Hallmarks of early sex-chromosome evolution in the dioecious plant *Mercurialis annua* revealed by *de novo* genome assembly, genetic mapping and transcriptome analysis

Kate E. Ridout[1,2,3,*], Paris Veltsos[1,*], Aline Muyle[4], Olivier Emery[1,5], Pasi Rastas[6], Gabriel A.B. Marais[4¶], Dmitry A. Filatov[3¶], and John R. Pannell[1¶**]

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

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

[3] RDM Nuffield Division of Clinical Laboratory Sciences, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom

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

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

[6] Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom.

* contributed equally to the research.

¶ supervised the research.

** Author for correspondence: john.pannell@unil.ch

Key words: dioecy, gene expression, sex chromosomes, whole genome sequencing, sex linkage, Y-chromosome degeneration, next generation sequencing, Pacific Biosciences

Running title: Sex chromosome evolution in *Mercurialis annua*

Abstract

The evolution of sex chromosomes involves the suppression of recombination around a sex-determining locus, and the subsequent divergence in DNA sequence between the two homologous sex chromosomes. Dioecious plants offer the opportunity to study independent early stages of this process, because of multiple, recent transitions between hermaphroditism and dioecy. Here, we present data from *de novo* genome assembly and annotation, genetic mapping and transcriptome analysis of the diploid dioecious herb *Mercurialis annua*, revealing several of the typical hallmarks of early sex-chromosome evolution. Until now only a single sex-linked PCR marker has been published. Our analysis identified a single linkage group, LG10, as the likely sex chromosome, with a region containing 69 sex-linked transcripts with a clearly lower male than female recombination, high X/Y divergence and multiple incidences of premature stop codons on the Y allele. We found many genes with sex-biased expression. Female-biased genes were randomly distributed across the genome, but male-biased genes were slightly enriched on the Y chromosome. Interestingly, Y-linked genes had reduced expression compared with X-linked genes, a pattern consistent with Y chromosome degeneration. *M. annua* has been a powerful model for the study of rapid sexual-system transitions in plants; our results here establish it as a model for the study of the early stages of sex-chromosome evolution.

Introduction

The evolution of separate sexes, or dioecy, from hermaphroditism has occurred repeatedly in angiosperms, and about half of all angiosperm families have dioecious members (Renner and Ricklefs 1995; Renner 2014). Dioecy ensures outcrossing, it sets the stage for the possible evolution of sex chromosomes (Charlesworth and Charlesworth 1978), and it establishes the possibility of the evolution of sexual dimorphism, i.e., the expression of different male versus female phenotypes (Charlesworth 1999; Geber 1999; Moore and Pannell 2011; Barrett and Hough 2013). Indeed, the leading model for sex-chromosome evolution invokes genetic linkage between loci involved in sex determination (Charlesworth and Charlesworth 1978) and loci implicated in sexual dimorphism (Rice 1987; Charlesworth, et al. 2005; Charlesworth 2015).

Sex chromosomes have evolved in a wide range of organisms with separate sexes (Bachtrog, et al. 2014; Beukeboom and Perrin 2014), and share features consistent with a model that invokes dimorphism between males and females (Charlesworth and Charlesworth 2005; Charlesworth, et al. 2005; Bergero and Charlesworth 2009; Charlesworth and Mank 2010; Bachtrog, et al. 2011; Charlesworth 2013; Beukeboom and Perrin 2014). Briefly, once single-locus genetic sex determination evolves, the associated sex chromosomes begin diverging by accumulating structural changes and repetitive elements (Wang, Na, et al. 2012), genes with differential effects on the fitness of males and females (Gibson, et al. 2002), and through the differential fixation of deleterious mutations on the sex chromosome associated with the heterogametic sex (Y or W, in species with XY or ZW sex-determination, respectively) (Charlesworth, et al. 2005; Charlesworth 2013). Eventually, the Y (or W) chromosome degenerates because selection favours the suppression of recombination between the sex chromosomes (Charlesworth 1991; Charlesworth, et al. 2005; Bergero and Charlesworth

2009); suppressed recombination should be selected when mutations with sex-specific fitness benefits (i.e., sexually antagonistic mutations) occur on the sex chromosome that spends most of its time in individuals of the sex to which the benefit applies (Rice 1987). Once recombination has ceased in the heterogametic sex, the affected genomic regions experience evolutionary forces leading to genetic degeneration, such as background selection (Charlesworth, et al. 1993a), selective sweeps (Maynard Smith and Haigh 1974) or Muller's Ratchet (Charlesworth, et al. 1993b). Analyses of genetic divergence between sex chromosomes have revealed that recombination suppression may occur in discrete 'strata' along the sex chromosomes, with a stepwise increase in the size of the non-recombining region. Evidence for this process comes from analysis of both animals (Lahn and Page 1999; Nam and Ellegren 2008) and plants (Bergero, et al. 2007; Wang, Na, et al. 2012), and the pattern is consistent with the theoretical expectation that the non-recombining region will expand as it captures additional linked sexually antagonistic alleles (Charlesworth 2015).

Notwithstanding many common elements, sex chromosome divergence and degeneration vary in important ways among lineages (Mank 2013; Bachtrog, et al. 2014), and specifically among plants (reviewed by Ming, et al. 2007; Chibalina and Filatov 2011; Charlesworth 2013, 2015; Vyskot and Hobza 2015). For instance, in some species, sex chromosomes have diverged in size substantially, with the Y being much larger than the X and all other chromosomes – largely due to the accumulation of repetitive sequences, e.g., *Silene latifolia* (Cermak, et al. 2008) and *Rumex acetosa* (Steflova, et al. 2013) and *Coccinia grandis* (Sousa et al. 2013). Heteromorphic sex chromosomes are also associated with the fixation of deleterious mutations on the Y (or W) chromosome, with a pattern indicative of evolutionary strata, and the loss of genes (Bergero, Qui, et al. 2015). In other species, sex is determined by a polymorphism on sex chromosomes that have not diverged in size. In these species, the sex-determining region is probably small, and recombination is retained in 'pseudoautosomal regions' (PAR) that span almost the whole chromosome, e.g., *Asparagus officinalis* (Telgmann-Rauber, et al. 2007)*, Spinacia oleracea* (Yamamoto, et al. 2014), *Diospyros lotus* (Akagi, et al. 2014), and *Fragaria chiloensis* (Tennessen, et al. 2016). Similar differences between homomorphic and heteromorphic sex chromosomes have been found among animal lineages (Bachtrog, et al. 2011; Beukeboom and Perrin 2014). For example, while sex-chromosome degeneration is common in mammals, drosophila and ratite birds (Pigozzi 2011), many amphibians with genetic sex determination (Schmid and Steinlein 2001; Eggert 2004; Stock, et al. 2011) have homomorphic sex chromosomes and do not show evidence of sex-chromosome degeneration.

It is still unclear why some plants show high levels of sex-chromosome differentiation while others do not. Limited sex-chromosome degeneration may reflect a recent origin of dioecy (Ming, et al. 2007; Charlesworth 2013). However, other forces must be at work, because in some lineages sex-chromosome degeneration has occurred very rapidly (e.g., Zhou and Bachtrog 2012; Papadopulos, et al. 2015), while other lineages, such as the Salicaceae family (willows and poplars, in which dioecy is approximately 45 My old; (Manchester, et al. 2006)) and the *Phoenix* genus (date palms, with dioecy being 50 My old; (Couvreur, et al. 2011)), retain apparently homomorphic sex chromosomes that have not degenerated much (Charlesworth 2013). One hypothesis to explain the lack of degeneration in some plant sex chromosomes invokes purifying selection acting on haploid male gametophytes (i.e., the pollen grains and pollen tubes, Mascarenhas 1990) in which many genes are expressed, retarding any loss of function for critical Y-linked genes (Chibalina and Filatov 2011). However, this hypothesis does not explain why other plants have undergone substantial sex-chromosome degeneration, including gene loss (Bergero, Qui, et al. 2015), at a rate similar to that found in many animals (Papadopulos, et al. 2015).

Another poorly understood feature of sex-chromosome evolution is variation between lineages in dosage compensation, i.e. the degree of compensation for genetic degeneration of one sex chromosome, by increased gene expression from the other (Charlesworth 1998; Mank 2013). While dosage compensation is an important feature of gene expression in many animal lineages, it is not ubiquitous (Mank 2013), e.g., chromosome-wide dosage compensation has not been found in birds, though it may be gene-specific (Zimmer, et al. 2016). In the dioecious plant species *Silene latifolia*, RNAseq analysis suggest that Y-chromosome gene loss might be modest (Bergero and Charlesworth 2011; Chibalina and Filatov 2011), and dosage compensation in this species has been controversial, with two studies reporting at least partial evidence for it (Muyle, et al. 2012, Papadopulos et al 2015) while others did not (Chibalina and Filatov 2011; Bergero, Qui, et al. 2015). The emerging consensus, based on partial sequencing of the *S. latifolia* genome, is that the Y chromosome is in fact highly degenerate, with many genes lost or not expressed, and with associated partial dosage compensation from X-linked homologues, including some genes with full compensation (Papadopulos, et al. 2015).

Males and females of dioecious plants may also show secondary sexual dimorphism, i.e., differences in vegetative phenotypes between the sexes, though it is usually less striking than in animals (Lloyd and Webb 1977; Moore and Pannell 2011; Barrett and Hough 2013). Most dioecious plants show differences between males and females in a wide range of morphological (Eckhart 1999), life-history (Delph 1999) and physiological traits (Dawson and Geber 1999). Ultimately, such phenotypic differences between the sexes must be associated with differential gene expression. One of the few examples of differential gene expression in dioecious plants is provided by *Silene latifolia* (Zluvova, et al. 2010; Muyle, et al. 2012), which shows sexual dimorphism in numerous phenotypic traits (Meagher 1994; Delph and Meagher 1995; Delph and Bell 2008). Identifying differentially expressed genes between males and females not only indicates the extent of transcriptomic sexual dimorphism, but also allows us to ask whether sex chromosomes are enriched for sex-biased genes compared with other regions of the genome. Such enrichment could be a response to degeneration, and would thus be indicative of existing suppressed recombination (reviewed in Mank 2009; Parsch and Ellegren 2013).

Here, we report evidence of relatively mild degeneration of the Y chromosome, and sex-biased gene expression for approximately 5% of genes of the dioecious herb *Mercurialis annua* (Euphorbiacae), based on *de-novo* assembly, annotation and population genetic analysis of its genome and transcriptome. Although sex determination in *M. annua* has been studied for many decades (reviewed in Russell and Pannell 2015), our study represents the first attempt to understand the implications of dioecy at the genomic and transcriptomic level. Whole-genome data for species with separate sexes are scarce in plants. Exceptions include *Carica papaya* (Liu, et al. 2004), *Vitis* (Fechter, et al. 2012), *Diospyros lotus* (Akagi, et al. 2014), and *Populus* (Tuskan, et al. 2006), which has largely homomorphic sex chromosomes (Filatov 2015; Geraldes, et al. 2015); see also Papadopulos et al. (2015).

Mercurialis annua is likely to be a revealing model for the study of sex-chromosome evolution, because the sex chromosomes appear homomorphic (Durand 1963, and P. Veltsos, personal observation), and might be in the early stages of degeneration, potentially much earlier than, for example, *S. latifolia* or the well-studied *Rumex* species (Hough, et al. 2014).

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 recent work has shown it to have a simple XY system (Khadka, et al. 2005; Russell and Pannell 2015). Not only are there no signs of heteromorphism in the karyotypes between males and females (P. Veltsos, R. Hobza, B. Vyskot and J.R. Pannell, unpublished data), but crosses between males with 'leaky' gender expression have revealed that YY males are viable but partially sterile (Kuhn 1939, and P. Veltsos, G. Cossard and J.R. Pannell, personal observation), pointing to the likelihood that the Y chromosome is still largely intact, but to some extent degenerate. Our genomic analyses presented here confirm this view.

To sequence the genome of diploid *M. annua*, we combined short read sequencing on the Illumina platform with long read technology developed by Pacific Biosciences. We first describe important details of the genome and compare its content with the six other sequenced members of the order Malpighiales, as well as several other more distantly related plant species. Using transcript segregation analysis, we construct a genetic map, identify non-recombining and sex-linked genes and scaffolds and perform a comparative analysis with the non-sex-linked regions. We then investigate the sex chromosomes with regard to evolutionary strata and degeneration, including the degree of fixed deleterious mutations on Y-linked sequences, and numbers of deleted genes. Finally, we examine sex-biased gene expression with an emphasis on the potential for dosage compensation in *M. annua*.

Results

M. annua genome assembly

We generated ~57.8 Gb of DNA-seq data from a male individual (M1) of *M. annua*, corresponding to ~90x coverage of the 640 Mb genome (2n=16), using a combination of short-read Illumina and long-read Pacific Biosciences sequencing. After filtering, genome coverage dropped to ~74x (Table 1; Figure S1). *De novo* assembly and scaffolding gave a final assembly of 89% of the genome (78% without gaps), 65% of which was distributed in scaffolds > 1 kb, with an N_{50} of 12,808 across 74,927 scaffolds (Table S4). Assembly statistics were consistent with other members of the Malpigiales (Table S2), 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 transcripts; the majority of the unassembled sequence data is therefore expected to be repetitive. We estimated the completeness of the genome assembly using BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão et al. 2015). Out of 956 genes in a plant-specific database, 29.2% were completely recovered, and the remaining were either duplicated (6.6%), fragmented (25.2%), or missing from the assembly (39%).

Repeat masking identified simple tandem repeats in over 10% of the assembly; given that microsatellites are particularly hard to assemble, this fraction is likely to be underestimated. DNA transposon and retrotransposon masking using homology information characterised 15% of the assembly, with an additional 33% comprising of 1,472 predicted novel transposable elements. The most frequent transposable repeat types annotated in the genome were the Gypsy LTR, Copia LTR, and L1 LINE retrotransposons (Table S3), similar to findings in other plant genomes (Chan, et al. 2010; Sato, et al. 2011; Wang, Wang, et al. 2012; Rahman, et al. 2013). Across all data, over 58% of the ungapped *M. annua* assembly was found to be repetitive (Table S3), corresponding to 44% of the 640 Mb total predicted genome size. Thus, given that the assembly covers 78% of the genome and assuming the missing fraction is entirely made of repeats, up to 66% of the *M. annua* genome could be repetitive. High AT-rich repeat content has been reported for other plant species, as has a similar number of unclassified repetitive elements (e.g., Chan, et al. 2010). We estimate that the genome of *Mercurialis annua* is around 240 Mb larger than that of *Ricinus communis* (see Table S2), likely reflecting on-going transposon activity following lineage splits among species in Malpigiales.

Table 1: Raw data before and after filtering of genomic and transcript libraries.

Sample	Technology	Library	Insert Size (bp)	Read number	Basecount (gb)	
					No filtering	After filtering
M1 Male	Illumina	Mate Pair	5000	50599202	5.0	2.2
	Illumina	Paired End	250	364231376	36.4	33.1
	Illumina	Paired End	500	130681342	13.0	11.1
	PacBio	Long read	N/A	859697	1.9	0.8
G1 Female	Illumina	Paired End	250	401657040	40.2	40.2
G2 Female	Illumina	Paired End	250	341596022	34.2	34.2
Transcripts						
M1 Male	Illumina	Paired End	250	32805020	3.3	3.3

Gene content and genome annotation for *M. annua*

Genome annotation was carried out using a single male individual (parent M1) with 3.3 Gb of RNAseq reads. The transcriptome was assembled into 49,809 transcripts. AUGUSTUS gene prediction revealed 31,604 coding gene models (including alternative isoforms). Thus, approximately 63% of transcripts are predicted to be protein-coding. The remaining transcripts comprise non-coding RNA that does not contain an open reading frame.

We reduced the 31,604 protein-coding transcripts to 27,770 genes by merging putative isoforms. The degree of splicing found in *M. annua* is slightly lower than in other plant species, though the number of alternative isoforms detected in *M. annua* will likely increase as more transcripts are sequenced. For example, Syed *et al.* (2012) revealed that over 60% of *Arabidopsis* genes with more than one intron display alternative splicing.

By combining the transcript library with *de novo* gene predictions (See Supplementary Methods), we identified a total of 28,417 protein-coding gene models, of which 87% (24,800) could be annotated, most of which referenced to *R. communis*. Indeed, the top ten represented gene-ontology (GO) terms in the Biological Process category (the most represented category) are identical between *M. annua* and *R. communis*.

A separate annotation was carried out for the expression analysis; in total, we collected RNAseq data from 30 females and 35 males from multiple families (all individuals; Supplementary Table S1). Mapping pooled samples to the genome yielded a total of 68,990 predicted genes, with 125,524 transcript isoforms.

Putative sex-linked sequences in *M. annua*

Khadka et al. (2002) identified a 1,562 bp region tightly linked to the male-determining region, and Russell and Pannell (2015) confirmed that this PCR marker is found only in males across diploid populations of the species range, indicating its presence on a putative Y chromosome with tight linkage to the sex-determining locus. We extended this sex-linked marker region to 8,899 bp by mapping to the error-corrected Pacific Biosciences reads. More than 6 kb of this extended sex-linked region showed strong homology to a non-functional Gypsy repeat element, with the remainder mapping to a novel repeated retrotransposon currently identified only in *M. annua*. Recent or ongoing proliferation is supported both by the detection of highly similar transcripts to both repeats in the transcriptome, as well as their high copy number in the genome (our analysis detected 100,000 Gypsy repeats). We were unable to extend the sequence of this sex-linked marker region further, likely due to the prevalence of these repeats.

To identify additional sex-linked sequences in *M. annua*, we used RNAseq to trace SNP haplotypes segregating from our parental individuals (Male M1, females G1, G2; families described in Table S1) through to the F_1 (20 individuals) and F_2 generations (39 individuals). Using the software SEX-DETector (Muyle, et al. 2016) on each family separately, we identified a total of 527 (188 supported by both families) X-linked transcripts with Y-linked homologues, and a single female-specific transcript with no Y-linked homologue. The degree of divergence between X- and Y-linked homologues varied continuously among the 527 X/Y pairs (715 alleles in total) from SEX-DETector (data not shown). For the markers that could be mapped, the degree of non-synonymous divergence peaked at about the middle of the sex-linked region (Figure 1).

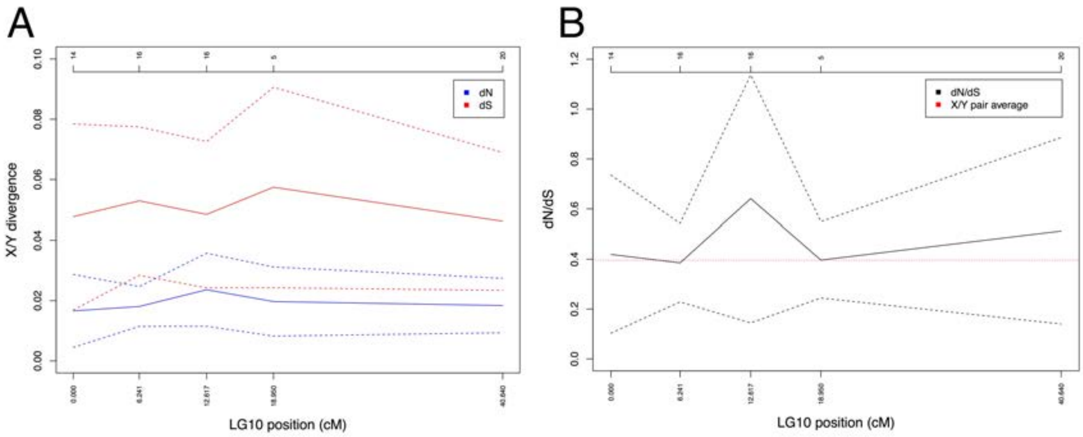

Figure 1. Average nucleotide divergence for each mapping position of the LG10 sex-linked markers. (A) Mean proportions of synonymous mutations at synonymous sites (dS) and non-synonymous mutations at non-synonymous sites (dN). (B) Ratio of dN/dS (average for all X/Y transcript pairs is shown as a red line). The ladder at the top of the panel indicates the number of sex-linked transcripts at each respective map position. For clearer presentation, one mapped location (6.343) containing one transcript was merged with the nearby location at 6.241 cM. Dotted lines indicate one standard deviation. The dN and dS values for two transcripts that mapped to two adjacent locations.

After aligning the 527 X/Y homologous pairs, we found 12 X-linked (2%) and 192 Y-linked loci (36%) containing stop codons, a difference that is highly significant (Fishers-exact test p $< 2.2e^{-16}$). Of the 12 X-linked pseudogenes, five were also pseudogenized on the Y chromosome. The remaining seven X-linked loci with pseudogenised alleles were also segregating for an alternative X-linked allele not containing premature stop codons.

Sex-linked and autosomal transcripts were mapped to the genome assembly using reciprocal best BLAST. The resulting genomic contigs were divided into four bins: X-linked contigs, XY-linked contigs (co-assembled X/Y genomic contigs and scaffolds; see Materials and Methods), Y-linked contigs and autosomal contigs; these mapped to 494, 218, 61, and 6302 transcripts from the segregation analyses, respectively. Details of this analysis are given in Table 2. Briefly, the X-linked bin comprised 706 genomic contigs containing 1,825 transcripts from the full genome annotation and a total of 9 Mb of sequence; the Y-linked bin comprised of 68 genomic contigs containing 105 transcripts and a total of 474 Kb sequence; the XY-linked bin (probably representing chimeric genome assembly in regions of low divergence) comprised 82 contigs containing 431 transcripts and a total of 1.6 Mb; and the autosomal bin comprised 8,858 genomic contigs containing 25,393 transcripts and 97 Mb sequence data.

Table 2. Summary statistics of the sex-linked bins with regard to the whole genome. Only genes supported by transcript data were used for sex-linked analyses (*de novo* predicted genes with no transcript support were excluded). Confidence intervals are one standard deviation.

	Whole genome	X-linked	XY-linked	Y-linked	Autosomal
Number of contig	720537	706	82	68	8858
Total base pairs	546375413	8900651	1629595	474637	97386675
% GC	34.7	34.3	34.5	35.8	34.3
Average contig length	15330 (±20028)	12607 (±21189)	19873 (±23575)	6979 (±16089)	10994 (±19097)
Tandem repeat density (%)	6.50	4.23	4.29	4.61	3.87
TE repeat density (%)	45.0	46.9	46.7	77.7	45.4
Coding gene number	27770	1350	321	77	18844
Average coding gene length	1575 (±1124)	1375 (±993)	1232 (±733)	1089 (±610)	1292 (±889)
Coding gene density	6.57%	9.09%	12.21%	4.97%	11.47%
Effective number of codons	51.8	52.1	52.5	51.8	52.0
Number of SNPs	200410	7779	2050	936	125490
Number of genes with SNPs	16315 (59%)	637 (47%)	160 (50%)	48 (62%)	8761 (46%)
Median nucleotide diversity (π) / kb	0.00212 (±0.00289)	0.00205 (±0.00262)	0.00229 (±0.00255)	0.00386 (±0.00329)	0.00292 (±0.00942)
Synonymous π (π_S)	0.00120	0.00116	0.00088	0.00186	0.00287
Non-synonymous π (π_N)	0.00022	0.00025	0.00025	0.00029	0.00041
Non coding transcript number	14349	475	110	28	6549
Average non coding transcript length	344 (±215)	517 (±400)	459 (±281)	421 (±190)	510 (±486)
Non coding transcript density	1.26%	1.97%	2.64%	1.98%	2.64

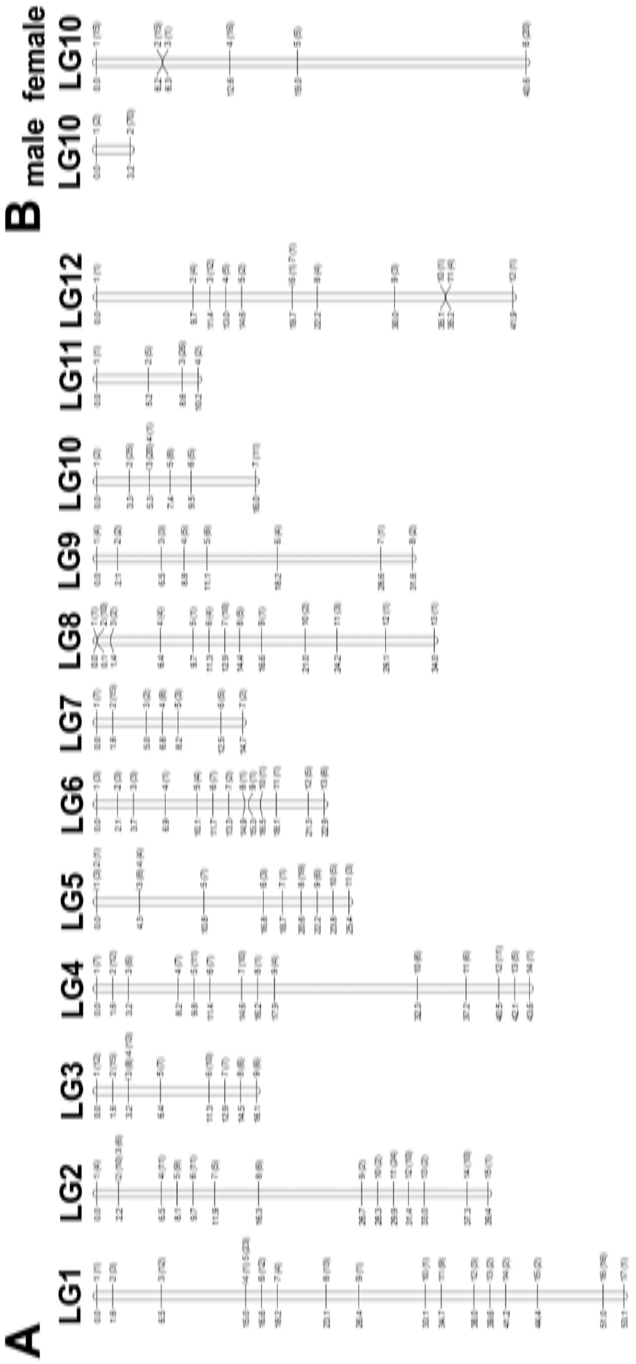

Figure 2. (A) Genetic map, averaged across males and females, showing the 12 largest linkage groups. (B) Sex-specific genetic maps for the inferred sex chromosome (LG10). Each uniquely mapped location is named sequentially, and the associated number of markers is indicated in parentheses. The male and female maps contain the same number of markers, and there is clearly more limited recombination in males. All marker names and their associated positions are available in the Supplementary Information.

Genetic map for diploid *M. annua*

Of 2,968 transcripts with 9,858 acceptable SNPs, 1,278 transcripts (4,551 SNPs) were mapped to 236 linkage groups (LGs). The largest twelve LGs included 678 transcripts, accounting for a total of 2,228 SNPs (Figure 2A). The family sizes were too small to clearly recover the expected eight haploid chromosomes of *M. annua* (Durand 1963), though it is likely they are represented amongst the twelve largest LGs. LG10 is entirely composed of sex linked transcripts, and the male recombination map showed no recombination (Figure 2B), clearly indicating it is part of the sex chromosome. Figure 3A decomposes sex linkage assignment by family for the transcripts mapped to LG10, with the centre of the map having the highest support for sex linkage (support consistent between both families). Figure 3B indicates the number of mapped transcripts to LG10 with premature stop codons on the Y variant, and also indicates that most degeneration is localised in the centre of the linkage map. However there are no significant differences in the proportion of premature stop codons across the length of LG10 (Figure 3B).

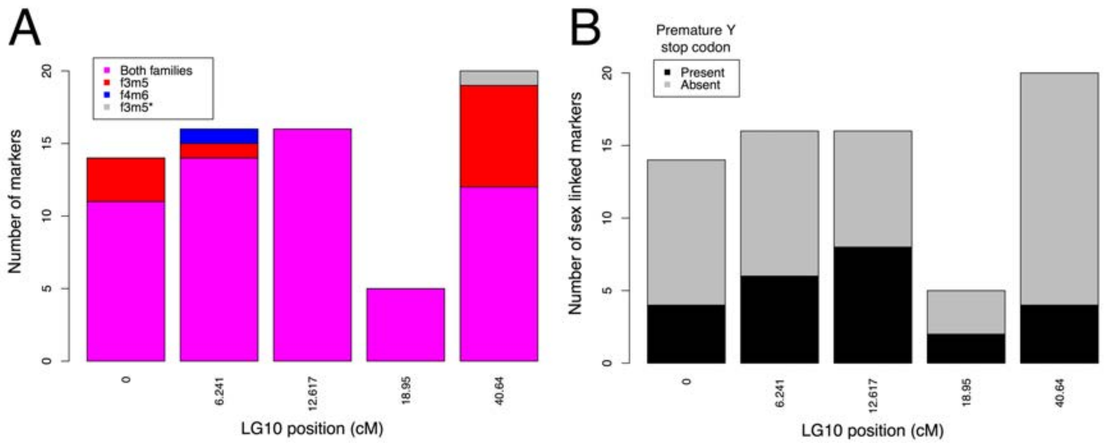

Figure 3. (A) Number of markers at different mapping positions (based on the female map), with the source of support in SEX-DETector indicated: the legend indicates the family supporting sex linkage, a star next to the family name indicates a transcript that was considered autosomal based on the other family. (B) Number of transcripts at each mapping position on LG10 with premature stop codons in the Y allele.

Variation in gene density and length

We examined genomic contigs that had been separated into either the autosomal or one of the sex-linked bins. Contigs from the Y-linked sequence bin were found to be significantly less rich in protein-coding genes those than from the autosomal bin (P < 0.001, Figure 4A). There was also an abundance of short genomic Y-linked contigs (and to a lesser extent, X-linked and autosomal contigs) that are entirely coding, probably because coding sequences are less complex to assemble.

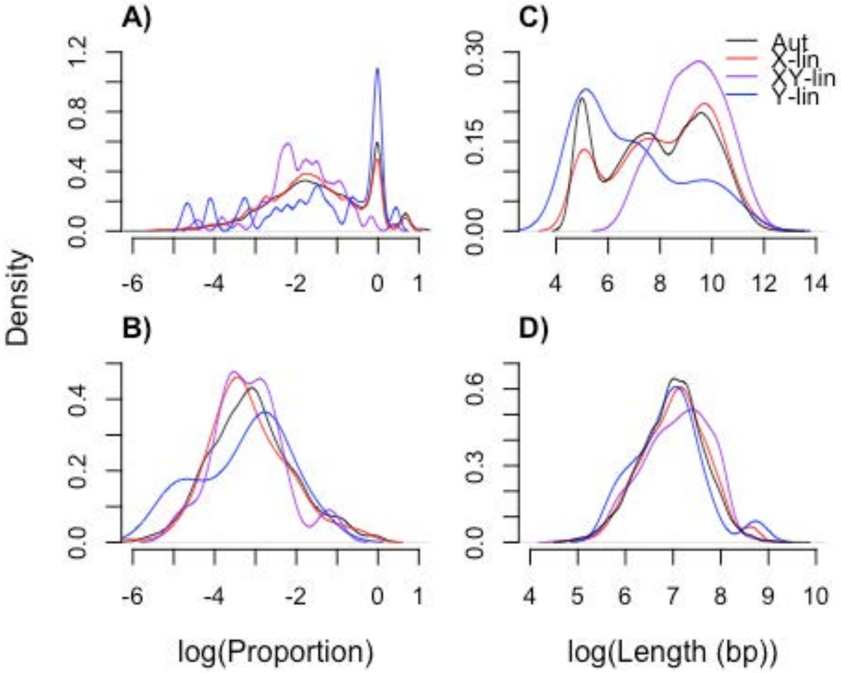

Figure 4. (A) Proportion of coding gene density, (B) Proportion of non-coding gene density, (C) Scaffold lengths, and (D) Gene lengths. Density was calculated using a Gaussian kernel, such that the area under each curve sums to 1.

Next, we investigated genes found within these bins. For non-protein-coding genes there was a slightly, though not quite significantly, greater proportion found in the Y-linked bin, when compared to the autosomal bin (P = 0.06; Figure 4B). X-linked and XY-linked protein-coding transcripts were found to be significantly longer than expected by chance (P = 0.005 and P = 0.004 respectively, Wilcoxon test; Figure 4D), when compared to the autosomal length distribution, whereas Y-linked protein-coding transcripts were somewhat shorter than expected, albeit not significantly ($P = 0.18$). Non-coding transcripts (genes with no open reading frame) were also significantly longer in X-linked contigs, when compared to autosomal contigs ($P = 0.02$).

Variation in nucleotide diversity and codon usage

We investigated nucleotide diversity (π) and the π_N/π_S ratio (ω) across transcripts mapping to the X-linked, Y-linked, XY-linked and autosomal contig bins, based on sequences from the six individuals from across the species range (individuals M1, M2, M3, G1, G2 and G3). The distribution of π/kb did not differ significantly between the Y-linked bin and the autosomal bin (Table 2; Figure S2). We further calculated synonymous (d_S) and non-synonymous (d_N) divergence across all X/Y gene pairs, as well as between *M. annua* and its dioecious sister species *M. huetti*, and between *M. annua* and its monoecious distant relative *R. communis* for autosomal and sex-linked genes (Table 3, Figure 5). For X/Y pairs (without in-frame stop codons), mean $d_N/d_S = 0.396$ (Figure 1, Table 3). For autosomal genes, $d_N/d_S = 0.161$ between the two *Mercurialis* species (*M. annua, M. huetii*) and 0.200 between *M. annua* and *R. communis* (Table 3, Figure 5). d_S was lower between X/Y gene pairs in *M. annua* than between orthologous autosomal genes in *M. annua* and *M. huetii* (Figure 5). Codon usage in *M. annua* did not differ significantly between X- and Y-linked genes and autosomal genes (Nc =52.1, Nc = 51.8, Nc = 52, respectively; Figure S3).

Table 3. Average counts of d_S (synonymous) and d_N (non-synonymous) mutations in coding genes from X/Y pairs identified using the SEX-DETector analysis, from SEX-DETector X-linked genes against *M.huetti* autosomes, and from alignments of 2,908 autosomal protein coding genes found in *M.annua, M.huetti,* and *R.communis*. Predicted coding genes only were used, excluding X/Y pairs containing stop codons.

	d_S	d_N	d_N/d_S
X/Y pairs	0.048	0.019	0.396
X / *M. huetti*	0.159	0.032	0.369
M.annua / M. huetti	0.118	0.019	0.161
M.annua / R. communis	0.520	0.104	0.200

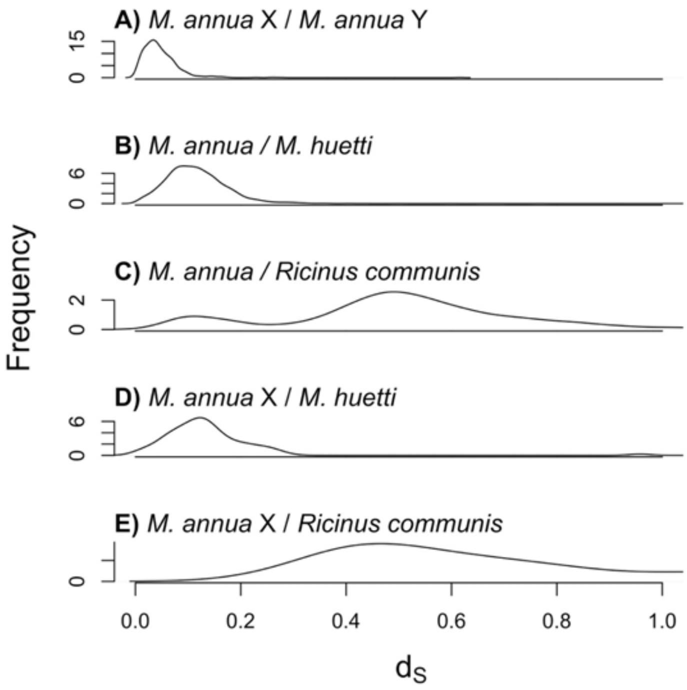

Figure 5. Density plots of synonymous substitutions at synonymous sites (d_S) in: all *M. annua* X/Y pairs, all *M. annua / M. huetti* one-to-one orthologous gene pairs, all *M. annua / Ricinus communis* one-to-one orthologous gene pairs, sex-linked *M. annua / M. huetti* one-to-one orthologous gene pairs, and sex-linked *M. annua / R. communis* one-to-one orthologous gene pairs.

Differences in gene expression between males and females

We examined patterns of gene expression using RNAseq data. Males showed higher gene expression than females over all genes (Figure S4). On the basis of inferred sex-linkage, we found that Y-linked genes were significantly less strongly expressed than X-linked genes (Wilcoxon test p-value = 1.799e-06; median Y/X expression ratio 0.9386; Figure 6). Significance was maintained even after the removal of Y-linked genes containing stop codons (p-value = 0.002; median Y/X expression ratio 0.952). Genes with female-biased expression in *M. annua* did not map preferentially to the sex-linked contigs, but genes with male-biased expression were significantly underrepresented on X-linked contigs (Table 4; P<0.01; Fisher's exact test), and were slightly (but not significantly) enriched on Y-linked contigs (P=0.08). A total of 10 male-biased genes mapped onto Y-linked contigs (whereas only three female-biased genes did).

Table 4. Numbers of genes in males and females that map to sex-linked contigs. "Unbiased genes" were not significantly differentially expressed between males and females. "Female/Male only" genes were expressed only in the given sex and not the other. "Female/Male overexp" are significantly overexpressed in the given sex relative to the other.

	Unbiased	Female only	Female-biased	Male only	Male-biased
Total transcripts	68990	35	440	436	2473
Longest transcript	26197	806	5091	1296	15001
Shortest transcript	25	51	58	51	62
N50	1429	129	1021	229	1252
Total bases	50721696	3742	237112	1252	1599388
Transcripts mapping to the genome	68054	35	438	433	2469
Transcripts mapping to the X-linked bin	2295 (3.4%)	1 (2.9%)	13 (3.0%)	12 (2.8%)	60 (2.4%)
Transcripts mapping to the XY-linked bin	433 (0.6%)	0	1 (0.2%)	4 (0.9%)	12 (0.49%)
Transcripts mapping to the Y-linked bin	131 (0.2%)	0	3 (0.7%)	3 (0.7%)	7 (0.28%)
Transcripts mapping to the autosomal bin	16165 (23.8%)	7 (20%)	46 (10.5%)	5 (1.2%)	275 (11.1%)

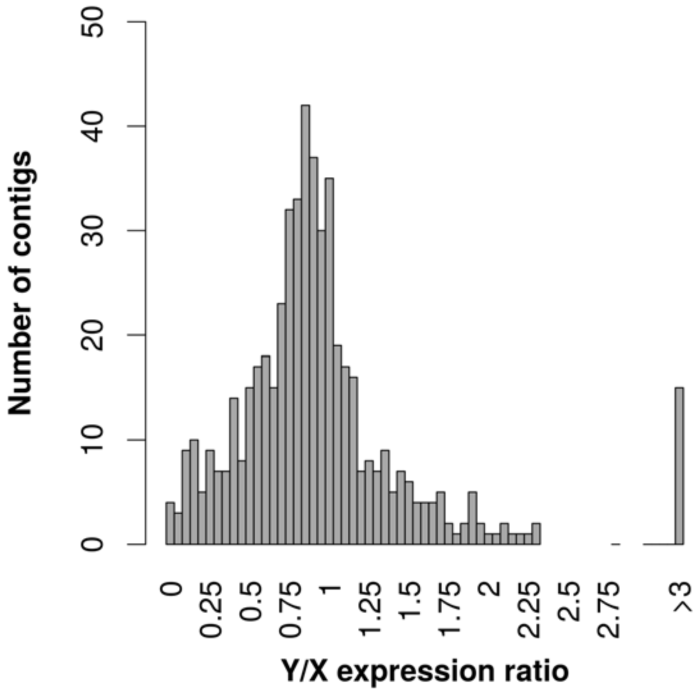

Figure 6. Distribution of Y over X mean expression levels in *M. annua* males, computed on 528 inferred X/Y transcripts. The median of the Y/X ratio is 0.9386.

Sex-biased (protein-coding) genes were significantly shorter across the entire transcript length than the unbiased genes: average gene length for unbiased and male-biased genes was 735 bp and 581 bp, respectively ($P < 2.2e^{-16}$); average female-biased gene length was 507 bp ($P = 5.196e^{-12}$). After excluding genes present in one sex only, the average gene lengths were 647 bp and 539 bp for males and females respectively, and both genes sets are still significantly shorter than unbiased genes ($P = 3.266\ e^{-07}$ and $P = 1.858\ e^{-08}$). Gene fragments expressed in only one sex were extremely short, with average gene lengths of 212 bp and 107 bp in male-only- and female-only-expressed genes, respectively.

Discussion

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

Our study provides the first draft assembly and annotation of the diploid dioecious species *Mercurialis annua* (2n = 16), a species that has proven to be a revealing model for the study of sexual-system transitions in plants (reviewed in Pannell, et al. 2008). The draft assembly is based on a post-filtering sequencing coverage of 74x and constitutes approximately 89% of the estimated 640 Mb total genome of diploid *M. annua*. Moreover, the majority of the 956 plant-specific BUSCO genes were not missing from the assembly (39% missing) (Simão et al. 2015), suggesting that it is reasonably complete. We estimate that up to 2/3 of the *M. annua* genome comprises repetitive sequences, mostly Gypsy LTR, Copia LTR, and L1 LINE retrotransposons, a finding in line with other plant genomes (Chan, et al. 2010; Sato, et al. 2011; Wang, Wang, et al. 2012; Rahman, et al. 2013).

RNAseq analysis suggests that 63% (>27,000) of *M. annua* transcripts are protein-coding, with almost 25,000 annotated gene models in common with the related *R. communis* draft genome; the remainder likely comprises non-coding RNA, expressed pseudogenes, and gene fragments. Gene content is thus comparable with that of other diploid species such as *Arabidopsis thaliana* (>27,000 genes)*,* or *R. communis* (>31,000 genes) rather than species that show evidence of past polyploidization events, such as diploid *Gossypium (Wang, Wang, et al. 2012, >40,000 genes).*

We identified a large linkage group (LG10) composed entirely of sex-linked transcripts. The region spanned approximately 40 cM in the female recombination map and only 3.2 cM in the male map (Figure 2b), clearly indicating limited recombination in males. Our map is fragmented into more than the eight linkage groups expected diploid *M. annua* (2n = 16;

Durand 1963). Analysis of larger families will be required to collapse different LGs that correspond to chromosones , as well as to identify a possible PAR on the sex chromosome, which recombines in both sexes. Nevertheless, the sex-linked region in our map contains a localised peak in both the dN/dS ratio (approximately 0.641 compared to average for XY comparisons of 0.396) and in the proportion of alleles on the Y chromosome with a premature stop codon (50% in the peak compared to an average of 21% for all sex-linked mapped transcripts; data not shown, see Figures 1 and 3). These patterns are suggestive of a history of suppressed recombination and relaxed selection on the Y chromosome.

The draft genome of *M. annua* will be useful for comparative genomic analysis in the Malpigiales, which is poorly sampled for fully sequenced species. It also provides a valuable resource for on-going study of sexual-system and sequence evolution in the *M. annua* species complex. For example, we are currently employing the *M. annua* draft genome to test hypotheses concerning sequence evolution during and following species range expansions, using exon capture (S. Gonzalez-Martinez, C. Roux and J.R. Pannell, unpublished data).

Sex-linked genes in *M. annua*

Our segregation analyses identified hundreds of new potentially sex-linked genes. Substantial divergence between the X- and Y-linked gametologs implies complete sex-linkage for many of these genes. We were able to map 69 of these genes into a single linkage group in which there was no evidence for recombination in males; this region is likely to contain the sex-determining gene. Interestingly, compared to all sex linked transcripts, these genes show greater dN/dS ratios, as well as a greater chance of containing a premature stop codon, though, the differences were not statistically significant. It remains to be seen how many of these genes are consistently found in a non-recombining region of the Y chromosome in further crosses for more families; future work based on sampled across the species range, will be able to address this question. Nevertheless, our analysis indicates that the non-recombining region of diploid *M. annua* is likely to be small, and it establishes a large number of candidate sex-linked sequences that may be useful in further investigations, especially those in the X-linked and Y-linked bins.

To date, only a single Y-linked marker has been characterised, notably a 1,562 bp SCAR marker (Khadka, et al. 2002). Russell and Pannell (2015) confirmed the marker to be found only in males across diploid populations of the species range, indicating its presence on a putative Y chromosome with tight linkage to the sex-determining locus. In an approximately 9 kb extension of this region, we found that most of the surrounding sequence was repetitive, with transposon affinities. The repetitive nature of the Y-linked SCAR marker in *M. annua* is consistent with the observation of highly repetitive sequence in other plant Y chromosomes (Charlesworth 2016), but has hitherto rendered it relatively useless for comparative cytogenetic or genomic analysis across the genus (unpublished data).

Limited gene loss and pseudogenation of Y-linked genes in *M. annua*

Overall, our results suggest that the Y chromosome of diploid *M. annua* is young compared to several other plants studied to date. Only one of the 528 X-linked genes (0.2%) did not have a Y-linked homologue, suggesting a low level of gene loss from the *M. annua* Y chromosome. This is perhaps not surprising if most identified sex-linked genes are in fact in the (recombining) pseudo-autosomal region of the Y chromosome. The missing transcript from the Y chromosome has either been deleted (Bergero, Qui, et al. 2015), or it was simply not expressed when RNA was sampled. Either way, gene loss on the *M. annua* Y chromosome appears to be much lower than in other plant species. For instance, recent RNAseq studies have found that up to 28% and 14.5% of genes on the Y-chromosome have probably been lost in *Rumex hastatulus* (Hough, et al. 2014) and *S. latifilia* (Bergero and Charlesworth 2011; Chibalina and Filatov 2011; Bergero, Qiu, et al. 2015), respectively. Studies of *S. latifolia* using BACs suggest a rate of Y-linked gene loss of 30% (Blavet, et al. 2015), and analysis based on a partially sequenced genome points to the loss of expression of as many as 45% of Y-linked genes and pseudogenisation through premature stop codons of 23% (Papadopulos, et al. 2015).

In contrast to the comparatively low proportion of genes lost from the Y chromosome of *M. annua*, pseudogenisation of its Y alleles (approximately 36%) was similar to that found in other species, such as *R. hastatulus* (28%, Hough, et al. 2014) and *S. latifilia* (14.5%-30%, Bergero, Qiu, et al. 2015; Blavet, et al. 2015). This pattern points to an earlier disruption of gene expression of Y-linked genes compared to their pseudogenisation, as found in both *S. latifolia* (Papadopulos et al 2015) and *Drosophila albomicans* (Zhou and Bachtrog 2012). In this context, it is interesting that *M. annua* YY males show signs of sterility but are otherwise viable (Kuhn 1939), suggesting Y-chromosome degeneration is indeed at an early stage, but has perhaps begun to lose the function of genes important for male fertility.

Lack of evidence for evolutionary strata on the *M. annua* Y chromosome

The distribution of divergence between X- and Y-linked transcripts on the genetic map of the putative sex chromosome (LG10; Figure 1) provides no evidence for the evolution of strata that might point to discrete expansions of a region of suppressed recombination (Lahn and Page 1999; Bergero, et al. 2007; Nam and Ellegren 2008; Wang, Na, et al. 2012). The mapping position at 12.617 cM is entirely composed of sex-linked markers supported by data from both mapping families (Figure 3a), contains the highest proportion of Y-linked sequences with premature stop codons (50%, Figure 3b), and has a peak of X/Y divergence, which is higher than the average of all sex linked transcrtips (Figure 1). It might therefore correspond to the sex-determining region itself. If so, then the sex-determining region would appear to be small and young. Measures of diversity for sequences on a higher density map and based on larger mapping families would potentially allow one to locate regions of lower and higher X/Y divergence, as was the case for *Silene latifolia* (Papadopulos et al 2015). At present, however, only *Silene latifolia* (Filatov 2005; Bergero, et al. 2013), *Carica papaya* (Wang, Na, et al. 2012) and *Rumex hastatulus* (Hough, et al. 2014) show any evidence for evolutionary strata in plants.

Nucleotide sequence variation in sex-linked sequences of *M. annua*

Our results from the SEX-DETector (Muyle, et al. 2016) analysis provide some evidence for relaxed selection on Y-linked sequences compared to X-linked or autosomal sequences, with a higher d_N/d_S ratio for X/Y pairs of genes without in-frame stop codons than for autosomal *M.annua/M. huetti and M. annua/R. communis* orthologues. This observation suggests that purifying selection in X- and Y-linked genes is more relaxed than between the orthologous gene pairs that are not sex-linked, perhaps reflecting weaker purifying selection on the Y chromosome. Nevertheless, we surprisingly found no evidence for a difference in the level of absolute nucleotide diversity between Y-linked and autosomal genes of *M. annua*. Decreased nucleotide diversity in non-recombining sex-linked regions has been reported for a number of species, such as *S. latifolia* (Filatov, et al. 2001; Qiu, et al. 2010) and humans (Hellborg and Ellegren 2004), and is likely due to the smaller effective population size for Y chromosomes and the additional effects on a non-recombining region of background selection (Charlesworth, et al. 1993a), selective sweeps (Maynard Smith and Haigh 1974) or Muller's Ratchet (Charlesworth, et al. 1993b), which all reduce genetic diversity. The lack of substantially reduced absolute nucleotide diversity on the Y chromosome of *M. annua* suggests that it has not been subject to greater effects of drift than other regions of the genome, perhaps because the majority of Y-linked genes have been recently recombining.

Nor did we find evidence for different levels of codon bias between the Y-linked and other sequences. Codon usage bias is expected to be lower for non-recombining regions of the genome in which purifying selection should be weaker (Hill and Robertson 1966). Recent investigations into *Rumex hastatulus* Y-linked genes revealed a shift towards less preferred codon usage, increasing in severity with time since the putative cessation of recombination between X and Y chromosomes (Hough, et al. 2014). This is thought to be a reflection of either rapid sequence evolution, or degeneration of the genes. Here, we do not see this reduction, perhaps again reflecting the possibility that the Y-chromosome of *M.annua* is still in the early stages of degeneration, or that codon usage is not under selection in *M annua*.

Synonymous site divergence and the age of *M. annua* sex chromosomes

We found that synonymous nucleotide site divergence between X- and Y-linked pairs of genes was lower than between orthologous autosomal genes in *M. annua* and its sister species *M. huetii*, which is also dioecious. This suggests that much of the non-recombining region of the *M. annua* Y chromosome stopped recombining with the sex-determining locus more recently than the *M. annua-M. huetii* split. It is possible that the apparent youth of the *M. annua* sex chromosomes has been maintained by some degree of sex-chromosome turnover or by rare cases of Y/Y recombination, as is seen in frogs (Perrin 2009; Stock, et al. 2011; Blaser, et al. 2014), or X-Y gene conversion, as seen in mammals (Iwase, et al. 2010). It is also possible that the Y chromosome has been somewhat protected from degeneration due to gene expression in male gametophytes (Chibalina and Filatov 2011; but see Bergero, Qiu, et al. 2015; Papadopulos, et al. 2015). If this were the case, we might expect to see more genes with male-biased expression on Y-linked scaffolds, and to some extent we do (see below), but it is not known whether these genes are expressed in the male gametophytes. A final possibility is that dioecy was lost and re-evolved in one or both of *M. annua* and *M. huetii* since their divergence from one another, with different sex-determining loci, or that both species evolved dioecy in parallel. This last possibility is unlikely, because the perennial species of *Mercurialis* from which *M. annua* and *M. huetii* evolved was almost certainly dioecious too (Krahenbuhl, et al. 2002; Obbard, Harris, Buggs, et al. 2006). However, shifts from dioecy to monoecy and back again may be more likely than previously thought (Kafer, et al. 2017), and this might apply especially in the annual clade of *Mercurialis*, where such transitions have occurred recently in its polyploid populations (Pannell, et al. 2008).

Because there is no fossil-calibrated molecular clock for *Mercurialis*, we estimated divergence times in *Mercurialis* by applying mutation rates inferred for *Arabidopsis (Koch, et al. 2000),* i.e., 1.4×10^{-8} to 2.2×10^{-8} substitutions per synonymous site per year. Given an average synonymous-site divergence between pairs of X- and Y-linked genes of 0.048 synonymous substitutions per synonymous site, we infer that recombination between the X- and Y-linked genes of *M. annua* may have ceased between 1.1 and 1.7 million years ago. If accurate, this age would be substantially younger than estimated for *Silene latifolia* sex chromosomes (between 5 and 10 million years; (Nicolas, et al. 2005)) and is of the same order of magnitude as *Fragaria* species that have evolved separate sexes (2 million years; (Njuguna, et al. 2013)). Many of the sex-linked genes identified in this study may in fact be on the pseudoautosomal region instead of the non-recombining region of the Y chromosome; our estimate may therefore be too young and will need to be verified with reference to divergence estimates within the sex-determining region itself.

Sex-biased gene expression in *M. annua*

Dioecy is well established in the genus *M. annua (Krahenbuhl, et al. 2002; Obbard, Harris, Buggs, et al. 2006)*, and males and females have diverged substantially in their phenotypes, i.e., they show sexual dimorphism (Hesse and Pannell 2011; Labouche and Pannell 2016). Our study indicates that dimorphism in *M. annua* is substantial at the level of gene expression, too. Although female-biased genes were randomly distributed among the different genome compartments, we find some evidence for an enrichment of genes with male-biased expression in the non-recombining parts of the Y chromosome. The candidate sex determining and other fertility genes are likely to be some of these genes, which provide a useful list for further study. The significant decrease in Y expression with regard to X found in males is particularly intriguing, as it suggests that there is scope for dosage compensation of the male X-linked copy, despite the presence of a Y copy (which might have degenerated). Y-linked genes were significantly less strongly expressed than X-linked genes, perhaps consistent with a certain degree of degeneration of Y-linked gene expression. Male-biased genes were significantly underrepresented on X-linked compared to Y-linked contigs; these genes are candidates for future investigations into sex determination and sexual antagonism.

Concluding remarks

The genome of the dioecious plant *Mercurialis annua* shows evidence for several hallmarks of the early stages of Y-chromosome degeneration. While the karyotypes of male and female individuals are indistinguishable (P. Veltsos, R. Hobza, B- Vyskot, and J.R. Pannell, unpublished data) and very few genes are entirely missing from the Y, there has been some gene loss from the Y through pseudogenisation. Our analysis has identified LG10 as the likely sex chromosome for diploid *M. annua.* More detailed genetic mapping is required, but it would seem that the sex-determining region of *M. annua* is small. Nevertheless, our analysis indicates that sex-linked transcripts harbour a greater number of amino-acid-changing mutations than other parts of the genome, pointing to the potential relaxation of purifying selection that might be associated with suppressed recombination. Moreover, a number of non-functional Y-linked genes that have apparently functional X-linked homologues are still being expressed, and total Y-linked expression is significantly reduced in comparison to the X.

Mercurialis annua shows outstanding variation in its sexual system and has become a valuable model for testing hypotheses for transitions between combined and separate sexes and for sex-ratio and sex-allocation theory (Pannell 1997; Obbard, Harris and Pannell 2006; Pannell, et al. 2014). The availability of an annotated genome of *M. annua* will be useful in understanding potential further links between sex determination and sexual dimorphism. Diploid *M. annua* appears to display an interesting combination of homomorphic sex chromosomes, moderate Y-chromosome degeneration, and substantial divergence in gene expression, physiology and morphology. It will now be interesting to trace sex determination and the sex chromosomes from the diploid lineage studied here into the related polyploid lineages that appear to have lost and then regained dioecy – with very similar morphological differences between males and females but perhaps with sex determined at different loci.

Methods

Genome and transcriptome sequencing and assembly

All samples sequenced were from diploid *Mercurialis annua* individuals sampled in north-western France. Plants were grown together in the glasshouse. Genomic samples were taken from a single male individual, M1. RNAseq samples were collected from this individual plus three females, G1, G2 and G3, and two males, M2 and M3, all of which were unrelated. F1 and F2 progeny were then produced by crossing G1xM1 and G2xM1 (Supplementary Table S1), which were also used for RNA extraction and transcriptome sequencing.

Genomic libraries were prepared from leaf tissue using the Qiagen Plant DNA kit for the male M1. Illumina paired-end and mate-pair sequencing was carried out by the Beijing Genomics Institute (BGI) using Illumina HiSeq 2000 technology (100bp reads). Pacific Biosciences long-read sequencing was performed on individual M1 by the Centre for Integrative Genomics hosted by the University of Lausanne (Table 1).

RNA was extracted from a mixture of flower buds and leaf tissues using Qiagen plant RNAeasy kit for a total of 30 females and 35 males. Individual libraries were prepared for all 65 individuals (Supplementary Table S1) and sequenced on three lanes of Illumina HiSeq 2000 at the Wellcome Trust Centre for Human Genetics.

Genomic read filtering was performed as follows: Sliding window trimming and adaptor removal was carried out using Trimmomatic v. 0.30 with default parameters (Lohse et al., 2012). 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 default parameters. Reads containing the character 'N' were then 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 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, fill gaps, and join scaffolds using PBJelly2 (English, et al. 2012). Additional scaffolding was carried out using SSPACE (Boetzer, et al. 2011), which revisits gaps using existing paired-end and mate-paired sequences, using default parameters. Finally, L_RNA_Scaffolder (Xue, et al. 2013) was used to bridge genomic scaffolds using the transcript assembly. Transcripts from the male M1 were assembled with Trinity (Grabherr, et al. 2011) assembler using default parameters.

Genome annotation

Transposable elements (TEs) and tandem repeats were predicted using a combination of Tandem Repeats Finder (TRF, Benson 1999), RepeatModeler and RepeatMasker (http://www.repeatmasker.org). Repeats libraries from *Mercurialis, Euphorbiacae* and *Vitis vinifera* were used for the masking. The masked genome was used for further analyses.

Assembled transcripts, annotated *Euphorbiacae* proteins (NCBI GeneBank), and the CEGMA (Core Eukaryotic Genes Mapping Approach, Parra, et al. 2007b) protein set were given to the gene predictor MAKER2 (Holt and Yandell 2011), following the GMOD (Generic Model Organism Database; http://gmod.org/wiki/MAKER) manual. The resultant genes were filtered for correct start site and frame, and used to train the *ab initio* gene predictors AUGUSTUS (Stanke, et al. 2006) and SNAP (Semi-HMM-based Nucleic Acid Parser, Korf 2004). A second run of MAKER2 combined the training genes, *Euphorbiacae* proteins, and the AUGUSTUS and SNAP *de novo* gene predictions.

AUGUSTUS was used to predict complete gene models from the assembled transcripts. Predicted genes were then mapped onto the genome using GMAP (Wu and Watanabe 2005), and the transcript genes and MAKER2 genes were combined to form the final annotation (Supplementary Methods). For sex-chromosome analysis, only genes from the transcript library and annotated with AUGUSTUS were used, to reduce the possibility of false positives.

All proteins were compared to the GenBank *Rosids* database using BLASTX with an e-value threshold of 1e^{-10}. All BLAST results were passed to BLAST2GO (Conesa, et al. 2005), where Gene Ontology (GO) terms (Ashburner, et al. 2000) and InterPro (Hunter, et al. 2009) annotations were assigned using the default settings.

Segregation analysis

To identify sex-linked genes, we employed a probabilistic model based on Bayesian inference and implemented in the software SEX-DETector (Muyle, et al. 2016). 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 rate of false positives, than those methods implemented 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 et al., 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) and including PFAM domain searches as ORF retention criteria; this was considered our reference transcriptome. The RNAseq reads from the parents and progeny were mapped onto the reference transcriptome using BWA (Li and Durbin 2009). The alignments we obtained were analysed 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).

We also followed an RNAseq-based segregation analysis approach (Bergero and Charlesworth 2011; Chibalina and Filatov 2011). The basic principle makes use of the fact that X-linked haplotypes that are passed only from fathers to daughters and not to sons, and so indicate X-linkage, while transmission from father to sons (and not to daughters) indicates Y-linkage. This principle was successfully used for sex-chromosome analysis of *Silene latifolia* (Bergero and Charlesworth 2011; Chibalina and Filatov 2011) and *Rumex hastatulus* (Hough, et al. 2014). We examined sequence variation among a total of 29 females and 33 males. These were produced by crosses performed between the male M1 and females G1 and G2, resulting in 10 female and 10 male F_1 progeny. F_2 individuals were then produced by crossing two F1 females (f3 and f4) and two F_1 males (m5, m6), yielding a total of 12 F_2 females and 21 F_2 males (Table S1).

<u>Genetic map construction</u>

Mapping was based on the segregation of transcripts in two different mapping families (f4m6: 12 males, 4 females; and f3m5: 9 males, 8 females), using data generated from SEX-DETector. The intermediate VCF files from the two families were filtered to include only biallelic SNPs for which we had no genotypes missing and for which there were at least 5 counts of the minor allele. In addition, for the sex-linked transcripts, SNPs that could not be confidently assigned to the X or Y haplotype were removed. Genetic mapping thus used the same SNPs as used by SEX-DETector, i.e., inferred on the basis of no recombination in males. The VCF files were combined and processed with custom scripts into a single file in pre-makeped LINKAGE format (Lathrop and Lalouel 1984), which was analysed in Lep-MAP2 (Rastas, et al. 2016). After an iteration of genetic mapping, transcripts mapping to different LGs, or mapping to positions further than 4 cM apart, were removed from the analysis to avoid chimeric transcripts or duplicated genes. The mapping was repeated using the remaining transcripts.

After the removal of SNPs with high segregation distortion (using the Filtering module of Lep-MAP2, with dataTolerance set to 0.001 and segregation distortion by chance set to 1:1000), 9858 informative SNPs remained. The SNPs were converted to haplotypes using the AchiasmaticMeiosis module of Lep_MAP2, given each transcript as linkage group (under the assumption that there is little or no recombination within each transcript) to make linkage files based on the hypothesis of chiasmatic meiosis in both sexes. The constructed maternal and paternal genotype files were subsequently filtered again with the Filtering module of Lep-MAP2, with dataTolerance set to 0.001 and missingLimit set to 4, and removing markers with > 4 missing genotypes per family. When SNPs from the same transcript were not informative in all families, they were replaced with informative ones from one of the families, using a custom script. This allowed transcripts that were originally informative in one family to be mapped to the same linkage group as transcripts that were informative in the other family. Information of recombination within the transcript was retained, allowing for the same transcript to be mapped to multiple positions if its SNPs supported it.

We used the module SeparateChromosomes with a LOD score limit of 8 and a size limit of 3. A total of 2,968 transcripts were mapped to 236 linkage groups (LGs). The largest 12 LGs comprised of 678 transcripts that accounted for 2,228 SNPs. Finally the markers were ordered using the module OrderMarkers for the 12 largest LGs. 3 transcripts were removed from the map ends after manual inspection because of excess recombination with nearby markers. The linkage maps were drawn with MapChart v3.2 (Voorips 2002).

Characterization of genomics scaffolds

All transcripts from the sex-linked pool were mapped onto the genome using reciprocal best BLASTN with an e-value threshold of 1e^{-50} and a minimum identity match of 98%. Each high-scoring segment pair (HSP) was allowed to match only once. When X- and Y-linked homologues mapped to separate genomic scaffolds, these scaffolds were separated into the X-linked bin and the Y-linked, non-recombining bin, respectively. When both the X- and Y-linked homologue matched the same genomic scaffold, these scaffolds were classed as undetermined sex-linked scaffolds (XY-linked). When an autosomal transcript mapped onto a scaffold, this scaffold was added to the autosomal bin.

One-to-one orthologues between *M. annua, M. huetti,* and *R. communis* were identified using reciprocal best BLASTP (e-value 1e^{-50}, culling limit 1), and considering genes with only a single hit on a single contig (i.e., no genes that were split across contigs).

SNP calling

Transcripts from three unrelated males and three unrelated females were used for SNP calling. Reads were aligned to the 27,770 annotated gene models from the M1 male transcriptome using Bowtie 2 (Langmead and Salzberg 2012), allowing up to 2 mismatches per read. Picard Tools (http://broadinstitute.github.io/picard/) were used to mark duplicate read pairs for use in the Genome Analysis ToolKit (GATK, DePristo, et al. 2011). Local realignment around insertions and deletions (indels) was performed with GATK followed by SNP calling on each individual using the Haplotype Caller module and finally joint genotyping. SNPs and indels were separated and filtered to produce two high-quality variant sets with the following parameters: '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.

Estimating divergence

Nucleotide diversity (π) within sex-linked and autosomal exons was calculated using vcftools (version 0.1.13, Danecek, et al. 2011). Non-synonymous (d_N) and synonymous (d_S) divergence values were calculated using the program SNAP (Korger 2000). π_N and π_S were calculated using DnaSP (version 5, Librado and Rozas 2009).

Expression analysis

Reads from the transcriptomes of all 30 females and 35 males were aligned individually to the reference genome, assembled into genes, and analyzed using the Tuxedo suite pipeline; TopHat (Trapnell, et al. 2009), Cufflinks (Trapnell et al., 2012), CuffDiff2 (Trapnell, et al. 2014). Differential expression analysis was performed using Cuffdiff2, and graphical representation of the data was produced with CummeRbund v.2.4.0 (Goff, et al. 2013) in R v.3.0.1 (R Development Core Team 2013). X and Y expression levels were studied in *M. annua* males using the SEX-DETector output. 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 the library size of the individual. X and Y expression levels were then averaged among individuals and the ratio of the means was computed.

Data access

All DNA and RNA sequencing data generated in this study have been submitted to NCBI under BioProject ID PRJNA369310.

Acknowledgements

We thank Santiago Gonzalez-Martinez, Guillaume Cossard, Wen-Juan Ma and Melissa Toups for valuable comments on the manuscript and Jos Käfer for help with the Galaxy workflow for the SEX-DETector analysis. KER and PV were supported by grants 31003A_163384 and 31003A_141052 to JRP from the Swiss National Science Foundation, and by the University of Lausanne. JRP and DAF acknowledge support from the NERC and BBSRC, UK, who funded the early stages of this project. The computations were performed at the server at the department of Plant Sciences, University of Oxford and the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. AM and GABM acknowledge support from ANR (grant number ANR-14-CE19-0021-01).

Author contributions

Conceived the project: JRP, DF; wet lab work: DF; genome assembly and annotation: KER; gene expression analysis: KER, OE; segregation analysis and mapping: PV, AM, PR, GM;

population genetic and divergence analysis: KER, PV; wrote the paper: JRP, KER, PV; commented and contributed to the final manuscript: all authors; managed the project: JRP.

Disclosure declaration

The authors declare no conflict of interest

References

Akagi T, Henry IM, Tao R, Comai L. 2014. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science* 346:646-650.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genet.* 25:25-29.

Au KF, Underwood JG, Lee L, Wong WH. 2012. Improving PacBio long-read accuracy by short-read alignment. *PLoS One* 7.

Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, Rice WR, Valenzuela N. 2011. Are all sex chromosomes created equal? *Trends Genet.* 27:350-357.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman TL, Hahn MW, Kitano J, Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing It? *PLoS Biol.* 12.

Barrett SCH, Hough J. 2013. Sexual dimorphism in flowering plants. J. Exp. Bot. 64:67-82.

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

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

Bergero R, Charlesworth D. 2011. Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr. Biol.* 21:1470-1474.

Bergero R, Forrest A, Kamau E, Charlesworth D. 2007. Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes. *Genetics* 175:1945-1954.

Bergero R, Qiu S, Charlesworth D. 2015. Gene loss from a plant sex chromosome system. *Curr. Biol.* 25:1234-1240.

Bergero R, Qiu S, Forrest A, Borthwick H, Charlesworth D. 2013. Expansion of the pseudo-autosomal region and ongoing recombination suppression in the *Silene latifolia* sex chromosomes. *Genetics* 194:673.

Bergero R, Qui S, Charlesworth D. 2015. Gene loss from a plant sex chromosome system. *Curr. Biol.* 25:1234–1240.

Beukeboom LW, Perrin N. 2014. *The Evolution of Sex Determination*. Oxford: Oxford University Press.

Blaser O, Neuenschwander S, Perrin N. 2014. Sex-chromosome turnovers: the hot-potato model. *Amer. Nat.* 183:140-146.

Blavet N, Blavet H, Muyle A, Kafer J, Cegan R, Deschamps C, Zemp N, Mousset S, Aubourg S, Bergero R, et al. 2015. Identifying new sex-linked genes through BAC sequencing in the dioecious plant *Silene latifolia*. *BMC Genomics* 16:8.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578-579.

Cermak T, Kubat Z, Hobza R, Koblizkova A, Widmer A, Macas J, Vyskot B, Kejnovsky E. 2008. Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. *Chrom. Res.* 16:961-976.

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al. 2010. Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotech.* 28:951-U953.

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

Charlesworth B. 1998. Sex chromosomes: evolving dosage compensation. *Curr. Biol.* 8:R931-R933.

Charlesworth B, Morgan MT, Charlesworth D. 1993a. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289-1303.

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

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

Charlesworth D. 2016. Plant sex chromosomes. *Ann. Rev. Pl. Biol.* 67, 397-420.

Charlesworth D. 1999. Theories of the evolution of dioecy. In: Geber MA, Dawson TE, Delph LF, editors. *Gender and Sexual Dimorphism in Flowering Plants*. Heidelberg: Springer. p. 33-60.

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

Charlesworth D, Charlesworth B. 2005. Sex chromosomes: evolution of the weird and wonderful. *Curr. Biol*. 15:R129-R131.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95:118-128.

Charlesworth D, Mank JE. 2010. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. *Genetics* 186:9-31.

Charlesworth D, Morgan MT, Charlesworth B. 1993b. Mutation accumulation in finite populations. *J. Hered*. 84:321-325.

Chibalina MV, Filatov DA. 2011. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* 17:1475-1479.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676.

Couvreur TLP, Forest F, Baker WJ. 2011. Origin and global diversification patterns of tropical rain forests: inferences from a complete genus-level phylogeny of palms. *BMC Biol.* 9:12.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.

Dawson TE, Geber MA. 1999. Sexual dimorphism in physiology and morphology. In: Geber MA, Dawson TE, Delph LF, editors. *Gender and Sexual Dimorphism in Flowering Plants*. Heidelberg: Springer. p. 175-215.

Delph DF. 1999. Sexual dimorphism in life history. In: Geber MA, Dawson TE, Delph LF, editors. *Gender and Sexual Dimorphism in Flowering Plants*. Heidelberg: Springer. p. 149-173.

Delph LF, Bell DL. 2008. A test of the differential-plasticity hypothesis for variation in the degree of sexual dimorphism in *Silene latifolia*. *Evol. Ecol. Res.* 10:61-75.

Delph LF, Meagher TR. 1995. Sexual dimorphism masks life history trade-offs in the dioecious plant *Silene latifolia*. *Ecology* 76:775-785.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* 43:491-+.

Durand B. 1963. Le complèxe *Mercurialis annua* L. *s.l.*: une étude biosystématique. *Ann. Sci. Nat. Bot. Paris* 12:579-736.

Durand B, Durand R. 1991. Sex determination and reproductive organ differentiation in *Mercurialis*. Pl. Sc. 80:49-66.

Durand B, Louis JP, Hamdi S, Cabre E, Yu LX, Guerin B, Teller G. 1987. Major regulator genes, phytohormone levels and specific gene expression for reproductive organogenesis in *Mercurialis annua* L. (2n=16). J. Cell. Biochem :18-20.

Eckhart VM. 1999. Sexual dimorphism in flowers and inflorescences. In: Geber MA, Dawson TE, Delph LF, editors. *Gender and Sexual Dimorphism in Flowering Plants*. Heidelberg: Springer. p. 123-148.

Eggert C. 2004. Sex determination: the amphibian models. *Reprod. Nut. Dev*. 44:539-549.

English AC, Richards S, Han Y, Wang M, Vee V, Qu JX, Qin X, Muzny DM, Reid JG, Worley KC, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 7: e47768.

Fechter I, Hausmann L, Daum M, Sorensen TR, Viehover P, Weisshaar B, Topfer R. 2012. Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol. Genet. Genom*. 287:247-259.

Filatov D, Laporte V, Vitte C, Charlesworth D. 2001. DNA diversity in sex-linked and autosomal genes of the plant species *Silene latifolia* and *Silene dioica. Mol. Biol.Evol.* 18:1442-1454

Filatov DA. 2005. Evolutionary history of *Silene latifola* sex chromosomes revealed by genetic mapping of four genes. *Genetics* 170:975-979.

Filatov DA. 2015. Homomorphic plant sex chromosomes are coming of age. *Mol. Ecol.* 24:3217-3219.

Geber MA. 1999. Theories of the evolution of sexual dimorphism. In: Geber MA, Dawson TE, Delph LF, editors. *Gender and Sexual Dimorphism in Flowering Plants*. Heidelberg: Springer. p. 97-122.

Geraldes A, Hefer CA, Capron A, Kolosova N, Martinez-Nunez F, Soolanayakanahally RY, Stanton B, Guy RD, Mansfield SD, Douglas CJ, et al. 2015. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Mol. Ecol.* 24:3243-3256.

Gibson JR, Chippindale AK, Rice WR. 2002. The X chromosome is a hot spot for sexually antagonistic fitness variation. *Proc. Roy. Soc. B.* 269:499-505.

CummeRbund: Visalization and Exploraton of Cufflinks High-throughput Sequencing Data. [R package version 2.4.0]. [Internet]. 2013. Available from: http://www.bioconductor.org/packages/release/bioc/vignettes/cummeRbund/inst/doc/cummeRbund-manual.pdf

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotech.* 29:644-U130.

Hellborg L, Ellegren H. 2004. Low levels of nucleotide diversity in mammalian Y chromosomes. *Mol. Biol. Evol.* 21:158-163.

Hesse E, Pannell JR. 2011. Sexual dimorphism in a dioecious population of the wind-pollinated herb *Mercurialis annua*: the interactive effects of resource availability and competition. *Ann. Bot*. 107:1039-1045.

Hill WG, Robertson A. 1966. The effect of linkage on the limits to artificial selection. *Genet. Res*. 8:269-294.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* 12: 491.

Hough J, Hollister JD, Wang W, Barrett SCH, Wright SI. 2014. Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. *Proc. Nat. Acam. Sci. USA* 111:7713-7718.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. *Genom. Biol.* 14.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. 2009. InterPro: the integrative protein signature database. *Nuc. Acids Res.* 37:D211-D215.

Iwase M, Satta Y, Hirai H, Hirai Y, Takahata N. 2010. Frequent gene conversion events between the X and Y homologous chromosomal regions in primates. *BMC Evol. Biol.* 10:11.

Kafer J, Marais GAB, Pannell JR. 2017. Why are separate sexes rare in flowering plants. *Mol. Ecol*. (in press).

Khadka DK, Nejidat A, Tal M, Golan-Goldhirsh A. 2002. DNA markers for sex: Molecular evidence for gender dimorphism in dioecious *Mercurialis annua* L. *Mol. Breed*. 9:251-257.

Khadka DK, Nejidat A, Tal M, Golan-Goldhirsh A. 2005. Molecular characterization of a gender-linked DNA marker and a related gene in *Mercurialis annua* L. *Planta* 222:1063-1070.

Koch M, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis* and related genera (Brassicaceae). *Mol. Biol. Evol.* 17:1483–1498.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinf.* 5:59.

Korger B. 2000. HIV signature and sequence variation analysis. In: Rodrigo AG, Learn GH, editors. *Computational analysis of HIV molecular sequences*. Dordrecht, Netherlands: Kluwer Academic Publishers. p. 55-72.

Krahenbuhl M, Yuan YM, Kupfer P. 2002. Chromosome and breeding system evolution of the genus *Mercurialis* (Euphorbiaceae): implications of ITS molecular phylogeny. *Pl. Syst. Evol.* 234:155-170.

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

Labouche AM, Pannell JR. 2016. A test of the size-constraint hypothesis for a limit to sexual dimorphism in plants. *Oecologia* 181: 873-884.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* 286:964-967.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Meth*. 9:357-U354.

Lathrop GM, Lalouel JM. 1984. Easy calculations of lod scores and genetic risks on small computers. *Amer. J. Hum. Genet.* 36:460-465.

Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451-1452.

Liu ZY, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu QY, Pearl HM, Kim MS, Charlton JW, Stiles JI, et al. 2004. A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* 427:348-352.

Lloyd DG, Webb CJ. 1977. Secondary sex characters in plants. *Bot. Rev.* 43:177-216.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

Manchester SR, Judd WS, Handley B. 2006. Foliage and fruits of early poplars (Salicaceae: Populus) from the eocene of Utah, Colorado, and Wyoming. *Int. J. Pl. Sci.* 167:897-908.

Mank JE. 2013. Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet.* 29:677-683.

Mank JE. 2009. Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. *Amer. Nat.* 173:141-150.

Mascarenhas JP. 1990. Gene activity during pollen development. *Ann. Rev. Pl. Phys. Pl. Mol. Biol.* 41:317-338.

Maynard Smith J, Haigh J. 1974. The hitchhiking effect of a favourable gene. *Genet. Res.* 219:23-35.

Meagher TR. 1994. The quantitative genetics of sexual dimorphism in *Silene latifolia* (Caryophyllaceae). 2. Response to sex-specific selection. *Evolution* 48:939-951.

Ming R, Wang JP, Moore PH, Paterson AH. 2007. Sex chromosomes in flowering plants. *Amer. J. Bot.* 94:141-150.

Moore JC, Pannell JR. 2011. Sexual selection in plants. Current Biology 21:R176-R182.

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comp. Biol.* 13:1028-1040.

Muyle A, Käfer J, Zemp N, Mousset S, Picard F, Marais G. 2016. SEX-DETector: a probabilistic approach to study sex chromosomes in non-model organisms. *Genom. Biol. Evol.* 8: 2530-2543.

Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GAB. 2012. Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS Biol.* 10: e1001308.

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

Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F. 2005. A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *PLoS Biol.* 3:47-56.

Njuguna W, Liston A, Cronn R, Ashman TL, Bassil N. 2013. Insights into phylogeny, sex function and age of Fragaria based on whole chloroplast genome sequencing. *Mol. Phyl. Evol.* 66:17–29.

Obbard DJ, Harris SA, Buggs RJA, Pannell JR. 2006. Hybridization, polyploidy, and the evolution of sexual systems in *Mercurialis* (Euphorbiaceae). *Evolution* 60:1801-1815.

Obbard DJ, Harris SA, Pannell JR. 2006. Sexual systems and population genetic structure in an annual plant: testing the metapopulation model. *Amer. Nat.* 167:354-366.

Pannell J. 1997. Variation in sex ratios and sex allocation in androdioecious *Mercurialis annua*. *J. Ecol.* 85:57-69.

Pannell JR, Dorken ME, Pujol B, Berjano R. 2008. Gender variation and transitions between sexual systems in *Mercurialis annua* (Euphorbiaceae). *Int. J. Pl. Sci.* 169:129-139.

Pannell JR, Eppley SM, Dorken ME, Berjano R. 2014. Regional variation in sex ratios and sex allocation in androdioecious *Mercurialis annua*. *J. Evol. Biol.* 27:1467–1477.

Papadopulos AST, Chester M, Ridout K, Filatov DA. 2015. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proc. Nat. Acad. Sci. USA* 112:13021-13026.

Parra G, Bradnam K, Korf I. 2007a. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061-1067.

Parra G, Bradnam K, Korf I. 2007b. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. *Bioinformatics* 23:1061-1067.

Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nat. Rev. Genet.* 14:83-87.

Perrin N. 2009. Sex reversal: a fountain of youth for sex chromosomes? *Evolution* 63:3043-3049.

Pigozzi MI. 2011. Diverse stages of sex-chromosome differentiation in tinamid birds: evidence from crossover analysis in *Eudromia elegans* and *Crypturellus tataupa*. *Genetica* 139:771-777.

Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D. 2010. Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes. *Proc. Roy. Soc. Lond. B* 277:3283-3290.

R Development Core Team X. 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rahman AYA, Usharraj AO, Misra BB, Thottathil GP, Jayasekaran K, Feng Y, Hou SB, Ong SY, Ng FL, Lee LS, et al. 2013. Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genom*. 14.

Rastas P, Calboli FCF, Guo BC, Shikano T, Merila J. 2016. Construction of ultradense linkage maps with Lep-MAP2: Stickleback F2 recombinant crosses as an example. *Genom. Biol. Evol.* 8:78-93.

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

Renner SS, Ricklefs RE. 1995. Dioecy and its correlates in the flowering plants. *Amer. J. Bot.* 82:596-606.

Rice WR. 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.

Russell JRW, Pannell JR. 2015. Sex determination in dioecious *Mercurialis annua* and its close diploid and polyploid relatives. *Heredity* 114:262-271.

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al. 2011. Sequence analysis of the genome of an oil-bearing tree, *Jatropha* curcas L. *DNA Res.* 18:65-76.

Schmid M, Steinlein C. 2001. Sex chromosomes, sex-linked genes, and sex determination in the vertebrate class amphibia. EXS. 143–176. In: Scherer G, Schmid M, editors. Genes and *Mechanisms in Vertegrate Sex Deteremination*. Basel: Birkhäuser Verlag.

Simão FA, Waterhouse RM, Ioannidis P Kriventseva EV, Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 1: btv351.

Smarda P, Bures P, Smerda J, Horova L. 2012. Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. *New Phytol.* 193:513-521.

Stanke M, Schoffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC *Bioinformatics* 7: 62.

Steflova P, Tokan V, Vogel I, Lexa M, Macas J, Novak P, Hobza R, Vyskot B, Kejnovsky E. 2013. Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY1Y2) in the dioecious plant *Rumex acetosa*. *Genom. Biol. Evol.* 5:769-782.

Stock M, Horn A, Grossen C, Lindtke D, Sermier R, Betto-Colliard C, Dufresnes C, Bonjour E, Dumas Z, Luquet E, et al. 2011. Ever-young sex chromosomes in European tree frogs. *PLoS Biol.* 9.

Syed NH, Kalyna M, Marquez Y, Barta A, Brown JWS. 2012. Alternative splicing in plants - coming of age. *Trends Pl. Sci.* 17:616-623.

Telgmann-Rauber A, Jamsari A, Kinney MS, Pires JC, Jung C. 2007. Genetic and physical maps around the sex-determining M-locus of the dioecious plant asparagus. *Mol. Genet. Genom.* 278:221-234.

Tennessen JA, Govindarajulu R, Liston A, Ashman TL. 2016. Homomorphic ZW chromosomes in a wild strawberry show distinctive recombination heterogeneity but a small sex-determining region. *New Phytol.* 211:1412–1423.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2014. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Prot.* 9:2513-2513.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596-1604.

Voorrrips, RE. 2002. MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J. Hered.* 93: 77-78.

Vyskot B, Hobza R. 2015. The genomics of plant sex chromosomes. *Pl. Sci.* 236:126-135.

Wang JP, Na JK, Yu QY, Gschwend AR, Han J, Zeng FC, Aryal R, VanBuren R, Murray JE, Zhang WL, et al. 2012. Sequencing papaya X and Y-h chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc. Nat. Acad. Sci. USA.* 109:13710-13715.

Wang KB, Wang ZW, Li FG, Ye WW, Wang JY, Song GL, Yue Z, Cong L, Shang HH, Zhu SL, et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* 44:1098.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859-1875.

Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, Sun XW. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genom. 14:604.

Yamamoto K, Oda Y, Haseda A, Fujito S, Mikami T, Onodera Y. 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.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in drosophila. *Science* 337:341-345.

Zhou Q, Bachtrog D. 2012. Chromosome-wide gene silencing initiates Y degeneration in *Drosophila*. *Curr. Biol.* 22: 522-525.

Zimmer F, Harrison PW, Dessimoz C, Mank JE. 2016. Compensation of dosage-sensitive genes on the chicken Z chromosome. *Genom. Biol. Evol.* 8:1233-1242.

Zluvova J, Zak J, Janousek B, Vyskot B. 2010. Dioecious *Silene latifolia* plants show sexual dimorphism in the vegetative stage. *BMC Pl. Biol.* 10: 208.

Supplementary Information: Methods

Genome assembly

It was clear from our assemblies that assembly software such as CLC Genomics Workbench 7.0 (https://www.qiagenbioinformatics.com/) and AllPaths-LG (Stanke, et al. 2006), both of which aggressively collapsed repeats within our data, were not efficient when combined with the Pacific Biosciences reads (data not shown).

Combining gene predictions

The gene set predicted from the transcript data was taken to be the most accurate, and the three gene sets were then merged with it as follows: (1) Annotated genes predicted by the first Maker run that did not overlap existing genes were added to the final gene set; (2) Annotated exons from the first Maker run that did not overlap existing exons were added to the final gene set; (3) Steps 1 and 2 were repeated for the second Maker run; (4) Genes from expression analysis predicted to have open reading frames using Trinity (Grabherr et al., 2011) that did not overlap existing genes were added.

Supplementary Information: Tables and Figures

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

	Sex	Individual	Parents
	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. The *de novo* assembly of *M. annua* compared to the assemblies of the other sequenced Malpighiales.

	Genome size (mb)	Scaffold number	Total assembled bases	Bases without gaps	Scaffold n50 (bp)
Mercurialis annua (Annual mercury)	640	720537	546375413 (89%)	479897493 (78%)	6398
Ricinus communis (Castor bean)	400	25763	350621860 (88%)	336959153 (84%)	496528
Populus trichocarpa (Poplar)	485	2514	417286671 (83%)	403899978 (81%)	18835763
Linum usitatissimum (Flax)	350	88420	318300000 (91%)	Unknown	693500
Manihot esculenta (Cassava)	760	12977	533000000 (70%)	Unknown	258000
Hevea brasiliensis (Rubber tree)	2000	608017	1100000000 (55%)	Unknown	2972
Jatropha curcas (Oil-bearing tree)	410	39277	297660620 (73%)	295313628 (72%)	15950

Table S3. 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	20380	5832769	2.09	1.22
	En-Spm	12522	3127655	1.12	0.65
	hAT-Ac	10042	1837516	0.66	0.38
	Other	12875	2051441	0.73	0.43
Retrotransposons	LTR/Gypsy	99988	31437770	11.25	6.55
	LTR/Copia	33933	11372184	4.07	2.37
	LTR/Caulimovirus	2115	782568	0.28	0.16
	LTR/Other	1404	348956	0.12	0.07
	LINE/L1	21313	9554789	3.42	1.99
	LINE/CRE	8876	2368568	0.85	0.49
	LINE/R1	3330	925137	0.33	0.19
	LINE/Other	825	141630	0.05	0.03
	SINE	1012	142744	0.05	0.03
Other	RC/Helitron	801	258178	0.09	0.05
	Small RNA	1316	298024	0.11	0.06
	Simple repeats	40827	13362428	4.78	2.78
TRF	Tandem Repeats	215433	35543829	12.72	7.41
RepeatModeler	Unknown	789194	160024517	57.27	33.35
Totals:		1276186	279410703	100	58.22

Table S4. Summary of the final genome and transcriptome *de novo* assemblies. 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	832132	526869874 (86%)	455924417 (74%)	335033	5903
PBJelly2	825889	547240256 (89%)	481594113 (78%)	329358	5642
SSPACE	726400	545659344 (89%)	479897493 (78%)	329358	6126
L_RNA_Scf	720537	546375413 (89%)	479897493 (78%)	329358	6398
Final over 1kb	74927	397889956 (65%)	332624619 (54%)	329358	12808
Transcripts					
Trinity	49809	56442052	56442052	19144	1821

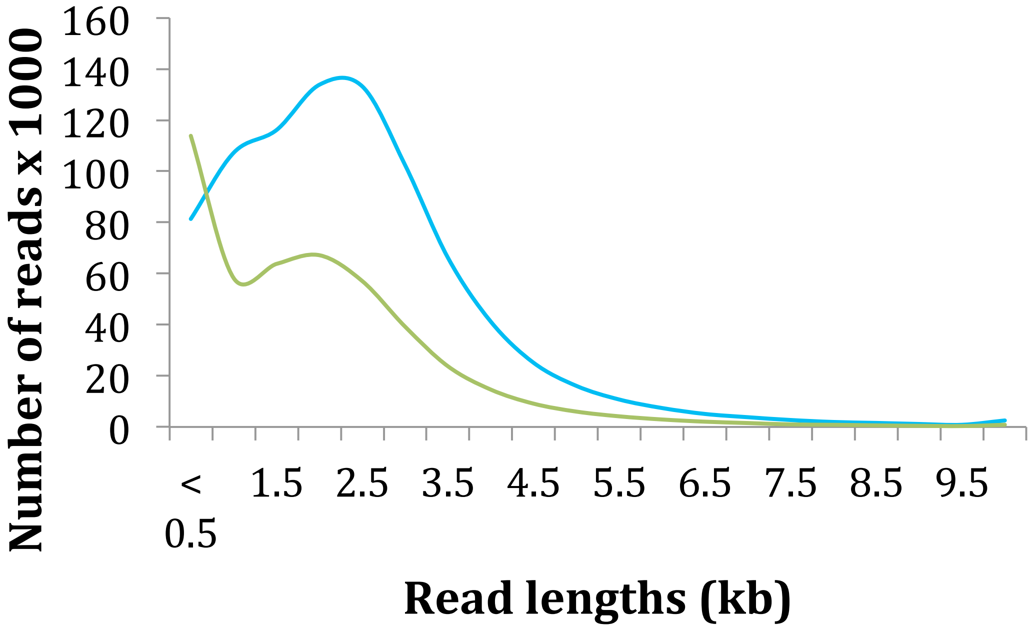

Figure S1. Distribution of Pacific Biosciences read lengths before (blue curve) and after error correction (green curve) in kilobases (kb). Read lengths are distributed in bins where X values represent the top end of each bin, for example bin 1.5 contains reads of 1-1.5kb length.

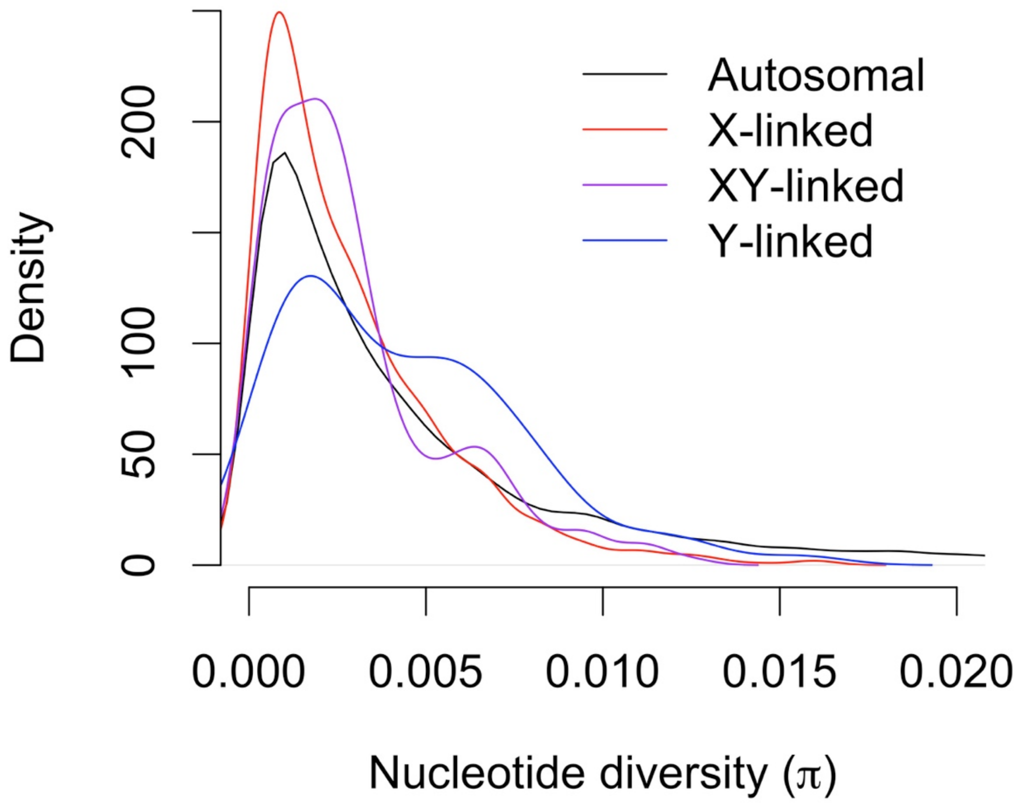

Figure S2: Density plot of nucleotide diversity (π) / kb across protein-coding genes from each of the four bins; X-linked, XY-linked, Y-linked and autosomal. Density is calculated using a Gaussian kernel, such that the area under each curve sums to 1.

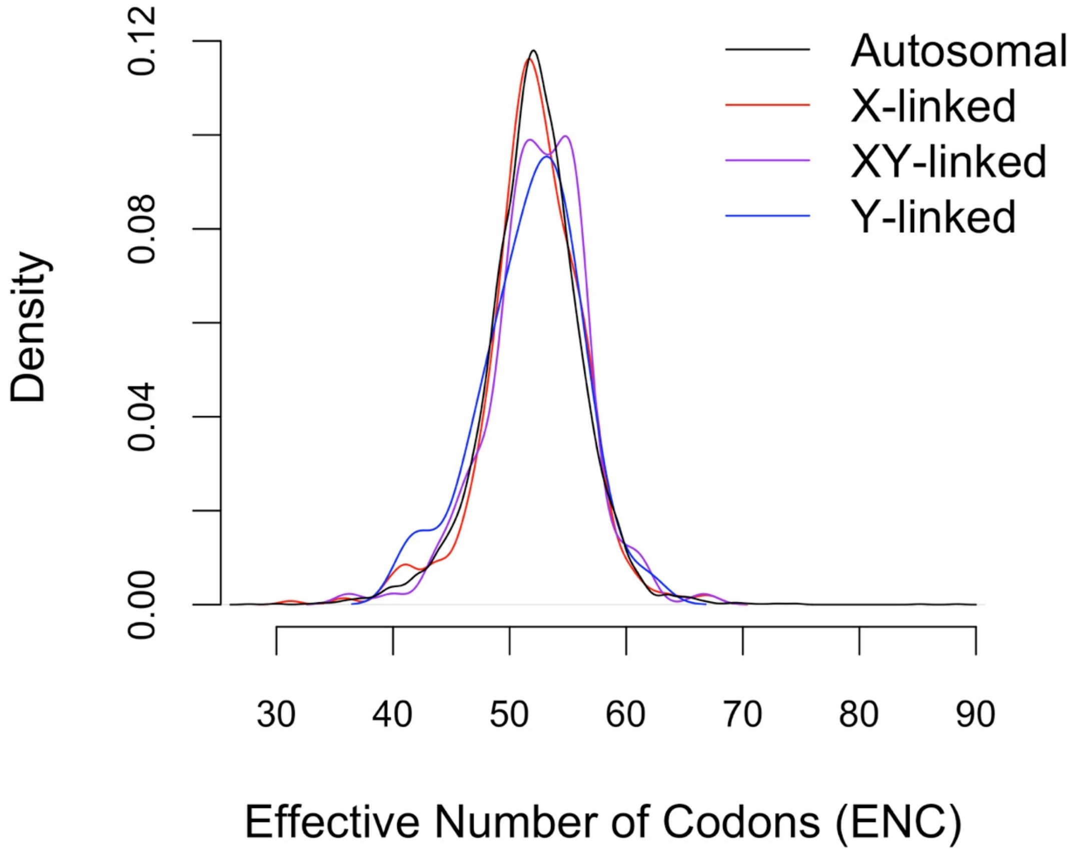

Figure S3. Density plots of the effective number of codons per gene in the X-linked, Y-linked, XY-linked and autosomal bins. Density is calculated as in Figure S3.

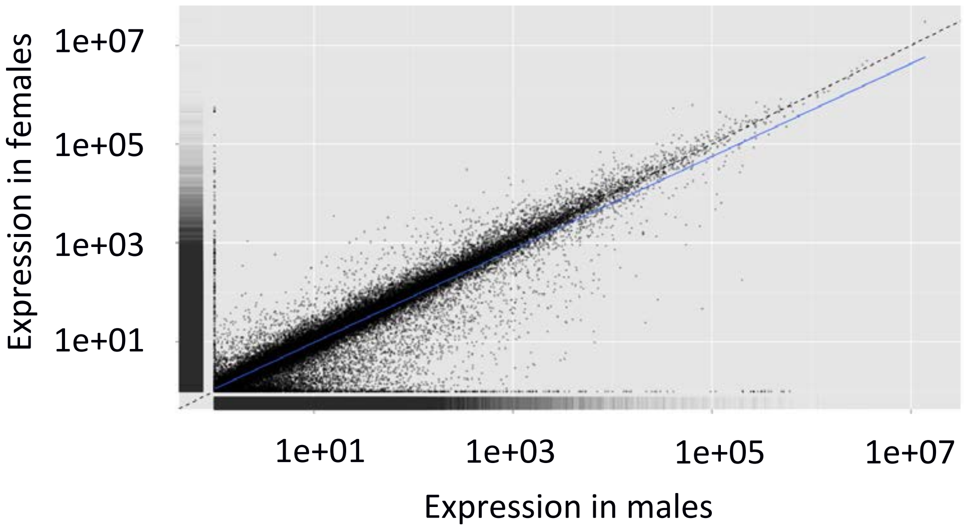

Figure S4: Scatterplot of male versus female gene expression; individual pairwise comparisons of average gene expression levels (in FPKM) between males and females were plotted on a log2-transformed scale. The smooth-fit regression line was drawn in blue. The dotted line represents equal expression between males and females.

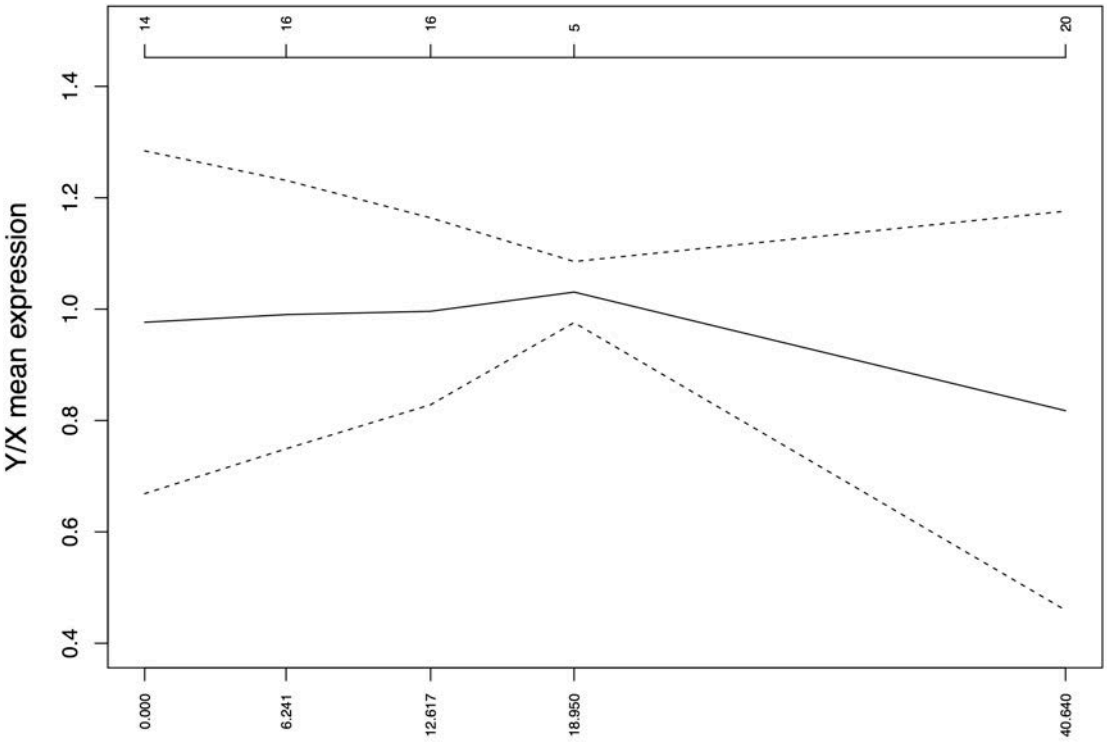

Figure S5. Mean expression difference between the Y and X allele per LG10 location. The ladder at the top of the panel indicates the number of sex-linked transcripts at each respective map position. For clearer presentation, one mapped location (6.343) containing one transcript was merged with the nearby location at 6.241 cM.

Further supplementary files

File S1: Pipeline that takes a vcf file from Sex-DETector and converts it to an input file for Lep-MAP2, and also formats the Lep-MAP2 output for plotting (pipeline maintained at http://parisveltsos.com/research/R/posts/2016/07/vcf-to-Lep-MAP/).

File S2: Detailed genetic map. First tab is at the transcript level and includes the family name that was used for sex linkage assignment (‘SEX-DETector column), second tab resolves the map to the SNP level. Column ‘common’ shows the combined map positions, columns ‘male’ and ‘female’ show the sex specific maps, columns ‘maleonly’ and ‘femaleonly’ show the sex specific maps after removing markers not informative in the opposite sex (InformativeMask 1 and 2 respectively in module OrderMarkers of Lepmap 2).

File S3: Fasta sequence of X- and Y- inferred sequences of the sex linked transcripts.