Maternal X chromosome upregulation in both sexes initiates dosage compensation evolution

Authors: Aline Muyle[1]*, Nicklaus Zemp[2], Cécile Fruchard[1], Radim Cegan[3], Jan Vrana[4], Clothilde Deschamps[5], Raquel Tavares[1], Franck Picard[1], Roman Hobza[3,4], Alex Widmer[2], Gabriel AB Marais[1]

Affiliations:

[1]Laboratoire "Biométrie et Biologie Evolutive", CNRS / Université Lyon 1, France.

[2]Institute of Integrative Biology, ETH Zurich, Switzerland.

[3]Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno.

[4]Center for Biotechnological and Agricultural Research in Hana, Olomouc, Czech Republic

[5]Center for Bioinformatics in Rhônes-Alpes (PRABI), Lyon, France

*Correspondence to: aline.muyle@univ-lyon1.fr

Abstract: When sex chromosomes evolve, Y genes become less expressed than their X counterpart. This degeneration is compensated through various mechanisms in different animal species, that reestablish ancestral expression levels in males or balance males and females. This phenomenon, called dosage compensation, has been observed for some genes in the plant *Silene latifolia*. However, the mechanism involved remains unknown. Using an outgroup without sex chromosomes as a reference, we show that the maternal X chromosome is hyperexpressed in both sexes in *S. latifolia*. This compensates for the lower Y expression in males. However, the paternal X chromosome in females maintained its ancestral expression levels, causing global hyperexpression of sex chromosomes in females, which is likely to be suboptimal. Because *S. latifolia* sex chromosomes have evolved fairly recently, our findings provide insights into the first steps of dosage compensation evolution, in addition to revealing a link between dosage compensation and imprinting.

One Sentence Summary: In *Silene latifolia*, Y chromosome expression degeneration is compensated by hyperexpression of the maternal X chromosome in both males and females.

Main Text

Sex chromosomes have repeatedly evolved from autosomes in different taxa (*1*). Y chromosomes initially have similar gene content compared to the X, but with time Y genes become less expressed and eventually get lost. This causes reduced dosage in XY males compared to XX females, with potential negative effects on fitness. In some species, a mechanism called dosage compensation evolved that reestablishes ancestral expression levels in males. In some cases only, this is accompanied by dosage balance, i.e. balanced expression between males and females (*2–4*). For example, in *Drosophila* the X chromosome is hyperexpressed specifically in males, resulting in complete dosage compensation along with dosage balance (*5*). In placental mammals, on the other hand, one X is inactivated in females, resulting in dosage balance between males and females without dosage compensation, except for a few genes which expression was doubled (*6–10*). The situation is similar in *Caenorhabditis elegans*, where both X chromosomes are downregulated in hermaphrodites and only a few genes have their expression doubled for complete dosage compensation (*11*). It is currently unknown how these mechanisms evolved, that is to say whether expression was doubled for a few genes first or whether X chromosome inactivation happened first. To answer this question, species with younger sex chromosomes must be studied.

The plant *Silene latifolia* is an ideal model to study early steps of sex chromosome evolution thanks to its X/Y pair that evolved ~5 Mya (*12*). In this species, only dosage balance has been studied so far. Equal expression levels were observed for males and females for some genes in spite of Y expression degeneration (*13–18*). However, the mechanism through which dosage balance is achieved and whether dosage compensation evolved in *S. latifolia* is unknown. In order to elucidate these questions, an outgroup without the sex chromosome pair must be used as an ancestral autosomal reference (*19*). This allows to determine whether X chromosome expression increased or decreased in *S. latifolia* compared to the outgroup. Also, there is a technical bias that was not appropriately corrected in previous analyses of dosage compensation

in *S. latifolia* (*19*). When RNA-seq reads are mapped on a transcriptome reference, there is a bias for reads to map better when they carry the same allele as the reference. This influences downstream expression level estimates. Here, we present an analysis of dosage compensation in *S. latifolia* using an outgroup species as a reference and correcting for reference mapping bias. We also develop a new statistical approach, as required to fully characterize dosage compensation (*19*).

Because only a third of the large and highly repetitive *S. latifolia* genome has so far been sequenced (*18*), we used an RNA-seq approach based on the sequencing of a cross (parents and progeny of each sex), to infer sex-linked contigs (i.e. contigs located on the non-recombining region of the sex chromosome pair) (*20*). X/Y contigs and X-hemizygous contigs (X-linked contigs without Y allele expression) were inferred separately for three tissues: flower buds, seedlings and leaves (Supplementary Table S2). About half of the inferred sex-linked contigs could be validated by independent data (Supplementary Table S2 and Materials and Methods). Contigs with sex-biased expression are not expected to be dosage compensated as they are likely to be involved in sex-specific functions. They were therefore removed for further analyses (Supplementary Table S2 and Materials and Methods). X-hemizygous contigs have previously been characterized as partially dosage balanced (*17*, *18*) and are discussed in Supplementary Text S1.

Paternal and maternal allele expression levels in males and females were estimated for sex-linked and autosomal contigs in *S. latifolia* after correcting for reference mapping bias (Materials and Methods). Then, the corresponding allelic expression levels in one or two closely related outgroups without sex chromosomes was used as a reference to orientate expression changes in *S. latifolia*. For autosomal contigs, expression levels remained statistically similar between *S. latifolia* and the outgroups (Figure 1). This is due to the close relatedness of *S. latifolia* and the outgroups, and validates the use of the outgroups as a reference for ancestral expression levels. We used the ratio of Y over X expression levels in males as a proxy for Y degeneration in order to categorize X/Y contigs. As the Y allele degenerates (paternal allele in blue in Figure 1), the expression of the X allele in males increases (maternal allele in red in Figure 1). This is the first evidence of dosage compensation in *S. latifolia*, i.e. ancestral expression levels are reestablished in males in spite of Y expression degeneration. In females, the maternal X allele expression increases with Y degeneration (gray bars in Figure 1), in a way

similar to the maternal X allele in males. The paternal X allele in females, however, maintained its ancestral expression level regardless of Y degeneration (black bars in Figure 1). Hence, dosage balance between males and females is not achieved in *S. latifolia* due to a global hyperexpression of sex-linked contigs in females compared to ancestral expression levels. All results were confirmed on two other tissues as well as on validated contigs only (Supplementary Figures S1-S6).

The maternal X allele is hyperexpressed both in males and females (Figure 1 and Supplementary Figures S1-S6), suggesting a link between dosage compensation and imprinting (the differential expression of alleles depending on the parent of origin). In order to statistically test this observation, we used a linear regression model with mixed effects (Materials and Methods). The outgroups were used as a reference and expression levels in *S. latifolia* were then analyzed while taking into account the variability due to contigs and individuals. The interaction between the effect of the parental origin and the degeneration level was estimated, which allowed to plot the difference in expression between the maternal and paternal alleles in females for different degeneration categories (Figure 2). In autosomal contigs, the maternal and paternal alleles are expressed in a similar way in females, showing a global absence of imprinting for these contigs. However, for X/Y contigs, the difference between the maternal and paternal X in females increases with Y degeneration, demonstrating the existence of a link between the evolution of dosage compensation and imprinting in *S. latifolia*. Results were confirmed on two other tissues as well as on validated contigs only (Supplementary Figures S7-S12).

Our findings provide the first characterization of a dosage compensation mechanism where sex chromosomes are hyperexpressed in females compared to ancestral expression levels (*4*). This suggests that underexpression of sex-linked genes in males is more deleterious than hyperexpression in females. The young age of *S. latifolia* sex chromosomes could explain this suboptimal pattern, and dosage balance between sexes might later evolve in this species. The results imply that hyperexpression of sex-linked genes in males and females evolves prior to X-inactivation or X downregulation, as hypothesized by Ohno for placental mammals (*21*).

Previous studies that showed dosage balance between sexes in *S. latifolia* were criticized as being a possible result of buffering mechanisms, where one copy of a gene is expressed at a higher level when haploid than when diploid due to higher availability of the cell machinery (*18*,

22, *23*). However, the observed hyperexpression of the X chromsome in *S. latifolia* males cannot be explained by buffering mechanisms alone, otherwise the maternal X in females wouldn't be hyperexpressed. Rather, our findings provide evidence for a strong link between dosage compensation evolution and imprinting. This link has previously been observed in marsupials, where the paternal X chromosome is systematically the one to be inactivated (*24*). This is also the case in mice placenta (*25*).

It would be interesting to understand how hyperexpression of the maternal X is achieved in *S. latifolia* males and females at the molecular level. Chromosome staining suggests that DNA methylation could be involved. Indeed, one arm of one of the two X chromosomes in females was shown to be hypomethylated, as well as the same arm of the single X in males (*26*). Given our results, it is very likely that this hypomethylated X chromosome is the maternal hyperexpressed X. Unfortunately, parental origin of the X chromosomes was unknown in this study (*26*). It would be extremely insightful in future to study DNA methylation patterns in *S. latifolia* paternal and maternal X chromosomes, along with a homologous pair in a closely related species without sex chromosomes.

References and Notes

1. D. Bachtrog, Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).

2. B. Charlesworth, The evolution of chromosomal sex determination and dosage compensation. *Curr. Biol. CB*. **6**, 149–162 (1996).

3. J. E. Mank, Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet. TIG*. **29**, 677–683 (2013).

4. L. Gu, J. Walters, Sex chromosome dosage compensation evolution in animals: a beautiful theory, undermined by facts and bedeviled by details. *Genome Biol. Evol.* (2017).

5. M. I. Kuroda, A. Hilfiker, J. C. Lucchesi, Dosage Compensation in Drosophila-a Model for the Coordinate Regulation of Transcription. *Genetics*. **204**, 435–450 (2016).

6. E. Pessia, J. Engelstädter, G. A. B. Marais, The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? *Cell. Mol. Life Sci. CMLS*. **71**, 1383–1394 (2014).

7. C. M. Disteche, Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* **46**, 537–560 (2012).

8. P. Julien *et al.*, Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol.* **10**, e1001328 (2012).

9. E. Pessia, T. Makino, M. Bailly-Bechet, A. McLysaght, G. A. B. Marais, Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5346–5351 (2012).

10. S. Ercan, Mechanisms of x chromosome dosage compensation. *J. Genomics*. **3**, 1–19 (2015).

11. S. E. Albritton *et al.*, Sex-biased gene expression and evolution of the x chromosome in nematodes. *Genetics*. **197**, 865–883 (2014).

12. A. Rautenberg, L. Hathaway, B. Oxelman, H. C. Prentice, Geographic and phylogenetic patterns in Silene section Melandrium (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. *Mol. Phylogenet. Evol.* **57**, 978–991 (2010).

13. M. V. Chibalina, D. A. Filatov, Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol. CB*. **21**, 1475–1479 (2011).

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

15. A. Muyle *et al.*, Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. *PLoS Biol.* **10**, e1001308 (2012).

16. Y. Kazama *et al.*, SlWUS1; an X-linked gene having no homologous Y-linked copy in Silene latifolia. *G3 Bethesda Md*. **2**, 1269–1278 (2012).

17. R. Bergero, S. Qiu, D. Charlesworth, Gene Loss from a Plant Sex Chromosome System. *Curr. Biol. CB* (2015), doi:10.1016/j.cub.2015.03.015.

18. A. S. T. Papadopulos, M. Chester, K. Ridout, D. A. Filatov, Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13021–13026 (2015).

19. A. Muyle, R. Shearn, G. A. Marais, The Evolution of Sex Chromosomes and Dosage Compensation in Plants. *Genome Biol. Evol.* **9**, 627–645 (2017).

20. A. Muyle *et al.*, SEX-DETector: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. *Genome Biol. Evol.* **8**, 2530–2543 (2016).

21. S. Ohno, Sex chromosomes and sex-linked genes. *Springer-Verl. Berl. Heidelb. N. Y.* (1967).

22. J. H. Malone *et al.*, Mediation of Drosophila autosomal dosage effects and compensation by network interactions. *Genome Biol.* **13**, r28 (2012).

23. R. A. Veitia, F. Veyrunes, S. Bottani, J. A. Birchler, X chromosome inactivation and active X upregulation in therian mammals: facts, questions, and hypotheses. *J. Mol. Cell Biol.* **7**, 2–11 (2015).

24. D. J. Whitworth, A. J. Pask, The X factor: X chromosome dosage compensation in the evolutionarily divergent monotremes and marsupials. *Semin. Cell Dev. Biol.* **56**, 117–121 (2016).

25. A. Wagschal, R. Feil, Genomic imprinting in the placenta. *Cytogenet. Genome Res.* **113**, 90–98 (2006).

26. J. Siroky, M. R. Castiglione, B. Vyskot, DNA methylation patterns of Melandrium album chromosomes. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* **6**, 441–446 (1998).

Figures

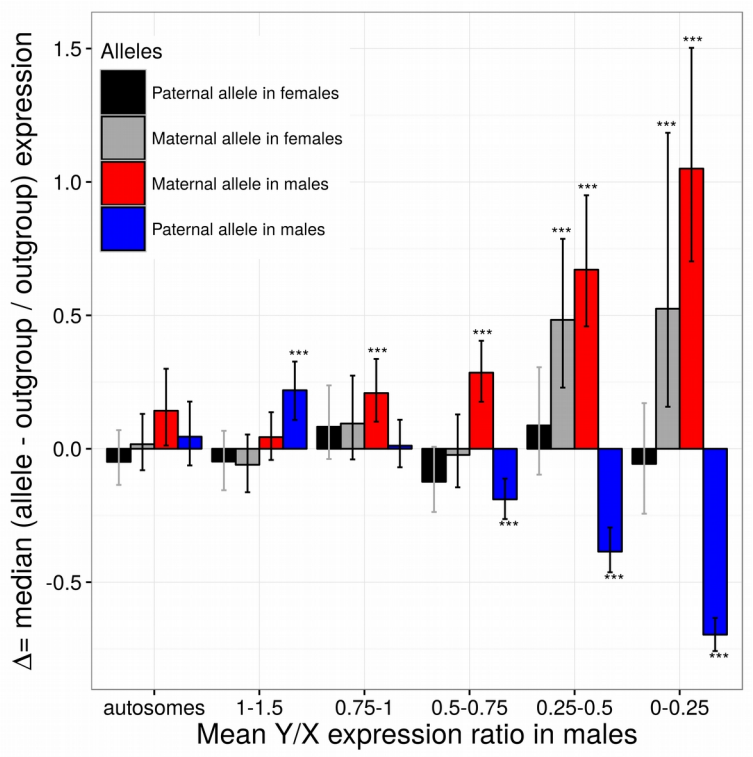


Fig. 1: Normalised difference in allelic expression levels between *S. latifolia* and the outgroup without sex chromosomes *S. vulgaris* (hereafter Δ), in autosomal and sex-linked contigs for the seedling tissue. Maternal and paternal allelic read numbers were summed at SNP positions and normalised for each individual separately, then averaged among individuals for each contig. Δ was computed as follows: Δ=(allelic expression in *S. latifolia* – allelic expression in the outgroup) / allelic expression in the outgroup). If Δ is lower, higher or equal to zero, then expression in *S. latifolia* is respectively lower, higher or equal to the outgroup. For all contig categories, Δ was compared to zero using a Wilcoxon test. The median Δ, confidence intervals and p-values adjusted for multiple testing using a

Benjamini and Hochberg correction are shown (***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.05). The Y/X ratio was computed in S. *latifolia* males and averaged among individuals to use as a proxy for Y degeneration. Contigs with sex-biased expression were removed, as well as contigs with Y/X expression ratios above 1.5. Sample sizes for the different contig categories are: autosomal:200; 1-1.5:148; 0.75-1:139; 0.5-0.75:160; 0.25-0.5:114; 0-0.25:79 (note that 200 autosomal contigs were randomly selected in order to have similar statistical power among gene categories). In the absence of dosage compensation, the single X in males should be expressed at levels similar to the outgroup that does not have sex chromosomes, in other words, without dosage compensation Δ should be close to zero for the maternal allele in males (red bars). Results show that the maternal allele is hyper-expressed in *S. latifolia* when the Y chromosome is degenerated, both in males and females.

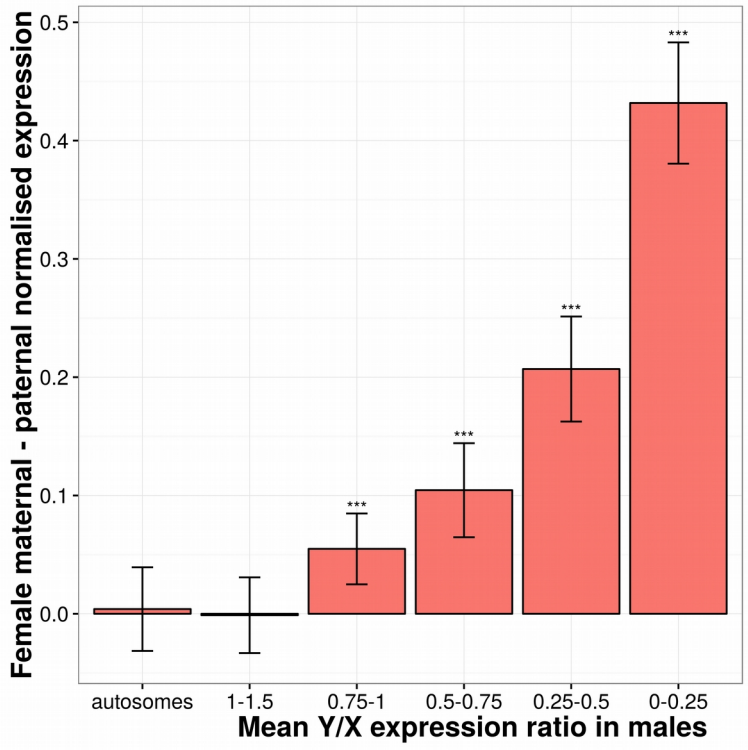


Fig. 2: Normalised expression difference between the maternal and paternal allele in *S. latifolia* females in autosomal and sex-linked contigs for the seedling tissue. The Y axis unit is the normalised allelic read count difference in log scale. A linear regression model with

mixed effects was used to study allelic expression in *S. latifolia* for every SNP position. In order to measure the changes in *S. latifolia* expression due to sex chromosomes evolution, the outgroup *S. vulgaris* that does not have sex chromosomes was used as a reference in the model (see Materials and Methods for details). The framework provided estimates for the normalised difference between the effect of paternal and maternal origin of alleles in interaction with the contig status (autosomal or sex-linked with various levels of Y degeneration), while accounting for inter-contig and inter-individual variability. See Fig. 1 legend for sample sizes for the different contig categories and statistical significance symbols. Results show that Y degeneration is linked to a significant expression difference between the paternal and maternal alleles in females, which is not observed in autosomal and non-degenerated sex-linked contigs.

Supplementary Materials

Supplementary Text 1: Dosage compensation in X-hemizygous genes

X-hemizygous contigs are inferred by SEX-DETector from X polymorphisms that segregate in the absence of Y allele expression in the RNA-seq data. For that reason, X-hemizygous contigs are harder to detect compared to X/Y contigs due to lower polymorphism (X/Y contigs can be detected both with X and X/Y polymorphisms). Another inherent bias to X-hemizygous contig inference comes from the assembly step. If the X and the Y copy are too divergent to be assembled together, the X contig will be detected as X-hemizygous by SEX-DETector. This bias was at least partly corrected in the analyses presented here (see Material and Method section 4). Also, previous work showed that only some X-hemizygous genes are dosage compensated in *Silene latifolia* (*1–5*). For all the above reasons, results on X-hemizygous contigs are analysed separately here. Poor dosage compensation of X-hemizygous contigs compared to X/Y contigs with high Y degeneration was observed across all tissues (Supplementary Figures 1 to 6). Also, the parental origin of the X chromosome has limited to no effect on female X expression levels for X-hemizygous contigs, unlike X/Y contigs (Supplementary Figures 7 to 12). A reason that could explain such a different pattern for X-hemizygous genes compared to X/Y genes is the possible dosage insensitivity of X-hemizygous genes. X-hemizygous genes could have lost their Y copy because dosage was not important for them and selection neither slowed down the loss of the Y copy nor selected for dosage compensation when degeneration inevitably occurred (*5*). A well described characteristic of dosage sensitive genes is that they tend to code proteins involved in large complexes (*6*). Using the GO-term analysis, X-hemizygous contigs were found to be significantly depleted in ribosomal protein coding genes, which is consistent with the global dosage insensitivity of X-hemizygous genes in *S. latifolia*.

Materials and Methods:

1) RNA-seq data

1.1) Plant material and sequencing

RNA-seq data were generated for a cross in the dioecious plant *S. latifolia* (which has sex chromosomes), a cross in the gynodioecious plant *S. vulgaris* (which does not have sex chromosomes) and for individuals of the hermaphrodite plant *S. viscosa* (which does not have sex chromosomes either). Plants were grown in the greenhouse from seeds collected in the wild. The following individuals and tissues were sequenced (see Supplementary Table 1 for library sizes):

- For flower buds for *S. latifolia*, both parents were sequenced (a male from a wild population, Leuk144-3_father, and a female from a ten-generation inbred line, U10_37_mother), along with their progeny (C1_01_male, C1_3_male, C1_04_male, C1_05_male, C1_26_female, C1_27_female, C1_29_female, C1_34_female). For flower buds for *S. vulgaris*, the hermaphrodite father came from a wild population (Guarda_1), the female mother from another wild population (Seebach_2) and their hermaphrodite (V1_1, V1_2, V1_4) and female (V1_5, V1_8, V1_9) progeny were sequenced. For flower buds for *S. viscosa*, three individuals were sequenced (Svi_BUL, Svi_KIR, SVi_RUS).

- For leaves for *S. latifolia* the same cross progeny individuals were sequenced (C1_01_male, C1_3_male, C1_04_male, C1_05_male, C1_26_female, C1_27_female, C1_29_female, C1_34_female). For leaves for *S. vulgaris* four hermaphrodites were sequenced (V2_10_L, V2_11_L, V2_12_L, V2_13_L). For leaves for *S. viscosa* two indivuduals were sequenced (Svi_CZE, SVi_SWE).

- For seedlings for *S. latifolia* different cross progeny individuals were sequenced (Seed_lati_female_1, Seed_lati_female_2, Seed_lati_female_3, Seed_lati_female_4, Seed_lati_male_1, Seed_lati_male_2, Seed_lati_male_3, Seed_lati_male_4). For seedlings for *S. vulgaris* four hermaphrodites were sequenced (Seed_vulg_herm_1, Seed_vulg_herm_2, Seed_vulg_herm_3, Seed_vulg_herm_4).

For the flower bud and leaf datasets, individuals were grown in a temperature-controlled greenhouse. The QiagenRNeasy Mini Plant extraction kit was used to extract total RNA two

times separately from four flower buds at developmental stages B1–B2 after removing the calyx. Samples were treated additionally with QiagenDNase. RNA quality was assessed with an Aligent Bioanalyzer (RIN.9) and quantity with an Invitrogen Qubit. An intron-spanning PCR product was checked on an agarose gel to exclude the possibility of genomic DNA contamination. Then, the two extractions of the same individual were pooled. Individuals were tagged and pooled for sequencing. Samples were sequenced by FASTERIS SA on an Illumina HiSeq2000 following an Illumina paired-end protocol (fragment lengths 150–250bp, 100 bp sequenced from each end). Seedlings were grown in a temperature controlled climate chamber in Eschikon (Switzerland) using the same conditions as in (*7*). The *S. latifolia* and *S. vulgaris* seedlings were collected without roots at the four leaf stage. The sexing of the *S. latifolia* seedlings was done using Y specific markers (*8*) that were amplified with the direct PCAR KAPA3G Plant PCR Kit (note male number 3 was later shown to be a female). High quality RNA (RIN > 7.5) was extracted using the total RNA mini kit of Geneaid. Twelve RNA-seq libraries were produced using the Truseq kit v2 from Illimina. Libraries were tagged individually and sequenced in two Illumina Hiseq 2000 channels at the D-BSSE (ETH Zürich, Switzerland) using 100 bp paired-end read protocol.

A normalised 454 library was generated for each sex of *S. latifolia* using bud extracts from 4 different developmental stages that were pooled.

Plants from an 11 generation inbred line were grown under controlled conditions in a greenhouse in Eschikon (Switzerland). One female (U11_01) and one male (U11_02) were randomly selected. From each plant, high quality RNA (RIN > 7.5) were extracted using the total RNA mini kit of Geneaid from very small flower buds, small and large flower buds, flowers before anthesis without calyces, rosette leaves and seedlings (4 leaves stage). Additionally, pollinated flowers (1 h after pollination) and developing fruits (5 day after pollination) were extracted from the female and pollen from the male. RNA of the different tissues was equally pooled for each sex and cDNA was produced using the Clontech SMARTer Kit. The two cDNA pools were then normalized using a duplex specific endonuclease of the Evrogen TRIMMER kit. For each sex two ranges were selected (1- 1.3 kb and 1.2 -2 kb) using the Pippin Prep (Sage Science). Four SMRTbell libraries were prepared using the C2 Pacific Biosciences (PacBio) chemistry and sequenced on a PacBio RS II at the Functional Genomic Center Zurich (FGCZ).

1.2) Reference trancriptome

A reference transcriptome was built from the *Silene latifolia* flower bud data and used for all tissue types and species afterwards. Adaptors, low quality and identical reads were removed. The transcriptome was then assembled using TRINITY (*9*) on the combined 10 individuals described previously as well as 6 individuals from (*10*) and the normalized 454 sequencing that was transformed to illumina using 454-to-illumina-transformed-reads. Isoforms were then collapsed using /trinity-plugins/rsem-1.2.0/rsem-prepare-reference. PolyA tails, bacterial RNAs and ribosomal RNAs were removed using ribopicker. ORFs were predicted with trinity transcripts_to_best_scoring_ORFs.pl. In order to increase the probability of X and Y sequences to be assembled in the same contig, ORFs were further assembled using home made perl scripts to run CAP3 (*11*) Version Date 10/15/07 inside of TRINITY components with parameter -p 70.

2) Inference of sex-linked contigs

Illumina reads from the 10 individuals of the cross were mapped onto the assembly using BWA (*12*) version 0.6.2 with the following parameters: bwa aln -n 5 and bwa sampe. The libraries were then merged using SAMTOOLS version 0.1.18 (*13*). The obtained alignments were locally realigned using GATK IndelRealigner (*14*) and were analysed using reads2snps (*15*) version 3.0 with the following parameters: -fis 0 -model M2 -output_genotype best -multi_alleles acc -min_coverage 3 -par false. This allowed to genotype individuals at each loci while allowing for biases in allele expression, and without cleaning for paralogous SNPs. Indeed, X/Y SNPs tend to be filtered out by paraclean, a program which removes paralogous positions (*16*). SEX-DETector (*17*) was then used to infer contig segregation types after estimation of parameters using an SEM algorithm. Contig posterior segregation type probabilities were filtered to be higher than 0.8. Because the parents were not sequenced for the leaf and seedling datasets, SEX-DETector was run using the flower bud data for the parents.

3) Reference mapping bias correction

In order to avoid biases towards the reference allele in expression level estimates, a second mapping was done using the program GSNAP (*18*) with SNP tolerant mapping option. A GSNAP SNP file was generated by home-made perl scripts using the SEX-DETector SNP detail output file. Shortly, for each polymorphic position of all contigs, the most probable posterior SNP type was used to extract the possible alleles and write them to the GSNAP SNP file. This

way, reference mapping bias was corrected for both sex-linked and autosomal contigs. Only uniquely mapped and concordant paired reads were kept after this. See Supplementary Table S1 for percentage of mapped reads. SEX-DETector was run a second time on this new mapping and the new inferences were used afterwards for all analyses (see Supplementary Table S2 for inference results).

4) Validation of sex-linked contigs

3.1) Detection of false X-hemizygous contigs

Erroneous inference of X-hemizygous contig can be due to a true X/Y gene which X and Y copies were assembled into different contigs. In order to detect such cases, X-hemizygous contigs were blasted (*19*) with parameter -e 1E-5 against RNA-seq contigs that have male-limited expression (see section 5 below for how male-limited contigs were inferred). These cases were removed from the analyses presented here.

3.2) Validation using data from literature

A few sex-linked and autosomal genes in *S. latifolia* have already been described in the literature (see Supplementary Table S3).

3.3) Validation using a genetic map

A genetic map was built and contigs from the X linkage group were used to validate SEX-DETector inferences.

Plant material and genotyping: A female individual from an interspecific *S. latifolia* cross (C1_37) was back crossed with a male from an 11 generation inbred line (U10_49). The offspring (hereafter called BC1 individuals) were grown under controlled conditions in a greenhouse in Eschikon (Switzerland). High quality RNA from flower buds as described in (*7*) was extracted from 48 BC1 individuals (35 females and 13 males). 48 RNA-seq libraries were produced using the Truseq kit v2 from Illimina with a median insert size of about 200 bp. Individuals were tagged separately and sequenced in four Illumna Hiseq 2000 channels at the D-BSSE (ETH Zürich, Switzerland) using 100bp paired-end read protocol. The parents used for this back cross had previously been sequenced in a similar way (*10*, *20*).

Linkage group identification: RNA-seq reads were mapped against the *S. latifolia* flower buds reference transcriptome using BWA (*12*) with a maximum number of mismatch equal to 5.

Libraries were merged and realigned using GATK (*14*) and SNPs were analysed using reads2snps (*15*). Using a customized perl script, SNP genotypes from the parents and the offspring as well as the associated posterior probabilities were extracted from the reads2snps output file. Informative SNPs according to CP and BC1 design and a posterior odd of 0.8 were then converted into a JoinMap format using a customized R script. If more than one informative SNP per contig was present, the SNP was used with less segregation distortion and less missing values. This led to 8,023 BC1 and 16,243 CP markers. Loci with more than 10 % missing values were excluded, which left 7,951 BC1 and 15,118 CP markers. Linkage groups were identified using the default setting of JoinMap 4.1 (*21*). Robustness of the assignment of the linkage groups was tested using LepMap (*22*). Blasting the contigs against known sex-linked genes allowed the identification of the X chromosome linkage group. Contigs could not be ordered along the linkage groups due to the too limited number of individuals that prevented the convergence of contig order. However, contigs were reliably attributed to linkage groups.

3.4) Validation using isolated Y chromosome DNA-seq data

Y chromosome DNA was isolated using flow cytometry. The samples for flow cytometric experiments were prepared from root tips according to (*23*) with modifications. Seeds of *S. latifolia* were germinated in a petri dish immersed in water at 25°C for 2 days until optimal length of roots was achieved (1 cm). The root cells were synchronized by treatment with 2mM hydroxyurea at 25°C for 18h. Accumulation of metaphases was achieved using 2.5µM oryzalin. Approximately 200 root tips were necessary to prepare 1ml of sample. The chromosomes were released from the root tips by mechanical homogenization using a Polytron PT1200 homogenizer (Kinematica AG, Littau, Switzerland) at 18,000rpm for 13 s. The crude suspension was filtered and stained with DAPI (2µg/ml). All flow cytometric experiments were performed on FACSAria II SORP flow cytometer (BD Biosciences, San José, Calif., USA). Isolated Y chromosomes were sequenced with 2x100bp PE Illumina HiSeq.

Reads were filtered for quality and Illumina adapters were removed using the ea-utils FASTQ processing utilities (*24*). The optimal kmer value for assembly was searched using KmerGenie (*25*). Filtered reads were assembled using soapdenovo2 (*26*) with kmer=49, as suggested by KmerGenie. The obtained assembly was highly fragmented, therefore RNA-seq data was used to join, order and orient the genomic fragments with L_RNA_scaffolder (*27*). The following RNA-seq reads were used (see section 1.1): one sample of male flower buds

sequenced by 454, 6 samples of male flower buds sequenced by Illumina paired-end, 4 samples of male leaves sequenced by Illumina paired-end and one sample of male pooled tissues sequenced by PacBio. The genomic assembly was successively scaffolded with L_RNA_scaffolder using RNA-seq samples one after the other, first 454 samples then Illumina and finally PacBio. The obtained contigs were filtered to be longer than 200pb.

3.5) Set of validated sex-linked and autosomal contigs

The three sources of data (litterature, genetic map and filtered Y sequence data) were compared to SEX-DETector inferred sex-linked RNA-seq contigs using BLAST (*19*) with parameter -e 1E-5. Blasts were filtered for having a percentage of identity over 90%, an alignment length over 100bp and were manually checked. If a sex-linked RNA-seq contig blasted against a sequence from one of the three data sources (literature, X genetic map or filtered Y DNA-seq) it was then considered as validated. See Supplementary Table S2 for numbers of validated sex-linked contigs.

4) Expression level estimates

4.1) whole contig expression levels

Whole contig mean expression levels were obtained for each individual using GATK DepthOfCoverage (*14*) as the sum of every position coverage, divided by the length of the contig. Normalised expression levels, in RPKM (*28*), were then computed for each individual by dividing by the value by the library size of the individual (total number of mapped reads), accounting for different depths of coverage among individuals. Whole contig mean male and female expression levels were then computed by averaging male and female individuals for each contig.

4.2) Allelic expression levels

In order to study separately X and Y allele expression levels in males and females, expression levels were studied at the SNP level. In *S. latifolia*, for each sex-linked contig expression levels were estimated using read counts from both X/Y and X-hemizygous informative SNPs. SNPs were attributed to an X/Y or X hemizygous segregation type if the according posterior probability was higher than 0.5. SNPs are considered informative if the father is heterozygous and has a genotype that is different from the mother (otherwise it is not

possible to tell apart the X from the Y allele and therefore it is not possible to compute X and Y expression separately). X/Y SNPs for which at least one female had over two percent of her reads belonging to the Y allele were removed as unlikely to be true X/Y SNPs.

Contigwise X, Y, X+X and X+Y normalised expression levels in RPKM (*28*) were computed by summing read numbers for each X-linked or Y-linked alleles for all SNPs of the contig and each individual separately and then normalised using the library size and the number of studied sex-linked SNPs in the contig:

$$E = r/ (n * l) \quad (1)$$

With E = normalised expression level for a given individual, r = sum of total read counts, n = number of studied SNPs, l = library size of the individual (number of mapped reads).

For contigs that only have X/X SNPs (SNPs for which the father's X is different to both Xs from the mother), Y expression level is only computed from the father as all males are homozygous in the progeny. Such contigs were therefore removed when having under 3 X/X SNPs to avoid approximations on the contig mean Y/X expression level (39 contigs removed in the flower buds dataset, 44 in the leaves dataset and 40 in the seedlings dataset).

In order to make *S. latifolia* expression levels comparable to *S. viscosa* and *S. vulgaris* for sex-linked contigs, *S. vicosa* ans *S. vulgaris* expression levels were estimated using only the positions used in *S. latifolia* (informative X/Y or X-hemizygous SNPs). The read count of every position in every contig and for every *S. viscosa* and *S. vulgaris* individual was given by GATK DepthOfCoverage (*14*). Only positions corresponding to informative X/Y or X-hemizygous SNPs in *S. latifolia* were used to compute the expression level for each contig and each individual as explained in equation (1).

Contigwise *S. latifolia* X, Y, X+X, X+Y allelic expression levels were then averaged among individuals. Autosomal normalised expression levels in the two outgroups (*S. vulgaris* and *S. viscosa)* were averaged together.

5) Identification of sex-biased contigs

The analysis was done separately for the three tissues (flower buds, seedling and rosette leaves) as in (*20*). Contigs were filtered to be sufficiently expressed in at least half of the female and/or male libraries using a cutoff of 1 count per million reads. Contigs with sex-biased

411 expression were identified using a common dispersion in edgeR (*29*) including the different
412 replicates for the contigs inferred to sex-linked and X-hemizygous separately. See
413 Supplementary Table S2 for numbers of sex-biased contigs removed in order to study dosage
414 compensation.

415 Male-limited expressed contigs were identified by calculating the mean expression values
416 (FPKM) in both sexes and selecting those which were exclusively expressed in males.

417 **6) Analysis of expression divergence between *S. latifolia* and the two outgroups**

418 The normalised difference in allelic expression between *S. latifolia* and the two outgroups
419 (hereafter Δ) was computed in order to study how sex chromosome expression levels evolved in
420 *S. latifolia* compared to autosomal expression levels in the two outroups: Δ is equal to zero if *S.*
421 *latifolia* and the outgroups have equal expression levels, Δ is positive if *S. latifolia* has higher
422 expression levels compared to the outgroups and Δ is negative otherwise (2).

423 Δ = (*S. latifolia* expression level - outgroup expression level) / (outgroup expression
424 level) (2)

425 Sex-linked contigs were grouped by categories of degeneration level using the average Y
426 over X expression ratio in males. 200 autosomal contigs were randomly selected in order to have
427 similar statistical power among gene categories. Δ values for each allele (maternal and paternal
428 in males and females) and each gene category were compared to zero using a Wilcoxon test. P-
429 values were corrected for multiple testing using a Benjamini and Hochberg correction. The
430 estimated median Δ, confidence intervals and adjusted p-values were then used to plot Figure 1
431 and Supplementary Figures S1 to S6.

432 **7) Analysis of expression level difference between maternal and paternal alleles**

433 Maternal and paternal alleles expression were compared in *S. latifolia* for autosomal and
434 sex-linked contigs. In order to deal with the difference in numbers of autosomal versus sex-
435 linked contigs (Supplementary Table S2), 200 autosomal contigs were randomly selected in order
436 to keep comparable powers of detection. Allelic expression levels in *S. latifolia* for each
437 individual at every SNP position of each contig were analysed using a linear regression model
438 with mixed effects (with the R package lme4, *30*). We assumed a normal distribution of the read
439 count data after log transformation. In order to account for inter-individual and inter-contig

variability, a random "individual" and a random "contig" effect were included in the model. The aim of this modeling framework was to estimate the joint effect of the chromosomal origin of alleles (paternal or maternal in males or females) and the status of the gene (autosomal or sex-linked with various levels of Y degeneration defined by the average Y over X expression ratio in males). Two fixed effects with interaction were therefore considered in the model (3). In order to estimate the changes in sex-linked gene expression levels since the evolution of sex chromosomes, we used the average of the two outgroup expression levels as a reference (offset) for every SNP position, divided by two in order to be comparable to *S. latifolia* allelic expression levels.

log(Expression+1) ~ Chromosome * Degeneration + (1|individual) + (1|Contig), offset = log(outgroup average expression/2 +1) (3)

All effects of the model (fixed or random) were proved highly significant (p-values < $2.2.10^{-16}$) using comparison of the fit of model (3) to simpler nested models (removing one effect at a time in model (3)). In order to statistically test whether there was a difference between the effects of paternal and maternal alleles in females in different degeneration categories we used the contrasts provided by the lmerTest package in R (*30*). This strategy provided estimates, confidence intervals and p-values of the difference between the two effects of paternal and maternal origin in females in interaction with degeneration levels, while normalising by the expression of the two outgroups. Moreover, the presence of random effects allows to account for inter-individual and inter-contig variability. Finally, p-values were corrected for multiple testing using a Benjamini and Hochberg correction. These values were used to plot Figure 2 and Supplementary Figures S7 to S12.

8) GO term analysis

All contigs of the reference transcriptome were blasted against the Uniprot database using blastx with an expectation value cutoff of 0.001, and an upper limit of hit results per sequence of 20. Significant blastx hits were obtained for 35,763 contigs. Gene Ontology terms were associated with the contigs using the Blast2GO PRO version V 2.7.2 (*31*) and the Data Base version b2g_sep14, with 41,136 GO terms and 4,097 Enzymes available. Blast hits were related to the functional GO annotation by using the "Mapping" function of Blast2GO. The complete annotation of the sequences was performed using Annotation, ANNEX (increasing annotation

step), GO-SLIM (using the goslim_plant.obo option to reduce the GO vocabulary) and InterPro (by choosing the whole set of InterPro annotation applications available: BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, HMMPanther, ProfileScan, HAMAP, PatternScan, SuperFamily, Gene3D, Phobius, Coils, SignalIPHMM and TMHMM). Annotation using the default parameters (E-Value-Hit-Filtrer 1.0E-6; Annotation CutOff 55; GO Weight 5; Hsp-Hit Coverage CutOff 0) yielded a total of 27,152 annotated contigs (after merging the InterPro results with the GO annotation). Tests for enrichment were performed on all GO terms associated with the sequences, using a two-tailed Fisher's exact test with correction for multiple tests by using a false discovery rate of 0.05. Double IDs (both on the test and on the reference set) were removed.

Supplementary material references

1. R. Bergero, S. Qiu, D. Charlesworth, Gene Loss from a Plant Sex Chromosome System. *Curr. Biol. CB* (2015), doi:10.1016/j.cub.2015.03.015.

2. A. S. T. Papadopulos, M. Chester, K. Ridout, D. A. Filatov, Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13021–13026 (2015).

3. M. V. Chibalina, D. A. Filatov, Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol. CB*. **21**, 1475–1479 (2011).

4. M. Toups, P. Veltsos, J. R. Pannell, Plant sex chromosomes: lost genes with little compensation. *Curr. Biol. CB*. **25**, R427-430 (2015).

5. A. Muyle, R. Shearn, G. A. Marais, The Evolution of Sex Chromosomes and Dosage Compensation in Plants. *Genome Biol. Evol.* **9**, 627–645 (2017).

6. E. Pessia, T. Makino, M. Bailly-Bechet, A. McLysaght, G. A. B. Marais, Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5346–5351 (2012).

7. N. Zemp, A. Minder, A. Widmer, Identification of internal reference genes for gene expression normalization between the two sexes in dioecious white Campion. *PloS One*. **9**, e92893 (2014).

8. R. Hobza, A. Widmer, Efficient molecular sexing in dioecious Silene latifolia and S. dioica and paternity analysis in F(1) hybrids. *Mol. Ecol. Resour.* **8**, 1274–1276 (2008).

9. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

10. A. Muyle *et al.*, Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. *PLoS Biol.* **10**, e1001308 (2012).

11. X. Huang, A. Madan, CAP3: A DNA Sequence Assembly Program. *Genome Res.* **9**, 868–877 (1999).

12. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

13. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

14. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

15. G. Tsagkogeorga, V. Cahais, N. Galtier, The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis. *Genome Biol. Evol.* **4**, 740–749 (2012).

16. P. Gayral *et al.*, Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genet.* **9**, e1003457 (2013).

17. A. Muyle *et al.*, SEX-DETector: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. *Genome Biol. Evol.* **8**, 2530–2543 (2016).

18. T. D. Wu, S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinforma. Oxf. Engl.* **26**, 873–881 (2010).

19. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

20. N. Zemp *et al.*, Evolution of sex-biased gene expression in a dioecious plant. *Nat. Plants*. **2**, 16168 (2016).

21. J. W. VAN Ooijen, Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet. Res.* **93**, 343–349 (2011).

22. P. Rastas, L. Paulin, I. Hanski, R. Lehtonen, P. Auvinen, Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinforma. Oxf. Engl.* **29**, 3128–3134 (2013).

23. J. Vrána, H. Simková, M. Kubaláková, J. Číhalíková, J. Doležel, Flow cytometric chromosome sorting in plants: the next generation. *Methods San Diego Calif.* **57**, 331–337 (2012).

24. E. Aronesty, *ea-utils: "Command-line tools for processing biological sequencing data"* (2011; http://code.google.com/p/ea-utils).

25. R. Chikhi, P. Medvedev, Informed and automated k-mer size selection for genome assembly. *Bioinforma. Oxf. Engl.* **30**, 31–37 (2014).

26. R. Luo *et al.*, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. **1**, 18 (2012).

27. W. Xue *et al.*, L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics*. **14**, 604 (2013).

28. A. Oshlack, M. D. Robinson, M. D. Young, From RNA-seq reads to differential expression results. *Genome Biol.* **11**, 220 (2010).

29. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).

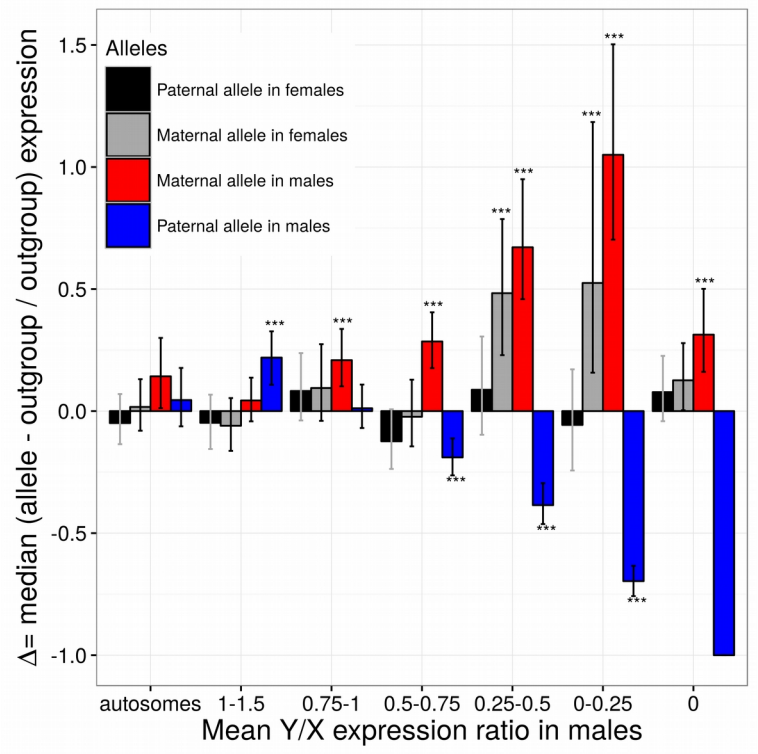
30. R Development Core Team, R: A language and environment for statistical computing, reference index version 2.15.0. *R Found. Stat. Comput. Vienna Austria* (2012).

31. A. Conesa *et al.*, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinforma. Oxf. Engl.* **21**, 3674–3676 (2005).

482

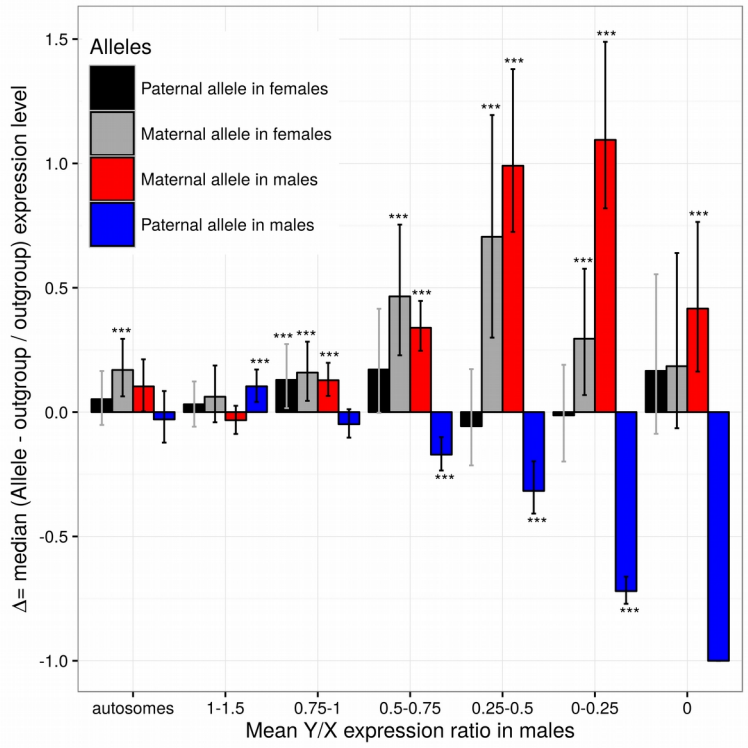
483

Supplementary Figures

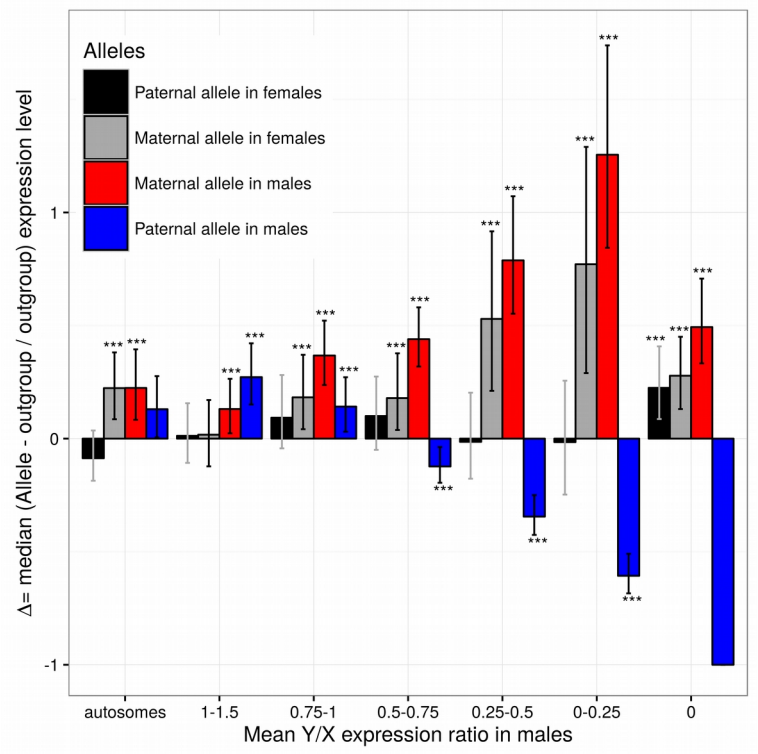


Supplementary Figure S1: Normalised difference in allelic expression levels between *S. latifolia* and the two outgroups without sex chromosomes *S. vulgaris* and *S. viscosa* (hereafter Δ), in autosomal and sex-linked contigs for the **seedling** tissue. Maternal and paternal allelic read numbers were summed at SNP positions and normalised for each individual separately, then averaged among individuals for each contig. Δ was computed as follows: Δ=(allelic expression in *S. latifolia* – allelic expression in the outgroup) / allelic expression in the outgroup). If Δ is lower, higher or equal to zero, then expression in *S. latifolia* is respectively lower, higher or equal to the outgroup. For all contig categories, Δ was compared to zero using a Wilcoxon test. The median Δ, confidence intervals and p-values adjusted for multiple testing using a Benjamini and Hochberg correction are shown (***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.05). The Y/X ratio was computed in S. *latifolia* males and averaged among individuals to use as a proxy for Y degeneration. Contigs with sex-biased expression were removed, as well as contigs with Y/X expression ratios above 1.5. Sample sizes for the different contig categories are: autosomal:200; 1-1.5:148; 0.75-1:139; 0.5-0.75:160; 0.25-0.5:114; 0-0.25:79; 0:205 (note that 200 autosomal contigs were randomly selected in order to have similar statistical power among

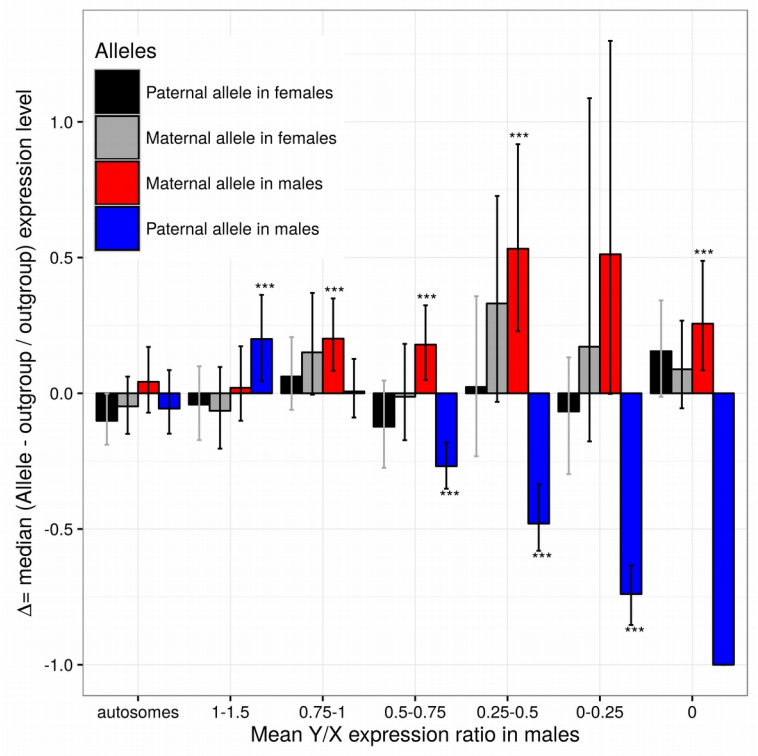
gene categories). In the absence of dosage compensation, the single X in males should be expressed at levels similar to the outgroup that does not have sex chromosomes, in other words, without dosage compensation Δ should be close to zero for the maternal allele in males (red bars). Results show that the maternal allele is hyper-expressed in *S. latifolia* when the Y chromosome is degenerated, both in males and females.



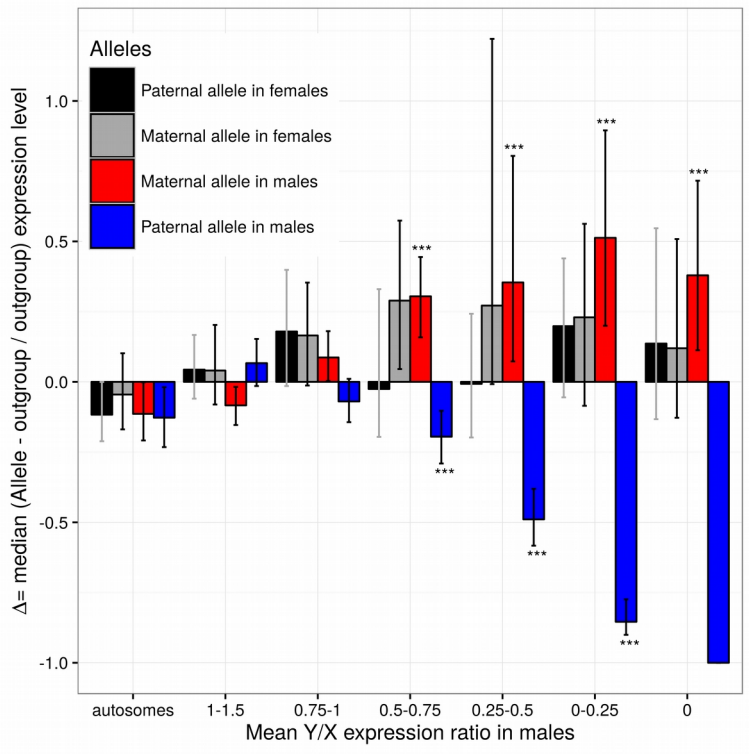
Supplementary Figure S2: Normalised difference in allelic expression levels between *S. latifolia* and the two outgroups without sex chromosomes *S. vulgaris* and *S. viscosa* (Δ), in autosomal and sex-linked contigs for the **flower bud** tissue. Same legend as Supplementary Figure 1 except for sample sizes for the different contig categories: autosomal:200; 1-1.5:95; 0.75-1:195; 0.5-0.75:203; 0.25-0.5:176; 0-0.25:116; 0:103.



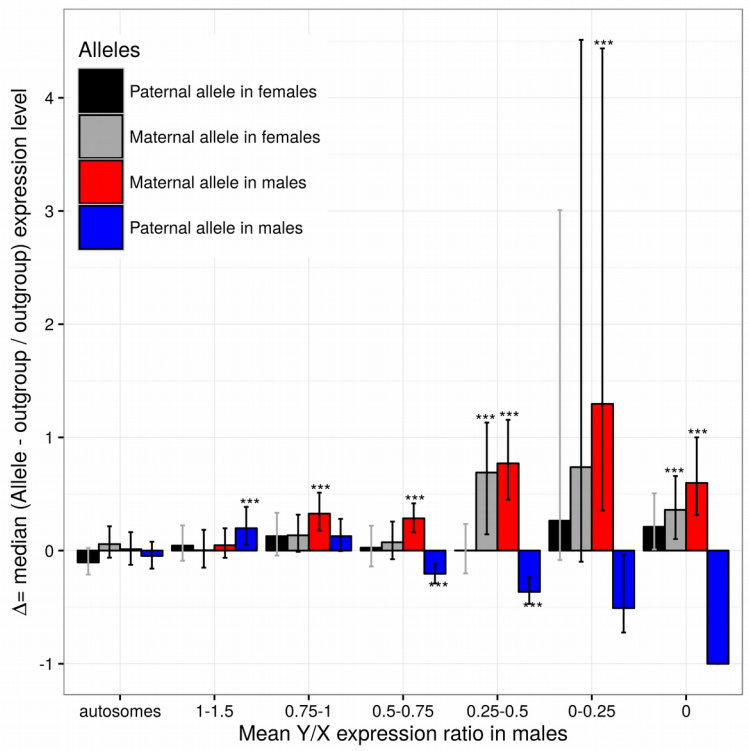
Supplementary Figure S3: Normalised difference in allelic expression levels between *S. latifolia* and the two outgroups without sex chromosomes *S. vulgaris* and *S. viscosa* (Δ), in autosomal and sex-linked contigs for the **leaf** tissue. Same legend as Supplementary Figure 1 except for sample sizes for the different contig categories: autosomal:200; 1-1.5:159; 0.75-1:132; 0.5-0.75:147; 0.25-0.5:126; 0-0.25:71; 0:275.



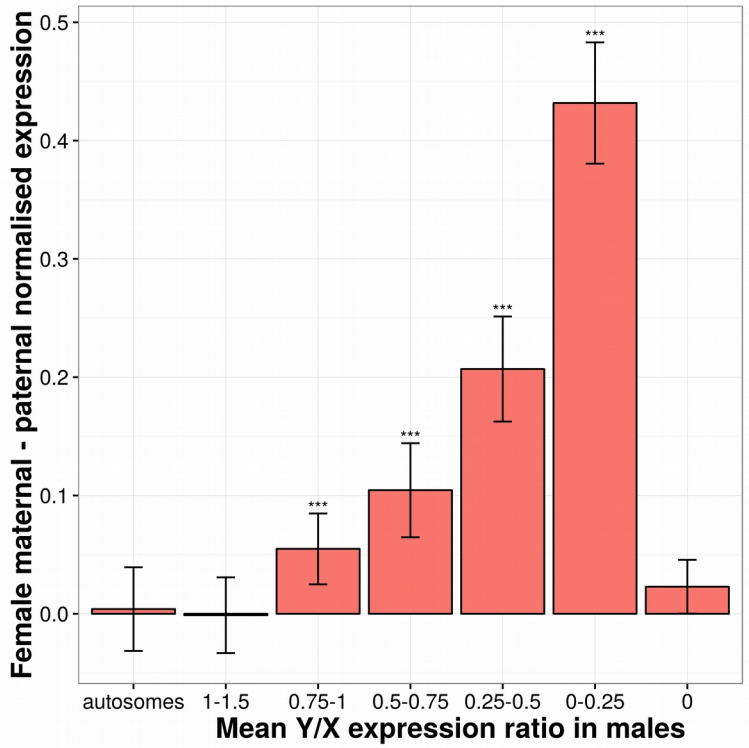
Supplementary Figure S4: Normalised difference in allelic expression levels between *S. latifolia* and the two outgroups without sex chromosomes *S. vulgaris* and *S. viscosa* (Δ), in autosomal and sex-linked contigs that were **validated** (see Materials and Methods), for the **seedling** tissue. Same legend as Supplementary Figure 1 except for sample sizes for the different contig categories: autosomal:77; 1-1.5:71; 0.75-1:82; 0.5-0.75:91; 0.25-0.5:44; 0-0.25:29; 0:89.



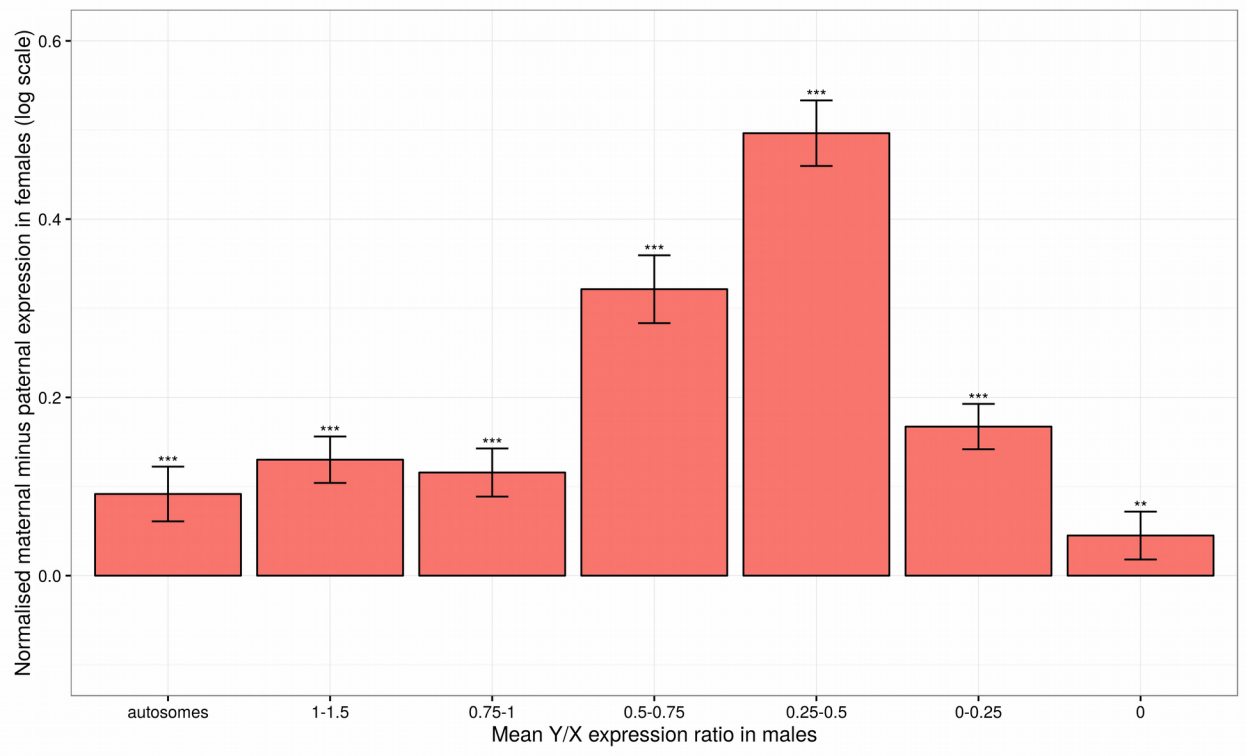
Supplementary Figure S5: Normalised difference in allelic expression levels between *S. latifolia* and the two outgroups without sex chromosomes *S. vulgaris* and *S. viscosa* (Δ), in autosomal and sex-linked contigs that were **validated** (see Materials and Methods), for the **flower bud** tissue. Same legend as Supplementary Figure 1 except for sample sizes for the different contig categories: autosomal:74; 1-1.5:86; 0.75-1:91; 0.5-0.75:67; 0.25-0.5:45; 0-0.25:31; 0:55.



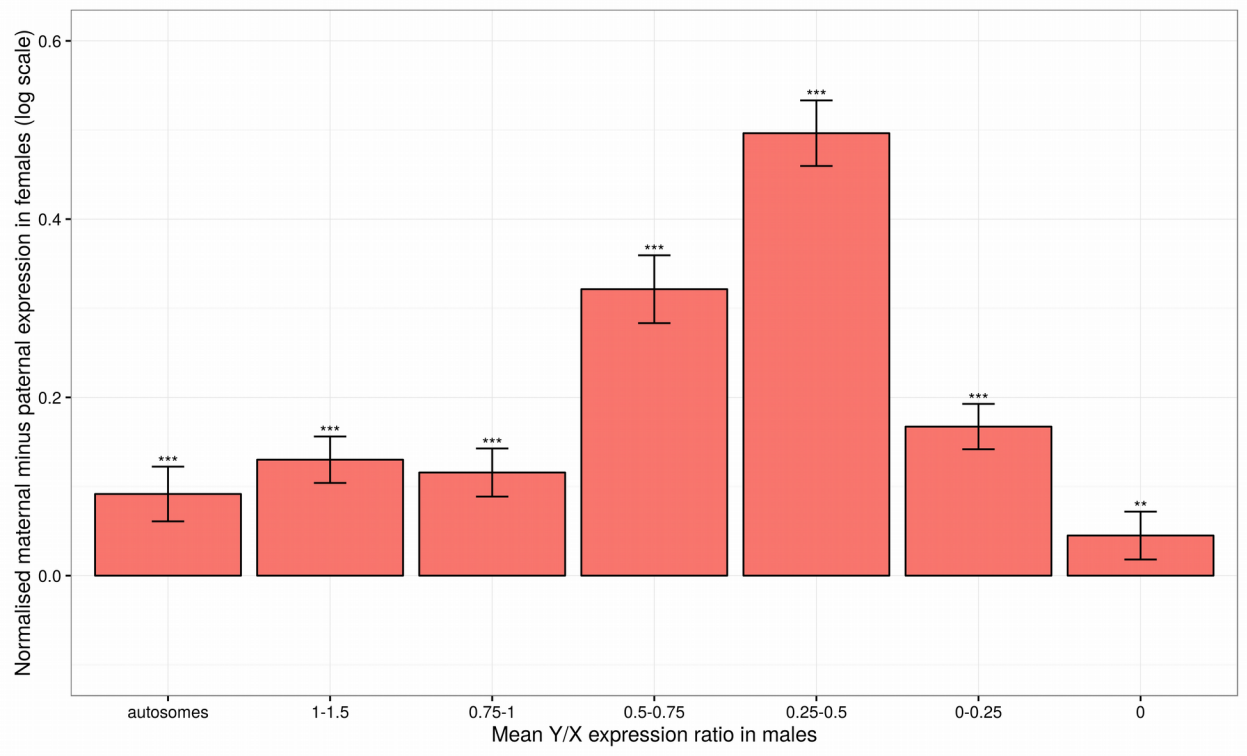
Supplementary Figure S6: Normalised difference in allelic expression levels between *S. latifolia* and the two outgroups without sex chromosomes *S. vulgaris* and *S. viscosa* (Δ), in autosomal and sex-linked contigs that were **validated** (see Materials and Methods), for the **leaf** tissue. Same legend as Supplementary Figure 1 except for sample sizes for the different contig categories: autosomal:79; 1-1.5:84; 0.75-1:74; 0.5-0.75:77; 0.25-0.5:52; 0-0.25:19; 0:119.



Supplementary Figure S7: Normalised expression difference between the maternal and paternal allele in *S. latifolia* females in autosomal and sex-linked contigs for the **seedling** tissue. The Y axis unit is the normalised allelic read count difference in log scale. A linear regression model with mixed effects was used to study allelic expression in *S. latifolia* for every SNP position. In order to measure the changes in *S. latifolia* expression due to sex chromosomes evolution, the outgroup *S. vulgaris* that does not have sex chromosomes was used as a reference in the model (see Materials and Methods for details). The framework provided estimates for the normalised difference between the effect of paternal and maternal origin of alleles in interaction with the contig status (autosomal or sex-linked with various levels of Y degeneration), while accounting for inter-contig and inter-individual variability. See Supplementary Figure S1 legend for sample sizes for the different contig categories and statistical significance symbols. Results show that Y degeneration is linked to a significant expression difference between the paternal and maternal alleles in females, which is not observed in autosomal and non-degenerated sex-linked contigs.



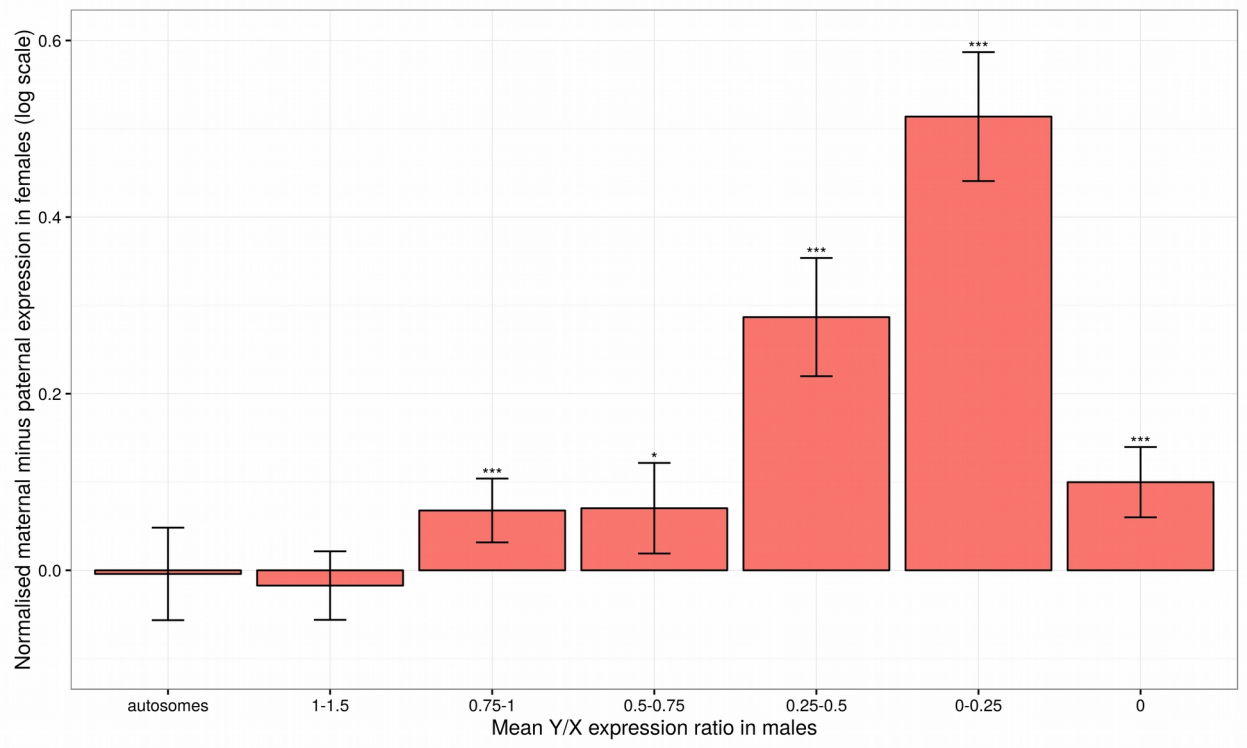
Supplementary Figure S8: Normalised expression difference between the maternal and paternal allele in *S. latifolia* females in autosomal and sex-linked contigs for the **flower bud** tissue. See supplementary Figure 7 for legend and Supplementary Figure S2 for sample sizes for the different contig categories.



570

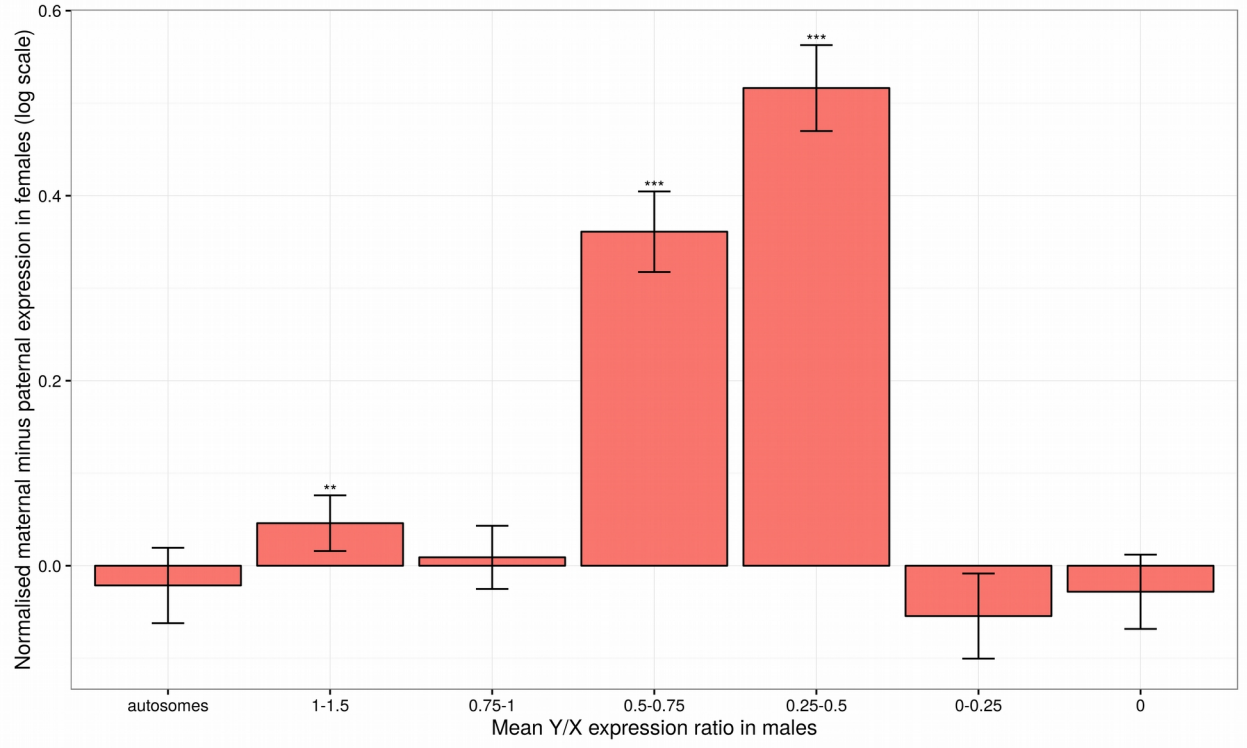
571 **Supplementary Figure S9**: Normalised expression difference between the maternal and paternal
572 allele in *S. latifolia* females in autosomal and sex-linked contigs for the **leaf** tissue. See
573 supplementary Figure 7 for legend and Supplementary Figure S3 for sample sizes for the
574 different contig categories.

575



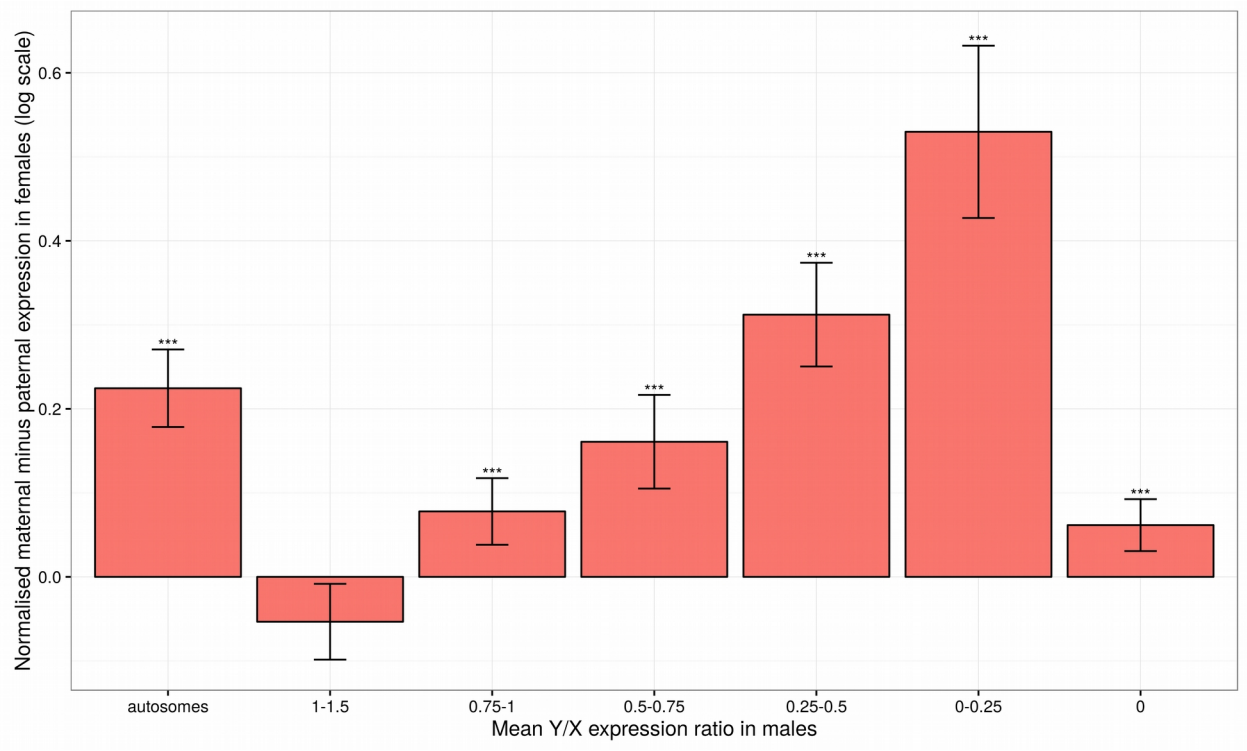
Supplementary Figure S10: Normalised expression difference between the maternal and paternal allele in *S. latifolia* females in autosomal and sex-linked **validated** contigs for the **seedling** tissue. See supplementary Figure 7 for legend and Supplementary Figure S4 for sample sizes for the different contig categories.

583



584 **Supplementary Figure S11**: Normalised expression difference between the maternal and
585 paternal allele in *S. latifolia* females in autosomal and sex-linked **validated** contigs for the
586 **flower bud** tissue. See supplementary Figure 7 for legend and Supplementary Figure S5 for
587 sample sizes for the different contig categories.

588



589

590 **Supplementary Figure S12**: Normalised expression difference between the maternal and
591 paternal allele in *S. latifolia* females in autosomal and sex-linked **validated** contigs for the **leaf**
592 tissue. See supplementary Figure 7 for legend and Supplementary Figure S6 for sample sizes for
593 the different contig categories.

594

595

Supplementary Tables

Supplementary Table S1: library sizes (number of reads) of each individual and mapping statistics.

Supplementary Table S2: Number of contigs after SEX-DETector inferences, removal of sex-bias and selection of validated contigs in the three tissues.

	Tissue type		
	flower buds	**leaves**	**seedlings**
number of ORFs	46178		
Unassigned	33172	33564	33781
Autosomal	11662	11558	11292
X/Y	1140	772	844
X-hemizygous	204	284	261
X/Y non sex-biased	901	733	732
X-hemizygous non sex-biased	103	275	205
X/Y non sex-biased validated	339	345	365
X-hemizygous non sex-biased validated	55	119	89
Autosomal validated	74	79	77

Supplementary Table S3: list of known sex-linked genes in *S. latifolia* and associated literature references.

Author contributions

Aline Muyle, Niklaus Zemp, Alex Widmer and Gabriel Marais conceived the study and experimental design. Niklaus Zemp and Alex Widmer prepared and sequenced the RNA-seq datasets and assembled the reference transcriptome. Aline Muyle ran SEX-DETector on the RNA-seq datasets for the three tissues (including mapping and genotyping steps), produced allelic expression levels, analysed the data, prepared Tables and Figures and wrote the Supplementary Material with inputs from other authors. Niklaus Zemp generated the X chromosome genetic map (with help from Aline Muyle for the mapping and genotyping part). Radim Cegan, Jan Vrana and Roman Hobza did the Y chromosome flow cytometry sorting and sequencing. Clothilde Deschamps did the first assembly of the sorted Y chromosome and

616 improved it with RNA-seq data with the help of Cecile Fruchard. Aline Muyle did the blasts to
617 validate the inferences of SEX-DETector. Raquel Tavares did the GO term analysis. Aline Muyle
618 and Frank Picard did the statistical analyses of the data. Gabriel Marais and Aline Muyle wrote
619 the main text of the manuscript with inputs from other authors.
620