Early sex-chromosome evolution in the diploid dioecious plant *Mercurialis annua*

Paris Veltsos[*,†], Kate E. Ridout[*, §,‡], Melissa A. Toups[*,**,††], Santiago C. González-Martínez[‡‡], Aline Muyle[§§], Olivier Emery[***], Pasi Rastas[†††], Vojtech Hudzieczek[§§§], Roman Hobza[§§§], Boris Vyskot[§§§], Gabriel A.B. Marais[§§], Dmitry A. Filatov[§], and John R. Pannell[*]

[*] Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland

[†] Current address: Department of Biology, Jordan Hall, 1001 East Third Street, Indiana University, Bloomington, IN 47405, United States of America

[§] Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, United Kingdom

[‡] Department of Oncology, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom

[**] Current address: Institute of Science and Technology Austria, Am Campus 1A, Klosterneuburg 3400, Austria

[††] Department of Integrative Biology, University of Texas, Austin, Texas, United States of America

[‡‡] BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France

[§§] Laboratoire Biométrie et Biologie Évolutive (UMR 5558), CNRS / Université Lyon 1, 69100, Villeurbanne, France.

[***] Current address: Department of Fundamental Microbiology, University of Lausanne, CH-1015 Lausanne, Switzerland.

[†††] University of Helsinki, Institute of Biotechnology, P.O.Box 56, 00014, Finland

[§§§] Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Kralovopolska 135, 61200, Brno, Czech Republic.

The first two authors contributed equally to the research.

The last three authors supervised the research.

Running title: Sex chromosome evolution in *Mercurialis annua*

Key words: sex chromosomes, whole genome sequencing, sex linkage, evolutionary strata, gene expression

Author for correspondence: John R. Pannell

Email: john.pannell@unil.ch

Telephone: +4121 692 41 70

Department of Ecology and Evolution

Le Biophore - Unil Sorge

University of Lausanne

Lausanne

1015

Switzerland

Abstract

Suppressed recombination around a sex-determining locus allows divergence between homologous sex chromosomes and the functionality of their genes. Here, we reveal patterns of the earliest stages of sex-chromosome evolution in the diploid dioecious herb *Mercurialis annua* on the basis of cytological analysis, *de novo* genome assembly and annotation, genetic mapping, exome resequencing of natural populations, and transcriptome analysis. Both genetic mapping and exome resequencing of individuals across the species range independently identified the largest linkage group, LG1, as the sex chromosome. Although the sex chromosomes of *M. annua* are karyotypically homomorphic, we estimate that about a third of the Y chromosome has ceased recombining, a region containing 568 transcripts and spanning 22.3 cM in the corresponding female map. Patterns of gene expression hint at the possible role of sexually antagonistic selection in having favored suppressed recombination. In total, the genome assembly contained 34,105 expressed genes, of which 10,076 were assigned to linkage groups. There was limited evidence of Y-chromosome degeneration in terms of gene loss and pseudogenization, but sequence divergence between the X and Y copies of many sex-linked genes was higher than between *M. annua* and its dioecious sister species *M. huetii* with which it shares a sex-determining region. The Mendelian inheritance of sex in interspecific crosses, combined with the other observed pattern, suggest that the *M. annua* Y chromosome has at least two evolutionary strata: a small old stratum shared with *M. huetii*, and a more recent larger stratum that is probably unique to *M. annua* and that stopped recombining about one million years ago.

Article summary Plants that evolved separate sexes (dioecy) recently are ideal models for studying the early stages of sex-chromosome evolution. Here, we use karyological, whole genome and transcriptome data to characterize the homomorphic sex chromosomes of the annual dioecious plant *Mercurialis annua*. Our analysis reveals many typical hallmarks of dioecy and sex-chromosome evolution, including sex-biased gene expression and high X/Y sequence divergence, yet few premature stop codons in Y-linked genes and very little outright gene loss, despite 1/3 of the sex chromosome having ceased recombination in males. Our results confirm that the *M. annua* species complex is a fertile system for probing early stages in

the evolution of sex chromosomes.

Introduction

Separate sexes, or dioecy, have evolved repeatedly from hermaphroditism in flowering plants, with about half of all angiosperm families having dioecious members (Renner and Ricklefs, 1995; Renner, 2014). The evolution of dioecy sets the stage for the possible evolution of sex chromosomes (Charlesworth and Charlesworth, 1978), which originate as a pair of homologous autosomes that may gradually diverge in sequence when recombination between them is suppressed. Such divergence is extreme in many animal species with old sex chromosomes, for which it can be difficult even to identify homologous sequences. In some plants, like *Silene latifolia* (Ono, 1939; Krasovec *et al.*, 2018) and *Rumex hastatulus* (Smith, 1955; Hough *et al.*, 2014), the homologous chromosomes have diverged sufficiently to become heteromorphic and distinguishable by the karyotype. In other plants, like *Asparagus officinalis* (Loeptien, 1979; Telgmann-Rauber *et al.*, 2007) and *Carica papaya* (Horovitz and Jiménez, 1967; Liu *et al.*, 2004), the sex chromosomes remain indistinguishable by the karyotype, and compromised gene function may be mild. Early sex-chromosome evolution usually involves the accumulation of repetitive sequences in a non-recombining region, but also loss of genes, which is presumably preceded by pseudogenization and loss of function.

Dioecious plants provide good models for studying the evolution of sex chromosomes because separate sexes have often evolved recently, potentially allowing us to observe the very earliest stages in sex-chromosome divergence (reviewed in Charlesworth (2016)). To date, the most comprehensive and detailed work on the evolution of plant sex chromosomes comes from the study of species with highly divergent heteromorphic sex chromosomes, such as *S. latifolia* (Kejnovsky and Vyskot, 2010) and *Rumex* species (Navajas-Pérez *et al.*, 2005), i.e., those that have already proceeded down a path of Y-chromosome degeneration and divergence from their X homologue. In these species, it has been possible to measure rates of loss of genes or their function (Krasovec *et al.*, 2018), differences in codon use between homologues (Qiu *et al.*, 2011), and differences in patterns of gene expression at sex-linked loci (Zemp *et al.*, 2016). Such measurements have generally confirmed the idea that non-recombining parts of sex chromosomes quickly degenerate in function, as expected by theory (Lynch *et al.*, 2016). For instance, degeneration may occur through structural genetic changes (Charlesworth *et al.*, 2005), or the accumulation

of repetitive elements (Wang *et al.*, 2012b) and deleterious mutations (Charlesworth *et al.*, 2005; Charlesworth, 2013), driven by processes such as background selection (Charlesworth *et al.*, 1993b), selective sweeps (Maynard Smith and Haigh, 1974) or Muller's Ratchet (Charlesworth *et al.*, 1993a).

Two hypotheses have been proposed to explain the suppression of recombination on the sex chromosomes of plants that have evolved dioecy from hermaphroditism. One hypothesis postulates that dioecy initially evolves through the spread of male- and female-sterility mutations, and that these mutations must become linked on opposite chromosomes to avoid the expression of either hermaphroditism, or both male and female sterility simultaneously (Charlesworth and Charlesworth, 1978). The main experimental support for this two-locus model comes from classic genetic studies in *Silene latifolia* (Westergaard, 1958) that demonstrated the presence of two sex-determining factors on the Y-chromosome, the stamen-promoting factor (SPF) and gynoecium suppression factor (GSF). More recent work mapped the location of SPF and GSF genes on the *S. latifolia* Y-chromosome (Kazama *et al.*, 2016), though the actual GSF and SPF genes have yet to be identified. Nevertheless, despite the likely importance of two loci in the evolution of dioecy, the two-locus model does not readily explain why non-recombining regions on sex chromosomes often expand greatly (Bergero and Charlesworth, 2009), probably well beyond the region harboring the original sex-determining genes.

The second hypothesis invokes expansion of the non-recombining region and divergence of the sex chromosomes with the accumulation of genes that have different effects on the fitness of males and females (Gibson *et al.*, 2002). This generates selection for reduced recombination between the sex-determining allele and nearby alleles with sex-specific effects (Charlesworth, 1991; Charlesworth *et al.*, 2005; Bergero and Charlesworth, 2009), so that alleles with sex-specific fitness benefits (i.e., sexually antagonistic alleles) occur on the sex chromosome that spends most of its time in individuals of the sex to which the benefits apply (Rice, 1987). Mutations that differentially affect gene expression between males and females are also expected to be enriched in the non-recombining region of the sex chromosomes compared with other regions of the genome, as found in *S. latifolia* (Zemp *et al.*, 2016). The suppression of recombination is expected to extend consecutively to generate linkage between the sex-determining locus and more sexually

antagonistic loci (Charlesworth, 2015), and these extensions can be identified as discrete 'strata', with greater genetic divergence between the sex chromosomes in strata that ceased recombining earliest. Evidence for such strata has been found both in animals (Lahn and Page, 1999; Nam and Ellegren, 2008) and plants (Bergero *et al.*, 2007; Wang *et al.*, 2012b).

If we are to make further progress in understanding the initial suppression of recombination associated with Y-chromosome degeneration, we need to consider systems in which the very earliest stages of sex-chromosome evolution might be observed, either in the youngest evolutionary strata of old (heteromorphic) sex chromosomes, or in sequences close to the sex-determining locus of young homomorphic sex chromosomes that have only just started to diverge. Some progress has been made in studies of several species with homomorphic sex chromosomes, e.g., *Asparagus officinalis* (Loeptien, 1979; Telgmann-Rauber *et al.*, 2007), *Spinacia oleracea* (Yamamoto *et al.*, 2014), *Diospyros lotus* (Akagi *et al.*, 2014), *Fragaria chiloensis* (Tennessen *et al.*, 2016), *Populus* (Geraldes *et al.*, 2015), *Carica papaya* (Horovitz and Jiménez, 1967; Liu *et al.*, 2004) and *Salix* (Pucholt *et al.*, 2015). However, our understanding of the early stages of sex-chromosome evolution remains poor.

Here, we report evidence for the very early evolution of sex chromosomes in the sexually dimorphic dioecious herb *Mercurialis annua* (Euphorbiacae), a plant in which separate sexes have been much studied in the past and that has also become a model to understand evolutionary transitions between combined and separate sexes (Obbard *et al.*, 2006), and the evolution of sexual dimorphism (Sánchez-Vilas *et al.*, 2010). Until recently, gender in *M. annua* was thought to be determined by allelic variation at three independent loci (Durand *et al.*, 1987; Durand and Durand, 1991), but it is now known to have a simple XY system (Khadka *et al.*, 2005; Russell and Pannell, 2015). The sex chromosomes in *M. annua* are homomorphic (Durand, 1963), and crosses between males with 'leaky' gender expression reveal that YY males are viable but partially sterile (Kuhn, 1939). It thus seems that the *M. annua* Y chromosome is in the early stages of degeneration.

Our study is based substantially on *de novo* assembly and annotation of the diploid *M. annua* genome (2n = 2x = 16). Whole-genome data for species with separate sexes are scarce in plants, with the few

exceptions being *Carica papaya* (Liu *et al.*, 2004), *Vitis* (Fechter *et al.*, 2012), *Diospyros lotus* (Akagi *et al.*, 2014), and *Populus* (Tuskan *et al.*, 2006), which have largely homomorphic sex chromosomes (Filatov, 2015). For the genome assembly of *M. annua*, we combined short-read sequencing on the Illumina platform with long-read technology developed by Pacific Biosciences. We also analyzed the karyotype of *M. annua* using current imaging technology in an attempt to identify heteromorphism at the cytological level, and we obtained SNPs from the transcriptomes of small families to construct a genetic map for *M. annua*, to identify non-recombining (sex-linked) genes and scaffolds, and to compare them with non-sex-linked regions. We then investigated the sex chromosomes with regard to evolutionary strata and degeneration, including the fixation pseudogenising mutations in the inferred fully Y-linked sequences, and the proportion of genes deleted from the Y. Finally, we examined sex-biased gene expression and assessed whether sex-biased genes might be enriched on the sex chromosomes, as expected by theory (Connallon and Clark, 2010; Meisel *et al.*, 2012). Our results suggest that the sex chromosomes of *M. annua* are about 1.5 million years old. Early stages of Y-chromosome degeneration are clearly apparent, but there are also signs that patterns of gene expression might have been affected by the accumulation of sexually antagonistic mutations.

Materials and Methods

Hairy root culture, chromosome preparation and cytogenetics

Mercurialis annua seeds were sterilized by incubation in 4% sodium hypochloride and subsequently washed in 50% ethanol and sterile water. Seeds were grown on MS medium and leaf discs were taken from 2-3 week old plants, which had been sexed using a previously developed Sequence Characterised Amplified Region (SCAR) marker (Khadka *et al.*, 2002). *Agrobacterium rhizogenes* strain ARqua1 (Quantdt *et al.*, 1993) was grown overnight at 28°C in LB medium supplemented with 300 µM acetosyringone to OD_600 = 0.6. The culture was resuspended in liquid 1/2 MS with 300 µM acetosyringone and used for direct inoculation of *Mercurialis annua* leaf discs. Explants were cocultivated at 28°C on MS medium with 300µM acetosyringone for 2 days. After cocultivation, explants were moved to MS medium supplemented with 300µg/L Timentin. Media were changed every 2 weeks.

In order to synchronize the hairy roots of *Mercurialis annua*, the DNA polymerase inhibitor aphidicolin was added for 12 h. Mitoses were then accumulated in protoplasts using oryzalin treatment, transferred on chromic acid washed slides by dropping, and stored at -20 ^{o}C until use. For FISH experiments, slides were denatured in 7:3 (v/v) formamide: 2xSSC for 2 min at 72°C, immediately dehydrated through 50%, 70%, and 100% ethanol (-20°C), and air dried. The hybridization mixture (30 µl per slide) consisted of 200 ng of labeled probe, 15 µl formamide, 6 µl 50% dextrane sulfate, and 3 µl of 20x SSC. The volume was brought to 30 µl by adding TE, pH 8. The probes were denatured at 70°C for 10 min, and slides hybridized for 18 h at 37°C in a humid chamber. Slides were analyzed using an Olympus Provis AX70 microscope, and image analysis was performed using ISIS software (Metasystems). DNA was labeled with Fluorolink Cy3-dUTP (Amersham Pharmacia Biotech; red labeling) using a Nick Translation mix (Roche) or with SpectrumGreen direct-labeled dUTP (Vysis; green labeling) and a Nick Translation kit (Vysis).

Genome and transcriptome sequencing

Diploid *Mercurialis annua* individuals from north-western France were grown together in the glasshouse. DNA for genomic sequencing was taken from leaf tissue of a single male, M1. RNA samples were collected from leaves and flower buds of this individual as well as from three females, G1, G2 and

G3, and two additional males, M2 and M3, all of which were unrelated. 65 F_1 and F_2 progeny (five families) were then produced by crossing G1xM1 and G2xM1 (F_1) as well as three pairs of G1xM1 progeny (F_2) (Table S1), which were also used for RNA extraction and transcriptome sequencing. DNA was extracted using a Qiagen Plant DNeasy kit (Qiagen, Hilden, Germany). Illumina paired-end and mate-pair sequencing was carried out by the Beijing Genomics Institute (BGI) using Illumina HiSeq 2000 technology (100 bp reads). Pacific Biosciences long-read sequencing was performed on the individual M1 by the Centre for Integrative Genomics hosted at the University of Lausanne. RNA was extracted from a mixture of flower buds and leaf tissues using the Qiagen plant RNAeasy kit (Qiagen, Hilden, Germany), and individual libraries were prepared for all 65 individuals (Table S1), which were sequenced on three lanes of Illumina HiSeq 2000 at the Wellcome Trust Centre for Human Genetics, Oxford.

Transcriptome assembly and annotation

The SEX-DETector pipeline (http://lbbe.univ-lyon1.fr/-SEX-DETector-.html?lang=fr) was used to produce the reference transcriptome (henceforth 'ORFs'), and obtain allele-specific expression and SNPs for X and Y haplotypes. Sex linkage was inferred from genetic mapping which used more data (see below). To identify sex-linked haplotypes, we employed a probabilistic model based on maximum-likelihood inference, implemented in SEX-DETector (Muyle *et al.*, 2016). SEX-DETector is based on the idea that X-linked and Y-linked haplotypes are passed only from fathers to daughters or from fathers to sons, respectively. SEX-DETector is embedded into a Galaxy workflow pipeline that includes extra-assembly, mapping and genotyping steps prior to sex-linkage inference, and has been shown to have greater sensitivity, without an increased false positive rate, than without these steps (Muyle *et al.*, 2016). First, poly-A tails were removed from transcripts using PRINSEQ (Schmieder and Edwards, 2011), with parameters -trim_tail_left 5 -trim_tail_right 5. rRNA-like sequences were removed using riboPicker version 0.4.3 (Schmieder and Edwards, 2011) with parameters -i 90 -c 50 -l 50 and the following databases: SILVA Large subunit reference database, SILVA Small subunit reference database, the GreenGenes database and the Rfam database. Transcripts were then further assembled within Trinity components using cap3 (Huang and Madan, 1999), with parameter -p 90 and custom Perl scripts. Coding sequences were predicted using Trinity TransDecoder (Haas *et al.*, 2013), including PFAM domain

searches as ORF retention criteria; these sequences were taken as our reference transcriptome. The RNAseq reads from parents and progeny were mapped onto the reference transcriptome using BWA (Li and Durbin, 2009). The alignments were analyzed using reads2snp, a genotyper for RNAseq data that gives better results than standard genotypers when X and Y transcripts have different expression levels (Tsagkogeorga *et al.*, 2012).

The SEX-DETector transcripts were mapped onto the genome using GMAP v2017-06-20 (Wu and Watanabe, 2005), disallowing chimeric genes, mapping 38,963 (99%) of them to 48,298 loci. We then collapsed the inferred ORFs into 34,105 gene models using gffread v0.9.9 (http://ccb.jhu.edu/software/stringtie/gff.shtml). In total, 26,702 (67.9%) of the ORFs could be annotated using fast blastx on the Swiss-Prot database and Blast2GO (Conesa *et al.*, 2005) in Geneious v9 (Kearse *et al.*, 2012). The resulting transcriptome annotation information, including GO terms, is presented in File S4. We characterized genes as candidate sex-determining genes by identifying those containing the following strings in their annotation: flower, auxin, cytokinin, pollen, kelch, ethylene, swi, retinol, calmodulin, short-root, jasmonic. This manual characterization allowed us to visualize the mapping locations, test for their enrichment in the non-recombining region, and identify whether any were among the outliers in various transcript-associated metrics.

Genome size estimation, assembly and annotation

The genome size was estimated from the distribution of k-mers (size 31) obtained from jellyfish v2.1.0 (Marçais and Kingsford, 2011) and analyzed by the web version of GenomeScope v1 (Vurture *et al.*, 2017), limiting max k-mer coverage to 1000.

Sliding-window trimming and adaptor removal were carried out using Trimmomatic v. 0.30, with the default parameters (Bolger *et al.*, 2014). Exact duplicate read pairs were collapsed using fastx-collapser from the Fastx-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Low complexity masking was carried out using DUST (Morgulis *et al.*, 2006), with the default parameters, and ambiguous reads were removed. Pacific Biosciences long reads were error-corrected using Bowtie2 version 2.1.0 (Langmead and Salzberg, 2012) in combination with LSC version 0.3.1 (Au *et al.*, 2012).

Filtered paired-end and mate-paired genomic reads were assembled using SOAPdenovo2 (Luo *et al.*, 2012), with k-mer values between 35 and 55 (odd values only). The best assembly was chosen using REAPR (Hunt *et al.*, 2013). GapCloser (Luo *et al.*, 2012) was run on the best assembly to correct false joins and fill gaps. Error-corrected Pacific Biosciences reads were then used to extend scaffolds, to fill gaps, and to join scaffolds using PBJelly2 (English *et al.*, 2012). Additional scaffolding was carried out using default parameters on SSPACE (Boetzer *et al.*, 2011), which revisits gaps using existing paired-end and mate-paired sequences. This step is intended to correct for any ambiguity introduced by the low-coverage of PacBio reads. Finally, L_RNA_Scaffolder (Xue *et al.*, 2013) was used to bridge genomic scaffolds using the transcript assembly.

Transposable elements (TEs) and tandem repeats were predicted using a combination of Tandem Repeats Finder (Benson, 1999), RepeatModeler and RepeatMasker (Smit *et al.*, 2013). Repeat libraries from *Mercurialis* (from RepeatModeler), Euphorbiacae and *Vitis vinifera* were used for masking the genome before further analyses.

Summary of SNP calling pipelines

We called SNPs using four different pipelines. SNPs were called on RNAseq data from the two largest families by SEX-DETector, to call X- and Y-specific SNPs, and infer haplotypes with premature stop codons and allele-specific expression. SNPs were also called from genome capture data of multiple diploid populations, for independent analysis of genomic regions of differing sex linkage as inferred from the genetic map. In addition, SNPs were called on all individuals from five families, to perform genetic mapping. Finally, SNPs were called on transcripts from the three unrelated males (M1, M2 and M3) and the three unrelated females (G1, G2 and G3) for comparisons between ORFs of differing sex linkage and comparisons of pairwise diversity between species. The subset of these for which there was evidence of separate X and Y copies, was analyzed in a phylogenetic framework to estimate the (tree-based) evolution rate of the Y sequences. Specific details of each pipeline are given in the corresponding sections below.

Genetic map construction

We called SNPs independently of the SEX-DETector pipeline, using all individuals available over the

three generations of our family crosses. We first mapped RNAseq reads to the reference transcriptome using Bowtie2 v2.3.1 (Langmead and Salzberg, 2012). The resulting sam files were converted to bam, sorted and converted to mpileup format using samtools v 1.3 (Li *et al.*, 2009). The resulting posterior file containing segregation information of transcripts in the five mapping families (>164,000 markers and their genotype likelihoods) was taken through the LepMap3 (LM3) pipeline (Rastas, 2017). First, we calculated the relatedness between individuals using the IBD module (of LM3) using a random subset of 3,000 markers. Three individuals were discarded because their relatedness to their putative father (M6) was < 0.2, i.e., they were likely the result of a different cross through contamination. The parental genotypes were then called using ParentCall2, with parameter halfSibs=1, to take into account the half-sib family structure, in addition to the genotypic information of offspring, parents and grandparents. The 158,000 remaining markers were further filtered using the Filtering2 module, with parameter dataTolerance=0.001, to remove markers segregating in a distorted fashion. We then combined markers of each transcript into pseudo-markers, as they were not informative enough on their own to separate into linkage groups. This was done by running OrderMarkers2 separately for each transcript without recombination (parameter values: recombination1=0, recombination2=0), and then obtaining genotype likelihoods for each transcript with parameter outputPhasedData=4. This resulted in 13,261 markers, each consisting of a single transcript, which we re-formatted as input for LM3. We defined linkage groups by running SeparateChromosomes2 with default parameters (lodLimit=10), adding additional markers to the resulting 8 linkage groups using JoinSingles2 with lodLimit=8 and lodDifference=2. We were able to assign 10,076 transcripts into eight linkage groups, which were ordered using OrderMarkers2. The linkage maps were drawn with R/qtl v1.41 (Broman *et al.*, 2003) running in R 3.4.2 (R Development Core Team, 2007).

Sequence divergence in *M. annua*, and in comparison with closely related species

We aligned reads from six unrelated individuals from France (M1, M2, M3, G1, G2, G3) to the 39,302 open reading frames from the reference transcriptome assembled via the SEX-DETector pipeline, based on the M1 male transcriptome and using the BWA-MEM algorithm of BWA v0.7.13 (Li and Durbin, 2009). Picard Tools v2.2.1 (http://broadinstitute.github.io/picard/) were used to mark duplicate read pairs. Local realignment around insertions and deletions (indels) was performed with GATK v3.7 (DePristo *et al.*,

2011), followed by SNP calling on each individual using the HaplotypeCaller module in GATK (McKenna *et al.*, 2010). Joint SNP calling was performed using GATK's GenotypeGVCFs module, overlapping SNPs and indels were filtered out with the VariantFiltration module, and SNPs and indels were separated and filtered to produce two high-quality variant sets with the following parameters in the VariantFiltration module: 'QUAL < 30' 'DP < 30' 'MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)' 'QD < 5.0'. The high-quality SNP set was used to perform variant quality score recalibration to filter the full SNP set. This set was used to calculate within-species nucleotide diversity (π) using vcftools v 0.1.13 (Danecek *et al.*, 2011).

Pairwise d_N and d_S values from *M. huetii* and *R. communis* were obtained using yn00 from PAML 4.9 (Yang, 2007). This analysis used 4,761 and 2,993 homologues from the *M. huetii* and *R. communis* (http://castorbean.jcvi.org/introduction.shtml) transcriptomes, based on a *de novo* assembly using default settings in Trinity (Haas *et al.*, 2013; File S5). Sequences were aligned with LASTZ (Harris, 2007). Effective codon usage was calculated using ENCprime (https://jnpopgen.org/software/).

To estimate the rate of evolution of Y sequences, we identified one-to-one orthologs between the *M. annua* sex linked genes and the closely related species *M. huetii* and *R. communis*, through reciprocal best BLASTP, with an e-value of $1e^{-50}$, a culling limit of 1, and considering only transcripts that were not split across contigs. The 98 orthologs identified were aligned with LASTZ (Harris, 2007) and analysed with PAML 4.9 (Yang, 2007) to obtain tree-based estimates of d_N and d_S values. We employed three models: the 'null model' (M_0), where all parts of the tree are assumed to have the same d_N/d_S values; the 'branch-specific' model (M_1), where each branch may have its own d_N/d_S value; and the 'two-ratio' model (M_2), where the branch of interest (Y) may have a d_N/d_S ratio that differs from the rest of the tree (Nielsen and Yang, 1998). Preliminary analysis revealed similar results for the M_1 and M_2 models (results not shown); we thus chose the M_2 as the simpler model for further analysis. We tested whether the Y branch evolved differently from the rest of the tree, using a likelihood ratio test (LRT) between the M_2 and M_0 models, with one degree of freedom and LRT = 2 x abs(l_2-l_0), where l_2 and l_0 are the likelihoods for M_2 and M_0 respectively.

Expression analysis

Reads from the RNA libraries of all 30 females and 35 males were pseudoaligned to the reference transcriptome using Kallisto v0.43.1 (Bray *et al.*, 2016). The resulting raw count data were compiled in a table, and differential expression analysis between male and female samples was conducted with edgeR v3.18.1 (Robinson *et al.*, 2010), running in R v.3.4.0 (R Development Core Team, 2007). Genes were filtered from the analysis if their average $\log_2$ count per million (as computed by edgeR's aveLogCPM function) was negative. This retained genes with an average count of 24 or more per sample. Two individual libraries were removed that were obvious outliers in the MDS plots of the filtered data. Libraries were normalized with the default (TMM) normalization. Dispersion was measured with default parameters using a negative binomial model. Sex-biased transcripts were defined statistically, i.e. only by a false discovery rate (FDR – Benjamini and Hochberg; 1995) of 5%. A statistical definition for sex-biased genes is justified because our RNAseq analysis was based on atypically many samples, and because the sex-biased categories (male-, female- and unbiased) used in our analysis capture biological information in that they differed statistically in various measurements. This would also ensure that our results for effects of sex bias should be robust, because our approach would dilute "real" sex-biased genes with unbiased genes. We note, however, that the lack of minimum $\log_2$FC in our definition of sex-bias means that we cannot differentiate between allometric differences in tissue composition or gene regulation differences between males and females (Montgomery and Mank, 2016).

SEX-DETector calculates X- and Y-allele expression from each male. X and Y read numbers were summed for each contig and individual separately and divided by the number of X/Y SNPs of the contig and adjusted for the library size of the respective individual. X and Y expression levels were then averaged among individuals, and the ratio of the means was computed. Gene ontology (GO) enrichment analysis was performed on the sex-biased genes, as well as separately for the male- and female-biased genes, using topGO with the weight01 algorithm, which accounts for GO topology (Alexa and Rahnenfuhrer, 2010).

Comparison of expression between sex-linked genes and autosomal genes

We categorized transcripts based on their genomic location into fully sex-linked (SL; transcripts in the non-recombining region of LG1 in males), pseudo-autosomal (PAR; transcripts in the rest of LG1 that

showed recombination in both sexes) and autosomal (Au; transcripts that mapped to the remaining seven linkage groups). We investigated the effects of sex linkage, sex-biased expression and their interactions through linear models using R v 3.4.2 on the following metrics: Absolute male/female $\log_2$FC (fold-change, i.e., overall sex-bias intensity); nucleotide diversity (π); d_N/d_S from pairwise comparisons with *M. huetii* and *R. communis*; and Y/X allele-specific $\log_2$FC obtained from SEX-DETector. The data were not normally distributed and were thus Box-Cox-transformed after adding $1e^{-11}$ to all values to allow inclusion of zero values and estimating the transformation parameter using the command boxcoxnc with the Anderson-Darling method in the R package AID v2.3 (Dag *et al.*, 2014). Variables Absolute male/female ($\log_2$FC) and π remained non-normal after transformation when unbiased genes were included in the analysis, because there were too many transcripts with ~ 0 values. Exclusion of these transcripts' values did not qualitatively affect the results (data not shown). We also analyzed the sex-biased genes on their own within the same statistical framework. This allowed for both a more similar distribution between the sex-bias categories, and for successful tests of normality after Box-Cox transformation. The summary tables from the models were generated using the R package sjPlot v 2.4 (Lüdecke, 2017).

We also compared the proportions of sex-biased/unbiased genes and male-biased/female-biased genes between pairs of genomic regions of differing sex linkage using chisq tests (or Fishers's exact tests when the gene numbers were small), employing the sjt.xtab function of the in R package sjPlot v 2.4 (Lüdecke, 2017). The pairs of genomic regions compared were: PAR-SL; Au-LG1; and Au-SL. Finally, we looked for enrichment of candidate genes in the same genomic region comparisons using the same methodology.

Divergence between males and females from natural populations

To assess divergence between males and females, we examined F_{ST} and sex-biased SNP density from individuals sampled from across the species' range. Seeds obtained from multiple *M. annua* populations (Table S2) were germinated and grown in a glasshouse at the University of Lausanne. Twenty individuals of each sex were selected for genomic analysis to achieve a balance between representation of many populations and sufficient common sequences of good quality. High-quality genomic DNA was extracted from about 100 mg of frozen leaf material using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany),

following standard protocols. Based on the successful probes from a previous gene-capture experiment (see González-Martínez *et al.*, 2017), about 21 Mbp of the *M. annua* diploid genome (175,000 120-mer probes) were sequenced, with sequence capture using SureSelect DNA Capture technology (Agilent Technologies, (Santa Clara, CA) followed by sequencing using Illumina HiSeq 2500, outsourced to Rapid Genomics (Gainesville, FL). Our gene-capture experiment included both genic and intergenic regions and avoided repetitive regions and organelle genomes. Raw data was provided by Rapid Genomics as FASTQ files.

Sequencing quality was assessed using FASTQC v.10.5 (www.bioinformatics.babraham.ac.uk/ projects/fastqc). Adapter removal and trimming was done using Trimmomatic v0.36 (Bolger *et al.*, 2014), with options ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:15 MINLEN:36. After filtering, we retained an average of 13 million paired-end sequences per sample. We aligned sequences to the *M. annua* scaffolds using BWA-MEM v.0.7.12 (Li and Durbin, 2009) and processed the alignments with Picard Tools v.1.141 (http://broadinstitute.github.io/picard/). We ran Samtools mpileup (Li *et al.*, 2009) with the probabilistic alignment disabled, and called SNPs using Varscan (Koboldt *et al.*, 2009) with a minimum variant allele frequency of 0.15 and a minimum threshold for homozygotes of 0.85. We required a minimum of 10 reads per site and a Phred quality score > 20.

The gene capture ORFs were aligned to the genomic scaffolds using BLAST v2.2.31 (Altschul *et al.*, 1990). We retained hits with a maximum e-value of e^{-20}, and a minimum identity of 95%. In the case of multiple ORFs mapping to the same genomic location, we selected the longest alignment, and removed all overlapping ORFs. This resulted in 21,203 transcripts aligning to the genome, 8,317 of which were also included in the linkage map. All scaffolds with a transcript were assigned a linkage group position in the map. 31 scaffolds contained transcripts assigned to more than one linkage group; in these cases, we assigned the scaffold to the linkage group with >50% of the transcripts (in the case of equal assignment, we excluded the scaffold from the analysis). In total, we were able to assign 6,458 scaffolds to linkage-group positions.

We used the BLAST output to construct a bed file that defined the coordinates of each transcript on a genomic scaffold. We then extracted genetic information for the transcripts specified in the bed file. Using

this information, we calculated two complementary measures of differentiation between males and females for transcripts exceeding 100 bp. First, we computed F_{ST} for each transcript using VCFtools v.0.1.15 (Danecek *et al.*, 2011). Similarly, we assessed sex-biased heterozygosity for each transcript (Toups *et al.*, 2019), which we defined as the $\log_{10}$ of the male:female SNP-density ratio. We expect this ratio to be zero for autosomal transcripts, and elevated on young sex chromosomes due to excess heterozygosity in males. For transcripts on older sex chromosomes, the X and Y alleles may have diverged substantially, so that the Y alleles would no longer map to the X allele reference. In this case, there will be inflated female heterozygosity, due to X-chromosome hemizygosity in males.

For both F_{ST} and sex-biased heterozygosity in the two sexes, we assessed differences between the autosomes, PAR, and sex-linked region using Wilcoxon signed rank tests. We then computed moving averages in sliding windows of 20 transcripts on all chromosomes using the rollmean function, from package zoo v1.8-0 in R v3.2.2. In order to identify regions of elevated F_{ST} and sex-biased heterozygosity, we computed 95% confidence intervals on the basis of an average of 30 consecutive transcripts on the autosomes calculated 1,000 times. Finally, for both metrics, we assessed whether the top 5% of transcripts were overrepresented on LG1 using chi-squared tests.

Data Availability

All raw DNA and RNA sequence data generated in this study have been submitted to NCBI under accession SRP098613 BioProject ID PRJNA369310. Scripts, supplementary files and intermediate files are available at https://osf.io/a9wjb/?view_only=fd59f4dbce5a4c4a88ba31036b51fb35 File S1 contains the genetic map data and other information on each mapped transcript. File S2 contains the Sanger sequence of clones containing the X and Y variants, where there is a premature stop codon on the Y, and the script used to make the phylogenetic tree. File S3 summarizes the GO results for sex-biased genes. File S4 contains transcriptome annotation information. File S5 contains the assembled transcriptome and X and Y haplotype predictions from SEX-DETector. File S6 contains the genome and predicted genes on it. File S7 is the *M. annua* repeat library.

Results

Male karyotype

A karyotype for a male individual is shown in Figure 1, and chromosome spreads from both sexes are shown in Figure S1. Idiogram data, chromosomal arm ratios and relative length measurements from 50 images (obtained from 5 individuals; Table S3) suggest that all *M. annua* chromosomes can be discriminated from one another. A partial distinction between the one large, two medium and five smaller chromosomes is possible on the basis of the two common ribosomal DNA probes (45S and 5S), with a secondary constriction often visible at the location of the 45S locus. Our analysis was unable to discern any heteromorphic chromosomes, confirming the similar conclusion reached by Durand (1963).

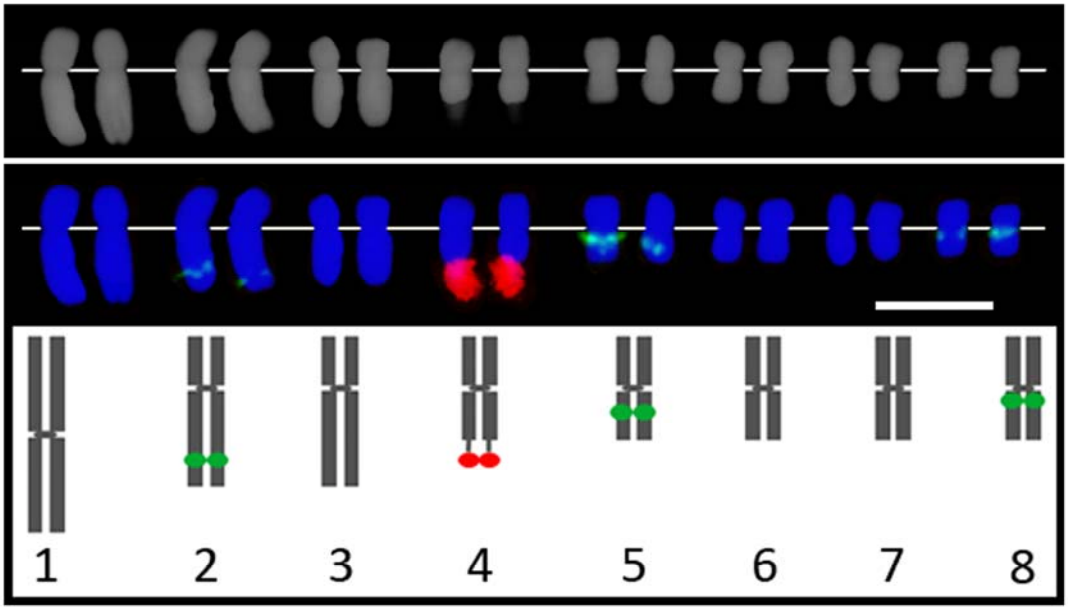


Figure 1: Basic karyotype for a *Mercurialis annua* male individual. The top row shows DAPI counterstain. The middle row shows chromosomes counterstained with DAPI (blue) after bicolor FISH with 45S rDNA (red) and 5S rDNA (green). The bottom row shows an idiogram of the male chromosomes. The white bar in the middle row represents 5 µm.

M. annua genome assembly

Using k-mer distribution, we estimate the haploid genome size to be 325 Mb (Figure S2, Table S4), which is similar to the previous estimate of 640 Mb for the diploid genome (Obbard *et al.*, 2006). Using a combination of short-read Illumina and long-read Pacific Biosciences DNA sequencing, we generated ~56.3 Gb of sequence data from the *M. annua* male M1, corresponding to ~86.6 coverage for a diploid

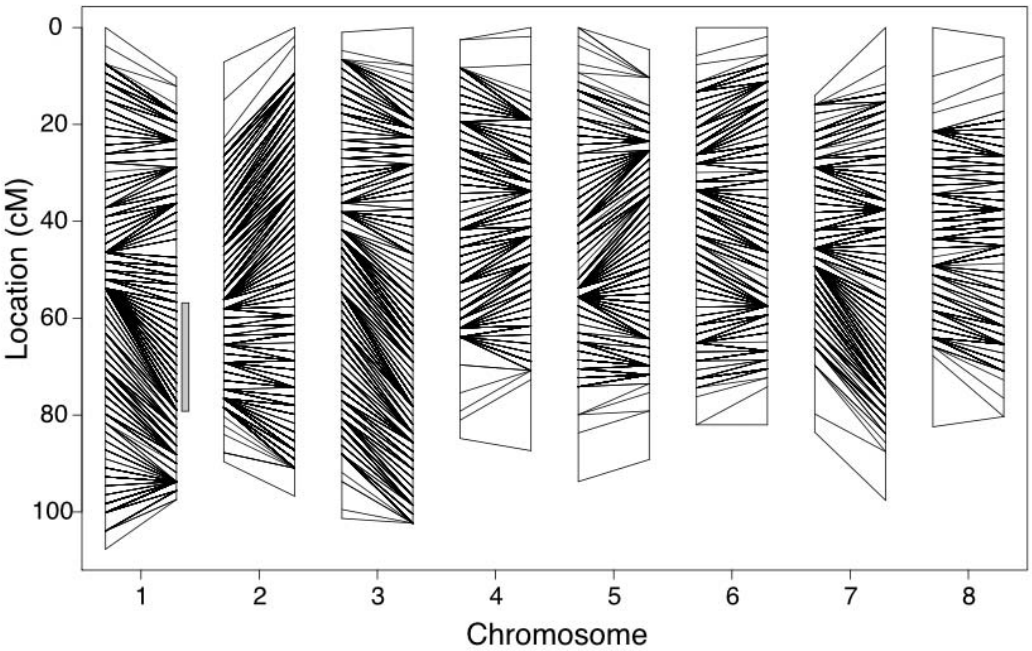
genome size of 650 Mb. After read filtering, genome coverage dropped to ~63.7x (Table S5). *De novo* assembly and scaffolding yielded a final assembly of 89% of the genome (78% without gaps), 65% of which was assembled in scaffolds > 1 kb, with an N_{50} of 12,808 across 74,927 scaffolds (Table S6). Assembly statistics were consistent with those for other species of the Malpighiales (Table S7), as was our estimate of total genomic GC content (34.7%; Smarda *et al.*, 2012). The *M. annua* assembly encompassed over 89% of the assembled transcriptome; most of the unassembled genome sequence data is therefore expected to be repetitive. Using BUSCO2 (Simão *et al.*, 2015), we recovered 76.1% of the 1,440 genes in the BUSCO embryophyte database (of which 3.9% were duplicated). A total of 10.3% of the genes were fragmented, and 13.6% were missing.

Repeat masking identified simple tandem repeats in over 10% of the assembly. This proportion is likely to be an underestimate, given the difficulty in assembling microsatellites. We characterized 15% of the assembly on the basis of homology information and DNA transposon and retrotransposon masking, with an additional 33% corresponding to 1,472 predicted novel transposable elements. The most frequent transposable repeats were *gypsy* LTR, *copia* LTR, and L1 LINE retrotransposons (Table S8), similar to findings for other plant genomes (Chan *et al.*, 2010; Sato *et al.*, 2011; Wang *et al.*, 2012a; Rahman *et al.*, 2013). Across all data, 58% of the ungapped *M. annua* assembly was repetitive (Table S8), corresponding to 44% of the genome. Given that our assembly covered 78% of the genome, and if we assume that the missing fraction comprises only repeats, up to 66% of the *M. annua* genome may be repetitive. This too would be consistent with what is known for other plants, which have a similarly high AT-rich repeat content and a similar number of unclassified repetitive elements (e.g. Chan *et al.*, 2010). The *M. annua* genome appears to be about 240 Mb larger than that of its relative *Ricinus communis* (see Table S7), perhaps reflecting on-going transposon activity.

Sex-linked transcripts and genetic map for diploid *M. annua*

Genetic mapping recovered eight linkage groups (LG1 through LG8), corresponding to the expected chromosome number for diploid *M. annua* (Durand, 1963). LepMap3 (LM3) identified 568 sex-linked (SL) transcripts, of which 365 were also identified as sex-linked by SEX-DETector and phased into X and Y

haplotypes. The 568 SL transcripts mapped to LG1, which is thus the sex chromosome. An additional 1,209 transcripts mapped to the two ends of LG1, and represent two putative pseudoautosomal regions (PAR1 + PAR2; which we refer to these as PAR and consider them together in all analyses); 8,299 transcripts mapped to the other (autosomal) linkage groups. In total, 641 markers could be resolved from one another and were thus informative. We inferred the sex-linked region to comprise of transcripts without recombination in males and in the same phase as that of their paternal grandfather. Assuming an equal transcript density across the genome, the sex-linked region represents 32.6% (about a third) of LG1 and 5.8% of the 320 Mb haploid genome (Obbard *et al.*, 2006), or 18.6 Mb. This Y-linked region maps to LG1 at male position 53.85 cM and spans 22.23 cM on the female recombination map (from positions 46.44 - 68.67 cM; Figure 2). The total male and female recombination maps were 700.43 and 716.06 cM,



respectively. All marker names and their associated positions and metrics are provided in File S1.

Figure 2: Genetic map of the 8 linkage groups (LG) for *M. annua*, with each LG represented by the male (left) and female (right) maps. Lines linking the two maps indicate the mapping position of groups of transcripts that segregated together (10,076 total transcripts). The sex-linked region is at 53.85 cM of chromosome 1 (LG1) on the male recombination map, and is highlighted by the gray vertical bar. The 2 pseudoautosomal regions (PAR) comprise the rest of LG1.

Characterization of contigs by sex linkage and ORF localization

We characterized the genome assembly by mapping all ORFs using GMAP and dividing genomic contigs into three groups: 'sex-linked', 'autosomal', and 'expressed ORF' contigs (containing all contigs that mapped to a transcript, regardless of its inclusion on the genetic map; note that PAR contigs were included in the autosomal bin). There were 548 sex-linked contigs containing 1,641 genes and a total of 8.3 Mb of sequence; 7,579 autosomal contigs containing 23,502 genes and 106 Mb sequence data; and 15,392 contigs with 34,105 expressed ORFs distributed across 176 Mb (Table 1).

Table 1. Summary statistics of contigs containing sex-linked, autosomal or all expressed ORFs. Only ORFs supported by expression data were used for analysis of sex linkage (*de novo* predicted genes were excluded). Confidence intervals are one standard deviation.

	Whole genome	Expressed ORFs	Sex-linked	Autosomal
Number of contigs	720,537	15,392	538	7,579
Total base pairs	546,375,413	176,423,112	8,271,542	105,835,810
% GC	34.7	34.5	34.5	34.3
Average contig length	15,330 (±20,028)	11,462 (±17,207)	15,374 (±21,615)	13,964 (±19,678)
Total repeat density (%)	51.5	18.7	18.4	18.3
Final coding transcript number	34,105	34,105	1,641	23,502
Average transcript length	887 (±754)	810 (±702)	1,265 (±878)	1,242 (±853)
Mean effective number of codons	53.7	53.7	52.2	51.9
Number of SNPs	181,567	180,925	5,274	102,881
Number of ORFs transcripts with SNPs	16,667 (49%)	16,572 (49%)	545 (33%)	7,714 (33%)
Median nucleotide diversity (π) / ORF	0.0035 (±0.0113)	0.0035 (±0.0113)	0.006647 (±0.008079)	0.007951 (±0.01016)

Sequence divergence, nucleotide diversity and codon usage of X and Y haplotypes

We used the SNP and haplotype calls from SEX-DETector to compare measures of nucleotide diversity and divergence between the X and Y ORFs within *M. annua*, as well as between X, Y and autosomal ORFs of *M. annua* and closely related species. The mean pairwise d_N/d_S between X and Y haplotypes within *M. annua* was 0.182 (Table 2). For sex-linked transcripts, we excluded the few SNPs that did not show clear sex-linked segregation, but we could not apply such error correction for autosomes. Accordingly, π might be underestimated for sex-linked genes compared to autosomes (see Figure S3, Table 1), so that metrics associated with sex-linked sequences should be regarded as conservative.

The pairwise d_S and d_N/d_S values for all possible orthologs between *M. annua* and its relatives *M. huetti* and *R. communis*, and the X and Y haplotypes, are presented in Table 2 and Figure 3. For autosomal genes,

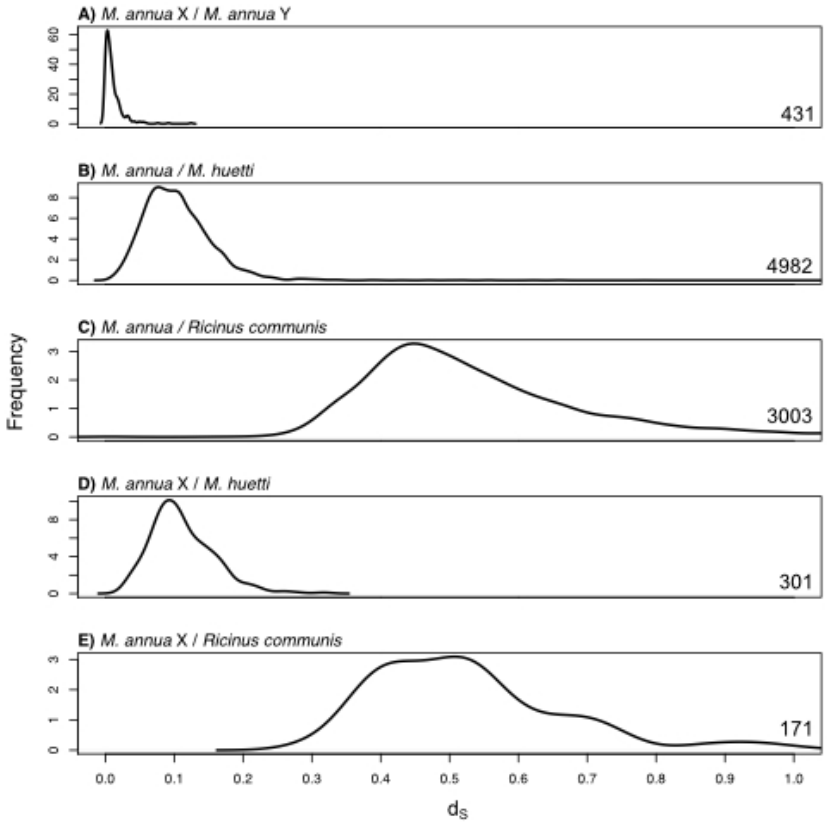
d_N/d_S between *M. annua* and *M. huetii* and between *M. annua* and *R. communis*, respectively, was 0.149 and 0.140. Pairwise d_S was lower between X/Y gene pairs within *M. annua* than between orthologous autosomal genes in *M. annua* and *M. huetii* or *R. communis*.

Codon usage in *M. annua* did not differ significantly between sex-linked, pseudoautosomal and autosomal (SL, PAR and Au) ORFs (Nc = 54, Nc = 53 and Nc = 53, respectively; Figure S4).

Table 2. Average pairwise d_S and d_N for ORFs of *M. annua* from X/Y haplotype comparison within *M. annua* or against *M. huetii* and *R. communis* sequences. Autosomal ORF comparisons with the other species are also reported. Haplotypes with premature stop codons were excluded.

	d_S	d_N	d_N/d_S
X/Y pairs	0.011	0.002	0.182
X / *M. huetii*	0.112	0.016	0.143
Y / *M. huetii*	0.113	0.017	0.150
X / *R. communis*	0.548	0.070	0.128
Y / *R. communis*	0.555	0.070	0.126
M. annua / *M. huetii*	0.106	0.014	0.149
M. annua / *R. communis*	0.565	0.079	0.140

Figure 3: Density plots of synonymous site divergence (d_S) for: (A) all *M. annua* X/Y haplotypes, (B) autosomal *M. annua* / *M. huetii* homologues, (C) autosomal *M. annua* / *R. communis* homologues, (D) X-linked *M. annua* / *M. huetii* homologues, and (E) X-linked *M. annua* / *R. communis* homologues. The numbers on the bottom right indicate the sample size for each plot.



We used PAML (Yang, 2007) to analyze the 98 sex-linked genes with orthologs in *M.annua*, *M. huetii*, and *R. communis*, using the 'two-ratio' M_2 model (which accommodates differences in d_N/d_S between the Y branch and other branches of the tree; Nielsen and Yang 1998). Our analysis identified 84 sequences with $0.001 < d_S < 2$ on both X and Y branches. Variation for 74 of these sequences was consistent with a tree topology corresponding to the species tree and recent X-Y divergence for *M. annua*, i.e. trees with topology (((*M. annua*-X, *M. annua*-Y), *M. huetii*), *R. communis*); we focused further analysis on these genes. For 52 of these 74 genes (21 after FDR correction at 5% level; Benjamini and Hochberg, 1995), the M_2 model provided a better fit than the M_0 model. Moreover, the Y lineage had more synonymous and more non-synonymous mutations compared to the X lineage, following divergence from *M. huetii*. Together, these results suggest that 28% of the Y-linked genes in our sample have been evolving faster than their corresponding X-linked orthologs as inferred from both tree-based d_N and d_S values (Wilcoxon Rank Sum =

392 and 568, respectively; $p < 0.001$ for both tests; Figure 4), as well as under relaxed selection as inferred from the tree-based d_N/d_S ratio (Wilcoxon Rank Sum = 267, $p < 0.001$; Figure 4). The seven sequences for which the Y variant clustered outside the clade (*M. annua*-X, *M. huetii*) were distributed widely across genetic map of the non-recombining region of the Y (Figure 5), i.e. they did not clearly indicate a potential ancestral/common sex-determining region between *M. annua* and *M. huetii*. We did not investigate these genes further.

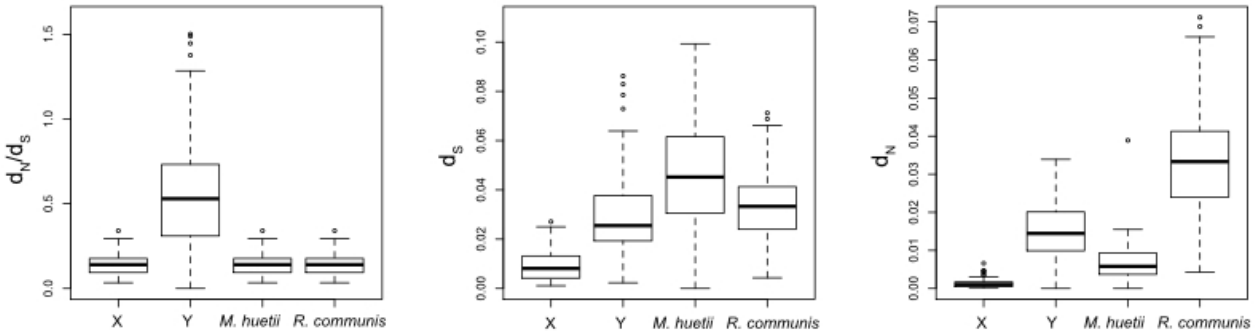


Figure 4: Boxplots of tree-based estimates of d_N/d_S, d_S and d_N for the X, Y, *M. huetii* and *R. communis* branches of the 72 gene trees that correspond to tree topology (((*M. annua*-X, *M. annua*-Y), *M. huetii*), *R. communis*). A Y-sequence d_N/d_S outlier of 2.7 and an *M. huetii*-sequence d_S outlier of 0.23 are not shown, for greater resolution.

Identification of a degenerated and duplicated sequence on the Y

The SEX-DETector pipeline detected putative premature stop codons in four Y and three X transcripts, although there was always a functional X haplotype for the same transcript. We confirmed one of the Y haplotype stop codons using PCR, cloning and sequencing from five male and nine female samples from individuals sampled across the geographic range of diploid *M. annua*. Briefly, a phylogenetic tree was constructed for this gene using RAxML v 8.2.10 using 1,000 bootstraps (Stamatakis, 2014) on sequences aligned with mafft v7.310 (Katoh and Standley, 2013) (Figure S5). Interestingly, two different Y-linked sequences could be obtained, sometimes from the same male, suggesting that the sequence has duplicated in at least some males. All males and females analyzed also contained a functional copy which had 89% identity with an F-box/kelch repeat protein in *Ricinus communis*. The transcript shows lower Y/X expression ratio and high π (marked "K_s" in Figure 5), suggesting that it represents a degenerated copy on the Y.

Identification and annotation of sex-biased genes

Expression analysis revealed more genes with male-biased than with female-biased expression (1,385 male-biased vs. 325 female-biased genes, based on 5% FDR). The relative proportion of male-biased genes increased when also using a minimum $\log_2$ fold-change ($\log_2$FC) threshold of one (1,141 vs. 140; Figure S6). An analysis of GO term enrichment is presented in File S3. Male-biased genes were enriched in biological functions such as anther-wall tapetum development, response to auxin, response to ethylene, cell-tip growth, pollen-tube growth, and floral-organ senescence. Female-biased genes were enriched for functions related to the maintenance of inflorescence meristem identity, jasmonic acid and ethylene-dependent systemic resistance, regulation of innate immune response, and seed maturation.

Identification of the ancestral evolutionary stratum on the Y

To identify a putative ancestral stratum in the non-recombining region, we investigated variation in the following metrics across the female recombination map (Figure 5): the magnitude of sex-biased expression; the nucleotide diversity (π) of each transcript; and the pairwise d_N/d_S ratio based on comparison with homologous genes in the closely related dioecious sister species *M. huetti*, and its monoecious more distant relative *R. communis* (for which we could identify 4,761 and 2,993 homologues, respectively); and the expression ratio of X and Y haplotypes inferred from SEX-DETector (Muyle *et al.*, 2016). We annotated Figure 5 with letters to indicate the position of candidate sex-determining genes, based on their blast hit descriptions. The sole obvious outlier for any of the metrics above was a transcript associated with auxin production, which had a high pairwise d_N/d_S value in the comparison with the close outgroup species, *M. huetii*.

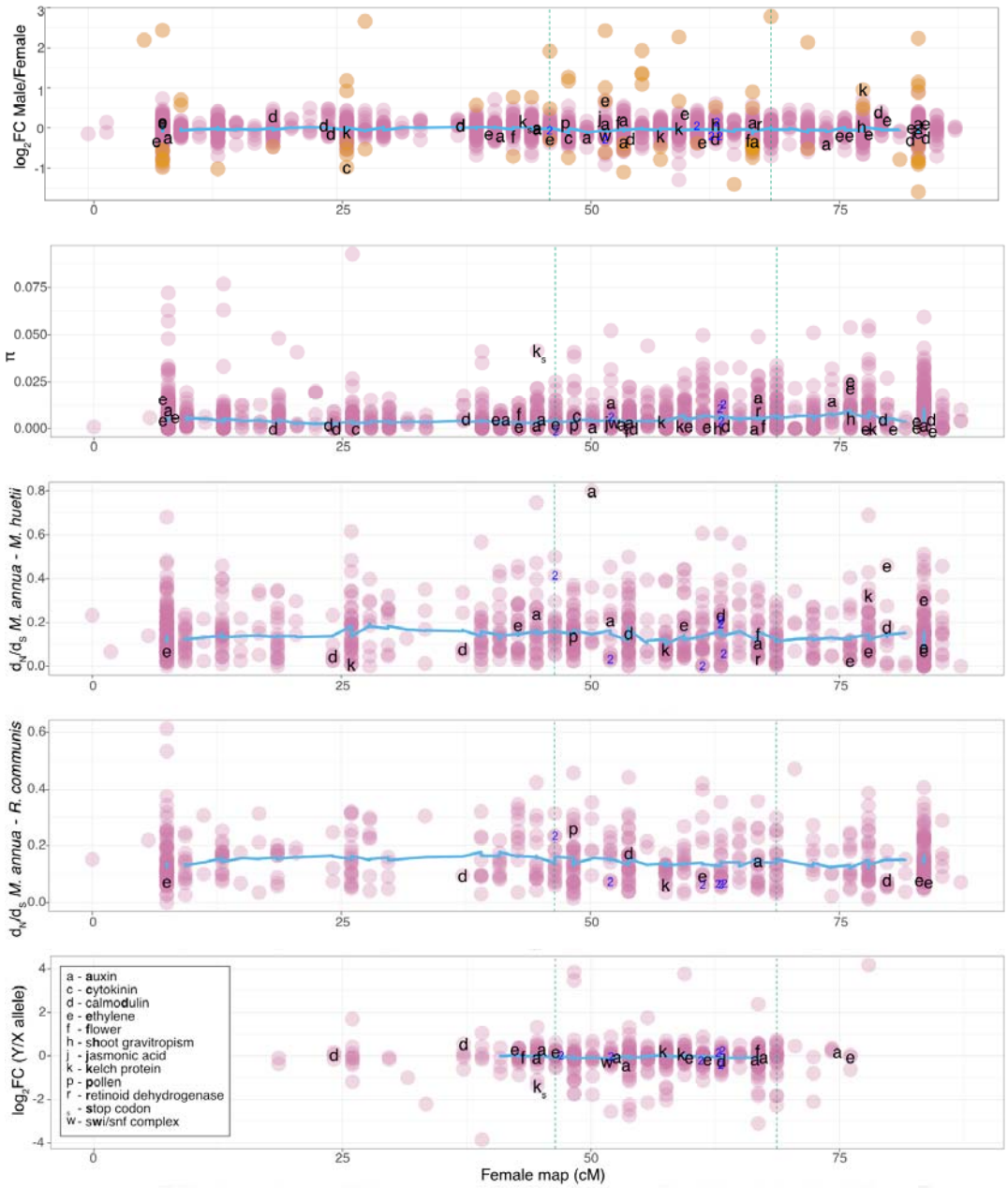


Figure 5: Summary of transcript metrics calculated for sequences across the female recombination map of the putative sex chromosome LG1, based on analysis of X and Y sequences combined. The non-recombining region in males is located between the vertical dotted lines. Orange points in the top panel indicate significant sex-bias (5% FDR). Potential candidate sex-determining genes are indicated by letters, with their putative function given in the inset legend. The only verified premature stop codon on the Y copy is indicated by k_s next to the non-recombining region (i.e., for a kelch protein gene). Transcripts where the Y sequence diverged before the *M. annua*/*M. huetii* species split are indicated with "2".

Comparison of sex-linked, autosomal and pseudo-autosomal ORFs

We examined the effects of sex-bias (male-, female-, and unbiased genes) and genomic location (SL, PAR or Au), as well as their interactions, on the five metrics plotted in Figure 5. We repeated the analyses

after excluding unbiased transcripts for the four metrics with sufficient gene numbers (sex bias $\log_2$FC, π, and pairise d_N/d_S using either *M. huetii* or *R. communis*) to directly compare male- and female-biased transcripts. The full models are presented in Tables S9-S11. Below, we summarize their significant results; Figure 6 shows the data distributions. The analyses revealed significantly higher differences in gene expression between the sexes (absolute $\log_2$FC) in the sex-linked region ($p = 0.037$; Table S9). These differences were largely due to significantly higher expression of male- than female-biased transcripts, i.e. in terms of the extent of expression difference between the sexes, the sex chromosomes are significantly masculinized (male-bias*SL interaction: $p=0.036$; Table S10).

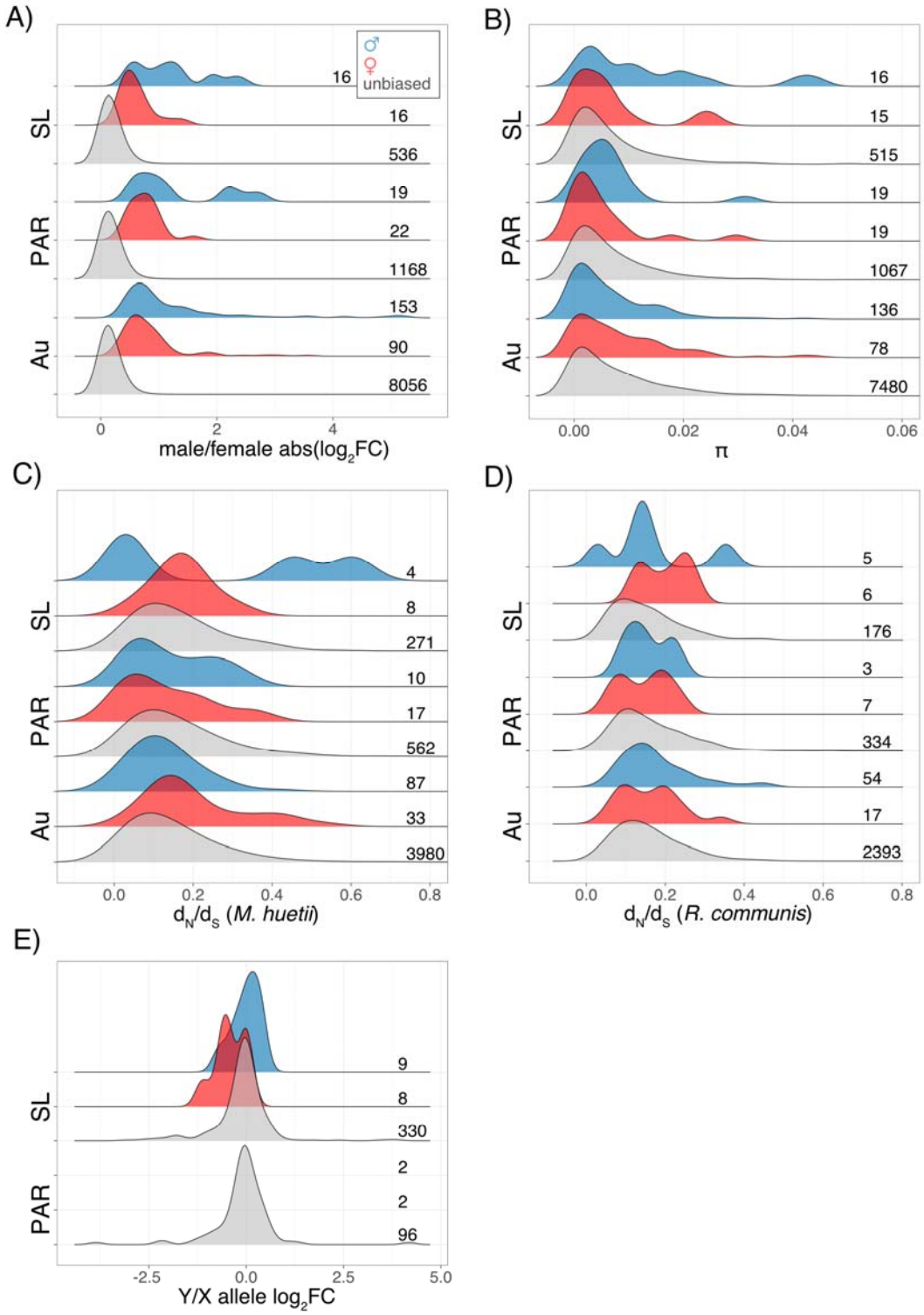


Figure 6: Graphical summary of the analysis of genes based on their sex-bias (blue: male-biased, red: female-biased, gray: unbiased) and sex linkage (SL: sex linked, PAR: remaining LG1, Au: LG 2-7). A) absolute $\log_2$FC expression difference between the sexes, B) nucleotide diversity (π), C) pairwise d_N/d_S compared to *M. huetii* homologues, D) pairwise d_N/d_S compared to *R. communis* homologues, E) relative expression of Y over X alleles. The numbers to the right indicate the number of genes in each category. Data are plotted prior to Box-Cox transformation.

The increased magnitude of male-biased gene expression in fully sex-linked genes is consistent with a recent masculinization as sex-linked genes had slightly lower nucleotide diversity (π) than those in the PAR or autosomal genes (p = 0.039; Table S9), but male-biased genes were not affected as much (p = 0.017; Table S9). The pairwise d_N/d_S compared to *M. huetii* was higher for sex-linked genes (p = 0.017; Table S9) and for female-biased genes localizing to the PAR (PAR*female-bias interaction: p = 0.005; Table S9). An analysis of the sex-biased genes only (excluding unbiased) revealed the PAR to have a lower π (PAR: p = 0.007; Table S10) and a lower pairwise d_N/d_S to *M. huetii* (p = 0.018; Table S10). Few patterns were significant for pairwise d_N/d_S with respect to to the more distantly related *R. communis* (it was higher for male-biased genes; p = 0.046; Table S9), possibly because of lower power due to fewer identified homologues (4,761 in *M. huetti* compared to 2,993 in *R. communis*). The Y-linked copy of female-biased genes had significantly lower expression than the X-linked counterpart (p = 0.035; Table S11).

Finally, we used chi-squared tests to compare the relative proportions of sex-biased versus unbiased and male- versus female-biased genes in different pairs of genomic regions (PAR-SL, Au-LG1, Au-SL; Table 3). The sex-linked region had significantly more sex-biased genes (compared to either PAR or Au regions). This enrichment was largely responsible for the significant sex-biased gene enrichment compared with the entire LG1 and the autosomes, and was largely due to more female-biased genes on LG1. The sex chromosomes thus appear to be feminized in terms of number of female-biased genes, in contrast to the genome as a whole, which was masculinized in terms of number of male-biased gene numbers, and in contrast to absolute expression, which was higher for male-biased genes. We did not detect any enrichment for candidate genes with functions likely to have sexually antagonistic effects in the sex-linked region (Table 3).

Table 3. Results of chi-squared tests for differences in the frequencies of sex-biased versus -unbiased genes, male- versus female-biased genes, and candidate sex-determining genes (CG) versus other genes (not CG) between the pseudoautosomal region (PAR), the sex-linked region (SL), linkage group 1 (LG1) and autosomes (Au).

	Sex-biased vs unbiased			Female-biased vs male-biased			Candidate gene vs other gene		
	Unbiased	Sex-biased	Total	Female biased	Male biased	Total	Not CG	CG	Total
PAR	1168	41	1209	22	19	41	1177	32	1209
	96.6 %	3.4 %	100 %	53.7 %	46.3 %	100 %	97.4 %	2.6 %	100 %

SL	536	32	568	16	16	32	546	22	568
	94.4 %	5.6 %	100 %	50 %	50 %	100 %	96.1 %	3.9 %	100 %
Total	1704	73	1777	38	35	73	1723	54	1777
	95.9 %	4.1 %	100 %	52.1 %	47.9 %	100 %	97 %	3 %	100 %
Test	χ^2=4.381 df=1 φ=0.053 **p=0.036**			χ^2=0.006 df=1 φ=0.036 p=0.941			χ^2=1.578 df=1 φ=0.033 p=0.209		
Au	8056	243	8299	90	153	243	8094	207	8301
	97.1 %	2.9 %	100 %	37 %	63 %	100 %	97.5 %	2.5 %	100 %
LG1	1704	73	1777	38	35	73	1723	54	1777
	95.9 %	4.1 %	100 %	52.1 %	47.9 %	100 %	97 %	3 %	100 %
Total	9760	316	10076	128	188	316	9817	261	10078
	96.9 %	3.1 %	100 %	40.5 %	59.5 %	100 %	97.4 %	2.6 %	100 %
Test	χ^2=6.326 df=1 φ=0.026 **p=0.012**			X^2=4.649 df=1 φ=0.129 **p=0.031**			χ^2=1.515 df=1 φ=0.013 p=0.218		
Au	8056	243	8299	90	153	243	8094	207	8301
	97.1 %	2.9 %	100 %	37 %	63 %	100 %	97.5 %	2.5 %	100 %
SL	536	32	568	16	16	32	546	22	568
	94.4 %	5.6 %	100 %	50 %	50 %	100 %	96.1 %	3.9 %	100 %
Total	8592	275	8867	106	169	275	8640	229	8869
	96.9 %	3.1 %	100 %	38.5 %	61.5 %	100 %	97.4 %	2.6 %	100 %
Test	χ^2=12.066 df=1 φ=0.038 **p<0.001**			χ^2=1.496 df=1 φ=0.085 p=0.221			χ^2=3.493 df=1 φ=0.021 p=0.062		

Sex-linked sequence analysis of individuals sampled across the species range

To confirm the lack of recombination in the sex-linked region, we estimated F_{ST} and sex-biased heterozygosity for genome capture data from 20 males and 20 females sampled from natural populations across the species' range (Table S2), based on 6,557 transcripts assigned to the linkage map. The sex-linked region showed clear differentiation between males and females using both metrics. F_{ST} between males and females was elevated in the sex-linked region relative to the PAR and the autosomes (Wilcoxon, $p = 4.79 \times 10^{-5}$, $p = 1.099 \times 10^{-7}$ respectively; Figure 7A), although the F_{ST} values are very small. Similarly, there was an excess of SNPs in males relative to females in the sex-linked region relative to the PAR and the autosomes (Wilcoxon, $p < 2.2 \times 10^{-16}$ in both; Figure 7C). We used sliding window analyses of F_{ST} and SNP density to test whether males and females are more strongly differentiated in the sex-linked region compared to the rest of the genome. These analyses identified a region of elevated F_{ST} at ~ 65 cM of LG1 in the female recombination map (Figure 7B) whose value exceeded that for any autosomal region. We also identified a region between 50 and 65 cM on LG1 (Figure 7D) with higher SNP density and higher heterozygote frequencies in males than females (Figure 7E). Finally, we tested for enrichment for these genome-wide metrics from many populations. Table S12 shows that LG1 is both enriched in F_{ST} and sex-biased heterozygosity in the sex-linked region for males than females, as expected for X and Y

chromosome divergence.

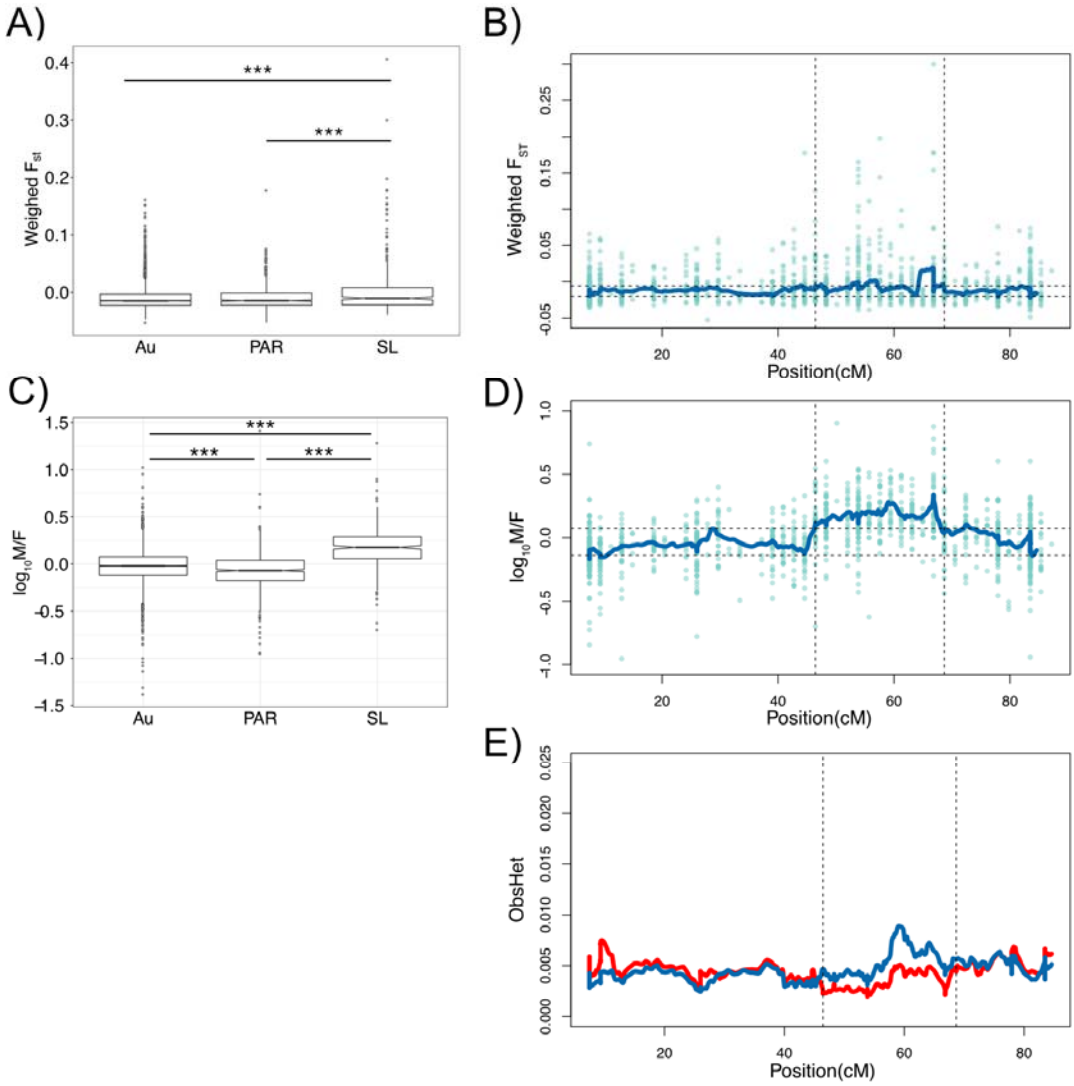


Figure 7: Summary of the effect of sex linkage on SNP metrics from genome capture data across the species range. (A) F_{ST} for sex linkage category, (B) F_{ST} distribution across the female recombination map of LG1, (C) ratio of male/female SNP density for sex linkage category, (D) ratio of male/female SNP density across the female recombination map of LG1, (E) observed heterozygosity in males (blue) and females (red) across LG1. Asterisks in A and C indicate significant Wilcoxon tests ($p < 0.001$), the sex-linked region is indicated by vertical lines in B, D and E.

Discussion

A genome assembly, annotation and genetic map for dioecious *M. annua*

Our study provides a draft assembly and annotation of diploid dioecious *M. annua* (2n = 16). Cytogenetic analysis has confirmed the haploid number of n = 8 chromosomes for diploid *M. annua* (Durand, 1963), with about 10,000 ORFs assigned to the corresponding eight linkage groups by genetic mapping. Similar to other plants, about 2/3 of the *M. annua* genome comprises repetitive sequences, mostly *gypsy* LTR, *copia* LTR, and L1 LINE retrotransposons (Chan *et al.*, 2010; Sato *et al.*, 2011; Wang *et al.*, 2012a; Rahman *et al.*, 2013). The number of genes in *M. annua* (34,105) is more similar to that of diploid species such as *Arabidopsis thaliana* (> 27,000 genes) and *Ricinus communis* (> 31,000 genes) and is thus compatible with having a long recent history under diploidy rather than polyploidy; for comparison, the gene content of diploid *Gossypium* is > 40,000 genes, pointing to a polyploid history (Wang *et al.*, 2012a). The draft annotated genome of *M. annua* will be useful for understanding potential further links between sex determination and sexual dimorphism, as well as the consequences of genome duplication and hybridization (Obbard *et al.*, 2006) for sexual-system transitions in polyploid lineage of the species complex (Pannell *et al.*, 2008).

The size of the non-recombining Y-linked region of *M. annua*

Patterns of SNP segregation in small families from 568 ORFs suggests that a substantial length (1/3) of the *M. annua* Y chromosome has ceased recombining with the X. Lack of recombination on the sex-linked region is further supported by the existence of a Y-specific non-functional gene duplication, which could be amplified in populations ranging from Israel to Britain, as well as additional population-genetic data from males and females sampled from across the diploid species' range. The sex-linked region had slightly, but significantly, higher F_{ST}, a higher male/female SNP ratio, and higher heterozygote frequency in males than females, indicating a history of low recombination. Previous mapping of bacterial artificial chromosomes (BACs) containing male-specific PCR products found them to be distributed over most of the sex-linked region, i.e., a region of between 52 and 66.82 cM, corresponding to 4.86% of the genome and implying a physical length of between 14.5 and 19 Mb for the non-recombining genome (Veltsos *et al.*, 2018). The large non-recombining region of the *M. annua* Y chromosome contrasts with that of other

plants with homomorphic sex chromosomes, which typically have < 1% of their Y not recombining, e.g., *Vitis vinifera* (Fechter *et al.*, 2012; Picq *et al.*, 2014), *Fragaria chiloensis* (Tennessen *et al.*, 2016) and *Populus* species (Paolucci *et al.*, 2010; Geraldes *et al.*, 2015).

Limited gene loss, pseudogenization, and mild purifying selection on the *M. annua* Y

We found only limited evidence for Y-chromosome degeneration within the non-recombining region of the *M. annua* Y chromosome, despite its substantial length. Very few genes have been lost or have become non-functional, and only one of the 528 X-linked genes corresponding to the non-recombining region of the Y (0.2%) did not have a Y-linked homologue. Our analysis may have overlooked genes that have been lost from the Y, because their detection from RNAseq data relies on the presence of polymorphisms on the X copy (see Blavet *et al.*, 2015) and requires their detectable expression in our sampled tissues. Nevertheless the absence of evidence for substantial gene loss is consistent with other signatures of only mild Y-chromosome degeneration in *M. annua* and is likely a reliable finding.

There was only slightly lower relative expression from the Y compared to the X allele. Allele-specific expression was inferred on the basis of mapping to the reference transcriptome, and we expect only limited mapping bias against expressed Y-specific genes. The fact that different levels of Y/X expression depended on sex-bias (it was reduced for female-biased genes) further suggests that the result is probably not an artifact of the analysis. We found some evidence for relaxed purifying selection on Y alleles compared to X alleles, as expected in genomic regions experiencing limited recombination (Charlesworth and Charlesworth, 2000). Specifically, pairwise comparisons of X and Y alleles (after excluding the few with in-frame stop codons) showed higher d_N/d_S than pairwise comparisons of *M. annua* orthologs with *M. huetii* or *R. communis*. Moreover, our tree-based analysis of common orthologs in all species indicated that 28% of the identified Y-linked alleles have experienced an accelerated rate of molecular evolution in the *M. annua* lineage since its divergence from its sister species *M. huetii*, which is also dioecious. However, we found no evidence for differential codon usage bias between the Y-linked and other sequences. Codon usage bias is expected to be lower for non-recombining regions of the genome experiencing weaker purifying selection (Hill and Robertson, 1966), and there appears to have been a shift towards less

preferred codon usage in *Rumex hastatulus* Y-linked genes, increasing in severity with time since the putative cessation of recombination between X and Y chromosomes (Hough *et al.*, 2014). Although our analysis involves multiple statistical tests of significance such that some results might represent type 1 error, taken together they point to a very recent cessation of recombination for much of the Y chromosome. In contrast, up to 28% of *Rumex hastatulus* Y-linked genes have been lost (Hough *et al.*, 2014) and about 50% of the *Silene latifolia* Y-linked genes are dysfunctional (Papadopulos *et al.*, 2015; Krasovec *et al.*, 2018).

Sex-biased gene expression in *M. annua*

We found substantial differences in gene expression between males and females of *M. annua*, consistent with the observation of sexual dimorphism for a number of morphological, life-history and defense traits in *M. annua* (Hesse and Pannell, 2011; Labouche and Pannell, 2016). There were at least four times more male-biased than female-biased genes; this is probably conservative because we did not use a minimum $\log_2$FC threshold in our analysis, which would have further increased the proportion of male-biased genes. The observed slight decrease in expression of Y- compared with X-linked alleles suggests that dosage compensation could have begun evolving. However, any such dosage compensation would likely be at a very early stage, not least because YY males (which lack an X chromosome) are known to be viable (Kuhn, 1939). Interestingly, sex-linkage sometimes influenced male- and female-biased genes differently, with male-biased genes having a higher nucleotide diversity (π) than female-biased genes when sex-linked, again perhaps reflecting a history of relaxed purifying selection. Male- and female-biased genes also differed in pairwise d_N/d_S relative to *M. huetii*, which was higher for female-biased genes, contrary to expectations under a hypothesis of faster male evolution (Wu and Davis, 1993).

Evidence for evolutionary strata on the *M. annua* Y?

We were unable to find any direct evidence for evolutionary strata with different levels of divergence on the *M. annua* Y chromosome, but phylogenetic comparisons point to the likely existence of at least two such strata on the *M. annua* Y. Dioecy evolved in a common ancestor to *M. annua* and *M. huetii* (Krähenbühl *et al.*, 2002; Obbard *et al.*, 2006), and both species share the same Y chromosome and sex-determining region, as revealed by crosses (Russell and Pannell, 2015) and the possession of common

male-specific PCR markers (Veltsos *et al.*, 2018). The phylogeny of sex-linked genes within the shared sex-determining region should have the topology ((*M. annua*-Y, *M. huetii*), *M. annua*-X). In contrast, our PAML analysis clearly indicates that the phylogeny of most genes in the non-recombining region of the *M. annua* Y have the topology ((*M. annua*-Y, *M. annua*-X), *M. huetii*), indicating divergence between the X and Y copies of these genes in *M. annua* more recently than the species split between *M. annua* and *M. huetii*. These results suggest that *M. annua* and *M. huetii* share an old (and probably small) non-recombining stratum that includes the sex-determining locus, and that a larger stratum has been added to the non-recombining region in the *M. annua* lineage. Another possible interpretation is gene conversion between X and Y copies within *M. annua*, as has been reported for humans and some plants (Rautenberg *et al.*, 2008; Wu *et al.*, 2010; Trombetta *et al.*, 2014). This would lead to an underestimate of the size of the ancestral stratum and may have erased evidence of a longer common shared history between the non-recombining regions of *M. annua* and *M. huetii*.

Although most Y-linked genes that we sampled in *M. annua* point to an X/Y divergence in *M. annua* that postdates the split from *M. huetii*, we found seven genes with a phylogeny consistent with the hypothesized old stratum shared with *M. huetii*. If these genes are indeed within an older stratum, we might expect them to colocalize in the same part of the non-recombining region. However, we found that they were quite scattered on the female recombination map. We have no explanation for this pattern, but we note that it would be consistent with a break of synteny between the *M. annua* Y and X chromosomes (and thus potentially also between the *M. huetii* and *M. annua* Y chromosomes) in the region corresponding to the non-recombining region.

Patterns of divergence in terms of tree-based d_S and d_N between the X and Y chromosomes of *M. annua* and their corresponding orthologs in *M. huetii* and *R. communis* were largely consistent with the expected phylogenetic relationships between these species (Krähenbühl *et al.*, 2002; Obbard *et al.*, 2006). The increased d_S for the Y branch is consistent with a higher mutation rate in male meiosis. Nevertheless, the d_S branch length for *R. communis* was lower than expected for a simple molecular clock across the phylogeny. The recent transition to an annual life cycle in the clade that includes *M. annua* and *M. huetii* may be

partially responsible for the larger number of mutations that have accumulated along their branches compared to the perennial *R. communis* (Smith and Donoghue, 2008).

Age of the non-recombining Y-linked region of *M. annua*

We may estimate the age of the *M. annua* non-recombining region on the basis of inferred mutation rates in plants. Ossowski *et al.* (2010) estimated the mutation rate for *Arabidopsis thaliana* as between 7×10^{-9} to 2.2×10^{-8} per nucleotide per generation. More recently, Krasovec *et al.* (2018) estimated the mutation rate for *S. latifolia* as 7.31×10^{-9}. The similarity of these estimates suggests that they might be broadly applicable. Accepting a mutation rate for plants of approximately 7.5×10^{-9} and using the synonymous site divergence between X- and Y-linked sequences of *M. annua* of 0.11 (Table 2), we infer that recombination between most of the X- and Y-linked genes in *M. annua* ceased as recently as 1.5 million generations ago (or even more recently if we adopt the upper end of the mutation-rate range estimated by Ossowski *et al.* (2010). Given that *M. annua* is an annual plant, the putative second non-recombining stratum of the Y chromosome of diploid *M. annua* could be less than a million years old. Krasovec *et al.* (2018) used their mutation-rate estimate to infer an age of 11 million years for the oldest sex-linked stratum of *S. latifolia*, i.e., an order of magnitude older than the expanded sex-linked region of *M. annua*. The few estimates for the time since recombination suppression in other plants with homomorphic sex chromosomes, and those with smaller non-recombining regions, range from 15 to 31.4 million years ago (reviewed in Charlesworth, 2016). *Mercurialis annua* thus appears to have particularly young sex chromosomes.

Concluding remarks

We conclude by speculating on a possible role for sexually antagonistic selection in favoring suppressed recombination on the *M. annua* Y. The canonical model for sex-chromosome evolution supposes that suppressed recombination originally evolves in response to selection to bring sexually antagonistic loci into linkage with the sex-determining locus, and that the Y chromosome begins to degenerate after recombination has ceased (Rice, 1987). Given this model, we might expect to see signatures of sexual antagonism in the sex-linked region before the onset of substantial Y-chromosome degeneration. While we

do not observe obvious signatures of sexually antagonistic selection, some of the observed patterns in the sex-linked region are difficult to interpret as the outcome of degeneration. For example, while the Y-linked alleles showed slightly lower expression than X-linked alleles, this was not true for male-biased genes. One possible explanation is that the female-biased Y-linked alleles in males might have been driven to lower expression by sexually antagonistic selection. If female-biased expression is beneficial in females (Parisi *et al.*, 2003; Connallon and Clark, 2010), the observed enrichment of female-biased genes on the X would be consistent with sexually antagonistic selection because X-linked alleles occur twice as often in females than in males. Similarly, the magnitude of sex-biased gene expression (absolute $\log_2$FC) was higher in the sex-linked region, as might be expected if levels of gene expression were the result of an arms race between the sexes.

Our study prompts investigations in several directions. The speculation that sexually antagonistic selection might have played a role in the evolution of suppressed recombination might be investigated further by seeking sex-linked QTL for sexually antagonistic traits (Delph *et al.*, 2010), or functional analysis of candidate genes. It would also be valuable to compare expression levels of sex-linked genes with those in a related species with ancestral levels of gene expression (see Zemp *et al.*, 2016). The orthologs of the *M. huetii* genes that are non-recombining only in *M. annua* would make for a good comparison. More generally, the *M. annua* species complex offers itself as a promising system for comparative analysis of diverging sex chromosomes within a small clade in which speciation, hybridization and genome duplication have contributed to diversification. We anticipate characterizing the sex-linked region of *M. huetii*, the sister species of *M. annua*, and to trace sex determination and the sex chromosomes from diploid *M. annua* into the related polyploid lineages that appear to have lost and then regained dioecy (Pannell *et al.*, 2004; Obbard *et al.*, 2006). These other lineages have very similar morphological differences between males and females, but may determine sex using different loci or have sex chromosomes with a different history.

Acknowledgements

We thank John Russell for setting up the crosses, Jos Käfer for help with the Galaxy workflow for the SEX-DETector analysis, Jerome Goudet for statistical advice, and Guillaume Cossard, Wen-Juan Ma, Deborah Charlesworth, Stephen Wright and anonymous reviewers for valuable comments on the manuscript. KER was supported by grants 31003A_163384 and 31003A_141052 to JRP from the Swiss National Science Foundation, and by the University of Lausanne. PV was supported by Sinergia grant 26073998 from the Swiss National Science Foundation to JRP, Nicolas Perrin and Mark Kirkpatrick. JRP and DAF acknowledge support from the NERC and BBSRC, UK, which funded the early stages of this project. The computations were performed at the server at the department of Plant Sciences, University of Oxford, the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics and the Texas Advanced Computing Center. AM and GABM acknowledge support from ANR (grant number ANR-14-CE19-0021-01). VH, RH, BV were supported by the Czech Science Foundation (grant 18-06147S). MAT was supported by the National Institute of Health grant R01GM116853 (to Mark Kirkpatrick) and the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement number 715257).

Author contributions

Conceived the project: JRP, DAF; wet lab work: DAF; cytogenetics work: VH, RH, PV; genome assembly and annotation: KER; gene expression analysis: KER, OE, PV; segregation analysis and mapping: PV, AM, PR, GM; population genetic and divergence analysis: KER, PV; analysis of natural populations: SCGM, MAT: wrote the paper: JRP, KER, PV; commented and contributed to the final manuscript: all authors.

Disclosure declaration

The authors declare no conflict of interest.

Supplementary Information: Tables

Table S1: Samples from wild accessions and genetic cross progeny used for RNAseq.

	Sex	Individual	Parents
P	Male	M1	
	Male	M2	
	Male	M3	
	Female	G1	
	Female	G2	
	Female	G3	
F1	Male	m5	G1xM1
	Male	m6	
	Male	m7	
	Male	m9	
	Male	m13	
	Female	f1	
	Female	f3	
	Female	f4	
	Female	f11	
	Female	f12	
	Male	m2_G2xM1	G2xM1
	Male	m4_G2xM1	
	Male	m6_G2xM1	
	Male	m11_G2xM1	
	Male	m13_G2xM1	
	Female	f1_G2xM1	
	Female	f3_G2xM1	
	Female	f5_G2xM1	
	Female	f6_G2xM1	
	Female	f8_G2xM1	
F2	Male	mA	f11xm6
	Female	fB	
	Female	fC	
	Female	fD	
	Female	fE	
	Female	fF	
	Male	mB	f3xm5
	Male	mC	
	Male	mD	
	Male	mE	
	Male	mF	
	Male	mG	
	Male	mH	
	Male	mK	
	Male	mM	
	Male	mN	
	Male	mO	
	Male	mp	
	Female	fA	
	Female	fI	
	Female	fJ	
	Female	fL	
	Male	mA	f4xm6
	Male	mC	
	Male	mD	
	Male	mE	
	Male	mF	
	Male	mG	
	Male	mI	
	Male	mM	

	Male	mR	
	Female	fB	
	Female	fJ	
	Female	fK	
	Female	fL	
	Female	fN	
	Female	fO	
	Female	fP	
	Female	fQ	

Table S2: Samples from wild accessions used for genome capture data.

Sample Name	Sex	Location Code	Location	Latitude	Longitude
WC05	M	AUS	Austria	48.2082° N	16.3738° E
WC06	M				
WC07	M				
WC08	F				
WC09	F				
WC10	F				
WC11	M	POZ	Poland	52.253931° N	16.534003° E
WC12	M				
WD01	M				
WD02	F				
WD03	F				
WD04	F				
WD05	M	CZE	Czech Republic	50.0755° N	14.4378° E
WD06	M				
WD07	M				
WD08	F				
WD10	F				
WD11	M	BRK	Brouck, Marange-Zondrange, France	49.131792° N	6.527163° E
WD12	F				
WE01	M	KOP	Corinth, Greece	37.50921° N	23.00136° E
WE03	F				
WE05	M	VOL	Volos, Greece	39.23511° N	22.5452° E
WE06	F				
WE07	M	ANT	Antalya, Turkey	36.894807° N	30.655359° E
WE08	F				
WE09	F	SOT	Southampton, UK	50.9097° N	1.4044° W
WE10	M	RIB	St Pere de Ribes to Vilafranca in Catalonia, Spain	41.31271° N	1.716517° E
WE11	F				
WE12	M	PAR	Paris, France	48.82174° N	2.440481° E
WF01	F				
WF06	M	Collserola	Collserola, close to Barcelona, Catalonia, Spain	41.43627° N	2.113471° E
WF07	F				
WF08	M				
WF09	F				
WF10	M				
WF11	F				
WF12	F	Sabadell	Sabadell, Catalonia, Spain	41.55728498° N	2.115934007° E
WG01	M				
WG04	F				
WG05	M				

Table S3: Summary of arm ratio (r=p/q) and relative chromosome length ((p+q)/absolute length * 100) based on 50 figures from 5 individuals. All measurements were done as in Lengerova *et al.* (2004).

Chromosome number	Arm ratio (r)	SD	Relative length
1	1.13	±0.05	20.17
2	1.93	±0.12	14.56
3	2.36	±0.40	14.53
4	1.53	±0.05	11.32
5	1.31	±0.12	10.64
6	1.25	±0.15	9.23
7	0.96	±0.00	10.24
8	1.14	±0.24	9.31

Table S4: Summary results from k-mer (size 31) analysis.

	M1		G1		G2	
	min	max	min	max	min	max
	1.85%	1.88%	1.82%	1.85%	1.83%	1.86%
Haploid length (Mb)	325.20	327.06	292.72	294.39	344.57	346.80
Repeat length (Mb)	126.45	127.17	101.09	101.67	142.62	143.55
Unique length (Mb)	198.75	199.89	191.63	192.72	201.95	203.26
Model fit	92.41%	98.17%	92.77%	97.68%	91.99%	98.27%
Read error rate	0.47%	0.47%	0.49%	0.49%	0.35%	0.35%

Table S5: Raw data before and after filtering of genomic and transcript libraries. RNA-seq was not filtered, as those reads would drop out during the assembly and ORF prediction.

Data type	Sample	Technology	Library	Insert Size (bp)	Read number	Basecount (gb)	
						No filtering	After filtering
DNA	M1 Male	Illumina	Mate Pair	5,000	50,599,202	5.0	2.6
		Illumina	Paired End	250	364,231,376	36.4	30.8
		Illumina	Paired End	500	130,681,342	13.0	7.2
		PacBio	Long read	N/A	859,697	1.9	0.8
	G1 Female	Illumina	Paired End	250	401,657,040	40.2	32.8
	G2 Female	Illumina	Paired End	250	341,596,022	34.2	29.6
RNA	M1 Male	Illumina	Paired End	250	32,805,020	3.3	NA
	M2 Male	Illumina	Paired End	250	35,160,712	3.5	NA
	M3 Male	Illumina	Paired End	250	33,420,774	3.3	NA
	G1 Female	Illumina	Paired End	250	37,234,320	3.7	NA
	G2 Female	Illumina	Paired End	250	43,082,244	4.3	NA
	G3 Female	Illumina	Paired End	250	44,688,600	4.5	NA

Table S6: Summary of the genome *de novo* assembly. Scaffold lengths are displayed in base pairs. Software was run sequentially from SOAP2 to L_RNA_scaffolder.

Software	Scaffold number	Total bases (% coverage)	Bases without gaps (% coverage)	Longest scaffold	Scaffold n50
SOAP2+GapCloser	832,132	526,869,874 (86%)	455,924,417 (74%)	335,033	5,903
PBJelly2	825,889	547,240,256 (89%)	481,594,113 (78%)	329,358	5,642
SSPACE	726,400	545,659,344 (89%)	479,897,493 (78%)	329,358	6,126
LRNA scaffolder	720,537	546,375,413 (89%)	479,897,493 (78%)	329,358	6,398
Final over 1kb	74,927	397,889,956 (65%)	332,624,619 (54%)	329,358	12,808

Table S7: The *de novo* assembly of *M. annua* compared to the assemblies of the other sequenced members of the Malpighiales.

	Genome size (mb)	Scaffold number	Total assembled bases	Scaffold n50 (bp)	Citation
Mercurialis annua (Annual mercury)	640	720,537	546,375,413 (85%)	6,398	(Obbard *et al*., 2006; this study)
Ricinus communis (Castor bean)	400	25,763	350,621,860 (88%)	496,528	(Chan *et al*., 2010)
Populus trichocarpa (Poplar)	485	2,514	417,286,671 (86%)	18,835,763	(Tuskan *et al*., 2006)
Linum usitatissimum (Flax)	350	88,420	318,250,901 (91%)	693,500	(Wang *et al*., 2012c)
Manihot esculenta (Cassava)	760	12,977	416,657,463 (55%)	180,909	(Prochnik *et al*., 2012)
Hevea brasiliensis (Rubber tree)	2,000	608,017	1,100,000,000 (55%)	2,972	(Tang *et al*., 2016)
Jatropha curcas (Oil-bearing tree)	410	39,277	297,660,620 (73%)	15,950	(Sato *et al*., 2011)
Salix suchowensis (Shrub willow)	429	103,144	303,767,511 (71%)	924,964	(Dai *et al*., 2014)

Table S8: Transposable elements in the *Mercurialis* genome. Tandem repeats were identified using Tandem Repeats Finder (TRF). Unknown transposable elements were identified using RepeatModeler for *de novo* transposable element prediction. All other repeats were predicted using Repeat Masker, and repeats from *Euphorbiacae* and *Vitis vinifera*.

Class	Type	Number of repeats	Total length of repeats (bp)	% total repeats	% genome
DNA transposons	MuDR	20,380	5,832,769	2.09	1.22
	En-Spm	12,522	3,127,655	1.12	0.65
	hAT-Ac	10,042	1,837,516	0.66	0.38
	Other	12,875	2,051,441	0.73	0.43
Retrotransposons	LTR/gypsy	99,988	31,437,770	11.25	6.55
	LTR/copia	33,933	11,372,184	4.07	2.37
	LTR/Caulimovirus	2,115	782,568	0.28	0.16
	LTR/Other	1,404	348,956	0.12	0.07
	LINE/L1	21,313	9,554,789	3.42	1.99
	LINE/CRE	8,876	2,368,568	0.85	0.49
	LINE/R1	3,330	925,137	0.33	0.19
	LINE/Other	825	141,630	0.05	0.03
	SINE	1,012	142,744	0.05	0.03
Other	RC/Helitron	801	258,178	0.09	0.05
	Small RNA	1,316	298,024	0.11	0.06
	Simple repeats	40,827	13,362,428	4.78	2.78
TRF	Tandem Repeats	215,433	35,543,829	12.72	7.41
RepeatModeler	Unknown	789,194	160,024,517	57.27	33.35
Totals:		1,276,186	27,9410,703	100	58.22

Table S9: Linear models on Box-Cox-transformed measures of sex-biased gene expression across all genes in *M. annua* (absolute $\log_2$FC difference between the sexes), nucleotide diversity (π), and pairwise dN/dS (with reference to *M. huetii* and *Ricinus communis)*, as a function of sex linkage (SL: sex-linked; PAR: pseudoautosomal, Au: autosomal), sex-bias (male: male-biased; female: female-biased) and their interactions.

Predictors		Dependent Variables							
		Abs($\log_2$FC)		π		dN/dS (*M. huetii*)		dN/dS (*R. communis*)	
		B (CI)	p	B (CI)	p	B (CI)	p	B (CI)	p
(Intercept)		-1.64 (-1.65 – -1.62)	**<.001**	-1.92 (-1.92 – -1.92)	**<.001**	-1.60 (-1.61 – -1.58)	**<.001**	-1.40 (-1.41 – -1.39)	**<.001**
Au	PAR	0.01 (-0.02 – 0.05)	.512	-0.01 (-0.01 – 0.00)	.105	0.03 (-0.01 – 0.07)	.120	0.01 (-0.02 – 0.04)	.447
	SL	0.05 (0.00 – 0.10)	**.037**	-0.01 (-0.02 – -0.00)	**.039**	0.07 (0.01 – 0.12)	**.017**	-0.02 (-0.06 – 0.02)	.335
SB	female	1.37 (1.25 – 1.49)	**<.001**	0.01 (-0.02 – 0.03)	.582	0.21 (0.06 – 0.36)	**.006**	0.03 (-0.10 – 0.17)	.642
	male	1.66 (1.57 – 1.75)	**<.001**	-0.02 (-0.04 – 0.00)	.066	-0.04 (-0.14 – 0.06)	.423	0.08 (0.00 – 0.15)	**.046**
Interactions	PAR: female	-0.07 (-0.33 – 0.20)	.633	-0.04 (-0.10 – 0.02)	.234	-0.38 (-0.64 – -0.11)	**.005**	-0.05 (-0.30 – 0.20)	.688
	SL:female	-0.31 (-0.61 – 0.00)	.052	-0.02 (-0.08 – 0.05)	.612	-0.13 (-0.48 – 0.22)	.463	0.16 (-0.11 – 0.43)	.237
	PAR:male	0.12 (-0.15 – 0.40)	.384	0.03 (-0.03 – 0.08)	.377	-0.02 (-0.31 – 0.28)	.903	-0.05 (-0.38 – 0.27)	.747
	SL:male	0.03 (-0.27 – 0.33)	.824	0.09 (0.03 – 0.15)	**.004**	0.13 (-0.32 – 0.58)	.562	-0.06 (-0.32 – 0.20)	.645
Observations		10076		9345		4761		2993	
R2 / adj. R2		.181 / .181		.002 / .001		.004 / .002		.003 / .000	
AIC		17367.804		-13367.118		5774.396		914.510	

Table S10: Linear models on Box-Cox-transformed measures of sex-biased gene expression for sex-biased genes of *M. annua* on their own (absolute $\log_2$FC difference between the sexes), nucleotide diversity (π), and pairwise d_N/d_S (with reference to *M. huetii* and *Ricinus communis)*, as a function of sex linkage (SL: sex linked; PAR: pseudoautosomal; Au: autosomal), sex-bias (male: male-biased; female: female-biased) and their interactions.

Predictors		Dependent Variables							
		Abs($\log_2$FC)		π		dN/dS (*M. huetii*)		dN/dS (*R. communis*)	
		B (CI)	p	B (CI)	p	B (CI)	p	B (CI)	p
(Intercept)		-0.44 (-0.56 – -0.32)	**<.001**	-3.25 (-3.35 – -3.15)	**<.001**	-1.15 (-1.25 – -1.05)	**<.001**	-1.52 (-1.68 – -1.37)	**<.001**
SB	male	0.28 (0.13 – 0.44)	**<.001**	-0.09 (-0.21 – 0.04)	.171	-0.20 (-0.32 – -0.08)	**<.001**	0.06 (-0.12 – 0.23)	.531
Au	PAR	0.01 (-0.26 – 0.28)	.947	-0.31 (-0.53 – -0.09)	**.007**	-0.21 (-0.37 – -0.04)	**.018**	-0.05 (-0.34 – 0.24)	.743
	SL	-0.33 (-0.64 – -0.01)	**.041**	-0.07 (-0.33 – 0.18)	.578	-0.05 (-0.28 – 0.17)	.644	0.18 (-0.13 – 0.48)	.249
Interactions	male:PAR	0.15 (-0.24 – 0.54)	.443	0.23 (-0.07 – 0.53)	.126	0.25 (0.00 – 0.51)	**.049**	0.00 (-0.48 – 0.48)	.995
	male:SL	0.47 (0.03 – 0.90)	**.036**	0.25 (-0.08 – 0.59)	.138	0.29 (-0.08 – 0.65)	.124	-0.29 (-0.72 – 0.13)	.177
Observations		316		224		159		92	
R2 / adj. R2		.098 / .083		.049 / .027		.088 / .059		.028 / -.028	
AIC		566.527		232.044		61.812		60.485	

Table S11: Linear models of Box-Cox-transformed measures of Y/X allelic expression and pairwise d_N/d_S between X- and Y-linked alleles over all genes, as a function of sex linkage (SL: sex linked, PAR: pseudoautosomal, Au: autosomal), sex-bias (male: male-biased; female: female-biased) and their interactions.

Predictors		Dependent Variables	
		Y/X allelic expression	
		B (CI)	p
	(Intercept)	-0.08 (-0.22 – 0.06)	.277
SB	female	1.09 (0.08 – 2.10)	**.035**
	male	0.16 (-0.85 – 1.17)	.758
Au	SL	-0.03 (-0.19 – 0.13)	.721
Interactions	female:SL	-1.39 (-2.52 – -0.26)	**.016**
	male:SL	-0.06 (-1.18 – 1.06)	.921
Observations		447	
R^2 / adj. R^2		.015 / .004	
AIC		983.892	

Table S12: Summary of chi-squared and Fisher exact tests on the proportion of 5% outliers in F_{ST}, male and female SNP density, and heterozygosity across the different linkage groups (focal being the LG shown in the second column), or within LG1 (focal being the sex-linked region). The p-values from comparisons between LGs have been adjusted for multiple testing and are indicated in bold when significant. SL = sex-linked.

Metric	Focal location	Focal Top 5%	Focal 95%	Non-focal Top 5%	Non-focal 95%	Top 5% Expected	% excess	chisq	chisq p	FET p
F_{ST}	**LG1**	102	1132	227	5096	62	**64.52**	32.82	**0**	**0**
	LG2	36	858	293	5370	45	-20.00	1.90	1	1
	LG3	44	914	285	5314	48	-8.33	0.33	1	1
	LG4	35	784	294	5444	41	-14.63	0.92	1	1
	LG5	26	697	303	5531	36	-27.78	3.12	0.619	0.562
	LG6	45	806	284	5422	43	4.65	0.09	1	1
	LG7	25	589	304	5639	31	-19.35	1.06	1	1
	LG8	16	448	313	5780	23	-30.43	2.24	1	0.975
	SL (vs PAR)	38	329	24	843	18	**111.11**	29.53	**0**	**0**
Male SNP density	**LG1**	125	1109	204	5119	62	**101.61**	82.04	**0**	**0**
	LG2	28	866	301	5362	45	**-37.78**	7.27	**0.056**	**0.039**
	LG3	46	912	283	5316	48	-4.17	0.06	1	1
	LG4	35	784	294	5444	41	-14.63	0.92	1	1
	LG5	14	709	315	5519	36	**-61.11**	15.47	**0.001**	**0**
	LG6	31	820	298	5408	43	-27.91	3.55	0.475	0.420
	LG7	21	593	308	5635	31	-32.26	3.27	0.566	0.514
	LG8	29	435	300	5793	23	26.09	1.33	1	1
	SL (vs PAR)	56	311	6	861	18	**211.11**	111.63	**0**	**0**
Female SNP density	LG1	70	1164	259	5064	62	12.90	1.20	1	1
	LG2	30	864	299	5364	45	-33.33	5.60	0.144	0.106
	LG3	44	914	285	5314	48	-8.33	0.33	1	1
	LG4	22	797	307	5431	41	**-46.34**	10.12	**0.012**	**0.005**
	LG5	47	676	282	5552	45	4.44	3.41	0.519	0.461
	LG6	54	797	275	5431	43	25.58	3.31	0.552	0.509
	LG7	40	574	289	5654	31	29.03	2.85	0.732	0.641
	LG8	22	442	307	5786	23	-4.35	0.03	1	1
	SL (vs PAR)	5	362	57	810	18	**-72.22**	13.61	**0**	**0**
Male heterozygosity	**LG1**	36	1198	293	5030	62	**-41.94**	13.53	**0.002**	**0.001**
	LG2	46	848	283	5380	45	2.22	0.01	1	1
	LG3	56	902	273	5326	48	16.67	1.42	1	1
	LG4	40	779	289	5449	41	-2.44	0.01	1	1
	LG5	36	687	293	5541	45	-20.00	0.00	1	1
	LG6	48	803	281	5425	43	11.63	0.65	1	1
	LG7	36	578	293	5650	31	16.13	0.83	1	1
	LG8	31	433	298	5795	23	34.78	2.54	0.890	0.777
	SL vs PAR	21	346	41	826	18	16.67	0.35	0.557	0.477
Female Heterozygosity	**LG1**	31	1203	298	5025	62	**-50.00**	19.38	**0.000**	**0.000**
	LG2	34	860	295	5368	45	-24.44	2.92	0.702	0.663
	LG3	64	894	265	5334	48	33.33	6.11	0.108	0.104
	LG4	33	786	296	5442	41	-19.51	1.69	1	1
	LG5	40	683	289	5545	45	-11.11	0.34	1	1
	LG6	56	795	273	5433	43	30.23	4.64	0.249	0.228
	LG7	46	568	283	5660	31	**48.39**	8.14	**0.035**	**0.038**
	LG8	25	439	304	5789	23	8.70	0.07	1	1
	SL vs PAR	9	358	53	814	18	**-50.00**	6.49	**0.011**	**0.006**

Supplementary Information: Figures

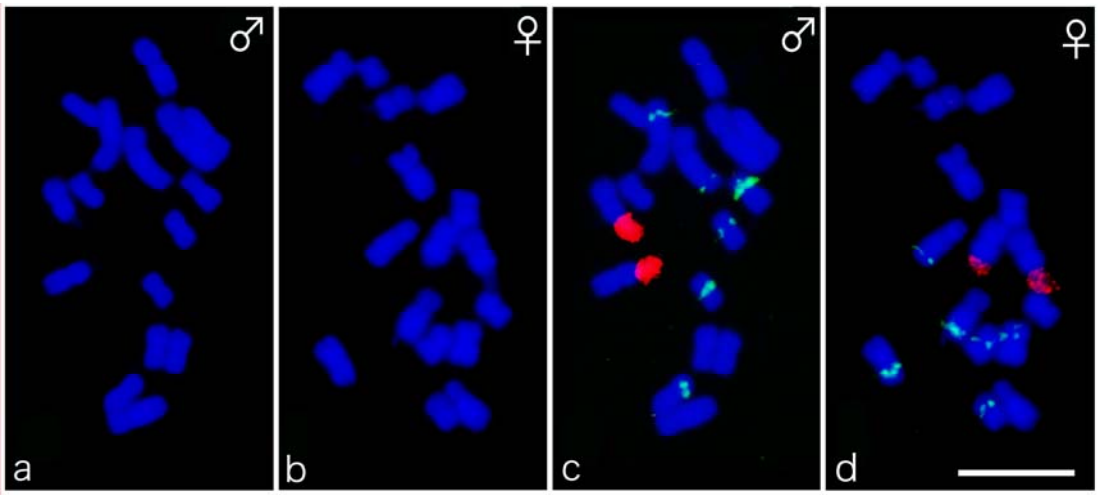


Figure S1: Male (a) and female (b) mitotic chromosomes of *Mercurialis annua* counterstained with DAPI (blue). Bicolor FISH of the same chromosome spreads is also shown (c and d). Hybridisation was performed with 45S rDNA (red) and 5S rDNA (green). Bar= 5 µm.

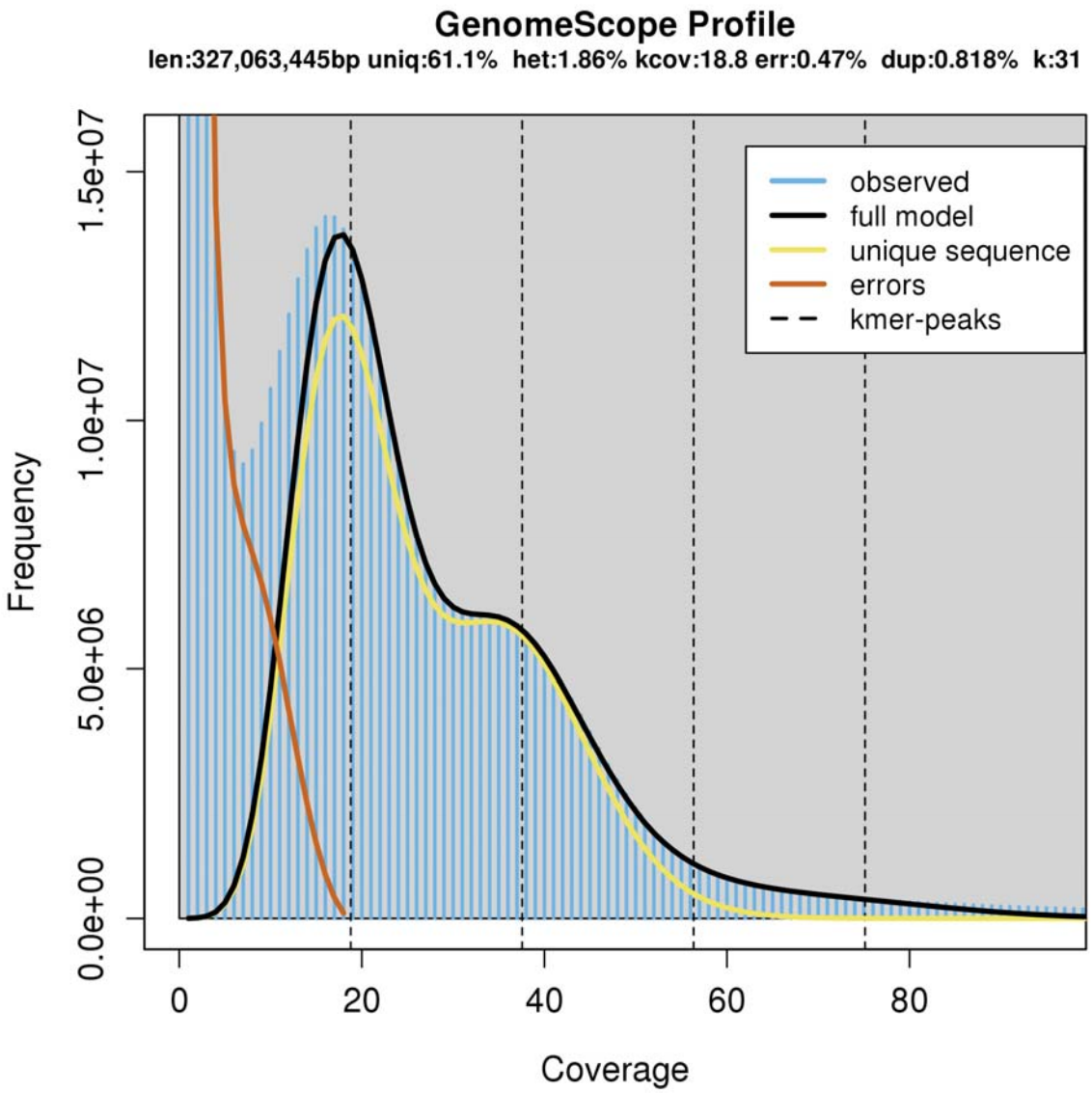


Figure S2: K-mer profile produced by GenomeScope.

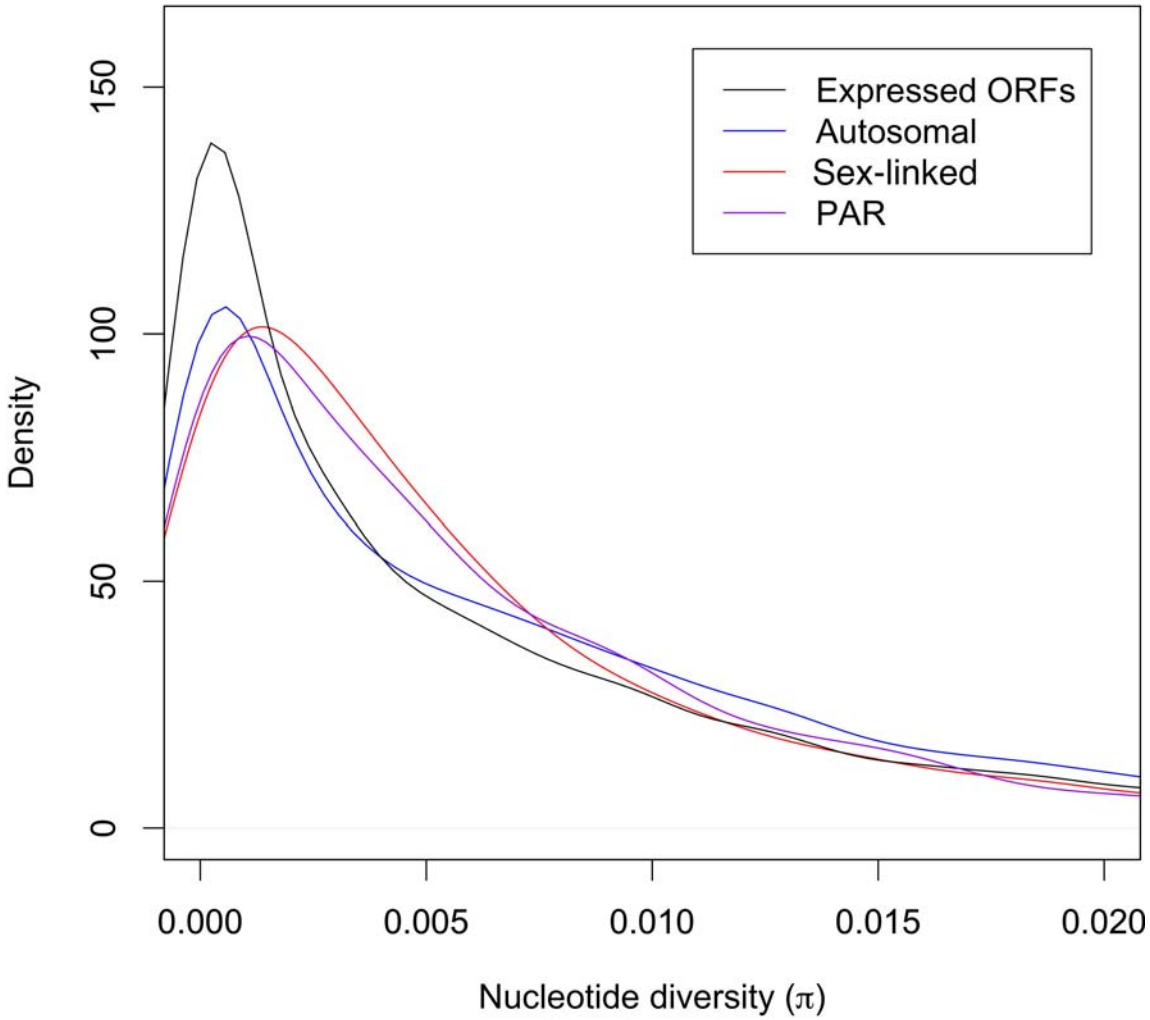


Figure S3: Density plot of nucleotide diversity (π) / kb across protein-coding genes for the X-linked, autosomal, PAR and expressed ORF bins. Density is calculated using a Gaussian kernel, such that the area under each curve sums to 1.

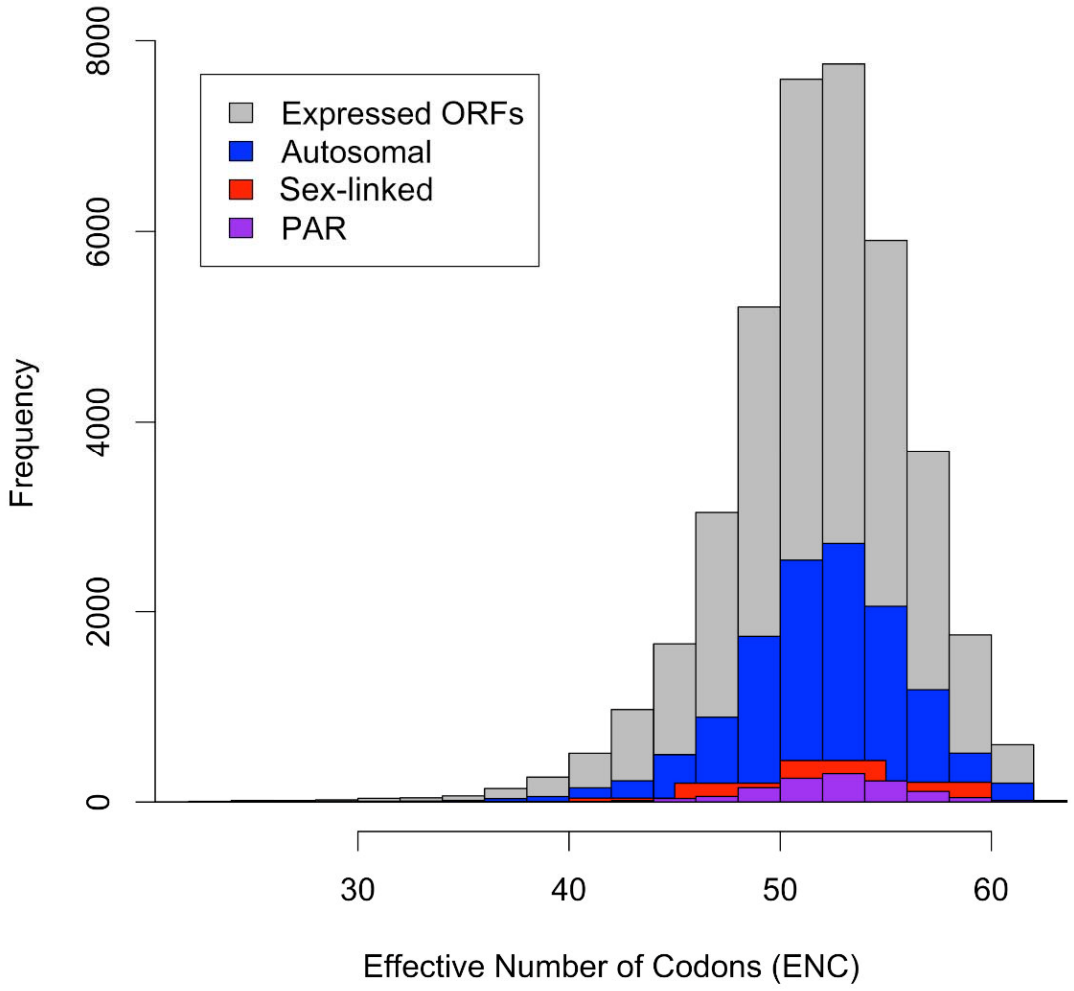


Figure S4: Histograms showing the effective number of codons per gene in the sex-linked, autosomal, PAR and expressed ORF bins.

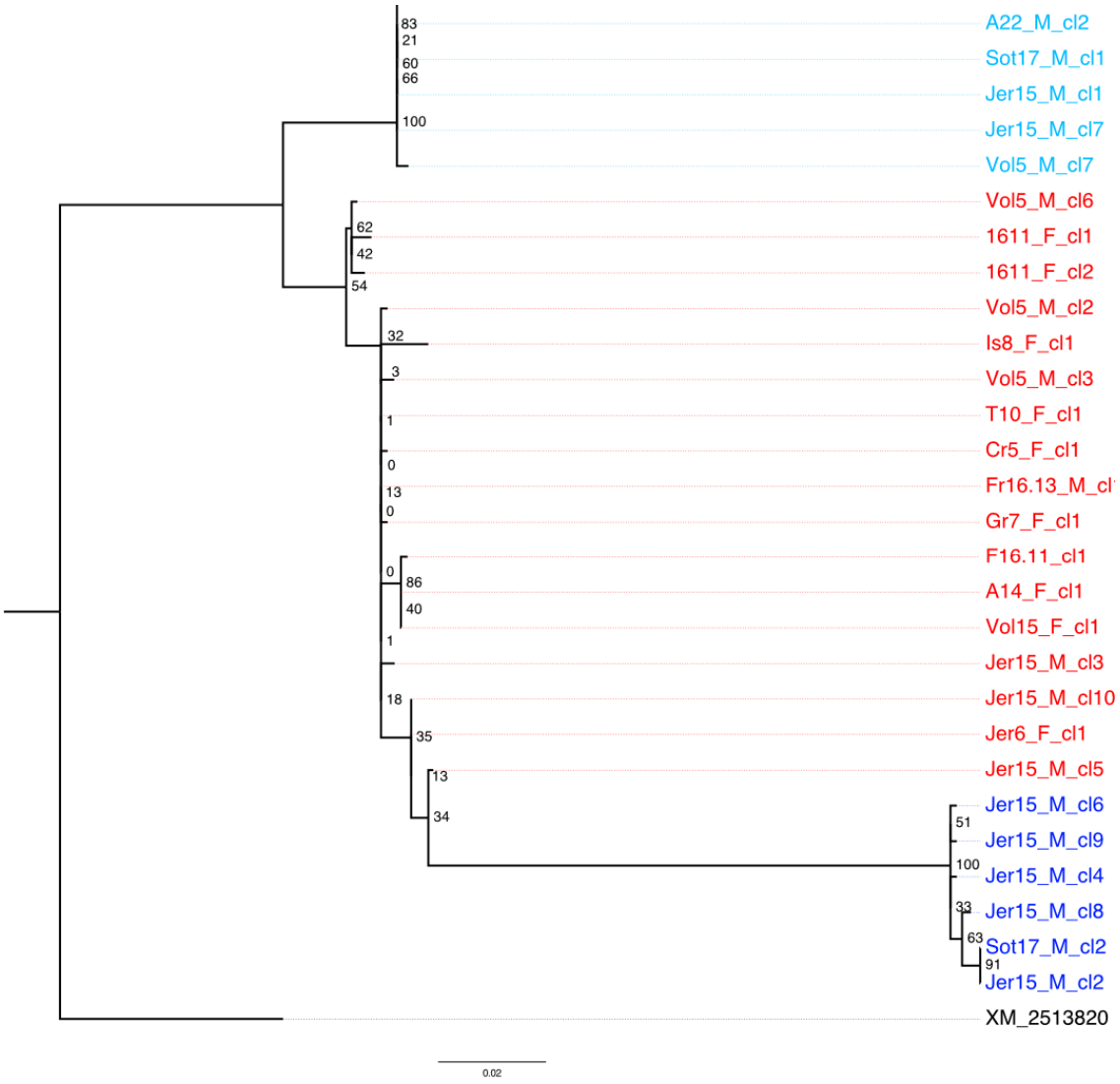


Figure S5: Phylogenetic tree of a sex-linked sequence inferred to have at least one premature stop codon in the Y-linked allele. Sequences marked in red amplify from both males and females, and are inferred to be on the X chromosome. The two groups marked in blue amplify from males only, sometimes from the same male indicating a duplication, and both have premature stop codons and other indels, which are more in the dark blue group. Numbers in black indicate node support from permutations. The outgroup is the *R. communis* homologue.

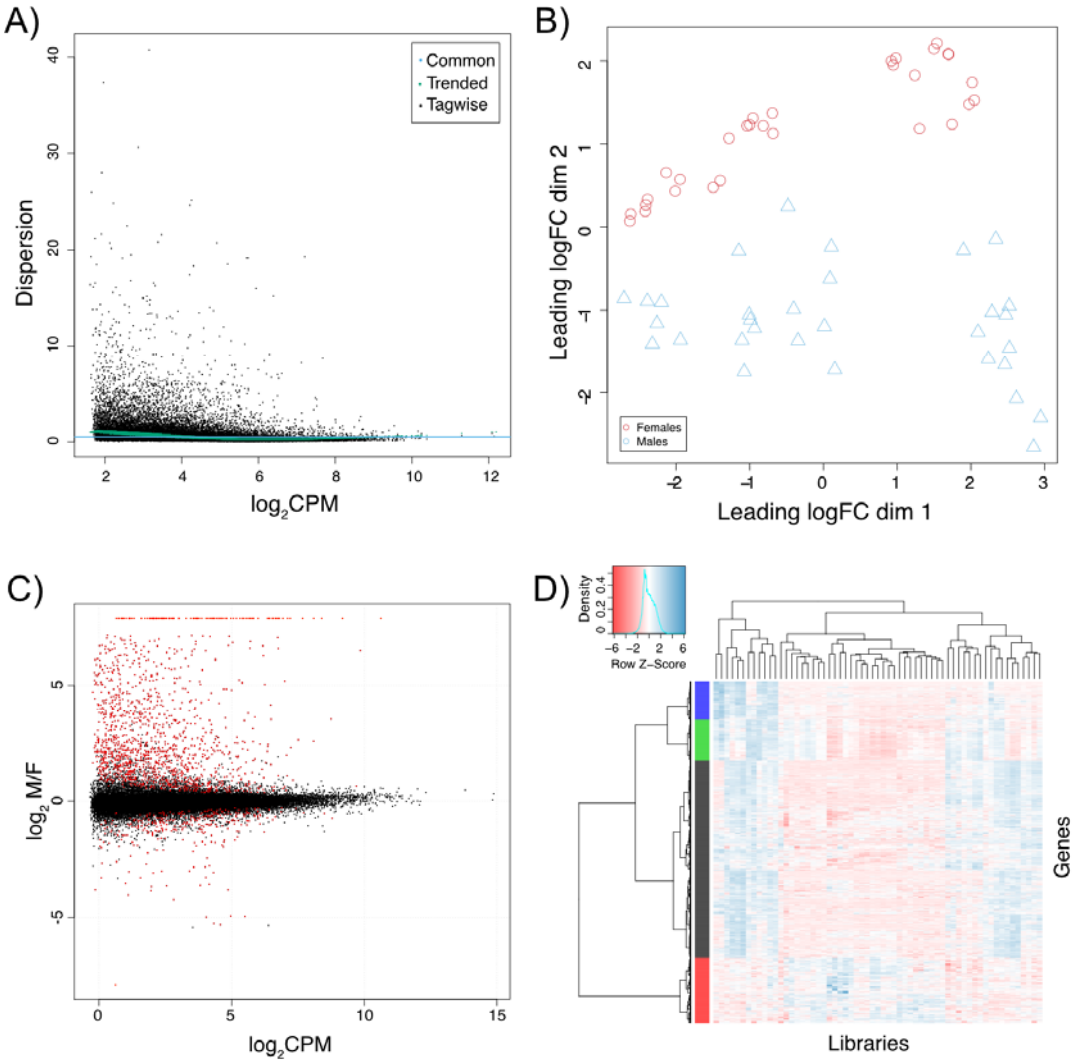


Figure S6: Summary of the differential expression analysis between males and females A) Dispersion plot, B) MDS plot with colour coding of male and female libraries, C) volcano plot indicating log fold change differences between males and females (differentially expressed genes are marked in red), D) heat map of the normalised counts in differentially expressed genes for each library.

Supplementary Information: Files

File S1: Information on the mapped transcripts. Tab1: Information on expression, sex-bias, grouping in relation to recombination in males (SL, PAR, Au), position in male and female recombination map, linkage group assignment, pairwise d_N, d_S compared to *M. huetii* and *R. communis* orthologs, first blastx hit name to swissprot, GO terms, relative expression of X and Y copy, nucleotide diversity. Tab2: Nucleotide diversity for pairs of transcripts between X/Y, *M. annua*/*M. huetii*, *M. annua*/*R. communis*. Tab3: Nucleotide diversity on transcripts with orthologs in both *M. huetii* and *R. communis,* used for estimating the Y branch length. Column names m1 and m2 refer to the model, lnL refers to their log likelihood. Topology 1 follows the species relationship, topology 2 is ((X, M. huetii)Y).

File S2: Fasta file of all sequences obtained from cloning and sequencing transcript 16827 from multiple individuals, and script used to generate the tree. The first letters in the sequence names are A: Antalya (TR), Jer: Jerusalem (IS), Vol: Volos (GR), 16: Montpellier (FR), T: Trabzon (TR), Sot: Southampton (UK). The following numbers identify the individual, and the sex is indicated by M (male) or F (female).

File S3: GO analysis summary for the sex-, male- and female- biased genes (one per tab). The first column abbreviates the main GO category - BP: Biological Process, CC: Cellular Component, MF: Molecular Function.

File S4: Full transcriptome annotation. The columns represent transcript name, best blastx hit to Swiss-Prot, and GO term annotation.

File S5: Transcriptome sequence. Includes the X and Y alleles, inferred from SEX-DETector, and the *M. huetii* transcriptome used.

File S6: Genomic contig sequence, and gff based on the transcriptome.

File S7: *M. annua* repeat library.

References

AKAGI, T., I. M. HENRY, R. TAO, and L. COMAI, 2014 A Y-chromosome–encoded small RNA acts as a sex determinant in persimmons. Science **346**: 646-650.

ALEXA, A., and J. RAHNENFUHRER, 2010 topGO: enrichment analysis for gene ontology. R package version **2**:

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215**: 403-410.

AU, K. F., J. G. UNDERWOOD, L. LEE, and W. H. WONG, 2012 Improving PacBio long read accuracy by short read alignment. PloS one **7**: e46679.

BACHTROG, D., 2013 Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet **14**: 113-124.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B **57**: 289-300.

BENSON, G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27**: 573-580.

BERGERO, R., and D. CHARLESWORTH, 2009 The evolution of restricted recombination in sex chromosomes. Trends Ecol Evol **24**: 94-102.

BERGERO, R., A. FORREST, E. KAMAU, and D. CHARLESWORTH, 2007 Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes. Genetics **175**: 1945-1954.

BLAVET, N., H. BLAVET, A. MUYLE, J. KÄFER, R. CEGAN, C. DESCHAMPS, N. ZEMP, S. MOUSSET, S. AUBOURG, R. BERGERO, D. CHARLESWORTH, R. HOBZA, A. WIDMER, and G. A.

B. MARAIS, 2015 Identifying new sex-linked genes through BAC sequencing in the dioecious plant *Silene latifolia*. BMC Genomics **16**: 546.

BOETZER, M., C. V. HENKEL, H. J. JANSEN, D. BUTLER, and W. PIROVANO, 2011 Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**: 587-579.

BOLGER, A. M., M. LOHSE, and B. USADEL, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114-2120.

BOUALEM, A., C. TROADEC, C. CAMPS, A. LEMHEMDI, H. MORIN, M. A. SARI, R. FRAENKEL-ZAGOURI, I. KOVALSKI, C. DOGIMONT, R. PERL-TREVES, and A. BENDAHMANE, 2015 A cucurbit androecy gene reveals how unisexual flowers develop and dioecy emerges. Science **350**: 688-691.

BRAY, N. L., H. PIMENTEL, P. MELSTED, and L. PACHTER, 2016 Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol **34**: 525-527.

BROMAN, K. W., H. WU, S. SEN, and G. A. CHURCHILL, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889-890.

CHAN, A. P., J. CRABTREE, Q. ZHAO, H. LORENZI, J. ORVIS, D. PUIU, A. MELAKE-BERHAN, K. M. JONES, J. REDMAN, G. CHEN, E. B. CAHOON, M. GEDIL, M. STANKE, B. J. HAAS, J. R. WORTMAN, C. M. FRASER-LIGGETT, J. RAVEL, and P. D. RABINOWICZ, 2010 Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol **28**: 951-956.

CHARLESWORTH, B., 1991 The evolution of sex chromosomes. Science **251**: 1030-1033.

CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH, 1993a The effect of deleterious mutations on neutral molecular variation. Genetics **134**: 1289-1303.

CHARLESWORTH, B., and D. CHARLESWORTH, 2000 The degeneration of Y chromosomes. Philosophical Transactions of the Royal Society B: Biological Sciences **355**: 1563.

CHARLESWORTH, B., and D. CHARLESWORTH, 1978 A model for the evolution of dioecy and gynodioecy. Am Nat **112**: 975-997.

CHARLESWORTH, D., 2015 Plant contributions to our understanding of sex chromosome evolution. New Phytol **208**: 52-65.

CHARLESWORTH, D., 2016 Plant sex chromosomes. Annu Rev Plant Biol **67**: 397-420.

CHARLESWORTH, D., B. CHARLESWORTH, and G. A. MARAIS, 2005 Steps in the evolution of heteromorphic sex chromosomes. Heredity **95**: 118.

CHARLESWORTH, D., M. T. MORGAN, and B. CHARLESWORTH, 1993b Mutation accumulation in finite populations. J Hered **84**: 321-325.

CHARLESWORTH, D., 2013 Plant sex chromosome evolution. J Exp Bot **64**: 405-420.

CONESA, A., S. GOTZ, J. M. GARCIA-GOMEZ, J. TEROL, M. TALON, and M. ROBLES, 2005 Blast2GO: a universal tool for annotation, visualisation and analysis of functional genomics research. Bioinformatics **21**:

CONNALLON, T., and A. G. CLARK, 2010 Sex linkage, sex-specific selection, and the role of recombination in the evolution of sexually dimorphic gene expression. Evolution **64**: 3417-3442.

DAG, O., O. ASAR, and O. ILK, 2014 A Methodology to Implement Box-Cox Transformation When No Covariate is Available. Communications in Statistics - Simulation and Computation **43**: 1740-1759.

DAI, X., Q. HU, Q. CAI, K. FENG, N. YE, G. A. TUSKAN, R. MILNE, Y. CHEN, Z. WAN, Z. WANG, W. LUO, K. WANG, D. WAN, M. WANG, J. WANG, J. LIU, and T. YIN, 2014 The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res **24**: 1274-1277.

DANECEK, P., A. AUTON, G. ABECASIS, C. A. ALBERS, E. BANKS, M. A. DEPRISTO, R. E. HANDSAKER, G. LUNTER, G. T. MARTH, S. T. SHERRY, G. MCVEAN, R. DURBIN, and G. P. A.

G. 1000, 2011 The variant call format and VCFtools. Bioinformatics **27**: 2156-2158.

DELPH, L. F., A. M. ARNTZ, C. SCOTTI-SAINTAGNE, and I. SCOTTI, 2010 The genomic architecture of sexual dimorphism in the dioecious plant *Silene latifolia*. Evolution **64**: 2873-2886.

DEPRISTO, M. A., E. BANKS, R. POPLIN, K. V. GARIMELLA, J. R. MAGUIRE, C. HARTL, A. A. PHILIPPAKIS, G. DEL ANGEL, M. A. RIVAS, M. HANNA, A. MCKENNA, T. J. FENNELL, A. M. KERNYTSKY, A. Y. SIVACHENKO, K. CIBULSKIS, S. B. GABRIEL, D. ALTSHULER, and M. J. DALY, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43**: 491-498.

DURAND, B., 1963 Le complèxe Mercurialis annua L. s.l.: Une étude biosystématique. Annales des Sciences Naturelles, Botanique, Paris **12**: 579-736.

DURAND, B., and R. DURAND, 1991 Sex determination and reproductive organ differentiation in *Mercurialis*. Plant science **80**: 49-66.

DURAND, B., J. P. LOUIS, S. HAMDI, E. CABRE, L. X. YU, B. GUERIN, and G. TELLER, 1987 Major regulator genes, phytohormone levels and specific gene-expression for reproductive organogenesis in *Mercurialis annua* L. (2n=16). J Cell Biochem 18-20.

ENGLISH, A. C., S. RICHARDS, Y. HAN, M. WANG, V. VEE, J. QU, X. QIN, D. M. MUZNY, J. G. REID, K. C. WORLEY, and R. A. GIBBS, 2012 Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One **7**: e47768.

FECHTER, I., L. HAUSMANN, M. DAUM, T. R. SÖRENSEN, P. VIEHÖVER, B. WEISSHAAR, and R. TÖPFER, 2012 Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. Molecular Genetics and Genomics **287**: 247-259.

FILATOV, D. A., 2015 Homomorphic plant sex chromosomes are coming of age. Mol Ecol **24**: 3217-3219.

GERALDES, A., C. A. HEFER, A. CAPRON, N. KOLOSOVA, F. MARTINEZ-NUÑEZ, R. Y. SOOLANAYAKANAHALLY, B. STANTON, R. D. GUY, S. D. MANSFIELD, C. J. DOUGLAS, and Q. C. CRONK, 2015 Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). Mol Ecol **24**: 3243-3256.

GIBSON, J. R., A. K. CHIPPINDALE, and A. M. RICE, 2002 The X chromosome is a hot spot for sexually antagonistic fitness variation. Proc R Soc Lond, B **269**: 499-505.

GONZÁLEZ-MARTÍNEZ, S. C., K. RIDOUT, and J. R. PANNELL, 2017 Range expansion compromises adaptive evolution in an outcrossing plant. Curr Biol **27**: 1-8.

HAAS, B. J., A. PAPANICOLAOU, M. YASSOUR, M. GRABHERR, P. D. BLOOD, J. BOWDEN, M. B. COUGER, D. ECCLES, B. LI, M. LIEBER, M. D. MACMANES, M. OTT, J. ORVIS, N. POCHET, F. STROZZI, N. WEEKS, R. WESTERMAN, T. WILLIAM, C. N. DEWEY, R. HENSCHEL, R. D. LEDUC, N. FRIEDMAN, and A. REGEV, 2013 *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc **8**: 1494-1512.

HARRIS, R. S., 2007 Improved pairwise alignment of genomic DNA. The Pennsylvania State University **PhD**:

HESSE, E., and J. PANNELL, 2011 Sexual dimorphism in a dioecious population of the wind-pollinated herb *Mercurialis annua*: the interactive effects of resource availability and competition.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res **8**: 269-294.

HOROVITZ, S., and H. JIMÉNEZ, 1967 Cruzamientos interespecíficos e intergenéricos en Caricaceas y sus implicaciones fitotécnicas. Agronomia Tropical **17**: 323-343.

HOUGH, J., J. D. HOLLISTER, W. WANG, S. C. H. BARRET, and S. I. WRIGHT, 2014 Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. Proc Natl Acad Sci USA **111**: 7713-7718.

HUANG, X., and A. MADAN, 1999 CAP3: A DNA sequence assembly program. Genome Res **9**: 868-877.

HUNT, M., T. KIKUCHI, M. SANDERS, C. NEWBOLD, M. BERRIMAN, and T. D. OTTO, 2013 REAPR: a universal tool for genome assembly evaluation. Genome Biol **14**: R47.

KATOH, K., and D. M. STANDLEY, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**: 772-780.

KAZAMA, Y., K. ISHII, W. AONUMA, T. IKEDA, H. KAWAMOTO, A. KOIZUMI, D. A. FILATOV, M. CHIBALINA, R. BERGERO, D. CHARLESWORTH, T. ABE, and S. KAWANO, 2016 A new physical mapping approach refines the sex-determining gene positions on the *Silene latifolia* Y-chromosome. Sci Rep **6**: 18917.

KEARSE, M., R. MOIR, A. WILSON, S. STONES-HAVAS, M. CHEUNG, S. STURROCK, S. BUXTON, A. COOPER, S. MARKOWITZ, C. DURAN, T. THIERER, B. ASHTON, P. MEINTJES, and A. DRUMMOND, 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**: 1647-1649.

KEJNOVSKY, E., and B. VYSKOT, 2010 *Silene latifolia*: The classical model to study heteromorphic sex chromosomes. Cytogenetic and Genome Research **129**: 250-262.

KHADKA, D. K., A. NEJIDAT, M. TAL, and A. GOLAN-GOLDHIRSH, 2002 DNA markers for sex: Molecular evidence for gender dimorphism in dioecious *Mercurialis annua* L. Mol Breed **9**: 251-257.

KHADKA, D. K., A. NEJIDAT, M. TAL, and A. GOLAN-GOLDHIRSH, 2005 Molecular characterization of a gender-linked DNA marker and a related gene in *Mercurialis annua* L. Planta **222**: 1063-1070.

KOBOLDT, D. C., K. CHEN, T. WYLIE, D. E. LARSON, M. D. MCLELLAN, E. R. MARDIS, G. M. WEINSTOCK, R. K. WILSON, and L. DING, 2009 VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics **25**: 2283-2285.

KRÄHENBÜHL, M., Y. M. YUAN, and P. KÜPFER, 2002 Chromosome and breeding system evolution of the genus *Mercurialis* (Euphorbiaceae): implications of ITS molecular phylogeny. Plant Syst Evol **234**: 155-169.

KRASOVEC, M., M. CHESTER, K. RIDOUT, and D. A. FILATOV, 2018 The mutation rate and the age of the sex chromosomes in *Silene latifolia*. Curr Biol **28**: 1832-1838.

KRYLOV, D. M., Y. I. WOLF, I. B. ROGOZIN, and E. V. KOONIN, 2003 Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res **13**: 2229-2235.

KUHN, E., 1939 Selbstbestäubungen subdiöcischer Blütenpflanzen, ein neuer Beweis für die genetische Theorie der Geschlechtsbestimmung. Planta **30**: 457-470.

LABOUCHE, A. M., and J. R. PANNELL, 2016 A test of the size-constraint hypothesis for a limit to sexual dimorphism in plants. Oecologia **181**: 873-884.

LAHN, B. T., and D. C. PAGE, 1999 Four evolutionary strata on the human X chromosome. Science **286**: 964-957.

LANGMEAD, B., and S. L. SALZBERG, 2012 Fast gapped-read alignment with Bowtie 2. Nature Methods **9**: 357-359.

LENGEROVA, M., E. KEJNOVSKY, R. HOBZA, J. MACAS, S. R. GRANT, and B. VYSKOT, 2004 Multicolor FISH mapping of the dioecious model plant, *Silene latifolia*. Theor Appl Genet **108**: 1193-1199.

LI, H., and R. DURBIN, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754-1760.

LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, G. ABECASIS, R. DURBIN, and G. P. D. P. S. 1000, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078-2079.

LIU, Z., P. H. MOORE, H. MA, C. M. ACKERMAN, M. RAGLBA, Q. YU, H. M. PEARL, M. S. KIM, J. W. CHARLTON, J. I. STILES, F. T. ZEE, A. H. PATERSON, and R. MING, 2004 A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature **427**: 348-352.

LOEPTIEN, H., 1979 Identification of the sex chromosome pair in asparagus (*Asparagus officinalis L.*). Zeitschrift fur Pflanzenzüchtung **82**: 162-173.

LÜDECKE, D., 2017 sjPlot: Data Visualization for Statistics in Social Science. R package version 2.4.0, https://CRAN.R-project.org/package=sjPlot.

LUO, R., B. LIU, Y. XIE, Z. LI, W. HUANG, J. YUAN, G. HE, Y. CHEN, Q. PAN, Y. LIU, J. TANG, G. WU, H. ZHANG, Y. SHI, Y. LIU, C. YU, B. WANG, Y. LU, C. HAN, D. W. CHEUNG, S.-M. YIU, S. PENG, Z. XIAOQIAN, G. LIU, X. LIAO, Y. LI, H. YANG, J. WANG, T.-W. LAM, and J. WANG, 2012 SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience **1**: 18.

LYNCH, M., M. S. ACKERMAN, J.-F. GOUT, H. LONG, W. SUNG, W. K. THOMAS, and P. L. FOSTER, 2016 Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet **17**: 704-714.

MARÇAIS, G., and C. KINGSFORD, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**: 764-770.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favourable gene. Genet Res **219**: 23-35.

MCKENNA, A., M. HANNA, E. BANKS, A. SIVACHENKO, K. CIBULSKIS, A. KERNYTSKY, K. GARIMELLA, D. ALTSHULER, S. GABRIEL, M. DALY, and M. A. DEPRISTO, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20**: 1297-1303.

MEISEL, R. P., J. H. MALONE, and A. G. CLARK, 2012 Disentangling the relationship between sex-biased gene expression and X-linkage. Genome Res 22: 1255-1265.

MONTGOMERY, S. H., and J. E. MANK, 2016 Inferring regulatory change from gene expression: the confounding effects of tissue scaling. Mol Ecol **25**: 5114-5128.

MORGULIS, A., E. M. GERTZ, A. A. SCHÄFFER, and R. AGARWALA, 2006 A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol **13**: 1028-1040.

MUYLE, A., J. KÄFER, N. ZEMP, S. MOUSSET, F. PICARD, and G. A. MARAIS, 2016 SEX-DETector: A probabilistic approach to study sex chromosomes in non-model organisms. Genome Biol Evol **8**: 2530-2543.

NAM, K., and H. ELLEGREN, 2008 The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. Genetics **180**: 1131-1136.

NAVAJAS-PÉREZ, R., R. DE LA HERRÁN, G. LÓPEZ GONZÁLEZ, M. JAMILENA, R. LOZANO, C. RUIZ REJÓN, M. RUIZ REJÓN, and M. A. GARRIDO-RAMOS, 2005 The evolution of reproductive systems and sex-determining mechanisms within *Rumex* (polygonaceae) inferred from nuclear and chloroplastidial sequence data. Mol Biol Evol **22**: 1929-1939.

NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**: 929-936.

OBBARD, D. J., S. A. HARRIS, R. J. A. BUGGS, and J. PANNELL, 2006 Hybridization, polyploidy, and the evolution of sexual systems in *Mercurialis* (Euphorbiaceae). Evolution **60**: 1801-1815.

ONO, T., 1939 Polyploidy and sex determination in *Melandrium*. I. Colcicine-induced polyploids of *Melandrium album*. Botanical Magazine **53**: 549-556.

OSSOWSKI, S., K. SCHNEEBERGER, J. I. LUCAS-LLEDÓ, N. WARTHMANN, R. M. CLARK, R. G. SHAW, D. WEIGEL, and M. LYNCH, 2010 The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327**: 92-94.

PÁL, C., B. PAPP, and L. D. HURST, 2001 Highly expressed genes in yeast evolve slowly. Genetics

158: 927-931.

PANNELL, J., M. E. DORKEN, B. PUJOL, and R. BERJANO, 2008 Gender variation and transitions between sexual systems in *Mercurialis annua* (Euphorbiaceae). Int J Plant Sci **169**: 129-139.

PANNELL, J. R., D. J. OBBARD, and R. J. A. BUGGS, 2004 Polyploidy and the sexual system: what can we learn from *Mercurialis annua*? Biol J Linn Soc **82**: 547-560.

PAOLUCCI, I., M. GAUDET, V. JORGE, I. BERITOGNOLO, S. TERZOLI, E. KUZMINSKY, R. MULEO, G. SCARASCIA MUGNOZZA, and M. SABATTI, 2010 Genetic linkage maps of *Populus alba* L. and comparative mapping analysis of sex determination across *Populus* species. Tree Genetics & Genomes **6**: 863-875.

PAPADOPULOS, A. S., M. CHESTER, K. RIDOUT, and D. A. FILATOV, 2015 Rapid Y degeneration and dosage compensation in plant sex chromosomes. Proc Natl Acad Sci USA **112**: 13021-13026.

PARISI, M., R. NUTTALL, D. NAIMAN, G. BOUFFARD, J. MALLEY, J. ANDREWS, S. EASTMAN, and B. OLIVER, 2003 Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science **299**: 697-700.

PICQ, S., S. SANTONI, T. LACOMBE, M. LATREILLE, A. WEBER, M. ARDISSON, S. IVORRA, D. MAGHRADZE, R. ARROYO-GARCIA, P. CHATELET, P. THIS, J. F. TERRAL, and R. BACILIERI, 2014 A small XY chromosomal region explains sex determination in wild dioecious *V. vinifera* and the reversal to hermaphroditism in domesticated grapevines. BMC Plant Biol **14**: 229.

PROCHNIK, S., P. R. MARRI, B. DESANY, P. D. RABINOWICZ, C. KODIRA, M. MOHIUDDIN, F. RODRIGUEZ, C. FAUQUET, J. TOHME, T. HARKINS, D. S. ROKHSAR, and S. ROUNSLEY, 2012 The Cassava Genome: Current Progress, Future Directions. Trop Plant Biol **5**: 88-94.

PUCHOLT, P., A. C. RÖNNBERG-WÄSTLJUNG, and S. BERLIN, 2015 Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* l.). Heredity **114**: 575-583.

QIU, S., R. BERGERO, K. ZENG, and D. CHARLESWORTH, 2011 Patterns of codon usage bias in Silene latifolia. Mol Biol Evol **28**: 771-780.

QUANTDT, H. J., A. PÜHLER, and I. BROER, 1993 Transgenic root nodules of *Vicia hirsuta*: A fast and efficient system for the study of gene expression in indeterminate-type nodules. Mol Plant-Microbe Interact **6**: 699-706.

R, development, Core, and Team, 2007 R: A Language and Environment for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. R-project.org

RAHMAN, A. Y., A. O. USHARRAJ, B. B. MISRA, G. P. THOTTATHIL, K. JAYASEKARAN, Y. FENG, S. HOU, S. Y. ONG, F. L. NG, L. S. LEE, H. S. TAN, M. K. SAKAFF, B. S. TEH, B. F. KHOO, S. S. BADAI, N. A. AZIZ, A. YURYEV, B. KNUDSEN, A. DIONNE-LAPORTE, N. P. MCHUNU, Q. YU, B. J. LANGSTON, T. A. FREITAS, A. G. YOUNG, R. CHEN, L. WANG, N. NAJIMUDIN, J. A. SAITO, and M. ALAM, 2013 Draft genome sequence of the rubber tree *Hevea brasiliensis*. BMC Genomics **14**: 75.

RASTAS, P., 2017 Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics **33**: 3726-3732.

RAUTENBERG, A., D. FILATOV, B. SVENNBLAD, N. HEIDARI, and B. OXELMAN, 2008 Conflicting phylogenetic signals in the SlX1/Y1 gene in *Silene*. BMC Evol Biol **8**: 299.

RENNER, S. S., 2016 Pathways for making unisexual flowers and unisexual plants:Moving beyond the “two mutations linked on one chromosome” model. Am J Bot **103**: 587-589.

RENNER, S. S., 2014 The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. Am J Bot **101**: 1588-1596.

RENNER, S. S., and R. E. RICKLEFS, 1995 Dioecy and its correlates in the flowering plants. Am J Bot **82**: 596-606.

RICE, W. R., 1987 The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution **41**: 911-914.

ROBINSON, M. D., D. J. MCCARTHY, and G. K. SMYTH, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139-140.

RUSSELL, J. R. W., and J. PANNELL, 2015 Sex determination in dioecious *Mercurialis annua* and its close diploid and polyploid relatives. Heredity **114**: 262-271.

SÁNCHEZ-VILAS, J., A. TURNER, and J. R. PANNELL, 2010 Sexual dimorphism in intra- and interspecific competitive ability of the dioecious herb *Mercurialis annua*. Plant Biology **13**: 218-222.

SATO, S., H. HIRAKAWA, S. ISOBE, E. FUKAI, A. WATANABE, M. KATO, K. KAWASHIMA, C. MINAMI, A. MURAKI, N. NAKAZAKI, C. TAKAHASHI, S. NAKAYAMA, Y. KISHIDA, M. KOHARA, M. YAMADA, H. TSURUOKA, S. SASAMOTO, S. TABATA, T. AIZU, A. TOYODA, T. SHIN-I, Y. MINAKUCHI, Y. KOHARA, A. FUJIYAMA, S. TSUCHIMOTO, S. KAJIYAMA, E. MAKIGANO, N. OHMIDO, N. SHIBAGAKI, J. A. CARTAGENA, N. WADA, T. KOHINATA, A. ATEFEH, S. YUASA, S. MATSUNAGA, and K. FUKUI, 2011 Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res **18**: 65-76.

SCHMIEDER, R., and R. EDWARDS, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**: 863-864.

SIMÃO, F. A., R. M. WATERHOUSE, P. IOANNIDIS, E. V. KRIVENTSEVA, and E. M. ZDOBNOV, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**: 3210-3212.

SMARDA, P., P. BUREŠ, J. SMERDA, and L. HOROVÁ, 2012 Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. New Phytol **193**: 513-521.

SMIT, A. F. A., R. HUBLEY, and G. P, 2013 RepeatMasker Open-4.0. http://www.repeatmasker.org.

SMITH, B., 1955 Sex chromosomes and natural polyploidy in dioecious *Rumex*. J Hered **46**: 226-232.

SMITH, S. A., and M. J. DONOGHUE, 2008 Rates of molecular evolution are linked to life history in flowering plants. Science **322**: 86-89.

STAMATAKIS, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312-1313.

TANG, C., M. YANG, Y. FANG, Y. LUO, S. GAO, X. XIAO, Z. AN, B. ZHOU, B. ZHANG, X. TAN, H. Y. YEANG, Y. QIN, J. YANG, Q. LIN, H. MEI, P. MONTORO, X. LONG, J. QI, Y. HUA, Z. HE, M. SUN, W. LI, X. ZENG, H. CHENG, Y. LIU, J. YANG, W. TIAN, N. ZHUANG, R. ZENG, D. LI, P. HE, Z. LI, Z. ZOU, S. LI, C. LI, J. WANG, D. WEI, C. Q. LAI, W. LUO, J. YU, S. HU, and H. HUANG, 2016 The rubber tree genome reveals new insights into rubber production and species adaptation. Nat Plants **2**: 16073.

TELGMANN-RAUBER, A., A. JAMSARI, M. S. KINNEY, J. C. PIRES, and C. JUNG, 2007 Genetic and physical maps around the sex-determining M-locus of the dioecious plant *Asparagus*. Mol Genet Genomics **278**: 221-234.

TENNESSEN, J. A., R. GOVINDARAJULU, A. LISTON, and T. L. ASHMAN, 2016 Homomorphic ZW chromosomes in a wild strawberry show distinctive recombination heterogeneity but a small sex-determining region. New Phytol **211**: 1412-1423.

TOUPS, M., N. RODRIGUES, N. PERRIN, and M. KIRKPATRICK, 2019 A reciprocal translocation radically reshapes sex-linked inheritance in the common frog. Mol Ecol doi: 10.1111/mec.14990.

TROMBETTA, B., D. SELLITTO, R. SCOZZARI, and F. CRUCIANI, 2014 Inter- and intraspecies phylogenetic analyses reveal extensive X-Y gene conversion in the evolution of gametologous sequences of human sex chromosomes. Mol Biol Evol **31**: 2108-2123.

TSAGKOGEORGA, G., V. CAHAIS, and N. GALTIER, 2012 The population genomics of a fast

evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biol Evol **4**: 740-749.

TUSKAN, G. A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV, U. HELLSTEN, N. PUTNAM, S. RALPH, S. ROMBAUTS, A. SALAMOV, J. SCHEIN, L. STERCK, A. AERTS, R. R. BHALERAO, R. P. BHALERAO, D. BLAUDEZ, W. BOERJAN, A. BRUN, A. BRUNNER, V. BUSOV, M. CAMPBELL, J. CARLSON, M. CHALOT, J. CHAPMAN, G. L. CHEN, D. COOPER, P. M. COUTINHO, J. COUTURIER, S. COVERT, Q. CRONK, R. CUNNINGHAM, J. DAVIS, S. DEGROEVE, A. DÉJARDIN, C. DEPAMPHILIS, J. DETTER, B. DIRKS, I. DUBCHAK, S. DUPLESSIS, J. EHLTING, B. ELLIS, K. GENDLER, D. GOODSTEIN, M. GRIBSKOV, J. GRIMWOOD, A. GROOVER, L. GUNTER, B. HAMBERGER, B. HEINZE, Y. HELARIUTTA, B. HENRISSAT, D. HOLLIGAN, R. HOLT, W. HUANG, N. ISLAM-FARIDI, S. JONES, M. JONES-RHOADES, R. JORGENSEN, C. JOSHI, J. KANGASJÄRVI, J. KARLSSON, C. KELLEHER, R. KIRKPATRICK, M. KIRST, A. KOHLER, U. KALLURI, F. LARIMER, J. LEEBENS-MACK, J. C. LEPLÉ, P. LOCASCIO, Y. LOU, S. LUCAS, F. MARTIN, B. MONTANINI, C. NAPOLI, D. R. NELSON, C. NELSON, K. NIEMINEN, O. NILSSON, V. PEREDA, G. PETER, R. PHILIPPE, G. PILATE, A. POLIAKOV, J. RAZUMOVSKAYA, P. RICHARDSON, C. RINALDI, K. RITLAND, P. ROUZÉ, D. RYABOY, J. SCHMUTZ, J. SCHRADER, B. SEGERMAN, H. SHIN, A. SIDDIQUI, F. STERKY, A. TERRY, C. J. TSAI, E. UBERBACHER, P. UNNEBERG, J. VAHALA, K. WALL, S. WESSLER, G. YANG, T. YIN, C. DOUGLAS, M. MARRA, G. SANDBERG, Y. VAN DE PEER, and D. ROKHSAR, 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**: 1596-1604.

VELTSOS, P., G. COSSARD, E. BEAUDOING, G. BEYDON, D. SAVOVA BIANCHI, C. ROUX, C. S. GONZÁLEZ-MARTÍNEZ, and J. R. PANNELL, 2018 Size and content of the sex-determining region of the Y chromosome in dioecious *Mercurialis annua*, a plant with homomorphic sex chromosomes. Genes (Basel) **9**: 277.

VURTURE, G. W., F. J. SEDLAZECK, M. NATTESTAD, C. J. UNDERWOOD, H. FANG, J.

GURTOWSKI, and M. C. SCHATZ, 2017 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics **33**: 2202-2204.

WANG, J., J. K. NA, Q. YU, A. R. GSCHWEND, J. HAN, F. ZENG, R. ARYAL, R. VANBUREN, J. E. MURRAY, W. ZHANG, R. NAVAJAS-PÉREZ, F. A. FELTUS, C. LEMKE, E. J. TONG, C. CHEN, C. M. WAI, R. SINGH, M. L. WANG, X. J. MIN, M. ALAM, D. CHARLESWORTH, P. H. MOORE, J. JIANG, A. H. PATERSON, and R. MING, 2012a Sequencing papaya X and Y-h chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci USA **109**: 13710-13715.

WANG, K., Z. WANG, F. LI, W. YE, J. WANG, G. SONG, Z. YUE, L. CONG, H. SHANG, S. ZHU, C. ZOU, Q. LI, Y. YUAN, C. LU, H. WEI, C. GOU, Z. ZHENG, Y. YIN, X. ZHANG, K. LIU, B. WANG, C. SONG, N. SHI, R. J. KOHEL, R. G. PERCY, J. Z. YU, Y. X. ZHU, J. WANG, and S. YU, 2012b The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet **44**: 1098-1103.

WANG, Z., N. HOBSON, L. GALINDO, S. ZHU, D. SHI, J. MCDILL, L. YANG, S. HAWKINS, G. NEUTELINGS, R. DATLA, G. LAMBERT, D. W. GALBRAITH, C. J. GRASSA, A. GERALDES, Q. C. CRONK, C. CULLIS, P. K. DASH, P. A. KUMAR, S. CLOUTIER, A. G. SHARPE, G. K. WONG, J. WANG, and M. K. DEYHOLOS, 2012c The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J **72**: 461-473.

WESTERGAARD, M., 1958 The mechanism of sex determination in dioeciou splants. Adv Genet **9**: 217-281.

WU, C. I., and A. W. DAVIS, 1993 Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. Am Nat **142**: 187-212.

WU, T. D., and C. K. WATANABE, 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**: 1859-1875.

WU, X., J. WANG, J. K. NA, Q. YU, R. C. MOORE, F. ZEE, S. C. HUBER, and R. MING, 2010 The origin of the non-recombining region of sex chromosomes in *Carica* and *Vasconcellea*. Plant J **63**:

801-810.

XUE, W., J. T. LI, Y. P. ZHU, G. Y. HOU, X. F. KONG, Y. Y. KUANG, and X. W. SUN, 2013 L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics **14**: 604.

YAMAMOTO, K., Y. ODA, A. HASEDA, S. FUJITO, T. MIKAMI, and Y. ONODERA, 2014 Molecular evidence that the genes for dioecism and monoecism in *Spinacia oleracea* L. are located at different loci in a chromosomal region. Heredity **112**: 317-324.

YANG, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**: 1586-1591.

ZEMP, N., R. TAVARES, A. MUYLE, D. CHARLESWORTH, G. A. MARAIS, and A. WIDMER, 2016 Evolution of sex-biased gene expression in a dioecious plant. Nat Plants **2**: 16168.