || INFO/MQBZ < -12.5 || RPBZ < -4 || COUNT(GT="mis") > 0.1 \* N\_SAMPLES’, (2) Quality-by-depth metric less than 5 (these were calculated according the documentation for GATK’s QualByDepth). Descriptive statistics were then generated again, and the modes of the sequencing depth distributions (based in the INFO/DP field in the VCF) was determined. A depth filter was applied to exclude variants with an average depth of less than 75% or more than 150% of the most common per-sample variant depth (determined as the depth at the distribution mode divided by the number of individuals). Filtered VCF files were then imported into PLINK<sup>113</sup> v1.9b7 and further filtered to exclude variants that deviated from Hardy-Weinberg-Equilibrium at  $p < 1e-8$ . Fully filtered variant sets were then subjected to variant pruning by linkage distance. For this, pairwise LD statistics were generated in PLINK v1.9b7<sup>113</sup> and visualized with the ggplot2 package<sup>111</sup> v3.5.1 in R<sup>112</sup> v4.4.2. Based on an LD decay plot, a cutoff was determined and variants were pruned to exclude variants with a correlation of  $r^2 > 0.2$  in 500kbp windows with PLINK<sup>113</sup> v1.9b7.

### *GWAS*

For GWAS, subsets of SNPs were generated from the original VCF file (without InDels) with bcftools<sup>95</sup> v1.21 to include only biallelic SNPs with a minor allele frequency of more than 0.05. A GWAS was performed with GEMMA<sup>114</sup> v0.98.5, using phenotypic sex as a binary variable and the first 6 axes of the previously generated PCA as covariates (option -c). To visualize GWAS results as Manhattan plots, the ggplot2<sup>111</sup> v4.0.0 package in R<sup>112</sup> v4.5.0 was used.

### *Analysis of differential mapping coverage*

The coverage of each set of read mappings (one for each haplotype) was computed in a sliding-window approach using the SAMtools<sup>95</sup> v1.22.1 function bedcov with default parameters, with windows that are 1000 bp long and a step size of 500 bp. Sex-specific coverage differences were filtered depending on the reference assembly that was used during the mapping process (e.g., for datasets based on mappings with a male reference assembly, only windows in which the average coverage of female samples was lower than the average coverage of male samples were retained). We checked the distribution of coverages in the remaining windows for adherence to a normal distribution and subsequently performed a one-sided Wilcoxon signed-rank test for each of the remaining windows using the wilcox.test() function in R<sup>112</sup> v4.5.0 to detect windows with significantly less coverage in samples of the non-reference sex, at a significance threshold of  $\alpha = 0.05$ . For all such windows, a volcano plot was generated using ggplot2<sup>111</sup> v4.0.0.

569

## 570 *Differential gene expression*

571 Read mapping and quantification of read counts were done with the STAR aligner<sup>115</sup> v2.7.11a against  
572 all scaffolds of the B236-h1 assembly. Differential gene expression analysis was done in R<sup>112</sup> v4.4.0  
573 with the bioconductor package DESeq2<sup>116</sup> v1.44.0 with the significance cutoff value alpha set to 0.05.  
574 For the purpose of determining candidate genes for different sex-determining systems, we required  
575 the RNA read counts between the sexes to fulfil the following conditions (see also Supplementary  
576 Table 6). For possible epiX/Y and epiZ/W systems: (1) The lowest individual read count in the group  
577 with expression must be greater than 20 and also (2) greater than 20 times the mean of the group  
578 without expression. (3) The highest individual read count in the group without expression must be  
579 lower than 1/20 of the mean count in the group with expression. For possible epiX/A or epiZ/A  
580 systems: (1) the expression ratio between the lower-expressed and higher-expressed group was 1:2  
581 (1:1.5 to 1:3) and (2) the lowest individual read count was greater than 10.

582

## 583 *Screen candidate genes for heterozygous DNA sequence variants and monoallelic RNA expression*

584 Duplicate reads were marked with Picard Tools<sup>110</sup> markDuplicates v3.1.0 and reads were split with  
585 GATK<sup>117</sup> SplitNCigarReads v4.4.0.0 followed by variant calling and filtering with bcftools<sup>95</sup> v1.22. All  
586 SNPs with QUAL > 20 and DP > 5 were kept. For each candidate gene based on differential gene  
587 expression, the number of heterozygous SNPs (defined as SNPs with allele frequencies between 25%  
588 and 75%) in the gene region was determined in the RNA and DNA (based on WGS, see above).  
589 Candidate genes were filtered out if any heterozygous SNPs were found in the RNA of female  
590 samples (epiW/A candidates), male samples (epiZ/A candidates) or either sex (epiZ/W and epiX/Y  
591 candidates). From the remaining candidates, those with only homozygous SNPs in the DNA were  
592 excluded.

593

## 594 *Genome-wide analysis of differential DNA methylation*

595 To generate methylation data, the raw subreads from PacBio sequencing (see above) were processed  
596 via ccs<sup>118</sup> v6.4.0 with the --hifi-kinetics flag enabled. HiFi CCS reads were then further processed with  
597 jasmine<sup>119</sup> v2.4.0 (<https://github.com/PacificBiosciences/jasmine>) to detect 5-Methylcytosine at CpG  
598 sites. The HiFi reads were mapped to the B236-h1 assembly with pbmm2<sup>120</sup> v1.17.0  
599 (<https://github.com/PacificBiosciences/pbmm2>) and methylation rates were determined with pb-  
600 CpG-tools<sup>121</sup> v3.0.0 (<https://github.com/PacificBiosciences/pb-cpg-tools>). Using R<sup>112</sup> v4.5.0, the count  
601 information of methylated and unmethylated cytosines at a given position for the female and male  
602 individual were used in a Fisher's exact test to determine differential methylation (FDR controlled at  
603 0.05) for each site that was analyzed in both individuals. Methylation rates in genomic space were  
604 visualized with the R package ggplot2<sup>111</sup> v4.0.1.

605

## 606 **ACKNOWLEDGEMENTS**

607 This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research  
608 Foundation) – project number 497528752 – in the scope of the TaxGen project (funding of DFG to  
609 B.K.; DFG grant no. KE 916/10 – 1) and supported by the DFG Research Infrastructure NGS\_CC (DcGC,  
610 project 407482635) as part of the Next Generation Sequencing Competence Network (DFG project  
611 423957469). NGS library preparation, data production and analyses were carried out at the DcGC

Dresden-concept Genome Center - a core facility of the CMCB and Technology Platform of the TUD Dresden University of Technology. D.L. was supported by the DFG project 545520056 awarded to N.A.M. Additional funding was received by the Institute of Dendrology, Polish Academy of Sciences. We thank K. Groppe, V. Kuhlenskamp, A. Schellhorn, I. Burau and gardeners from the Thünen Institute of Forest Genetics for technical assistance, and M. Fladung for helpful comments and discussions. We thank the following contributors from the MPI-CBG and the DcGC for wet lab work: N. Gscheidel (for gDNA preparation and PacBio sequencing), D. Pache (for PacBio sequencing), W. Tan (for HiC). We are grateful to M. Krause and J. Gusson Roscito from DcGC for coordination of Illumina sequencing. We thank S. Rust, J. Schöttler (both: Loki Schmidt Botanical Garden Hamburg, Germany) and S. Petersen (Botanical Garden Kiel, Germany) for providing access to the related botanical gardens, and M. Seho (Bavarian Office for Forest Genetics, Teisendorf, Germany) for providing additional samples. We thank T. de Jong (Leiden University, Netherlands) for helpful discussions.

#### DATA AVAILABILITY

All sequencing runs, HiFi reads, genome assemblies and annotations are deposited in the European Nucleotide Archive (ENA), available at <https://www.ebi.ac.uk/ena/browser/home> under PRJEB103025.

#### CODE AVAILABILITY

All scripts used to assemble the genomes, analyze the data and visualize the results are deposited in [https://github.com/daniel-bross/Taxus\\_genome](https://github.com/daniel-bross/Taxus_genome)

#### AUTHOR CONTRIBUTIONS

B.K., D.B., N.A.M. and J.M. conceptualized the project.

B.K., S.W. and N.A.M. supervised the project and conceived the methodology.

D.B., B.K., E.P.K. and H.S. collected samples.

S.W. coordinated the sequencing steps.

M.P. and L.U. performed the genome assemblies, M.P. performed the manual curation of assemblies and D.B. annotated the assemblies. D.L. performed the comparative annotation analysis.

D.B., J.M., S.K., M.P., M.M., L.U., N.A.M, D.L. and B.K. analyzed the sequencing data.

S.K. and B.K. supervised *k*-mer analyses.

H.S. advised SSR-marker analyses.

D.B., D.L., N.A.M. and M.P. conducted the visualization.

D.B. wrote the manuscript, and all authors read, edited and approved the manuscript.

B.K., S.W., N.A.M. and E.P.K. acquired the funding.

#### EXTENDED DATA

**Extended Data Figure 1. Hi-C contact heatmap of all four sequenced haplotypes show successful scaffolding of both haplotypes of the female and male individuals into chromosome-level assemblies.** Assembled chromosomes are shown in order of size and corresponding chromosomes are placed next to each other: B236-h1, B236-h2, B346-h1 and B346-h2. The plot was generated with HiGlass.

**Extended Data Figure 2. Riparian plot of synteny between the four assemblies based on orthologs in the structural annotation with GENESPACE shows similar results to the alignment-based synteny analysis with SyRI in Fig. 1a.**

**Extended Data Figure 3. Overview of gene density, repeat element (RE) fraction and SNP density in the remaining three assemblies.** Circos plots of assemblies B236-h2 (a), B346-h1 (b) and B346-h2 (c) show similar patterns across 12 chromosomes in regards to gene density, RE fraction and SNP density.

**Extended Data Figure 4. Sampling locations of *T. baccata* individuals used in this study.** Autochthonous populations are marked with orange points, while all other locations are marked with pink points. Numbers in the points indicate the number of samples taken from that population.

**Extended Data Figure 5. PCA of the unfiltered dataset show extreme distance between two *T. × media* individuals (on the left) and all other *T. baccata* individuals.**

**Extended Data Figure 6. GWAS results are consistent across haplotypes, showing no sex-associated SNPs.** The pairs of Manhattan plot and Q–Q plot were generated from variants in B236-h2 (a,b), B346-h1 (c,d) B346-h2 (e,f).

## REFERENCES

1. Leite Montalvão, A.P., Kersten, B., Fladung, M. & Müller, N.A. The Diversity and Dynamics of Sex Determination in Dioecious Plants. *Front. Plant Sci.* **11**:580488(2021).
2. Renner, S.S. & Müller, N.A. Sex determination and sex chromosome evolution in land plants. *Phil. Trans. R. Soc. B* **377**, 20210210 (2022).
3. Smith, B.W. The Mechanism Of Sex Determination In *Rumex Hastatulus*. *Genetics* **48**, 1265–1288 (1963).
4. Akagi, T. *et al.* Evolution and functioning of an X–A balance sex-determining system in hops. *Nat. Plants* **11**, 1339–1352 (2025).
5. Iwasaki, M. *et al.* Identification of the sex-determining factor in the liverwort *Marchantia polymorpha* reveals unique evolution of sex chromosomes in a haploid system. *Curr. Biol.* **31**, 5522–5532.e7 (2021).

- 687 6. Walas, Ł., Mandryk, W., Thomas, P.A., Tyrąła-Wierucka, Ż. & Iszkuło, G. Sexual systems in  
688 gymnosperms: A review. *Basic Appl. Ecol.* **31**, 1–9 (2018).
- 689 7. Leslie, A.B., Beaulieu, J.M., Crane, P.R. & Donoghue, M.J. Explaining the distribution of  
690 breeding and dispersal syndromes in conifers. *Proc. R. Soc. B.* **280**, 20131812 (2013).
- 691 8. Renner, S.S. The relative and absolute frequencies of angiosperm sexual systems: Dioecy,  
692 monoecy, gynodioecy, and an updated online database. *Am. J. Bot.* **101**, 1588–1596 (2014).
- 693 9. Schaefer, H. & Renner, S.S. A three-genome phylogeny of *Momordica* (Cucurbitaceae)  
694 suggests seven returns from dioecy to monoecy and recent long-distance dispersal to Asia.  
695 *Mol. Phylogenet. Evol.* **54**, 553–560 (2010).
- 696 10. Akagi, T. & Sugano, S.S. Random epigenetic inactivation of the X-chromosomal HaMStR gene  
697 causes sex ratio distortion in persimmon. *Nat. Plants* **10**, 1643–1651 (2024).
- 698 11. Charlesworth, D. & Harkess, A. Why should we study plant sex chromosomes? *Plant Cell* **36**,  
699 1242–1256 (2024).
- 700 12. Renner, S.S. & Ricklefs, R.E. Dioecy and its correlates in the flowering plants. *Am. J. Bot.* **82**,  
701 596–606 (1995).
- 702 13. Gong, W. & Filatov, D.A. Evolution of the sex-determining region in *Ginkgo biloba*. *Phil. Trans.*  
703 *R. Soc. B* **377**, 20210229 (2022).
- 704 14. Liu, Y. *et al.* The Cycas genome and the early evolution of seed plants. *Nat. Plants* **8**, 389–401  
705 (2022).
- 706 15. Sun, J.-J. *et al.* Genomic divergence and sex-linked region of *Welwitschia mirabilis* revealed  
707 by whole-genome re-sequencing. *Cell Rep.* **45**(2026).
- 708 16. Ohri, D. & Rastogi, S. Sex determination in gymnosperms. *Nucleus* **63**, 75–80 (2020).
- 709 17. World Flora Online. *Taxus* L. (Published on the Internet, 2026). accessed on: 21.02.2026
- 710 18. Anderson, E.D. Reproductive biology of Pacific yew (*Taxus brevifolia*) University of Victoria  
711 (B.C.) (2001).
- 712 19. DiFazio, S.P. The reproductive ecology of the pacific yew (*Taxus brevifolia* Nutt.) under a  
713 range of overstory conditions in western Oregon. Master Thesis, Oregon State University  
714 (1995).
- 715 20. Hogg, K.E., Mitchell, A.K. & Clayton, M.R. Confirmation of cosexuality in Pacific yew (*Taxus*  
716 *brevifolia* Nutt.). *BYU ScholarsArchive* **56**, 13 (1996).
- 717 21. Allison, T.D., Shimizu, T., Ohara, M. & Yamanaka, N. Variation in sexual reproduction in *Taxus*  
718 *cuspidata* Sieb. & Zucc. *Plant Species Biol.* **23**, 25–32 (2008).
- 719 22. Iszkuło, G. & Jasińska, A. Variation in sex expression in Polish and Ukrainian populations of  
720 *Taxus baccata* L. *Dendrobiology* **52**, 29–32 (2004).
- 721 23. Thomas, P.A. & Polwart, A. *Taxus baccata* L. *J. Ecol.* **91**, 489–524 (2003).
- 722 24. Zhang, S. *et al.* Climate change-driven threats to *Taxus* L. survival and strategies for  
723 sustainable Taxol resource management. *Ind. Crops Prod.* **235**, 121725 (2025).

- 724 25. Thakur, A. & Kanwal, K.S. Assessing the global distribution and conservation status of the  
725 *Taxus* genus: An overview. *Trees, Forests and People* **15**, 100501 (2024).
- 726 26. Li, Z. *et al.* Phased high-quality genome of the gymnosperm Himalayan Yew assists in  
727 paclitaxel pathway exploration. *GigaScience* **14**, giaf026 (2025).
- 728 27. Tomasino, M.P., Gennaro, A., Simeone, M.C., Schirone, B. & Ceoloni, C. New insights into the  
729 *Taxus baccata* L. karyotype based on conventional and molecular cytogenetic analyses.  
730 *Caryologia* **70**, 248–257 (2017).
- 731 28. Zarek, M. Preliminary studies on the molecular identification of sex in *Taxus baccata* L. *Leśne*  
732 *Prace Badawcze* **77**, 68–75 (2016).
- 733 29. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality,  
734 completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 1–27  
735 (2020).
- 736 30. Bross, D., Schroeder, H., Pers-Kamczyc, E. & Kersten, B. DNA markers targeting three cellular  
737 genomes for the discrimination between *Taxus baccata*, *T. cuspidata* and their hybrid. *Tree*  
738 *Genet. Genomes* **21**, 13–14 (2025).
- 739 31. Bachtrog, D. *et al.* Are all sex chromosomes created equal? *Trends Genet.* **27**, 350–357  
740 (2011).
- 741 32. Müller, N.A. X-specific methylation distorts sex. *Nat. Plants* **10**, 1619–1620 (2024).
- 742 33. Groh, J.S. *et al.* Ancient structural variants control sex-specific flowering time morphs in  
743 walnuts and hickories. *Science* **387**, eado5578 (2025).
- 744 34. Renner, S.S., Beenken, L., Grimm, G.W., Kocyan, A. & Ricklefs, R.E. The Evolution of Dioecy,  
745 Heterodichogamy, and labile sex expression in Acer. *Evol.* **61**, 2701–2719 (2007).
- 746 35. Blake-Mahmud, J. & Struwe, L. Time for a change: patterns of sex expression, health and  
747 mortality in a sex-changing tree. *Ann. Bot.* **124**, 367–377 (2019).
- 748 36. Cheng, J. *et al.* Chromosome-level genome of Himalayan yew provides insights into the origin  
749 and evolution of the paclitaxel biosynthetic pathway. *Mol. Plant* **14**, 1199–1209 (2021).
- 750 37. Song, C. *et al.* *Taxus yunnanensis* genome offers insights into gymnosperm phylogeny and  
751 taxol production. *Commun. Biol.* **4**, 1–8 (2021).
- 752 38. Xiong, X. *et al.* The *Taxus* genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**,  
753 1026–1036 (2021).
- 754 39. Liao, Q. *et al.* The genomic architecture of the sex-determining region and sex-related  
755 metabolic variation in *Ginkgo biloba*. *Plant J.* **104**, 1399–1409 (2020).
- 756 40. Carey, S.B. *et al.* ZW sex chromosome structure in *Amborella trichopoda*. *Nat. Plants* **10**,  
757 1944–1954 (2024).
- 758 41. Grayson, P., Wright, A., Garroway, C.J. & Docker, M.F. SexFindR: A computational workflow  
759 to identify young and old sex chromosomes. Preprint at  
760 <https://www.biorxiv.org/content/10.1101/2022.02.21.481346v1> (2022).

- 761 42. Charlesworth, D. Why and how do Y chromosome stop recombining? *J. Evol. Biol.* **36**, 632–  
762 636 (2023).
- 763 43. Charlesworth, B. & Charlesworth, D. A Model for the Evolution of Dioecy and Gynodioecy.  
764 *Am. Nat.* (1978).
- 765 44. Pannell, J.R. Plant Sex Determination. *Curr. Biol.* **27**, R191–R197 (2017).
- 766 45. Wang, W. *et al.* Genomic imprinting-like monoallelic paternal expression determines sex of  
767 channel catfish. *Sci. Adv.* **8**, eadc8786 (2022).
- 768 46. Zhang, X. *et al.* Xist in X chromosome inactivation: mechanisms and disease relevance. *Cell*  
769 *Commun. Signal.* **23**, 531 (2025).
- 770 47. Collins, D., Mill, R.R. & Möller, M. Species separation of *Taxus baccata*, *T. canadensis*, and *T.*  
771 *cuspidata* (Taxaceae) and origins of their reputed hybrids inferred from RAPD and cpDNA  
772 data. *Am. J. Bot.* **90**, 175–182 (2003).
- 773 48. Dubreuil, M. *et al.* Isolation and characterization of polymorphic nuclear microsatellite loci in  
774 *Taxus baccata* L. *Conserv. Genet.* **9**, 1665–1668 (2008).
- 775 49. Huang, C.-C., Chiang, T.-Y. & Hsu, T.-W. Isolation and characterization of microsatellite loci in  
776 *Taxus sumatrana* (Taxaceae) using PCR-based isolation of microsatellite arrays (PIMA).  
777 *Conserv. Genet.* **9**, 471–473 (2008).
- 778 50. Ueno, S., Wen, Y. & Tsumura, Y. Development of EST-SSR markers for *Taxus cuspidata* from  
779 publicly available transcriptome sequences. *Biochem. Syst. Ecol.* **63**, 20–26 (2015).
- 780 51. Krautwurst, M. *et al.* High-molecular-weight DNA extraction for broadleaved and conifer tree  
781 species. *Silvae Genet.* **73**, 85–98 (2024).
- 782 52. Baid, G. *et al.* DeepConsensus improves the accuracy of sequences with a gap-aware  
783 sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2023).
- 784 53. Bruegmann, T., Fladung, M. & Schroeder, H. Flexible DNA isolation procedure for different  
785 tree species as a convenient lab routine. *Silvae Genetica* **71**, 20-30 (2022).
- 786 54. Ziegenhagen, B., Guillemaut, P. & Scholz, F. A procedure for mini-preparations of genomic  
787 DNA from needles of silver fir (*Abies alba* mill.). *Plant Mol. Biol. Report.* **11**, 117–121 (1993).
- 788 55. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for  
789 diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
- 790 56. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo  
791 assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- 792 57. Zhou, C., McCarthy, S.A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*  
793 **39**, btac808 (2023).
- 794 58. Karlicki, M., Antonowicz, S. & Karnkowska, A. Tiara: deep learning-based classification system  
795 for eukaryotic sequences. *Bioinformatics* **38**, 344–350 (2022).
- 796 59. Uliano-Silva, M. *et al.* MitoHiFi: a python pipeline for mitochondrial genome assembly from  
797 PacBio high fidelity reads. *BMC Bioinf.* **24**, 1–13 (2023).

798 60. Laetsch, D.R. & Blaxter, M.L. BlobTools: Interrogation of genome assemblies. *F1000Research*  
799 **6**, 1287 (2017).

800 61. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction  
801 maps. *Genome Biol.* **19**, 1–12 (2018).

802 62. Chandra, G., Vasimuddin, M., Misra, S. & Jain, C. Accelerating whole-genome alignment in  
803 the age of complete genome assemblies. Preprint at  
804 <https://www.biorxiv.org/content/10.1101/2024.11.25.625328v1> (2024).

805 63. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and  
806 local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 1–13 (2019).

807 64. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements  
808 between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).

809 65. Lovell, J.T. *et al.* GENESPACE tracks regions of interest and gene copy number variation  
810 across multiple genomes. *eLife* **11**(2022).

811 66. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped  
812 cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).

813 67. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*  
814 **34**, W435–W439 (2006).

815 68. Flynn, J.M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element  
816 families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020).

817 69. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large  
818 genomes. *Bioinformatics* **21**, i351–i358 (2005).

819 70. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*  
820 **27**, 573–580 (1999).

821 71. Bao, Z. & Eddy, S.R. Automated De Novo Identification of Repeat Sequence Families in  
822 Sequenced Genomes. *Genome Res.* **12**, 1269–1276 (2002).

823 72. Ou, S. & Jiang, N. LTR\_retriever: A Highly Accurate and Sensitive Program for Identification of  
824 Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

825 73. Wheeler, T.J. Large-Scale Neighbor-Joining with NINJA. in *Algorithms in Bioinformatics* 375–  
826 389 (Springer, Berlin, Germany, 2009).

827 74. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation  
828 sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

829 75. Nakamura, T., Yamada, K.D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale  
830 multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).

831 76. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: A Comprehensive Software Library for  
832 Efficient Processing of Structured Genome Annotations. *IEEE/ACM Trans. Comput. Biol.*  
833 *Bioinf.* **10**, 645–656 (2013).

834 77. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–  
835 4574 (2021).

836 78. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100  
837 (2018).

838 79. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein  
839 evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).

840 80. Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic  
841 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein  
842 database. *NAR Genomics Bioinf.* **3**, lqaa108 (2021).

843 81. Hoff, K.J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with  
844 BRAKER. in *Gene Prediction: Methods and Protocols*, Vol. 1962 65–95 (Springer, New York,  
845 NY, USA, 2019).

846 82. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised  
847 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**,  
848 767–769 (2016).

849 83. Brůna, T., Gabriel, L. & Hoff, K.J. Navigating Eukaryotic Genome Annotation Pipelines: A  
850 Route Map to Using BRAKER, Galba, and TSEBRA. in *Insect Genomics: Methods and Protocols*,  
851 Vol. 2935 67–107 (Springer US, 2025).

852 84. Olsson, S. *et al.* De novo assembly of English yew (*Taxus baccata*) transcriptome and its  
853 applications for intra- and inter-specific analyses. *Plant Mol. Biol.* **97**, 337–345 (2018).

854 85. Tegenfeldt, F. *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider  
855 sampling of genomes. *Nucleic Acids Res.* **53**, D516–D522 (2025).

856 86. Consortium, T.U. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.*  
857 **53**, D609–D617 (2025).

858 87. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and  
859 genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

860 88. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy  
861 of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).

862 89. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat.*  
863 *Methods* **12**, 59–60 (2015).

864 90. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an  
865 extended version of Spaln that incorporates additional species-specific features. *Nucleic*  
866 *Acids Res.* **40**, e161–e161 (2012).

867 91. Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences  
868 onto genomic sequence. *Nucleic Acids Res.* **36**, 2630–2638 (2008).

869 92. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2.  
870 *Genome Biol.* **20**, 1–13 (2019).

871 93. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).

872 94. Gabriel, L., Hoff, K.J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for  
873 BRAKER. *BMC Bioinf.* **22**, 1–12 (2021).

- 874 95. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
- 875 96. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P. & Marth, G.T. BamTools: a C++  
876 API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
- 877 97. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols*  
878 *in Bioinformatics* **47**, 11.12.1–11.12.34 (2014).
- 879 98. Shumate, A. & Salzberg, S.L. Liftoff: accurate mapping of gene annotations. *Bioinformatics*  
880 **37**, 1639–1643 (2021).
- 881 99. Conway, J.R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of  
882 intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
- 883 100. Hart, A.J. *et al.* EnTAP: Bringing faster and smarter functional annotation to non-model  
884 eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020).
- 885 101. Haas, B. TransDecoder. (2023). <https://github.com/TransDecoder/TransDecoder/>
- 886 102. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically  
887 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*  
888 **47**, D309–D314 (2019).
- 889 103. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-  
890 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the  
891 Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
- 892 104. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics.  
893 *Bioinformatics* **33**, 2759–2761 (2017).
- 894 105. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence  
895 data. *Bioinformatics* **30**, 2114–2120 (2014).
- 896 106. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. (2010).  
897 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- 898 107. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for  
899 multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
- 900 108. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of  
901 BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed*  
902 *Processing Symposium (IPDPS)* 20–24 (IEEE, 2019).
- 903 109. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality  
904 control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
- 905 110. Institute, B. Picard Toolkit. (Broad Institute, GitHub repository, 2019).
- 906 111. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, (Springer-Verlag New York, Cham,  
907 Switzerland, 2016).
- 908 112. R Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for  
909 Statistical Computing, Vienna, Austria, 2024). <https://www.R-project.org/>

910 113. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based  
911 Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

912 114. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies.  
913 *Nat. Genet.* **44**, 821–824 (2012).

914 115. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

915 116. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for  
916 RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

917 117. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing  
918 next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

919 118. Wenger, A.M. *et al.* Accurate circular consensus long-read sequencing improves variant  
920 detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

921 119. PacificBiosciences. Jasmine. (2026). <https://github.com/PacificBiosciences/jasmine>

922 120. PacificBiosciences. pbmm2. (2026). <https://github.com/PacificBiosciences/pbmm2>

923 121. PacificBiosciences. pb-CpG-tools. (2026). <https://github.com/PacificBiosciences/pb-cpg-tools>  
924