Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae)

Ran Zhou[1], David Macaya-Sanz[1], Eli Rodgers-Melnick[1], Craig H. Carlson[2], Fred E. Gouker[2], Luke M. Evans[1], Jeremy Schmutz[3,4], Jerry W. Jenkins[3], Juying Yan[4], Gerald A. Tuskan[4,5], Lawrence B. Smart[2], Stephen P. DiFazio[1]

[1] Department of Biology, West Virginia University, 53 Campus Drive, Morgantown, WV 26506-6057, USA
[2] Horticulture Section, School of Integrative Plant Science, Cornell University, New York State Agricultural Experiment Station, Geneva, NY 14456, USA
[3] HudsonAlpha Institute of Biotechnology, 601 Genome Way Northwest, Huntsville, AL 35806, USA
[4] Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
[5] Biosciences Division, Oak Ridge National Lab, Oak Ridge, USA

Dioecy has evolved numerous times in plants, but heteromorphic sex chromosomes are apparently rare. Sex determination has been studied in multiple *Salix* and *Populus* (Salicaceae) species, and *P. trichocarpa* has an XY sex determination system on chromosome 19, while *S. suchowensis* and *S. viminalis* have a ZW system on chromosome 15. Here we use whole genome sequencing coupled with quantitative trait locus mapping and a genome-wide association study to characterize the genomic composition of the non-recombining portion of the sex determination region. We demonstrate that *Salix purpurea* also has a ZW system on chromosome 15. The sex determination region has reduced recombination, high structural polymorphism, an abundance of transposable elements, and contains genes that are involved in sex expression in other plants. We also show that chromosome 19 contains sex-associated markers in this *S. purpurea* assembly, along with other autosomes. This raises the intriguing possibility of a translocation of the sex determination region within the Salicaceae lineage, suggesting a common evolutionary origin of the *Populus* and *Salix* sex determination loci.

Keywords: Sex, *Salix*, Genome, Suppressed recombination, Dioecy

Introduction

Nearly 90% of flowering plants are hermaphroditic (containing both male and female floral parts in the same flower), and less than 6% are dioecious (separate male and female individuals) (Renner 2014). In angiosperms, dioecy has independently evolved hundreds of times from hermaphroditic progenitors (Renner 2014). Evolutionary pathways to dioecy include gynodioecious, heterostylous, and monoecious intermediates (Lloyd 1979; Ainsworth 2000; Charlesworth 2006), but monoecious intermediates tend to be the most common mechanism in woody angiosperms (Olson et al. 2017). Evolutionary factors favoring dioecy include inbreeding avoidance and the ability to maximize reproductive output through unisexual resource partitioning (Charlesworth and Charlesworth 1978; Charnov 1982; Ashman 2006). The molecular mechanisms of sex determination in plants have only been uncovered for a few species, and this manuscript seeks to add to this body of research by providing an analysis of the genomic region associated with sex determination in the purple osier willow, *Salix purpurea* L. (Salicaceae).

Trait divergence between females and males can be facilitated by the presence of sex chromosomes, as these are the only genomic regions that consistently differ between the sexes (Rice 1984; Mank 2009; Barrett and Hough 2013). Chromosomes harboring a sex determination region (SDR) usually have suppressed recombination and increased haplotype divergence due to independently accumulating mutations, leading to the development of sexual dimorphism at the sequence level (i.e., regions that consistently differ between males and females). The SDR may comprise a majority of the chromosome or only a small portion (Bergero and Charlesworth 2009). Heterogametic SDRs may confer either maleness (XY system), as in *Silene latifolia, Carica papaya, Phoenix dactylifera, Diospyros lotus*, and *Populus trichocarpa*; or femaleness

(ZW system), as in *Fragaria chiloensis, Silene otites*, and *Pistacia vera* (reviewed in Charlesworth 2016; Vyskot and Hobza 2015). Sex chromosomes also contain pseudoautosomal regions (PAR) where sex chromosomes recombine freely and may often show elevated recombination (Nicolas et al. 2005; Otto et al. 2011). Many plant sex chromosomes are homomorphic, exhibiting no strong morphological differences, suggesting that these chromosomes are at an early stage of development (Westergaard 1958; Ming and Moore 2007).

The Salicaceae family is an excellent model system for exploring the ecological and evolutionary dimensions of dioecy and sexual selection in plants. Widely distributed across temperate, boreal, and arctic regions of the globe, these genera represent a diverse assemblage of catkin-bearing trees and shrubs (Karp et al. 2011). There are approximately 30 *Populus* species, most of which are trees that grow in the northern hemisphere (Slavov and Zhelev 2010). In contrast, there are approximately 500 *Salix* species, most of which are shrubs (Dickmann and Kuzovkina 2014). Nearly all species in *Salix* and *Populus* are dioecious, but none have obvious heteromorphic sex chromosomes (Peto 1938). *Salix* is primarily insect pollinated (Karrenberg et al. 2002), and produces complex volatiles and nectar rewards (Füssel et al. 2007). In contrast, *Populus* is almost exclusively wind-pollinated. Furthermore, both lineages share a well-preserved whole genome duplication (Tuskan et al. 2006; Hou et al. 2016) and both show an ongoing propensity toward polyploid formation (Mock et al. 2012; Serapiglia et al. 2015), thus facilitating exploration of the relationship between polyploidy and sex chromosome evolution (Ashman et al. 2013; Glick et al. 2016).

There has been considerable work on characterizing sex determination in *Populus* over the past decade. The SDR has been mapped to the proximal telomeric end of chr 19 in *P. deltoides* and *P. nigra*, both of which are from section *Aigeiros* (Gaudet et al. 2008; Yin et al. 2008) and to a

pericentromeric region of chr 19 in *P. tremuloides, P. tremula*, and *P. alba*, all of which belong to section *Populus* (Pakull et al. 2009; Paolucci et al. 2010; Kersten et al. 2014). In both *P. deltoides* and *P. alba*, the SDR was mapped on a female genetic map but not on a male genetic map, possibly supporting female heterogamety (Yin et al. 2008; Paolucci et al. 2010). In *P. tremuloides* and *P. nigra*, the SDR was mapped on the male genetic map and not on the female genetic map, suggesting male heterogamety (Gaudet et al. 2008; Kersten et al. 2014). Recently, a genome-wide association study (GWAS) on 52 *P. trichocarpa* and 34 *P. balsamifera* found 650 SNPs significantly associated with sex. These sex-associated markers were nearly fixed heterozygous in males and homozygous in females, which is consistent with an XY sex determination system (Geraldes et al. 2015). However, the significant marker associations were not confined to chr 19 but were scattered throughout the genome, possibly due to problems with assembly of the structurally complex SDR (Geraldes et al. 2015).

In contrast to *Populus*, the SDR has been mapped to chr 15 in *S. viminalis* (subgenus *Vetrix*, section *viminella*) and *S. suchowensis* (subgenus *Vetrix*, section *Helix*) (Temmel et al. 2007; Hou et al. 2015; Pucholt et al. 2015). Furthermore, there is a preponderance of female heterozygosity in the SDR of these species, indicating a ZW sex determination system, in contrast to *Populus* (Hou et al. 2015; Pucholt et al. 2015). However, neither study identified candidate genes in the *Salix* SDR that were orthologous to genes in the SDR of *Populus* (Hou et al. 2015; Pucholt et al. 2015). Thus, it is unclear whether *Salix* and *Populus* have different sex determination mechanisms or sex-determining genes, or whether there is a common origin of dioecy in these two genera. In this study, we sought to explore the SDR in an additional Salicaceae species, *Salix purpurea* (subgenus *Vetrix*, section *Helix*). Using robust genome-wide linkage and association analyses and whole genome sequencing, we show that the principal SDR is on chr 15, and that

the genotype configuration in this region is consistent with a ZW system of sex determination. Furthermore, we present evidence that chr 19 is a potential source of the SDR on chr 15 in *Salix*.

Materials and methods

Genome assembly

This work is based on v1.0 of the *S. purpurea* genome (available at http://phyto zome.jgi.doe.gov). In brief, a female diploid genotype of *Salix purpurea* (clone 94006) was collected from the banks of the Fish Creek River in Upstate New York in 1994 (43.2168 N, − 75.6333 W). This clone has been an important parent in *Salix* breeding programs, and is also the source of the reference genome that has been developed by the Joint Genome Institute and a consortium of researchers. ALLPATHS-LG was used to assemble sequences representing ~ 140× coverage of Illumina paired-end sequences, as well as a set of mate-pair libraries (4.5 Kb, 5.3 Kb, 6.5 Kb), producing contigs with an L50 = 46 kb and scaffolds with L50 = 191 kb. The ALLPATHS-LG assembly has a total length of 348 Mb and a total span of 392 Mb (including gaps) but is still relatively fragmented due to a high level of heterozygosity (1 SNP per 120 bp, or 0.8%) and extensive structural variation. Assessment of the assembly quality against willow BACs and transcripts suggested that ~ 78–85% of the willow genome is captured in the current assembly. Gene annotations were accomplished using the Phytozome pipeline (Goodstein et al. 2012). The Repeat-Modeler (v1.0.8) package (http://www.repea tmask er.org) was used to identify and mask repetitive elements.

Genetic mapping and pseudomolecule assembly

An F_1 mapping population was produced by crossing two *S. purpurea* accessions, clone 94006 (female) and clone 94001 (male), and intercrossing two of the resulting progeny (female 'Wolcott' and male 'Fish Creek') to produce over 500 F_2 progeny (referred to as Family 317). The parents and progeny, were genotyped via "Genotyping by Sequencing" (GBS) using *Eco*T221 and *Ape*KI restriction enzymes, and 96-fold multiplexed sequencing on an Illumina HiSeq Genome Analyzer (Elshire et al. 2011). SNPs were identified using the reference-based pipeline of TASSEL (Glaubitz et al. 2014) using the *S. purpurea* v1.0 reference genome. SNPs were also called using the *de novo* UNEAK pipeline from TASSEL (Glaubitz et al. 2014). SNPs were filtered using the following parameters: -hetFreq 0.75 -mnTCov 0.01 -mnSCov 0.2 -mnMAF 0.05 -hLD -mnR2 0.2 -mnBonP 0.005, and < 40% missing data. A total of 8531 informative GBS markers were used to construct genetic maps for 411 F_2 progeny. Markers following expected Mendelian segregation ratios were divided into three groups based on parental genotypes: male backcross (n = 2623), female backcross (n = 2211), and intercross (n = 3697). Each of these marker sets were placed in draft linkage groups based on observed recombinations using an LOD cutoff of 6, as calculated with custom Python scripts. MSTmap (http://www.mstma p.org/) was used to determine initial marker orders, and positions were subsequently refined using the R/qtl Ripple command with the obligate crossover count as an optimality criterion and a window size of 5 (Arends et al. 2010). Final genetic distances were estimated using the Lander–Green algorithm as implemented in R/qtl. These three genetic maps were integrated with the reference genome assembly using custom Python scripts to produce a combined map on which 276 Mb (70%) of sequence scaffolds were anchored, with intervening gaps that were proportional to distances between mapped markers. The remaining unplaced scaffolds contained another 116 Mb of sequence. The assembly was compared to the *Populus*

trichocarpa v3.0 reference genome with LASTZ (v1.03.66), using parameters to exclude alignments between paralogous segments derived from the most recent shared whole genome duplication (gapped, chain, transition, maxwordcount = 4, exact = 100, step = 20).

As an indicator of recombination rate, we calculated the ratio of physical-to-genetic distance between marker pairs using linkage groups with > 30 markers. For each linkage group, pairwise distances were calculated between every N loci, where N was 10% of the total number of loci on the linkage group. For example, if the linkage group had 100 markers, the distance was calculated between all pairs of loci that were separated by 10 loci. Negative and extreme values (ratio > 15) were removed for the purpose of visualization.

SDRs and centromeres are both expected to have suppressed recombination. To differentiate these, we identified approximate locations of centromeres using a two-stage process. First, approximate boundaries of centromeres were defined as areas of low recombination (high physical:genetic distance ratio) on chromosomes. Then, the abundance of different repeat elements was estimated within these intervals, and the ten most abundant elements with significant enrichment (based on Fisher's exact test) were identified as pericentromeric repeats. Finally, based on empirical adjustment of thresholds, we identified centromeres as 100 kb windows with physical:genetic distance ratios of at least 0.22, with centromeric repeats comprising at least 3% of the interval. Windows within 2 Mb of one another were merged to determine the final centromere intervals.

Identification of the sex determination region

Sex was scored for F_2 progeny by repeated observations during the spring of 2012, 2013, and 2015 in common gardens at the New York State Agricultural Experiment Station (Cornell

University) in Geneva, NY. Quantitative Trait Locus (QTL) mapping was performed using the R/qtl package in R with a binary phenotype model (Arends et al. 2010). Logarithm of odds (LOD) support intervals or approximate Bayesian credible intervals were calculated using R/qtl. QTL mapping was performed for all three genetic maps (female backcross, male backcross, and intercross).

We also performed a Genome-Wide Association Study (GWAS) on the sex trait using a population of unrelated individuals collected from the wild. A population of 112 *Salix purpurea* individuals was collected from upstate New York, Pennsylvania, Connecticut, and Vermont and planted in common gardens at Cornell University in Geneva, NY and at West Virginia University in Morgantown, WV. Sex was scored in the spring of 2013 and 2014 for six clonal replicates at each site. The population was genotyped using GBS with the *Ape*KI restriction enzyme and 48-fold multiplex sequencing on an Illumina HiSeq Genome Analyzer. SNPs were called and filtered as described above, yielding 85,543 SNPs for analysis. A kinship matrix was calculated using the scaled Identity-by-State (IBS) method implemented in the EMMAX package (Kang et al. 2010). Clonal ramets were identified based on pairwise IBS values in comparison to pairwise IBS of the F_2 population described above (Fig. S1). This resulted in removal of 37 ramets belonging to 9 clonal groups. Fifteen individuals with inconsistent sex phenotypes across replicates were also excluded from this analysis. Repeated phenotyping failed to detect true hermaphrodites among most of this group. Furthermore, inclusion of the hermaphrodites with an intermediate phenotype in the QTL analysis did not substantively change the results of the association analysis, so we elected to drop them from the analysis. This left a total of 38 females and 22 males. To control for the influence of population structure, a principal components analysis (PCA) was performed using smartPCA in the Eigenstrat package (Price et

al. 2006). GWAS for sex was performed with the first two principal components and the kinship matrix as covariates using a mixed linear model implemented in the EMMAX package (Kang et al. 2010). We controlled for multiple testing using a Bonferroni correction with an alpha value of 0.05. We defined the physical SDR intervals based on all GWAS loci that passed the Bonferroni correction. Significant loci that occurred within 1 Mb on the same chromosome were merged into the same interval.

Characterization of the W chromosome in the SDR

Given that the reference genome was derived from a female clone, and that closely related *Salix* species show female heterogamety (Hou et al. 2015; Pucholt et al. 2015), we expected to see strong evidence of haplotype divergence in the *S. purpurea* SDR. Since ALLPATHS-LG generates genome assemblies that consist of chimeras of the two haplotypes from a heterozygous diploid genome (Gnerre et al. 2011), we expected the SDR to include segments of Z and W chromosomes. Ideally these female-specific segments would be identified based on the presence of female-specific alleles in the association population. If the W and Z chromosomes are divergent enough to prevent alignment of short-read sequences, markers derived from such alignments should be apparently homozygous (but actually hemizygous) in females, and null in males. However, due to the relatively low density of the GBS markers, this analysis is likely to miss intervals and unmapped scaffolds derived from the W chromosome that happen to lack GBS markers. We therefore used relative depth of coverage of female and male sequences as a complementary approach for identifying divergent W-derived sequences. For Z portions of the reference genome, male coverage should be roughly double that of the female for divergent portions of the SDR, whereas for W portions of the reference, coverage should be approximately

0.5X compared to the rest of the genome for the female, and there should be very low coverage in males.

To perform this depth-based assessment, we resequenced clone 94006 (the reference) and her male offspring, clone 'Fish Creek' (also father of the F2 mapping family) using 2 × 250 bp reads on an Illumina HiSeq sequencer. This yielded 106,305,281 paired reads (53 Gb) and 92,077,639 paired reads (46 Gb), respectively, for expected depth of 135× and 117×, respectively. These were aligned to the 94006 reference genome using Bowtie2 with the parameters -D 15 -R 2 -N 0 -L 20 -i S,1,0.75. SNPs were identified using the mpileup function of samtools, followed by bcftools with the parameters -g 1 -O v -m. We evaluated depth of coverage for the female reference and the male offspring using raw output from the samtools mpileup command.

We used polymorphisms identified from these alignments to construct representative female-specific reference sequences using alleles that occur in the female clone 94006 but which were absent in male clone Fish Creek. Although not explicitly phased, these approximations of the W haplotypes represent the maximum possible divergence between Z and W alleles for these individuals. Coding sequences containing female-specific polymorphisms (here called "W-type") were created using the FastaAlternateReferenceMaker module of the GATK package (DePristo et al. 2011). Genes with nonsense and frameshift mutations were then removed as possible pseudogenes. Finally, synonymous polymorphisms were estimated for all pairs of predicted transcripts using the '*yn00*' module in the PAML package (Yang 2007). The reference genome transcripts were compared to those containing female-specific polymorphisms as well as to those containing all alternative alleles.

All predicted proteins in the *S. purpurea* reference genome annotation were compared to the UniProt database (http://www.unipr ot.org/) using blastp and against the Pfam database (http://pfam.xfam.org/) using HMMER, with default parameters. Protein mapping results were submitted to Argot^2 (Falda et al. 2012) to obtain Gene Ontology (GO) annotations, using a stringent cutoff (Total Score = 1500) to filter Type I errors. We used Fisher's Exact Test to identify overrepresented GO terms for candidate genes in the SDR. All orthologs between *S. purpurea* and *P. trichocarpa* were retrieved from Phytozome (https ://phyto zome.jgi.doe.gov/). Synonymous (dS) and nonsynonymous (dN) substitution frequencies were estimated for each pair of primary transcripts from each species using the '*yn00*' module in the PAML package (Yang 2007). Pairs with dS > 0.4 were dropped, assuming they were incorrectly defined as orthologs. In total, 33,789 ortholog pairs were compared, including 27,118 genes from *S. purpurea* and 24,000 genes from *P. trichocarpa*.

Gene expression was evaluated using RNA sequencing for actively growing shoot tips for five male and five female progeny from the family used for QTL analysis. Detailed methods are described in Carlson et al. (2017). In brief, total RNA was extracted using the Spectrum™ Total Plant RNA Kit. Libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit. Libraries were sequenced on the Illumina HiSeq platform (1 × 100 bp) yielding an average of 17.9 million mapped reads per sample. Reads were mapped to the *S. purpurea* reference genome v1.0 using the CLC Genomics Workbench, and differential expression analyses were performed using EdgeR.

Results

Localization of the SDR to chromosome 15

Among the 396 phenotyped and genotyped individuals in the F_2 family, there were 234 females and 162 males. This ratio is significantly skewed toward females (*F:M* = 1.44; χ^2 = 13.1; *df* = 1; *P* < 0.001). QTL mapping identified sex-associated markers principally on chr 15 for all three maps (Fig. 1; Table S1). On the female map, 125 markers were linked to sex, 105 of which were on chr 15, spanning from 225.42 to 240.17 cM (Table 1). On the male map, only five markers were linked to sex, four of which were in the interval from 326.48 to 347.17 cM on chr 15 (Fig. 1; Table 1). An additional 50 markers were linked to sex on the intercross map, covering an interval of about 2.6 cM, all on chr 15 (Fig. 1; Table 1). Based on anchoring mapped markers to physical positions in the *S. purpurea* genome assembly, the potential SDR can be mapped to two regions on chr 15 ranging from ~ 0.4 to 1.9 Mbp and from ~ 10.9 to ~ 15.1 Mbp.

One additional sex-linked marker was located at the proximal end of chr 19 on the male map, with a LOD score of 4.68 (Fig. 1; Table S1). However, mapping failed entirely for chr 19 for female backcross markers, the only chromosome for which this was the case. Chromosome 19 had the lowest density of GBS markers in the genome (Table S2). Furthermore, this chromosome had the lowest proportion of markers in a female backcross configuration, and the highest proportion of markers with severe segregation distortion (Fig. S2; Table S2).

To confirm the location of the SDR in a diverse population, a GWAS for sex was performed using naturalized *S. purpurea* accessions collected from northeastern North America. Of the 60 genets that were unambiguously phenotyped for sex, 38 were female and 22 were male, which is a significantly female-biased sex ratio (*F:M* = 1.73; χ^2 = 4.3; *df* = 1; *P* = 0.02). Of the 85,543 SNP markers that passed filtering, 72 were significantly associated with sex ($P < 5.85 \times 10^{-7}$, Fig. 2; Fig. S3). Among these markers, 41 were located on chr 15, from 10.7 to 15.3 Mb, and four were located at the distal portion of chr 15 (1.9 Mbp). Thus, the primary SDR identified by

GWAS overlaps with those mapped by QTL in the F_2 family (Fig. 3). In addition, six markers from chr 19 at ~ 69 kb were also significantly associated with sex (Fig. 2), which also corresponds with the QTL results. In addition, there were minor peaks on chrs 1, 2, 3, and 5, and there were six scaffolds containing a total of 13 significant sex-associated markers that were not anchored to the genetic maps (Table S3).

To evaluate whether these secondary chromosomal peaks could have been due to assembly errors, we aligned these SDR sequences to the *S. purpurea* reference genome using blastn. None of these chromosomal loci shared homology with the chr 15 SDR (Table S4). The peak on scaffold1293 did match chr 15, and three of the chromosomal regions matched other unplaced scaffolds (Table S4). This would be expected if the aligned sequences were derived from divergent haplotypes that were not included in the main genome assembly (e.g., sequences derived from W haplotypes). We also compared these SDR sequences to the *Populus trichocarpa* v3.0 reference genome using blastn. The SDRs on chrs 1,2, and 5 had best hits to the same chromosomes in *P. trichocarpa*. However, the SDRs on chrs 3 and 19 had best hits to scaffold_25 in *P. trichocarpa* (Table S4). Because the SDR is known to be poorly assembled in the *P. trichocarpa* v3.0 assembly (Geraldes et al. 2015), we aligned scaffold_25 to the *P. trichocarpa* v1.0 assembly and found that it matched primarily to chr 19, positions 751 to 1040 kb, which coincides with the main *P. trichocarpa* SDR (Geraldes et al. 2015). Therefore, the QTL and GWAS results both indicate that sequences homologous to the *P. trichocarpa* SDR retain evidence of sex dimorphism in *S. purpurea*.

Salix purpurea *has a ZW system of sex determination*

Under Mendelian segregation, the frequency of heterozygotes should be 0.5 for both male and female F_2 progeny. However, sex-associated markers were heterozygous in 64% of female progeny on average, but only in 12% of male progeny (Table S1; Fig. 3). Similarly, sex-associated SNP loci were heterozygous in 79% of females in the association population on average, but only in 5% of males for these same loci (Fig. 4, Table S3, Fig. S4). This difference was significant based on a *t* test ($P < 2.2 \times 10^{-16}$). Both observations are consistent with a female heterogametic (ZW) system of sex determination, where females should be nearly fixed heterozygous for female-specific portions of the SDR, while males should be homozygous for those same loci. This is due to the typically biallelic nature of SNP polymorphisms, where polymorphic alleles from the W chromosome are identical by descent and therefore only occur in females. The discrepancy between the observed values and the expected fixed heterozygosity in females is likely due to null alleles caused by allele dropout and/or inadequate sequencing depth for the GBS markers (Andrews et al. 2016).

Since our reference sequence was derived from a female, we expected that the assembly could contain hemizygous or highly divergent portions of the W chromosome. We used two complementary approaches to determine the size and extent of these regions: the presence of female-specific alleles at the GBS markers in the association population, and relative depth of sequence coverage in the female reference and her male progeny (see Methods). Candidate W segments contained a large proportion of GBS markers that were homozygous in females and mostly lacking genotype calls (i.e., double null markers) in males in the association population (Fig. S5). We identified 231 of these W-type markers (0.27%) (Fig. 4; Table S5). Of these, 51 occurred on chr 15, another 158 occurred on 20 unanchored scaffolds, and the remaining 22 occurred on small segments of chrs 3, 5, and 7. On average, 80% of females were apparently

homozygous for these markers (presumably due to hemizygosity or divergence of W segments), whereas 85% of males had null alleles at these loci (Fig. 4, Table S5). The putative W haplotypes were interspersed along chr 15, suggesting that the genome assembly is a chimeric representation of the Z and W haplotypes (Fig. 4; Table S5).

We also identified putative hemizygous W chromosome segments in the reference genome based on depth of coverage of a male and female individual. If females are heterogametic and the non-recombining regions of the SDR are sufficiently diverged, then there should be regions in the female reference that are not covered by reads from a male individual. Aligning paired 250 bp Illumina sequences from a male offspring ('Fish Creek') of clone 94006 back to the female reference assembly, yielded a very high alignment rate of 95.19% compared to 96.67% when clone 94006 was aligned to itself. Nevertheless, after excluding known repeats and gaps, there were 22,733 regions totaling 7.69 Mb on chromosomes and another 6.87 Mb of unanchored scaffolds that had coverage in the female but lacked coverage in the male (Table 2; Fig. S6). These analyses identified 222 scaffolds comprised of > 30% female-specific sequences (Table S5). Some of these are likely caused by insertion/deletion polymorphisms that are not sex-specific. However, we identified 11 scaffolds that were also identified as putative W segments based on allelic configurations (see above). Portions of five of these scaffolds had high sequence similarity to chr 15, supporting the contention that these are alternate haplotypes from the SDR. For example, Scaffold0265 is 298 kb in length and contains 38.9% female-specific sequence and 20 W-type GBS markers (Table S6). This scaffold also contains three sex-associated markers identified in the GWAS. Cumulatively, these 11 scaffolds covered 1.04 Mb, which is a reasonable lower limit for the size of the divergent portions of the SDR.

The SDR is highly repetitive, has repressed recombination, and is divergent from the **Populus** *SDR*

The largest SDR on chr 15 of *S. purpurea* (10.7–15.3 Mb) overlaps with a large region (9.8–16.2 Mb) with elevated physical-to-genetic distance ratio of 0.867 Mb/cM, compared to the genome-wide average of 0.172 Mb/cM (Fig. 5), which indicates reduced recombination. This interval contained high repeat abundance relative to the rest of the genome (Fig. S7). To differentiate the SDR from the centromere, we identified centromeric intervals based on physical:genetic distance and abundance of centromere-associated repeats. All chromosomes except 10 and 14 showed centromeric regions based on these criteria (Fig. 5; Fig. S8). As expected, these intervals contained high repeat abundance and low gene content relative to the rest of the genome (Fig. S9). The SDR on chromosome 15 largely overlapped with the centromere, so these regions cannot be readily differentiated. However, there were several large stretches within the chromosome 15 SDR that have high gene density and low repeat abundance (Fig. 5), suggesting that the SDR contains euchromatic sequence as well as heterochromatic centromeric sequence.

A portion of the SDR in *S. purpurea* is homologous to the SDR in *S. suchowensis*. The *S. suchowensis* SDR primarily occurs on scaffold64, an ~ 900 kb scaffold that maps to chr 15 (Hou et al. 2015). Aligning this sequence to the *S. purpurea* genome with lastz, we observed homology from 6.2 to 7.3 Mb and from 14.1 to 15.1 Mb on *S. purpurea* chr 15 (Fig. S10). The latter sequence overlaps with a portion of the *S. purpurea* SDR. In contrast, the *S. viminalis* SDR matches from 5.9 to 8.4 Mb on *S. purpurea* chr 15, which is outside the *S. purpurea* SDR (Pucholt et al. 2017b).

P. trichocarpa is another member of the Salicaceae and has a fairly well characterized XY system of sex determination (Geraldes et al. 2015). In general, *S. purpurea* and *P. trichocarpa* have high synteny at the chromosome scale (Fig. 6), but chr 15 in *S. purpurea* stands out in several ways. First, the SDR on chr 15 of *S. purpurea* is not syntenic with chr 15 or any other chromosome of *P. trichocarpa* (Fig. 6). Second, the proportion of repeats is significantly elevated in the *S. purpurea* SDR, with an average of 37% repeat composition, compared to the genome-wide average of 24.8% (Welch's Two-Sample $T = -4.6$ $P5948, < 0.0001$; Table S7; Fig. S7). Chr 19, which contains the SDR in *P. trichocarpa*, also had the highest average repeat content in *S. purpurea* (33.5%, compared to 25.1% genome-wide average) (Table S7).

Gene content of the SDR

We identified 251 protein-coding genes within the *S. purpurea* SDR (Table S8). A GO enrichment analysis based on 203 genes annotated with GO terms identified four significantly enriched terms (Bonferroni adjusted $P < 2.45 \times 10^{-4}$), all of which were related to microtubule functions. These include microtubule-based movement (GO:0007018), microtubule motor activity (GO:0003777), and microtubule binding (GO:0008017), as well as kinesin complex (GO:0005871) (Table 3). This enrichment is partly due to two pairs of tandemly duplicated kinesin-like genes in the SDR (Table S8).

The SDR contains 20 genes that have > 70% female-specific sequence (read coverage in the female, but not the male), and many of these genes also show sex-biased expression in developing stem tissue in *S. purpurea* (Table S8; Carlson et al. 2017). These include an extracellular calcium-sensing receptor (SapurV1A.0301s0080), an auxin response factor (SapurV1A.0718s0100), a peptidase M50B-like protein (SapurV1A.0475s0170), a zinc finger

C3hC4 type transcription factor (SapurV1A.0301s0170), and a reticulon-like protein (SapurV1A.0530s0130). Among these, only the reticulon-like protein showed an elevated dN/dS ratio when compared to *P. trichocarpa* (0.687, versus a genome-wide average of 0.406). Of the 14 genes that showed significant female-biased expression in the SDR, only one lacked female-specific sequence (SapurV1A.1386s0030, a small heat shock protein). No genes showed significant male-biased expression after Bonferroni correction.

Multiple other chromosomes showed sex associations, but the sex-associated region of chr 19 is of particular interest, since it overlaps with the SDR of *P. trichocarpa*. This region spans approximately 10 kb in the current assembly, and harbors three small genes. SapurV1A.1005s0060 contains a Small MutS-Related (SMR) domain. A second gene, SapurV1A.1005s0050, is a calcium-dependent kinase with two EF-Hand domains. The third gene, SapurV1A.1005s0070, encodes a hypothetical protein (Table S8). None of these genes have sex-biased expression or unusual dN/dS ratios compared to *Populus* (Table S8).

We attempted to estimate the relative age of the region of suppressed recombination based on synonymous coding sequence polymorphisms of W alleles compared to Z alleles in the SDR. Calculated this way, the frequency of Z-W synonymous polymorphisms within the SDR was 0.00343 substitutions per synonymous site, while the frequency calculated the same way outside of the SDR was 0.00151. These differences were statistically significant ($t = -4.099$; $df = 249$; $P = 5.63e{-}05$). To test whether this difference was due to higher overall polymorphism in the SDR, we calculated the frequency of all observed polymorphisms based on these two individuals (i.e., including those that were polymorphic within the male as well). Genes within the SDR showed similar overall frequency of synonymous polymorphisms (0.00616 substitutions per synonymous site) compared to genes outside the SDR (0.00607), and the difference was not

significant ($t = -0.077$; $df = 235$; $P = 0.938$). There was no evidence of evolutionary strata in the SDR based on lack of clustering of genes with similar dS values.

Discussion

The S. purpurea *SDR is similar to other* Salix *species and divergent from* Populus

In all three of the *Salix* species studied thus far, *S. viminalis* (Pucholt et al. 2015), *S. suchowensis* (Hou et al. 2015; Chen et al. 2016), and now *S. purpurea*, the largest SDR is on chr 15, and shows clear female heterogamety. Furthermore, the *S. suchowensis* SDR overlaps with a portion of the *S. purpurea* SDR, but the *S. viminalis* SDR does not. This may reflect the evolutionary distinctness of *S. viminalis* from the other two taxa. Based on morphological characters, *S. viminalis* belongs to section *Viminella*, which is strongly differentiated from section *Helix*, which contains *S. purpurea* (Argus 1997) and *S. suchowensis* (Dickmann and Kuzovkina 2014). This is similar to the situation in *Populus*, where the location of the sex determination region varies across different sections of the genus, though all are located on chr 19 (Gaudet et al. 2008; Pakull et al. 2009, 2014; Paolucci et al. 2010; Tuskan et al. 2012; Kersten et al. 2014; Geraldes et al. 2015). Comparison of the sequence composition of the *Salix* SDRs and the *P. trichocarpa* SDR revealed no extensive stretches of homology, suggesting a largely independent evolution of these genome regions (Hou et al. 2015; Pucholt et al. 2017a). Clearly, the SDR is highly dynamic within this family, and it is also important to point out that relatively short but nevertheless important stretches of shared homology may be missed due to the fragmentary assemblies of these structurally complex genome regions.

The alternative peaks from the GWAS analysis on chrs 1, 2, 3, and 5 were not upheld by the QTL analysis, and mainly consisted of isolated markers. This is unlikely to represent a case of multi-locus sex determination (Moore and Roberts 2013), as the evidence is weak since there is little other corroborating information. The peaks on chrs 2, 3, and 5 consisted of solitary markers, while that on chr 1 included 5 markers that occurred within a 1 kb interval. Our results are similar to those in *P. trichocarpa*, which also contained multiple secondary GWAS peaks in a sex determination GWAS (Geraldes et al. 2015). While some of the secondary *Populus* peaks appear to be assembly and/or alignment artifacts (Geraldes et al. 2015), we found no evidence of assembly errors in these regions for *S. purpurea* based on examining the sequence assembly itself as well as the underlying genetic map. Problems with assembly of SDRs are common, presumably due to strong haplotype divergence and high repeat composition, which impede assembly of short-read sequence data (Miller et al. 2010). Furthermore, the suppressed recombination in these regions inhibits map-based assembly methods.

An alternative explanation for the secondary peaks is recent translocation by duplication from autosomes to the SDR in *S. purpurea*. If the portions of the W haplotype are not represented in the reference genome assembly, then the reads derived from the recently translocated regions could align to their original locations and be incorrectly scored as polymorphisms (Qi et al. 2014). Short-read sequence aligners such as Bowtie2 do not handle repetitive sequences well, and commonly misalign reads derived from such regions (Lian et al. 2016). We believe that this is the most likely explanation for the sex-associated peaks occurring at loci outside of the main SDR on chr 15. It is much less parsimonious to assume that multi-locus sex determination is

occurring in this species, given the expected evolutionary instability of such a system (Beukeboom and Perrin 2014).

Nevertheless, the GWAS peak on chr 19 is especially interesting because it coincides with the position of one of the SDRs in *Populus*. This peak also has more corroborating evidence than the other secondary peaks because it had one of the lowest observed P values, and it is recapitulated in the QTL analysis. Furthermore, the peak on chr 3 best matches a scaffold from the SDR region of *Populus* on chr 19, so at least two independent association results point to sex-specific genotypes in genomic segments with homology to the *Populus* SDR. If these represent recent translocations, then this could be a clue to the origin of the chr 15 SDR in the *Salix* lineage.

Recombination suppression and relative age of the SDR

Reduced recombination is a crucial component of sex chromosome evolution which ensures that male and female sterility factors do not co-occur in the zygote (Bergero and Charlesworth 2009; Ming et al. 2011). As expected, we observed reduced recombination across most of the SDR in *S. purpurea* (Fig. 5). This could be caused by large-scale structural polymorphisms and reinforced by the accumulation of nonhomologous sequences in the female-specific haplotype (Ming et al. 2011; Charlesworth 2015). The SDR also shows a higher proportion of repetitive elements, as expected in regions with reduced recombination. Similar features are also apparent within the SDR of *S. suchowensis* and *S. viminalis* (Hou et al. 2015; Pucholt et al. 2015; Chen et al. 2016), but are not as apparent for the *P. trichocarpa* SDR, which is estimated to be quite small (Geraldes et al. 2015). If this is accurate, it could indicate that the *P. trichocarpa* region has not yet developed these features, or that it is highly dynamic. In the case of *S. purpurea*, the SDR is quite large, with a lower limit of 1.04 Mb (based on the cumulative length of female-

specific scaffolds), and an upper limit of approximately 5 Mb, based on suppressed recombination and the occurrence of SNPs that are significantly associated with sex. It is possible that the SDR overlaps with the centromere on chr 15, and this could contribute to the large apparent size of the region of suppressed recombination. However, the SDR does not contain any of the tandem minisatellite repeats that are apparently characteristic of the *S. purpurea* centromeres, as identified in a previous study (Melters et al. 2013). It remains to be seen if the lack of these repeats is due to poor assembly, or if the centromere is located elsewhere on this chromosome.

Divergence between Z and W transcripts in the *S. purpurea* SDR is relatively low, suggesting that suppression of recombination is incomplete or recently established. This is similar to the SDRs of *P. trichocarpa* (Geraldes et al. 2015) and *S. viminalis* (Pucholt et al. 2017b), which also show low divergence of sex-specific sequences. Furthermore, we saw no evidence of the presence of evolutionary strata within or around the *S. purpurea* SDR. Such features occur due to the establishment of regions of suppressed recombination at different times during sex chromosome evolution (Charlesworth 2016). Evolutionary strata are apparent in well-established SDRs of other plants, including *Silene latifolia* (Bergero et al. 2007) and *Carica papaya* (Wang et al. 2012). However, no such regions were detected in *S. suchowensis* (Pandey and Azad 2016). Given the low divergence, lack of strata, and the frequent movement of the SDR within the family, it is reasonable to conclude that the SDR is highly dynamic in this family, and that sex determination loci frequently translocate to new positions and/or are superseded by other loci on autosomes, as predicted by theoretical models of SDR movement (van Doorn and Kirkpatrick 2007, 2010).

Candidate genes and their function

The SDRs are genomic regions that are statistically associated with gender. This association must be due to the presence of loci that control sex determination, but the regions also likely harbor loci that are under sexually antagonistic selection (van Doorn and Kirkpatrick 2007; Bachtrog et al. 2014). The gene content of these regions could, therefore, provide insights about mechanisms of sex determination as well as sex dimorphism. We identified 251 protein-coding genes in the SDRs of *S. purpurea* (Table S8). Most have not been functionally annotated, but clues can be inferred based on conserved domains and their predicted function in model organisms. It is also important to note that the assembly problems mentioned previously have probably prevented full enumeration of the gene content of the SDRs. This problem may be particularly challenging for female-specific portions of the W chromosome (Pucholt et al. 2015). Nevertheless, there are several genes in this region that could plausibly be involved in floral development and sex-specific regulation that are worthy of consideration.

Since floral morphology is the most striking difference between the sexes, it is reasonable to expect that genes involved in floral development would be located in the SDRs. Indeed, the SDR contains SapurV1A.0718s0010, an ortholog of WUSCHEL-related homeotic genes (e.g., *WOX1*). Orthologs in other species, including STF in *Medicago truncatula*, LAM1 in *Nicotiana sylvestris*, and MAW in *Petunia*, are key regulators of the lateral outgrowth of leaf blades and floral organs (Lin et al. 2013). This gene showed slightly elevated expression in male shoot tips compared to female shoot tips (Table S7).

Several genes in the SDR may be involved specifically with male development and function. For example, our analysis of GO term over-representation highlighted the presence of seven genes containing the kinesin motor domain (PF00225), which is involved in microtubule-based movement or organelles, including during pollen tube growth (Cai and Cresti 2009). For

example, loss-of-function mutants of the closest homolog of SapurV1A.0530s0110 in *Arabidopsis thaliana* (*NACK*1) showed reduced growth and prematurely terminated petals, pistils, and stamens (Nishihama et al. 2002). Since there is only one homolog of these kinesin-like genes in *P. trichocarpa*, it appears that this expansion occurred after the divergence of the two genera, a scenario supported by high sequence conservation between the tandem duplicates (Fig. S11).

The SDR on chr 19 deserves special attention due to its shared homology with the *Populus* SDR. One particularly interesting gene in this region is SapurV1A.1005s0060, which contains a Small MutS-Related (SMR) domain and a domain of unknown function (DUF1771). These domains frequently occur together in eukaryotes, but the function of DUF1771 has yet to be characterized (Fukui and Kuramitsu 2011). Proteins with the SMR domain, such as MutS2, can suppress (Fukui et al. 2007; Fukui and Kuramitsu 2011) or promote (Burby and Simmons 2017) homologous recombination by endonucleolytic digestion, and are involved in mismatch repair in diverse prokaryotes (Kunkel and Erie 2005). The roles of the SMR domain in plants are not fully characterized, but when coupled with the pentatricopeptide repeat motif, the SMR domain shows sequence-specific RNA endonuclease activity and affects chloroplast function (Zhou et al. 2017). Due to its potential roles in recombination, mismatch repair, and regulation of organellar function, this gene is an intriguing candidate in the context of sex determination as well as mediation of the female-biased sex ratios that are commonly observed in *Salix* (Alliende and Harper 1989; Alstrom-Rapaport et al. 1998; Ueno et al. 2007; Pucholt et al. 2017a), including in *S. purpurea*, as reported here.

Sex chromosome evolution in the Salicaceae

Populus and *Salix* are closely related genera that share many key characteristics, the most notable of which is that they are both nearly fixed for dioecy. *Populus* first appears in the fossil record between 40 and 60 MYA, apparently slightly earlier than *Salix* (Boucher et al. 2003). However, *Populus* and *Salix* exhibit much less divergence in nucleotide sequence and chromosome structure than expected, presumably due to long average generation times (Sterck et al. 2005; Hou et al. 2016). It may therefore seem surprising that the chromosomal location and gene content of the SDRs are so different, and that they have different heterogametic configurations (Hou et al. 2015; Pucholt et al. 2015). In fact, movement of sex determination loci and transitions between XY and ZW systems are well-known in organisms that lack strongly differentiated, heteromorphic sex chromosomes (Bachtrog et al. 2014).

A striking finding of this study is the existence of multiple loci with strong associations with sex, one of which is on chr 15 and shared with other *Salix* species (Pucholt et al. 2015; Chen et al. 2016), and one on chr 19, which harbors the SDR of multiple *Populus* species (Tuskan et al. 2012; Kersten et al. 2014; Geraldes et al. 2015). It is difficult to support a multi-locus model of sex determination in a primarily dioecious species, as this arrangement is likely to be evolutionarily unstable (Bull and Charnov 1977). The locus mapped to chr 19 is, therefore, likely to be an assembly or alignment artifact. This could be caused by a recent translocation from chr 19 to the W haplotype of chr 15, which would result in incorrect alignment of GBS reads to the original chr 19 locus if the W haplotype is not in the main genome assembly. However, because the locus matches a portion of the SDR of chr 19 in *Populus*, and the gene content of these regions is similar between the taxa, this finding would still provide valuable clues about sex determination and/or sex dimorphism in this family even if it is caused by a recent translocation. It is also noteworthy that the *S. purpurea de novo* genome assembly did not use the *P.*

trichocarpa genome assembly as a reference to guide placement of scaffolds in pseudomolecules, so the results reported here are not caused by carryover of biases or errors from the original *P. trichocarpa* assembly.

Unfortunately, a definitive comparison of the Salicaceae sex chromosomes is not possible with the currently available genome sequences. The SDRs of *Salix* and *Populus* are typical in that they have complex structural polymorphisms, high repeat content, and low recombination rates, all of which contribute to fragmentary and erroneous genome assemblies (Geraldes et al. 2015). Furthermore, the genomic analyses of Salicaceae SDRs reported to date have been based on genome sequences for the homogametic sex [a female in *P. trichocarpa* (Geraldes et al. 2015) and a male in *S. viminalis* (Pucholt et al. 2017b)], or on highly fragmented genome assemblies (Hou et al. 2015), so this is the first effort to fully reconstruct the non-recombining SDR in this family. Efforts are underway to fully assemble the W and Y chromosomes using long read sequencing and dense genetic mapping in multiple pedigrees. This will facilitate analyses that can date the origin of these regions based on differentiation of sex-specific haplotypes in the non-recombining portions of the SDR (Otto et al. 2011). Furthermore, elucidation of the sex determination system in additional Salicaceae taxa should help to determine the ancestral state. This family should therefore be instrumental in advancing our knowledge of the evolution and ecological significance of sex chromosomes as genetic and genomic resources continue to accumulate.

We have shown that sex is determined by a relatively large portion of chromosome 15 in *S. purpurea*. The sex-associated loci are nearly fixed heterozygous in females and are overwhelmingly homozygous in males, demonstrating that this species has a ZW sex determination system. The SDR is characterized by suppressed recombination and high repeat

content, as is expected for a plant SDR. Furthermore, the region appears to be relatively young based on the small number of synonymous substitutions that have occurred between Z and W alleles in that region. Comparison with the *Populus* SDR reveals homology over a short stretch, a finding that is recapitulated by the alignment of sex-associated markers to that chromosomal region in *S. purpurea*. We hypothesize that a translocation of that portion of the SDR has occurred between Chr15 and Chr19 in the Salicaceae lineage. The region contains several promising sex determination candidate genes, which are worthy of further functional analysis.

Acknowledgements

We are grateful to Matt Olson for helpful comments on the manuscript.

Funding

This work was supported by grants from the USDA-NIFA CAP program (4705-WVU-USDA-9703), the DOE JGI Community Sequencing Program, and the NSF Dimensions of Biodiversity Program (DEB-1542509). Sequencing was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, was supported by the Office of Science of the U.S. Department of Energy under Contract no. DE-AC02-05CH11231.

Data availability

All raw sequencing data are available from the NCBI Sequence Read Archive (accessions SRP003908, SRP086434, and SRP086435) and the genome assembly and annotation are available from Phytozome (https ://phyto zome.jgi.doe.gov).

1. Ainsworth C (2000) Boys and girls come out to play: the molecular biology of dioecious plants. Ann Bot 86:211–221.

2. Alliende MC, Harper JL (1989) Demographic studies of a dioecious tree. I. Colonization, sex and age structure of a population of *Salix cinerea*. J Ecol 77:1029–1047.

3. Alstrom-Rapaport C, Lascoux M, Wang YC et al (1998) Identification of a RAPD marker linked to sex determination in the basket willow (*Salix viminalis* L.). J Hered 89:44–49.

4. Andrews KR, Good JM, Miller MR et al (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92.

5. Arends D, Prins P, Jansen RC, Broman KW (2010) R/qtl: high-throughput multiple QTL mapping. Bioinformatics 26:2990–2992

6. Argus GW (1997) Infrageneric classification of *Salix* (Salicaceae) in the new world. Syst Bot Monogr 52:1–121

7. Ashman T-L (2006) The evolution of separate sexes: a focus on the ecological context. In: Harder LD, Barrett SCH (eds) Ecology and evolution of flowers. Oxford University Press, Oxford, p 370

8. Ashman T-L, Kwok A, Husband BC (2013) Revisiting the Dioecy-Polyploidy Association: alternate pathways and research opportunities. Cytogenet Genome Res 140:241–255.

9. Bachtrog D, Mank J, Peichel CL et al (2014) Sex determination: why so many ways of doing it? PLoS Biol 12:e1001899.

10. Barrett SCH, Hough J (2013) Sexual dimorphism in flowering plants. J Exp Bot 64:67–82.

11. Bergero R, Charlesworth D (2009) The evolution of restricted recombination in sex chromosomes. Trends Ecol Evol 24:94–102

12. 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.

13. Beukeboom LW, Perrin N (2014) The evolution of sex determination. Oxford University Press, New York

14. Boucher LD, Manchester SR, Judd WS (2003) An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. Am J Bot 90:1389–1399.

15. Bull JJ, Charnov EL (1977) Changes in the heterogametic mechanism of sex determination. Heredity 39:1–14.

16. Burby PE, Simmons LA (2017) MutS2 promotes homologous recombination in *Bacillus subtilis*. J Bacteriol 199:e00682–e00616.

17. Cai G, Cresti M (2009) Organelle motility in the pollen tube: a tale of 20 years. J Exp Bot 60:495–508.

18. Carlson CH, Choi Y, Chan AP et al (2017) Dominance and Sexual Dimorphism Pervade the *Salix purpurea* L. Transcriptome. Genome Biol Evol 9:2377–2394.

19. Charlesworth D (2006) Evolution of plant breeding systems. Curr Biol 16:726–735.

20. Charlesworth D (2015) Plant contributions to our understanding of sex chromosome evolution. New Phytol 208:52–65.

21. Charlesworth D (2016) Plant sex chromosomes. Annu Rev Plant Biol 67:397–420.

22. Charlesworth D, Charlesworth B (1978) Population genetics of partial male-sterility and the evolution of monoecy and dioecy. Heredity 41:137–153.

23. Charnov EL (1982) The theory of sex allocation. Monogr Popul Biol 18:1–355.

24. Chen Y, Wang T, Fang L et al (2016) Confirmation of single-locus sex determination and female heterogamety in willow based on linkage analysis. PLoS One 11:e0147671.

25. DePristo M, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

26. Dickmann DI, Kuzovkina J (2014) Poplars and willows of the world, with emphasis on silviculturally important species. In: Poplars and willows: trees for society and the environment. CABI, Wallingford, pp 8–91.

27. Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379.

28. Falda M, Toppo S, Pescarolo A et al (2012) Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. BMC Bioinformatics 13:S14.

29. Fukui K, Kuramitsu S (2011) Structure and function of the small MutS-related domain. Mol Biol Int 2011:1–9.

30. Fukui K, Kosaka H, Kuramitsu S, Masui R (2007) Nuclease activity of the MutS homologue MutS2 from *Thermus thermophilus* is confined to the Smr domain. Nucleic Acids Res 35:850–860.

31. Füssel U, Dötterl S, Jürgens A, Aas G (2007) Inter- and intraspecific variation in floral scent in the genus *Salix* and its implication for pollination. J Chem Ecol 33:749–765.

32. Gaudet M, Jorge V, Paolucci I et al (2008) Genetic linkage maps of *Populus nigra* L. including AFLPs, SSRs, SNPs, and sex trait. Tree Genet Genomes 4:25–36.

33. Geraldes A, Hefer CA, Capron A et al (2015) Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). Mol Ecol 24:3243–3256.

34. Glaubitz JC, Casstevens TM, Lu F et al (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e0090346.

35. Glick L, Sabath N, Ashman TL et al (2016) Polyploidy and sexual system in angiosperms: is there an association? Am J Bot 103:1223–1235.

36. Gnerre S, MacCallum I, Przybylski D et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518.

37. Goodstein DM, Shu S, Howson R et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186.

38. Hou J, Ye N, Zhang D et al (2015) Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci Rep 5:e9076.

39. Hou J, Ye N, Dong Z et al (2016) Major chromosomal rearrangements distinguish willow and poplar after the ancestral "Salicoid" genome duplication. Genome Biol Evol 8:1868–1875.

40. Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354.

41. Karp A, Hanley SJ, Trybush SO et al (2011) Genetic Improvement of Willow for Bioenergy and Biofuels. J Integr Plant Biol 53:151–165.

42. Karrenberg S, Kollmann J, Edwards PJ (2002) Pollen vectors and inflorescence morphology in four species of *Salix*. Plant Syst Evol 235:181–188.

43. Kersten B, Pakull B, Groppe K et al (2014) The sex-linked region in *Populus tremuloides* Turesson 141 corresponds to a pericentromeric region of about two million base pairs on *P. trichocarpa* chromosome 19. Plant Biol 16:411–418.

44. Kunkel T, Erie D (2005) DNA mismatch repair. Annu Rev Biochem 74:681–710.

45. Lian S, Liu T, Gong K et al (2016) A complete and accurate short sequence alignment algorithm for repeats. J Biosci Med 04:144–151.

46. Lin H, Niu L, McHale N et al (2013) Evolutionarily conserved repressive activity of WOX proteins mediates leaf blade outgrowth and floral organ development in plants. Proc Natl Acad Sci USA 110:366–371.

47. Lloyd DG (1979) Evolution towards dioecy in heterostylous populations. Plant Syst Evol 131:71–80.

48. Mank JE (2009) Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. Am Nat 173:141–150.

49. Melters DP, Bradnam KR, Young H et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10.

50. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for nextgeneration sequencing data. Genomics 95:315–327.

51. Ming R, Moore PH (2007) Genomics of sex chromosomes. Curr Opin Plant Biol 10:123–130.

52. Ming R, Bendahmane A, Renner SS (2011) Sex chromosomes in land plants. Annu Rev Plant Biol 62:485–514.

53. Mock KE, Callahan CM, Islam-Faridi MN et al (2012) Widespread triploidy in western North American aspen (*Populus tremuloides*). PLoS One 7:e48406.

54. Moore EC, Roberts RB (2013) Polygenic sex determination. Curr Biol 23:R510–R512.

55. Nicolas M, Marais G, Hykelova V et al (2005) A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. PLoS Biol 3:e4.

56. Nishihama R, Soyano T, Ishikawa M et al (2002) Expansion of the cell plate in plant cytokinesis requires a kinesin-like protein/MAPKKK complex. Cell 109:87–99.

57. Olson MS, Hamrick JL, Moore RC (2017) Breeding systems, mating systems, and gender determination in angiosperm trees. In: Groover A, Cronk QCB (eds) Comparative and evolutionary genomics of angiosperm trees. Springer International Publishing, Switzerland, pp 139–158

58. Otto SP, Pannell JR, Peichel CL et al (2011) About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet 27:358–367.

59. Pakull B, Groppe K, Meyer M et al (2009) Genetic linkage mapping in aspen (*Populus tremula* L. and *Populus tremuloides* Michx.). Tree Genet Genomes 5:505–515.

60. Pakull B, Kersten B, Lüneburg J, Fladung M (2014) A simple PCR-based marker to determine sex in aspen. Plant Biol 17:256–261.

61. Pandey RS, Azad RK (2016) Deciphering evolutionary strata on plant sex chromosomes and fungal mating-type chromosomes through compositional segmentation. Plant Mol Biol 90:359–373.

62. Paolucci I, Gaudet M, Jorge V et al (2010) Genetic linkage maps of *Populus alba* L. and comparative mapping analysis of sex determination across Populus species. Tree Genet Genomes 6:863–875.

63. Peto FH (1938) Cytology of poplar species and natural hybrids. Can J Res 16:446–455

64. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

65. Pucholt P, Rönnberg-Wästljung A-C, Berlin S (2015) Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). Heredity 114:575–583.

66. Pucholt P, Hallingbäck HR, Berlin S (2017a) Allelic incompatibility can explain female biased sex ratios in dioecious plants. BMC Genom 18:251.

67. Pucholt P, Wright AE, Conze LL et al (2017b) Recent sex chromosome divergence despite ancient Dioecy in the Willow, *Salix viminalis*. Mol Biol Evol 22:522–525.

68. Qi J, Chen Y, Copenhaver GP, Ma H (2014) Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. Proc Natl Acad Sci 111:10007–10012.

69. Renner SS (2014) The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. Am J Bot 101:1588–1596.

70. Rice WWR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution 38:1416–1424.

71. Serapiglia MJ, Gouker FE, Hart JF et al (2015) Ploidy level affects important biomass traits of novel shrub Willow (*Salix*) hybrids. BioEnergy Res 8:259–269.

72. Slavov GT, Zhelev P (2010) Salient biological features, systematics, and genetic variation of *Populus*. In: Jansson S, Bhalerao RP, Groover A (eds) Genetics and genomics of *Populus*. Springer, New York, pp 15–38

73. Sterck L, Rombauts S, Jansson S et al (2005) EST data suggest that poplar is an ancient polyploid. New Phytol 167:165–170.

74. Temmel NA, Rai HS, Cronk QCB (2007) Sequence characterization of the putatively sex-linked Ssu72 -like locus in willow and its homologue in poplar. Can J Bot 85:1092–1097.

75. Tuskan GA, DiFazio S, Jansson S et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604.

76. Tuskan GA, DiFazio S, Faivre-Rampant P et al (2012) The obscure events contributing to the evolution of an incipient sex chromosome in *Populus*: a retrospective working hypothesis. Tree Genet Genomes 8:559–571.

77. Ueno N, Suyama Y, Seiwa K (2007) What makes the sex ratio femalebiased in the dioecious tree *Salix sachalinensis*? J Ecol 95:951–959.

78. van Doorn GS, Kirkpatrick M (2007) Turnover of sex chromosomes induced by sexual conflict. Nature 449:909–912.

79. van Doorn GS, Kirkpatrick M (2010) Transitions between male and female heterogamety caused by sex-antagonistic selection. Genetics 186:629–645

80. Vyskot B, Hobza R (2015) The genomics of plant sex chromosomes. Plant Sci 236:126–135.

81. Wang J, Na J, Yu Q et al (2012) Sequencing papaya X and Y h chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci 109:13710–13715.

82. Westergaard M (1958) The mechanism of sex determination in dioecious flowering plants. Adv Genet 9:217–281.

83. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591.

84. Yin T, DiFazio SP, Gunter LE et al (2008) Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. Genome Res 18:422–430.

85. Zhou W, Lu Q, Li Q et al (2017) PPR-SMR protein SOT1 has RNA endonuclease activity. Proc Natl Acad Sci USA 114:E1554–E1563.

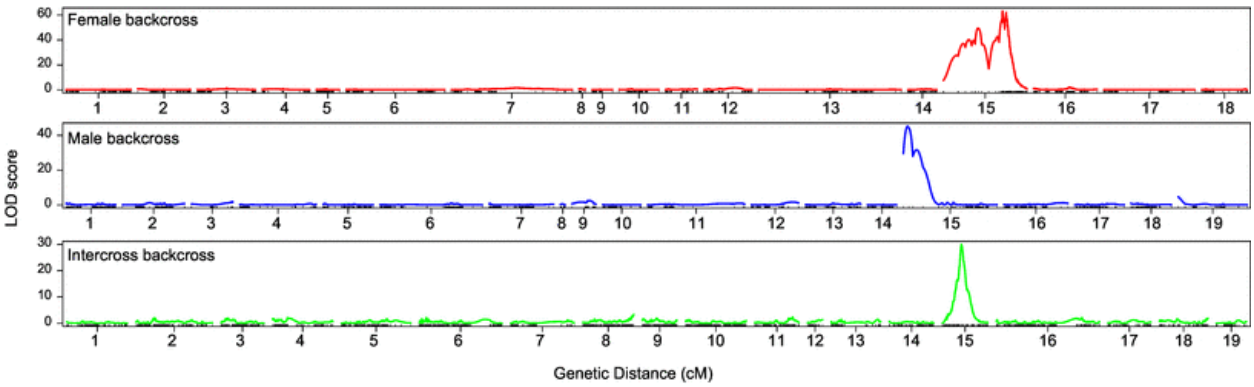

Figure 1. QTL for sex in an F_2 *S. purpurea* cross. From top to bottom are LOD scans for female backcross (red), male backcross (blue), and intercross (green) markers across the 19 major *S. purpurea* linkage groups. Chromosome 15 has a very strong QTL sex in all three maps, and the male backcross also shows a weak peak on chromosome 19 (LOD = 4.68; Table 1)

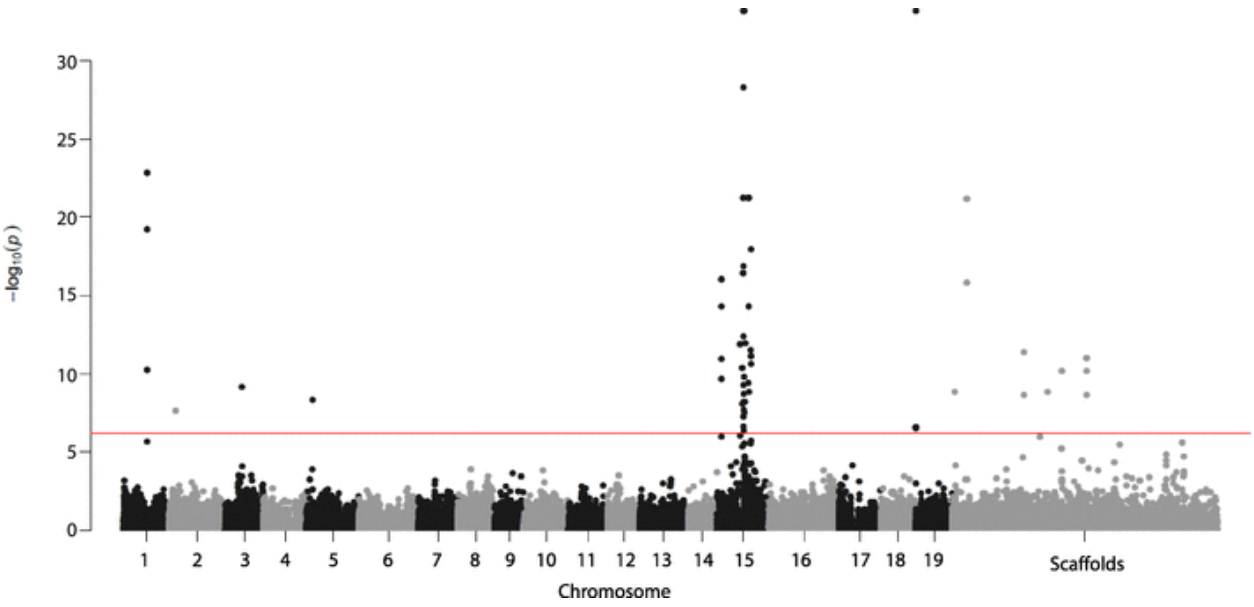

Figure 2. Manhattan plot derived from genome-wide association analysis for sex determination. The *y*-axis shows the strength of association ($-\log_{10}(P$ value)) for each SNP ordered by chromosome and SNP position (*x*-axis). The horizontal line indicates significance after a Bonferroni correction for multiple testing

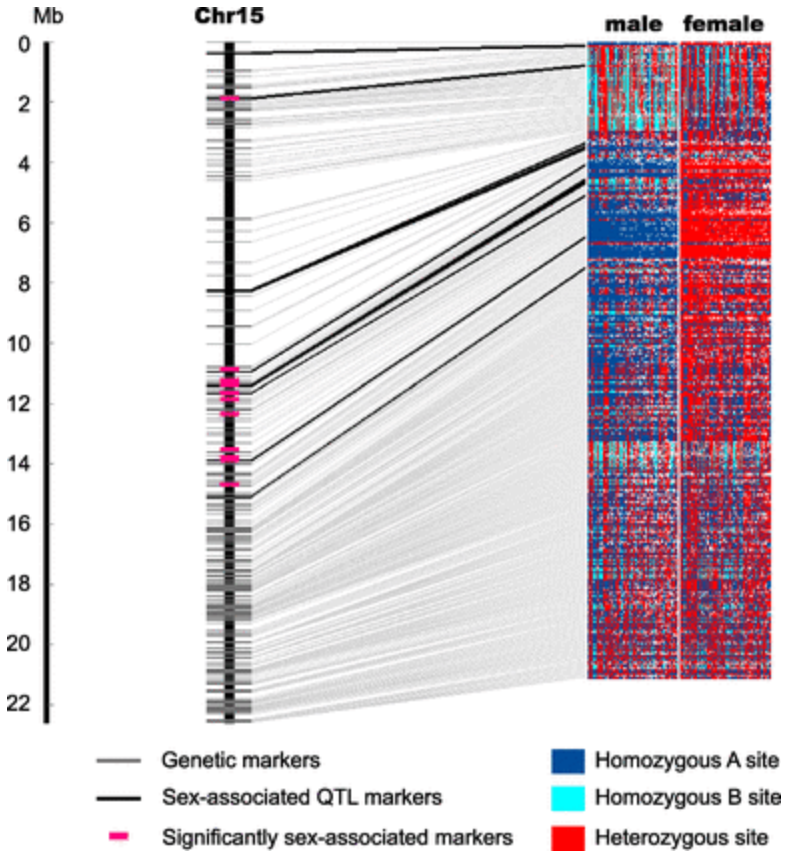

Figure 3. Genotype configuration of chromosome 15 in males and females from the F_2 family. Markers from all three genetic maps are shown as horizontal lines corresponding to their physical positions on the chromosome 15 physical assembly. Markers with top LOD scores in each map are colored as black. Significantly associated markers from the GWAS analysis with $P < 1 \times 10^{-7}$ are indicated by fuschia marks on the physical map. Each marker is connected between physical map and its genotype configurations with 100 selected progeny of each sex. Genotypes of QTL markers are colored according to their homozygosity or heterozygosity

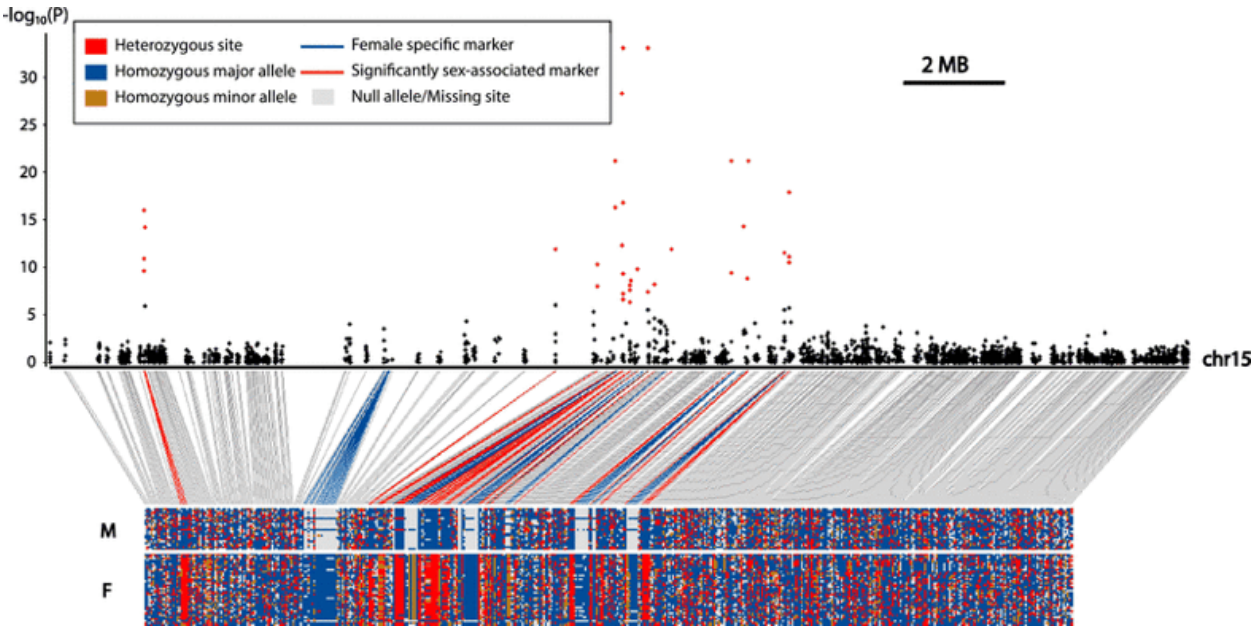

Figure 4. Genotype configurations of markers on chromosome 15 from the *S. purpurea* association population. The top is a blowup of chromosome 15 from the Manhattan plot in Fig. 2, with significantly sex-associated markers colored red. The bottom shows the genotype configurations in the association population, where each row represents an individual. "Major alleles" are those with higher frequency in males, shaded blue where homozygous; homozygotes for male minor alleles, gold; heterozygous sites, red; and missing data, light gray. Lines connect each plotted marker to its physical position. Red lines indicate that markers are significantly associated with sex while blue lines indicate the markers were identified as female-specific (putatively derived from the W haplotype)

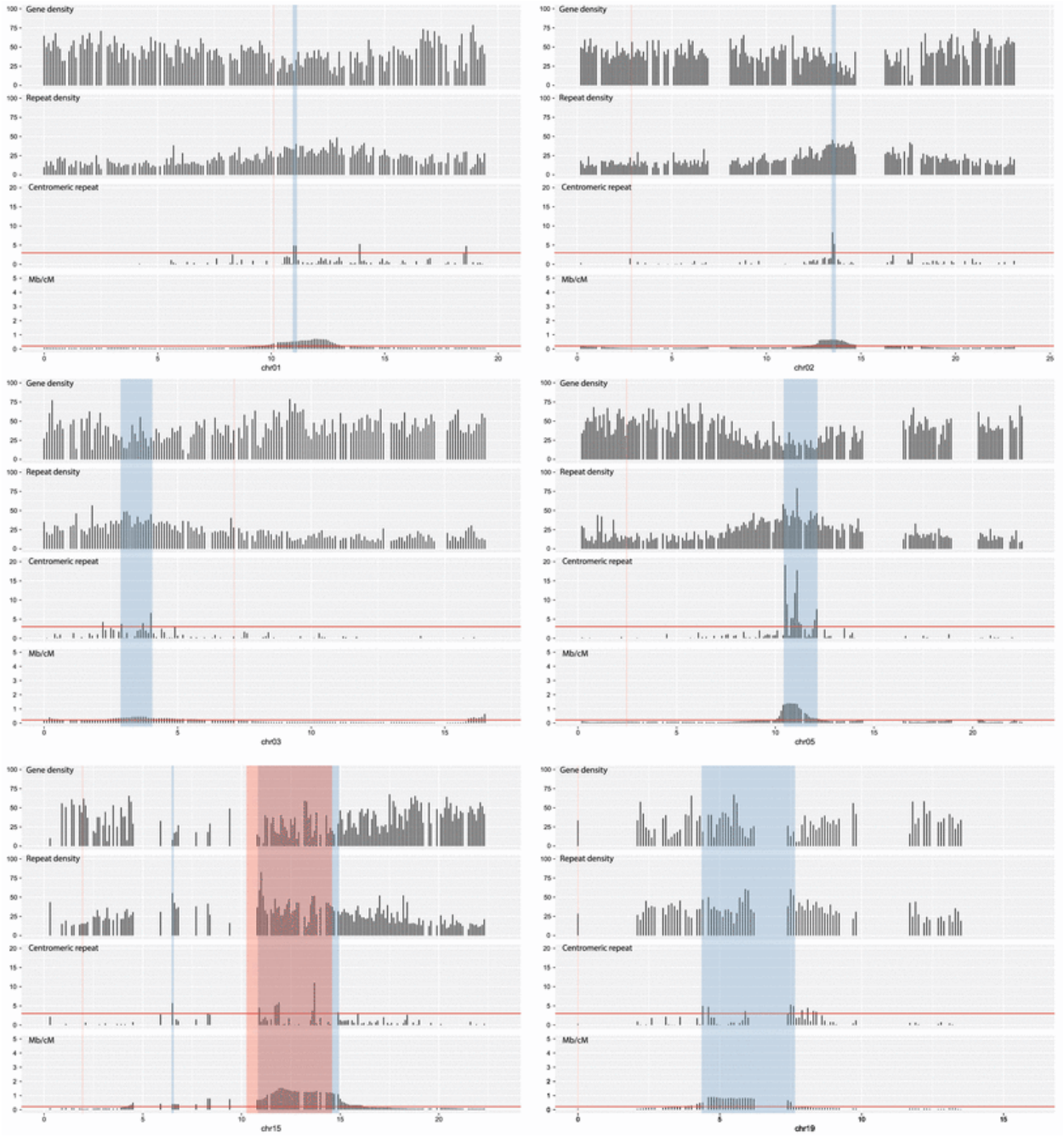

Figure 5. Delineation of putative centromeres relative to the SDRs. Bar plots represent, from the top, gene density, repeat density, density of centromeric repeats, and physical:genetic distance ratio (Mb/cM) in 100 kb windows. Blue shading shows positions of putative centromeres, as defined by empirical thresholds represented by horizontal red lines, and red shading represents the SDR

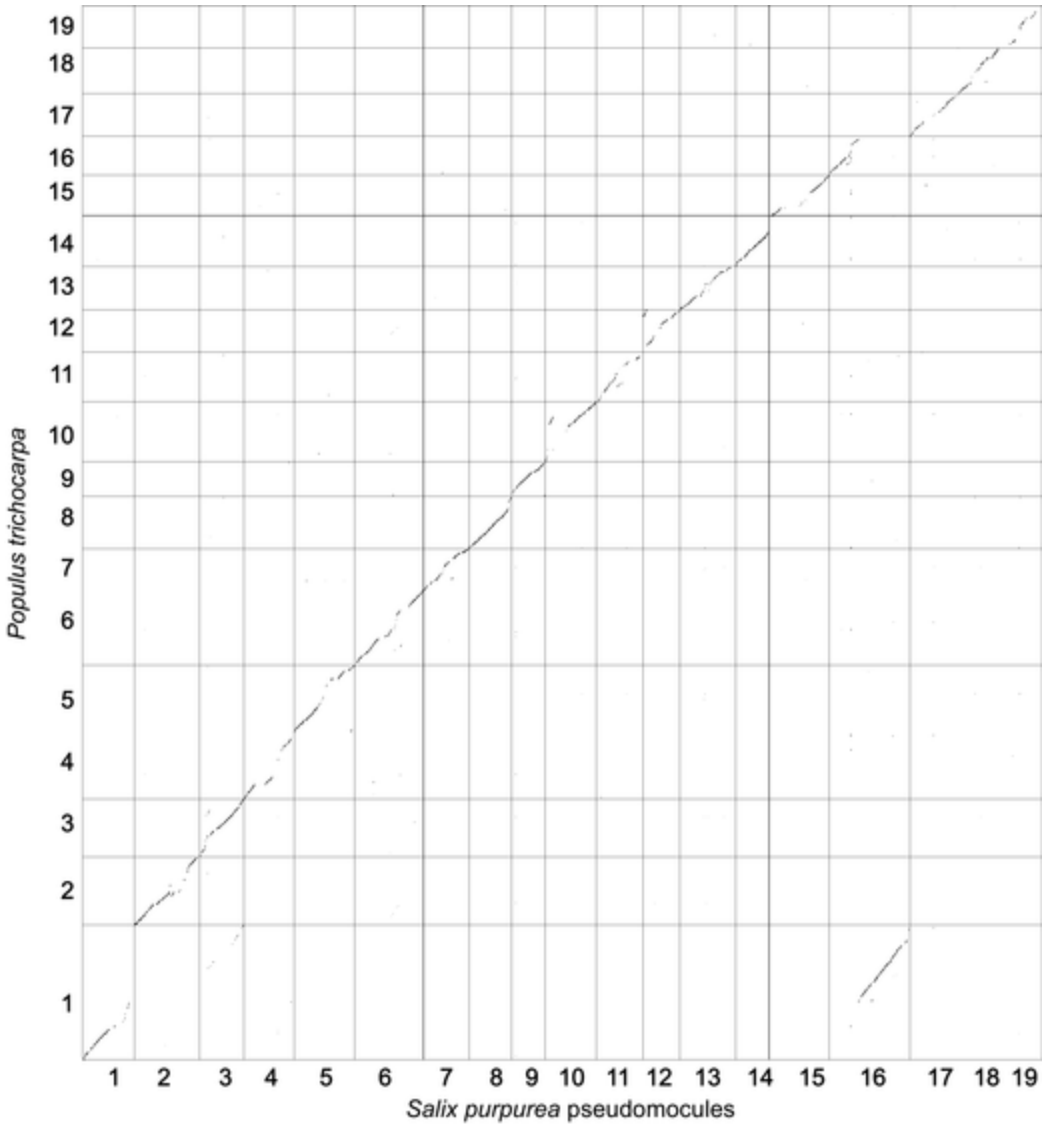

Figure 6. Comparison between the *S. purpurea* (*x*-axis) and *P. trichocarpa* (*y*-axis) genomes, with parameters set to exclude paralogous segments derived from the most recent whole genome duplication