A derived ZW chromosome system in Amborella trichopoda, representing the sister lineage to all other extant flowering plants

Jos Käfer, Adam Bewick, Amélie Andres-robin, Garance Lapetoule, Alex Harkess, José Caïus, Bruno Fogliani, Gildas Gâteblé, Paula Ralph, Claude Depamphilis, et al.

To cite this version:

Jos Käfer, Adam Bewick, Amélie Andres-robin, Garance Lapetoule, Alex Harkess, et al.. A derived ZW chromosome system in Amborella trichopoda, representing the sister lineage to all other extant flowering plants. New Phytologist, Wiley, In press, 233, pp.1636-1642. 10.1111/nph.17662. hal-03370272

HAL Id: hal-03370272
https://hal.archives-ouvertes.fr/hal-03370272
Submitted on 8 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A derived ZW chromosome system in *Amborella trichopoda*, representing the sister lineage to all other extant flowering plants

Jos Käfer¹, Adam Bewick²,³, Amélie Andres-Robin⁴, Garance Lapetoule¹, Alex Harkess⁵,⁶, José Caïus⁷, Bruno Fogliani⁸,⁹, Gildas Gâteblé⁸, Paula Ralph¹⁰, Claude W. dePamphilis¹⁰, Franck Picard¹, Charlie Scutt⁴,⁵, Gabriel AB Marais¹,¹¹,⁶, James Leebens-Mack²,⁶

¹ Université de Lyon; Université Lyon 1; CNRS; UMR 5558; Laboratoire de Biométrie et Biologie Evolutive; F-69622, Villeurbanne, France; ² University of Georgia, Department of Plant Biology, Athens, GA 30602-7271, USA; ³ Bayer Crop Science, Chesterfield, MO 63017, USA; ⁴ Ecole Normale Supérieure de Lyon, CNRS, Laboratoire Reproduction et Développement des plantes UMR 5667, F-69364 Lyon, France; ⁵ Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, AL 36849; ⁶ HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806; ⁷ Univ. Paris-Saclay, Institute of Plant Sciences, Plateforme transcriptOmine de l’IPS2 (POPS), F-91190 Gif-sur-Yvette, France; ⁸ Institut Agronomique néo-Calédonien (IAC), BP 73 Port Laguerre, 98890 Païta, New Caledonia; ⁹ Institute of Exact and Applied Sciences (ISEA), Université de la Nouvelle-Calédonie, BP R4, Nouméa Cedex, 98851 New Caledonia; ¹⁰ Department of Biology and Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA; ¹¹ LEAF-Linking Landscape, Environment, Agriculture and Food, Instituto Superior de Agronomia, Universidade de Lisboa, 1349-017 Lisbon, Portugal

* These authors contributed equally to this work.

Authors for correspondence:

Jos Käfer, Tel.: +33-472447932, Email: jos.kafer@univ-lyon1.fr;

James Leebens-Mack, Tel.: +1-706-583-5573, Email: jleepensmack@uga.edu

• Main text total number of words (including figure legends, excluding summary, references, acknowledgements and supporting information): 2924

• Number of words in the Introduction: 429

• Number of words in the Materials and methods: 467

• Number of words in the Results: 1247

• Number of words in the Discussion: 530

• Number of figures in the main text: 2

• Supporting information documents: 4
Abstract

• The genetic basis and evolution of sex determination in dioecious plants is emerging as an active area of research with exciting advances in genome sequencing and analysis technologies. As the sole species within the sister lineage to all other extant flowering plants, *Amborella trichopoda* is an important model for understanding the evolution and development of flowers. Plants typically produce only male or female flowers, but sex determination mechanisms are unknown for the species.

• Sequence data derived from plants of natural origin and an F1 mapping population were used to identify sex-linked genes and the non-recombining region.

• *A. trichopoda* has a ZW sex determination system. Analysis of genes in a 4 Mb non-recombining sex-determination region reveals recent divergence of Z and W gametologs, and few Z- and W-specific genes.

• The sex chromosomes of *A. trichopoda* evolved less than 16.5 My ago, long after the divergence of the extant angiosperms.

Introduction

Dioecy (separate male and female plants) is one of the many sexual systems exhibited by flowering plant species. Dioecy has evolved repeatedly in angiosperms and c. 6% of all species (i.e. more than 15,000 species) are dioecious (Renner, 2014). Though detailed knowledge of the molecular mechanisms of sex determination in dioecious plants is rather scarce, purely environmental sex determination seems to be rare (Renner, 2014; Charlesworth, 2016), and most dioecious species are therefore thought to have genetic sex determination. However, only c. 50 species with homomorphic or heteromorphic sex chromosomes have been reported, either using light microscopy or classical genetic inheritance studies, and more recently with genomic approaches (Ming et al., 2011; Muyle et al., 2017; Renner and Müller, 2021), and the exact sex-determining genes have been identified in even fewer species (Akagi et al., 2014; Harkess et al., 2017; Akagi et al., 2019; Harkess et al., 2020; Müller et al., 2020). Because dioecy evolved many times independently, different dioecious clades could have different mechanisms of sex determination, so these need to be identified in each clade separately. Very often, the sex chromosomes are only mildly diverged and homomorphic (with some notable exceptions; cf Renner and Müller, 2021).

*Amborella trichopoda* is an understory shrub endemic to New Caledonia and the sole extant species on the sister lineage to the clade including all other angiosperms, which form the most diverse division of plants. It seems clearly established that the most recent common ancestor of all living angiosperms had bisexual flowers, and that the transition to dioecy in the *A. trichopoda* lineage occurred after it diverged from the last common ancestor of all other flowering plants (Endress and Doyle, 2009; Sauquet et al., 2017). Despite the independent origin of dioecy within the *A. trichopoda* lineage, unraveling the mechanism of sex determination in this species could help to shed light on the evolution and development of unisexual flowers in other angiosperms and the evolution of sex chromosomes more generally.

*A. trichopoda* plants grown from seeds collected in the wild exhibit a sex ratio of nearly 50:50 (Anger et al., 2017) making genetic sex determination in the species plausible. However, there are also reports of individual plants exhibiting shifts sex over time (Buzgo et al., 2004; Anger et al., 2017). In fact, the reference genome sequence (Amborella Genome Project, 2013) was derived from a genotype (Santa Cruz Arboretum accession no. 75.3) that was first documented as male but cuttings produced female flowers (Buzgo et al., 2004). Here we
test for a genetic basis for sex determination in *A. trichopoda*, and go on to characterize the sex determination region in the *Amborella* genome.

**Materials and methods**

For the initial inference of sex-linkage, we obtained RNA-seq data of 20 individuals (10 females, 10 males; see Supporting Information Table S1 and Supporting Information, Materials and Methods S1) grown from seeds collected in the wild (hereafter referred to as the “wild population”. Sequence reads were aligned with STAR (Dobin et al., 2013) to the draft chromosomal assembly of the *A. trichopoda* genome (V6.1, available on CoGe, genome ID 50948), consisting of 13 pseudomolecules containing ~95% of the genomic scaffolds. Single Nucleotide Polymorphisms (SNPs) were inferred with reads2snp (Gayral et al., 2013) and analyzed with SDpop, a probabilistic framework based on the modeling of allele and genotype frequencies in populations (Käfer et al., 2021). These RNAseq data from flower buds (between stages 5 and 6 according to Buzgo et al. 2004) were also used to infer sex-biased expression with edgeR (Robinson et al., 2010).

We furthermore used shotgun DNA seq data from an F1 mapping population: parents along with 16 female and 17 male F1 offspring; these data will be referred to as the “F1 mapping population”, see Supporting Information Table S2. These data were also aligned to the V6.1 draft chromosomal assembly (using BWA; Li and Durbin 2009). Genotypes were inferred with freebayes (Garrison and Marth, 2012). Putative W- and Z-specific regions in the assembly were identified using coverage differences in genomic windows of 20 Kpb between females and males. These were used to construct masks of the W- and Z-specific sequences for remapping the sequencing data on the W- and Z-masked genomes (see Supporting Information, Materials and Methods S1).

Using the mappings of both datasets on the masked genome assemblies, we used SDpop and SEX-DETeector++ (Muyle et al., 2016; Badouin et al., 2020) to infer the Z- and W-haplotypes of the detected gametologous genes. Synonymous divergence ($dS$) between these haplotypes was calculated with codeML (Yang, 2007). Because of high sampling variance in gene-wise divergence values (Takahata and Nei, 1985), average $dS$ values per genes using both datasets were calculated. To estimate the time since recombination suppression, we used the maximum average $dS$ value in two ways. First, based on population genetics results, the divergence between two sequences that do not recombine is $2\mu t + \pi$ (Nei and Li, 1979; Takahata and Nei, 1985), where $\pi$ is the nucleotide diversity, $t$ the time since separation and $\mu$ the per-base mutation rate, which is again linked to $\pi$ and the effective population size $N_e$ by $\pi = 4N_e\mu$. We used the estimates of population genetic parameters and generation time obtained by the Amborella Genome Project (2013). Second, we compared the obtained $dS$ value to the estimated divergence of orthologous genes between *Amborella* and *Nuphar* by the One Thousand Plant Transcriptomes Initiative (2019), and scaled this difference using estimates of the crown age of the angiosperms (Salomo et al., 2017; Sauquet and Magallon, 2018).

**Results**

We applied SDpop (Käfer et al., 2021) to the wild population RNA-seq data, and found Z/W segregation of sex-linked SNPs. SDpop inferred 49 ZW gametologs and nine putatively Z-hemizygous genes (Supporting Information, Results S1 and Supporting Information Table S3), which represents 0.28% of all transcribed genes. Of the ZW gametologs, 33 (67%) were located in a single region of c. 4 Mb on chromosome 9 of the reference genome assembly (Fig. 1). Twelve of the genes inferred as ZW occurred on scaffolds that were not assigned
Fig. 1 Sex-linkage inferred from sequencing data in the whole Amborella trichopoda genome (a) and on chromosome 9 (b, c, d). Orange (panel a, outer graphs; panel b): SDpop posterior scores for the ZW segregation type per gene. The black line in (b) indicates the posterior score threshold (0.8) for classification as sex-linked. Green: (panel a, middle graphs; panel c): difference in average nucleotide diversity per gene between females and males. Grey (panel a, inner graphs; panel d): proportion of reads from males for each gene. The black lines in (d) indicate the thresholds (0.30 and 0.63) beyond which the genes were considered putatively sex-linked. In (a), the black lines represent the average in a sliding window of 50 genes.

to 13 chromosomal pseudomolecules, suggesting that they may reside on alternative haplotypes for the non-recombining region of chromosome 9 in the primary chromosomal assembly.

Using the 33 genotyped and phenotyped progeny and their parents from the F1 mapping population, we first assessed whether 671 female-specific SNP variants identified in the wild population transcripts were only present in the seed parent and female progeny. Indeed, males in the mapping population were found to have no more than five of the putatively female-specific variants, whereas females contained at least 465 of the female-specific SNP variants identified in the wild-collected population (Supporting Information Table S2). Thus, the divergence between Z and W genes is conserved among A. trichopoda populations.

We then used the mapping population to estimate the frequency of the sites that were heterozygous in females and homozygous in males, and found a clear excess of female-heterozygous SNPs in the sex-determining region identified by SDpop (Fig. 1). Furthermore, the male-versus-female coverage ratio showed a remarkable pattern in this region, with some genes having equal coverage for both sexes, others no male coverage, and
yet others for which two thirds of the reads were from males. As the existence of sex chromosomes (or any structural heterozygosity) was not considered in the haploid assembly process, the assembled sex determination region placed on chromosome 9 is a chimeric mix of fragments derived from the W and Z chromosomes likely together with regions of consensus sequence for the W and Z haplotypes. Accordingly, divergent regions of the W and Z haplotypes not included in the chromosomal pseudomolecules were assembled in unassigned scaffolds. For the inference of divergence and sex-biased gene expression, we thus obtained two masks for the assembly based on the coverage, one discarding W-specific regions, and the other Z-specific regions of the chromosome 9 assembly and scaffolds that were not included in the chromosomal pseudomolecules (see Supporting Information, Results S1; Supporting Information Fig. S1 and Supporting Information Fig. S2; and Data Files S1 and S2).

Realigning the wild population RNAseq data on these masked assemblies allowed detection of slightly more sex-linked genes using SDpop (Supporting Information Table S3). As these RNAseq data only contain expressed genes (c. 50% of all genes), we used the DNA-reseq data from the mapping population to expand our inference of putatively sex-linked genes. First, we analysed the data with SEX-DETector (Muyle et al., 2016; Badouin et al., 2020) that assigns genes as sex-linked or not based on the transmission of alleles from parents to progeny. When the non-recombining region is small, as is the case here, many genes are in linkage disequilibrium with this region due to the limited number of crossovers in one generation, and SEX-DETector overestimates the size of the non-recombining sex-determination region (Supporting Information Fig. S3). We thus used the divergence between the predicted Z and W haplotypes as a second criterion for inference of sex-linkage: genes with \( dS \) values higher than 0.005 (i.e., 2.5 times higher than the genome-wide heterozygosity) were considered as potential ZW gametologs. Second, we used the coverage data from the mapping-population sequencing: we considered as potentially sex-linked the genes that (Fig. 1d and Supporting Information Fig. S2). Using these criteria, we identified 211 potentially sex-linked genes, of which 140 are on chromosome 9 and 69 on the unassigned scaffolds (Supplementary Data File 3). Among these, stringent analyses of the wild-population and F1 mapping populations data implicated four gene models as W-specific (AmTr_v6.0_c9.20670.1, AmTr_v6.0_c9.20870.1, and AmTr_v6.0_s3897.A.10.1) and three as Z-specific (AmTr_v6.0_c9.20390.1, AmTr_v6.0_c9.20460.1, and AmTr_v6.0_s2752.A.10.1; Data File S3). These assignments are conservative and it is likely that more W- and Z-specific genes may be identified in the future as high-fidelity long read data are generated and used to construct a fully phased diploid assembly.

We calculated the synonymous divergence (\( dS \)) using the mapping on the two masked assemblies, and using the RNAseq as well as the DNA-reseq data (Fig. 2). As the sampling variance of these values is predictably high, we used the average of the two sequence datasets as our best estimate for the \( dS \) value of each gene (cf Takahata and Nei, 1985). The maximum \( dS \) value is between 0.05 and 0.06, depending on the masking of the reference. We calculated the time since recombination suppression based on these values in two ways, either using a population genetics-based (i.e., comparing gametolog divergence to the average level of nucleotide polymorphism) or phylogenetically based (i.e. comparing gametolog divergence to divergence estimates with other species; see Supporting Information, Results S1). The phylogeny-based estimate yields a wider range of divergence-time estimates (3.2 to 16.5 million years ago) then the population genetics-based estimate (7.2 to 8.7 million years ago), but both estimates rely heavily on their respective underlying assumptions (e.g. molecular clock, generation time, past population size) which cannot be directly inferred nor tested.

Using our RNAseq data (from the wild population), we quantified sex-related expression biases. Whereas the number of female-biased genes is approximately equal to the number of male-biased genes genome-wide, male-
These genes could have a role in female sex determination, but their function is currently not known.

Supporting Information Fig. S4). Of the sex-linked female-biased genes, only two almost completely lacked the 59 Mb sex chromosome. The maximum synonymous divergence of current haploid genome assembly, the non-recombining region is estimated to comprise only 0.005 above which genes were considered as sex-linked in the mapping population data.

The mapping population data showed another intriguing pattern on chromosome 9: male offspring were more often heterozygous than female offspring in a region from c. 10 to c. 20 Mb (Fig. 1). Upon closer inspection, it appears that this occurs particularly at SNP positions for which the mother, but not the father, was heterozygous (Supporting Information Fig. S5). This pattern is not typical of pseudoautosomal transmission in regions closely linked to the non-recombining region, nor can it be explained by allele loss (e.g. through lethality) or by increased heterozygosity in the mother (Supporting Information, Results S1; Supporting Information Fig. S6). An obligate recombination event between the left and the right arms of chromosome 9 during meiosis in the mother would be able to cause such a pattern, although this also would imply that the less frequent (minor) alleles all occur on the same copy of chromosome 9. Although a weak deviation from purely autosomal allele-genotype equilibria was also detected in the wild population (Fig. 1), more data (e.g. using different parents for an F1) are needed before a precise mechanistic hypothesis for this pattern can be formulated and tested.

Discussion

We here show that A. trichopoda has sex chromosomes of the ZW type, with female heterogamy. Based on the current haploid genome assembly, the non-recombining region is estimated to comprise only c. 4 Mb long of the 59 Mb sex chromosome. The maximum synonymous divergence of c. 5 to 6% indicates that recombination
suppression occurred rather recently, at least much more recently than the divergence of Amborella and other living angiosperms. Although it might be possible that the first genes for which recombination was suppressed are not expressed or have been lost, the fact that Z and W haplotype sequences were collapsed in the current assembly and that we were not able to unambiguously identify more than a handful of Z- or W- specific genes also suggests that the A. trichopoda sex chromosomes are rather young.

Sex change is well documented in A. trichopoda, in wild as well as in cultivated plants (Buzgo et al., 2004; Grosse-Veldmann et al., 2011; Anger et al., 2017). Our results show that the individual used for reference genome sequencing (Amborella Genome Project, 2013) is genetically female with a Z and W gametologs, although it had produced male flowers in the past. Sex changes are not uncommon in dioecious angiosperms, even in those with clearly heteromorphic sex chromosomes (Ainsworth, 2000). The identification of the non-recombining region should allow researchers to determine the genetic sex for comparison with the phenotypic sex, and help gain insight in the genes involved in sex determination.

Most plant sex chromosomes described so far are of the XY type (Ming et al., 2011; Renner and Müller, 2021), and this observation is considered by some as support for the hypothesis that dioecy evolves from gynodioecy through the occurrence of a recessive female-determining gene and a dominant male-determining gene (Charlesworth and Charlesworth, 1978). However, few sex determination mechanisms have been described at the gene level (cf Renner and Müller, 2021), and there seem to be many different developmental and evolutionary pathways to the production of unisexual flowers (Diggle et al., 2011; Golenberg and West, 2013).

Our inference that the ZW sex determination system is a recently derived trait in A. trichopoda underscores the fact that the species is not ancestral nor “basal”. The Amborella lineage and the ancestral lineage of all other flowering plants diverged from their most recent common ancestor (MRCA) in the early Cretaceous or even late Jurassic (Magallón et al., 2015; Silvestro et al., 2021). Thus, the biology of extant Amborella populations has been molded by more than 130 million years of evolution, since the early diversification of the extant flowering plants. However, A. trichopoda has retained some characteristics that are hypothesized to be shared with the MCRA of living angiosperms (e.g. free floral organs, including an undifferentiated perianth of tepals, and carpels that contain an apical, secretion-filled canal for pollen-tube growth; Endress and Igersheim 2000; Sauquet et al. 2017), and no genome duplication has occurred in the Amborella lineage since diverged from the other angiosperms (Amborella Genome Project, 2013). Thus, further study of the genes involved in sex determination in A. trichopoda might shed light on the regulatory pathways that underlie development of unisexual flowers in other angiosperms.

Acknowledgements

We thank the Environmental Service of the Northern Province of New Caledonia for permission for seed collection and the staff of the IAC for horticultural assistance, John Bowers and the Amborella Genome Project contributors for work on the publicly available chromosomal assembly of the Amborella trichopoda genome used in this study, Nicolas Lartillot and Elise Lucotte for discussions, as well as Susanne Renner and two anonymous reviewers for comments on the manuscript. Amborella genome mapping and analyses were supported by U.S. NSF Plant Genome Research Program (grant 0922742 to CWdP, JLM and co-PIs). This work was also supported by funding from the Agence Nationale de la Recherche (grant number ANR-14-CE19-0021-01 to GABM and CS).
Author contributions


Data availability

The data that support the findings of this study are openly available in SRA, bioprojects PRJNA748676 and PRJNA212863. We used the annotated V6.1 draft genome of *Amborella trichopoda* available on CoGe, genome ID 50948 (genomevolution.org/coge/GenomeInfo.pl?gid=50948).

References


Supporting Information

Additional supporting information may be found in the online version of this article.

Materials and Methods S1 Additional information about the data and analyses of sex linkage.

Table S1 RNAseq statistics for the twenty individuals grown from seeds collected in the wild.

Table S2 Read depths for the individuals from the F1 family sequenced using DNA-reseq data.

Results S1 Detailed description of the results of the analyses.

Table S3 Summary of sex-linkage inference using the RNAseq data from a wild population.

Fig. S1 Masking of putative W- and Z-specific regions in the non-recombining region of chromosome 9.

Fig. S2 Re-seq coverage per gene using mapping on the unmasked and masked genome assemblies.

Fig. S3 Results of the sex-linkage analysis in the mapping population data.

Table S4 Sex-biased gene expression analysis.

Fig. S4 Sex-biased gene expression inferred from RNAseq data from flower buds.

Fig. S5 Distribution of heterozygosity between parents and offspring in the mapping population.

Fig. S6 Allele transmission on chromosome 9.

Data File S1 bed file indicating the putative W-specific regions for masking.

Data File S2 bed file indicating the putative Z-specific regions for masking.

Data File S3 Table (xlsx format) with all putatively sex-linked gene models and gene-level statistics.