2020

Evolutionary genomics of dynamic sex chromosomes in the Salicaceae

Ran Zhou
West Virginia University, razhou@mix.wvu.edu

Recommended Citation

Zhou, Ran, "Evolutionary genomics of dynamic sex chromosomes in the Salicaceae" (2020). *Graduate Theses, Dissertations, and Problem Reports*. 7649.
https://researchrepository.wvu.edu/etd/7649

Evolutionary genomics of dynamic sex chromosomes in the Salicaceae

Ran Zhou

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
In
Biology

Stephen DiFazio, Ph.D., Chair
Craig Barrett, Ph.D.
Jennifer Hawkins, Ph.D.
Matthew Olson, Ph.D.
Amy Welsh, Ph.D.

Morgantown, West Virginia
2020

Keywords: evolution; sex chromosome; sex-determination region; *Salix*; *Populus*; genome;

Abstract
Evolutionary genomics of dynamic sex chromosomes in the Salicaceae

Ran Zhou

Identifying the sex-determination region (SDR) and other genomic features of sex chromosomes are of great importance in the studies of the evolution of sex. However, the process of accurately identifying the size and location of the SDR is often difficult, even when a genomic sequence is available. This usually is hindered by large repetitive elements and a lack of recombination in the SDR. In this thesis, I assemble sex chromosomes with whole genomic sequencing data, identify SDRs and explore their genomic features in two sister species from the Salicaceae family. I also develop an interpretation of the lability of the sex configuration in the two species. In Chapter 2, I use quantitative trait locus mapping and a genome-wide association study to characterize the genomic composition of the SDR in a reference genome derived a female *Salix purpurea* clone. I show that the SDR in *S. purpurea* has a female heterogametic (ZW) system on chromosome 15. The SDR is inferred to be between 5 to 7 Mb long and overlapping with the centromere. This SDR has several classic features like reduced recombination and high structural polymorphism. Intriguingly, chromosome 19 contains sex-associated markers, which raises the possibility of a translocation of the SDR within the Salicaceae lineage. In Chapter 3, I improve the quality of assembly of sex chromosomes in *S. purpurea* with long-reads sequencing data and a modified map. Using an improved assembly of the SDR, I show that two consecutive palindromes span over a region of 200 kb, with conspicuous 20 kb stretches of highly conserved homologous sequences among the four arms in the female-specific regions of the SDR. Comparison to the genome of a closely related species *S. suchowensis* provides evidence for gene conversion occurring among the palindrome arms. The hypothesis of the translocation of the SDR within the Salicaceae could not be rejected. In Chapter 4, I use a similar strategy from Chapter 3 to study the SDR of a male *Populus trichocarpa* clone. I show that the SDR in *P. trichocarpa* has a male heterogametic (XY) system on chromosome 19. A cluster of inverted repeats that are homologous with a response regulator gene is present in the male-specific region in the SDR. This research provides important genomic

resources for futures studies in these two species as well as the evolution of SDRs in the Salicaceae.

Acknowledgements

The fantastic journey would not have been possible without extensive support from friends, family, collaborators, and mentors. First and foremost, I want to thank my advisor Dr. Stephen DiFazio for his mentoring and patience through my program. He is thoughtful and provides numerous plans and ideas for my projects. Although not all these ideas ended up with being in the projects, they served as necessary steps to enjoy the processes of scientific research. I am forever grateful for his commitment to my projects and career development. With his devotion to his students' interests, the work in chapter 2 has been published and the work in chapter 3 is accepted in an excellent journal while I'm writing this. He made these publications of my important findings in my projects feasible. He is the reason why I'm here today. I also want to express my appreciation to my committee members, thank you for your time and service in my committee. I would also like to thank my former and current committee members, Dr. James McGraw, Dr. Dana Huebert Lima, Dr. Jennifer Hawkins, Dr. Craig Barrett, Dr. Matthew Olson, and Dr. Amy Welsh. Their guidance and encouragement during my preliminary examination studies and their comments on my research proposals have helped me through these milestones, which make this manuscript feasible.

I am thankful for the support and friendship of former and current lab members of the DiFazio lab, especially: Dr. Luke Evans, Dr. Eli Rodgers-Melnick, Dr. Hari Chhetri, Rose Strickland-Constable, Sandra Simon, and Roshan Abeyratne. It was a great pleasure to spend time with these talented and nice people. Not only the help you generously provide to me, but both your openness and acceptance to the cultural differences also make my life here very enjoyable. I would like to extend a special thank you to Dr. David Macaya-Sanz. It is of great fortune for me to work with him. His insight ideas on the population genetics and genomic skills have aided me greatly since my third year of doctoral research. Not only is he an excellent colleague, but also such a great friend in life. He provides a great model for young scientists with his work ethic and a positive attitude that I will carry with me throughout my scientific career.

I would like to sincerely thank my numerous collaborators involved in this work, whose data, leadership, and valuable insight made this work possible. I would like to thank Dr. Lawrence Smart, Dr. Craig Carlson and Dr. Fred Gouker for their support in the materials and data from their large collection of *Salix* clones I used for chapter 2 and 3. I am also especially

grateful for the advice of our Chinese collaborator Dr. Jianquan Liu and Dr. Tao Ma and discussions with them. I also want to thank Dr. Jianquan Liu particularly for his support and encouragement during pursuing my Ph.D. degree overseas.

Thank you to my family and friends who encouraged me to pursue my dream overseas and provided me with their care as much as they can. I want to thank my parents for their unconditional love and support for my decisions in the past twenty years. In front of their love, any words seem to be plain. Thank you to my grandparents. I also thank a lot of friends I met in the Biology department for social time and professional activities. Thank you to everybody in the Department of Biology who has helped me with my teaching and research tasks in the past six years. Lastly, a great thanks to many great friends I have made these years in China and the US. I always want to thank you for being such great friends in my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

GENERAL INTRODUCTION

Understanding how sex evolves is a fundamental yet interesting mystery to biologists. The flowers of angiosperms are largely cosexual, meaning that each individual has both sex functions. Some cosexual species have hermaphroditic flowers, and some are monoecious where pistils and stamens are present on different flowers within the same individual. Dioecy refers to the case where pistillate and staminate flowers are in different individuals. This is usually achieved in floral development as an arrest of sex organ formation (Vyskot & Hobza 2015). Dioecious species represent about 5% of plants (Renner 2014). This does not mean, however, that dioecy is rare. Instead, it occurs across many angiosperm phyla (Renner 2014; Henry et al. 2018). In dioecious species, control of sex expression could be either environmental or genetic (Vyskot & Hobza 2015). Most of the 15,600 dioecious species of angiosperms in the latest compilation by Renner (2014) probably have genetic sex determination. However, cytogenetic data are available from fewer than 100 angiosperm species (Charlesworth 2016). Among these, sex chromosomes have been only identified in 40 species and heteromorphic sex chromosomes have been revealed in just half of them (Ming et al. 2011; Renner 2014; Hobza et al. 2017). Apart from the limited information about sex chromosomes in plants, sex determination itself is a complex and dynamic process, and not yet fully understood (Beukeboom &Perrin 2014). Both of these reasons make understanding how sex is genetically determined a difficult but important task in the study of sex in plants.

Unlike heteromorphic sex chromosomes commonly found in animals, where X and Y chromosomes are often different at the cytological level, only some dioecious species have heteromorphic sex chromosomes, such as *Silene latifolia* (Delph et al. 2010). In other species, sex chromosomes may appear to be homomorphic sex chromosomes but are heteromorphic at the molecular level, such as sex chromosomes in papaya, where the loss of gene content on the Y chromosome is sufficiently extensive to cause lethality of the YY genotype (Charlesworth & Charlesworth 2000).

In the family Salicaceae, about 1,000 species are uniformly woody (trees or shrubs) in approximately 55 genera (Cronk et al. 2015). *Salix* and *Populus* are closely related sister genera where nearly all species are dioecious. *Salix* are generally insect-pollinated, whereas *Populus* are wind-pollinated (Cronk et al. 2015). Both genera are known to contain a palaeotetraploidization of the genome, with a haploid base number of 11 to 22 (Sterck et al. 2005), and then followed by reduction events to n = 19 (Cronk et al. 2015). Two reference genomes of *Populus trichocarpa* and *Salix purpurea* used extensively in my projects are both n=19. Thus, the Salicaceae family provides an excellent model system for studying sex chromosomes and its evolution under polyploidization background.

Populus trichocarpa is a dioecious woody plant with an identified male heterogametic system (Tuskan et al. 2012). In contrast to many animal groups, the sex of an individual cannot usually be determined in *Populus* before flowering without sex-specific genetic markers (Pakull et al. 2011; Kersten et al. 2014; Pakull et al. 2014). A generally applicable diagnostic marker, allowing sex determination in non-flowering trees without any additional knowledge of the genotype background, could be very useful for research and breeding purposes without waiting for flowing (Pakull et al. 2014). Thus lack of a completely assembled sex chromosome (Y chromosome) is a problem for both the fields of evolution and breeding. Although sex determination has been mapped to Chr19 in *Populus,* Chr19 is not the only chromosome containing sex-specific markers in sex association analysis (Geraldes et al. 2015). The inconsistent location of the SDR on multiple chromosomes in *Populus* is conspicuous compared to the consistent identification of SDRs around the center of Chr15 in several *Salix* species (Hou et al. 2015; Pucholt et al. 2015; Zhou et al. 2018). Multiple locations of sex-specific markers in *Populus* were proposed to be associated with the erroneous assembly of portions of the SDR in the reference genome (Geraldes et al. 2015). Alternatively, this could be evidence that an unprecedented multilocus sex determination system might exist in plants. However, without fully addressing the proper assembly of the SDR and the sex chromosome in the reference, testing these hypotheses remains out of reach. The conflicting location and content of the SDRs highlights a major gap in our knowledge about how sex evolved in *Populus*, but also a critical need for a complete Y chromosome assembly in both fields of plant sex evolution and breeding. In my projects, we aim to provide a superior assembly via long-reads sequencing technology to improve the continuity in the SDR. We also attempt to provide useful and high-quality data to answer questions about how the genetic architecture of sex shapes and is shaped by evolutionary processes.

In contrast to *Populus*, the sex determination systems in *Salix* have been mapped to chromosome 15. Work by Alstrom-Rapaport et al. (1998) and Gunter et al. (2003) showed that sex-linked markers were associated with femaleness or a female-specific locus in several families with certain genetic background in *Salix viminalis*. But not until recently, the sex-determination system was confirmed to be female heterogametic (ZW/ZZ) in *S. viminalis* (Pucholt et al. 2015), as well as in *Salix suchowensis* (Hou et al. 2015). However, without the complete assembly of the SDR in the sex chromosome, they could not provide an answer to the hypothesis about the shared orthologues between *Salix* and *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. To tackle this problem, we chose a willow species, *Salix purpurea*, which is a diploid perennial shrub species belonging to the subgenus Vetrix (Zsuffa 1990; Lin et al. 2009).

Short-read sequencing approaches provide lower-cost, high-accuracy data that are useful for population-level research (Goodwin et al. 2016). However, the short reads cannot fully span over repetitive regions of the genome, resulting in tens of thousands of fragmented assemblies and collapsed contigs (Li et al. 2018). To overcome this weakness, researchers have tried different strategies during genome assembly. For example, the genome of *Populus euphratica* is generated from short-read sequencing with a fosmid-pooling strategy to improve the quality of assembly (Ma et al. 2013). The initial release of the genome of *P. trichocarpa* was constructed from whole-genome shotgun sequencing data using Sanger sequencing (Tuskan et al. 2006). The assembly has been improved by merging the outbred haplotypes and by attempting to remove contaminating sequences. Because SDRs are typically enriched in transposable elements and ampliconic genes, biologically informative sequences are still undiscovered in the SDR (Bachtrog 2013).

In contrast to short-read sequencing, long-read sequencing technology provides high-quality genomes assemblies, such as *Oropetium thomaeum* (245 Mb), *Chenopodium quinoa* (1500 Mb), *Zea mays* (2300 Mb), and *Helianthus annuus* (sunflower, 3000 Mb) (reviewed by Li et al. 2018). Although raw reads of current PacBio systems can have sequencing error rates of up to 15% , high-quality error-corrected sequences can be produced with sufficient coverage (Jayakumar & Sakakibara 2017).This can be achieved through assemblers like CANU (Koren et al. 2017). Alternatively, a draft of assembly can be achieved through assemblers, and then consensus polishing can be applied to the assembly by QUIVER (Chin et al. 2013) with more accurate Illumina data (Jayakumar & Sakakibara 2017). Although long-read sequencing can provide better continuity of the assembly, it does not provide phase information (i.e. Y-linked or X-linked) about contigs. As shown in Harkess et al. (2017), it remains necessary to scrutinize all assembled contigs to determine if there are sex-linked pieces. The advantage is obvious with long-reads sequencing, including that sex-linked haplotypes are not collapsed during the assembly, and homologous scaffolds from W and Z are well separated after the primary assembly. We also show that when the reference genome is derived from a homogametic (ZZ) individual, the sex-associated markers are on several chromosomes along with the main sex-associated peak on the sex chromosome. On the contrary, when the reference genome contains the heterogametic W chromosome, there is only one single peak on the sex chromosome. This shows how the presence of heterogametic Y (or W) chromosome in the reference could influence the sex-association analysis. Given that the individual used for generating the *P. trichocarpa* reference genome was a female (Tuskan et al. 2006), we hypothesize that the multiple peaks on several chromosomes observed in sex association are artifacts due to the absence of a Y chromosome in the reference. By sequencing a male individual of *P. trichocarpa*, once the Y chromosome is assembled, we will be able to test this hypothesis by performing sex-association analysis with a newly assembled genome with the presence of Y chromosome. With our success in assembling the W chromosome in a female individual of *S. purpurea* with long-read sequencing data, we have confidence in assembling the Y chromosome in the male individual with our approach. Thus we will provide the first Y chromosome assembled with long-reads in Salicaceae family, which will benefit studies of the evolution of sex in plants and design of sex-specific markers for breeding projects.

In animals, Y chromosomes are different from X chromosomes in several ways, e.g. the Y chromosome is substantially smaller, the Y has fewer genes and more transposable elements, and the Y contains large palindromic (inverted) repeats containing genes that do not occur on the X (Bachtrog 2013). How the Y chromosome differentiated from the X chromosome is still not fully understood. Studying ancient Y chromosomes (for example those in mammals) inhibits our ability to reconstruct the early stages of sex chromosome evolution. In *Populus*, sex chromosomes appear to be quite young (Geraldes et al. 2015; McKown et al. 2017) and therefore provide a unique opportunity to study the initial stages of evolution of sex chromosomes (Charlesworth 2016; Hobza et al. 2017). The SDR in *P. trichocarpa* was inferred to be small and compact with less than 20 genes spanning ~100 kbp on chromosome 19 (Geraldes et al. 2015). There has not been a comprehensive comparison between X and Y at a chromosomal level for *Populus* because there is no publicly available *Populus* Y chromosome. For example, Hou et al. (2015) could not find evidence for the existence of homologous genes between the SDR of the Y chromosome (which they assumed was in the reference) and the one that they found on Chr15 in *S. suchowensis*. This analysis was flawed because genes in the SDR from the Y chromosome obviously are not present in the female reference genome. Similarly, a list of candidate genes for sex was given based on sex-association studies in Geraldes et al. (2015), but the authors could not provide further details about the Y specific genes in *P. trichocarpa*. Thus, further evaluation of these candidates and the evolutionary history of sex chromosomes is hindered because of the lack of sequence data from the Y chromosome. Upon completion of the assembly of the Y chromosome in this proposal, we will provide a comprehensive comparison with the X chromosome based on gene content and structural variation and other available features. Thus, a complete description of genomic features of the Y chromosome will provide valuable data and tools for understanding how Y chromosomes can differentiate from X chromosomes in an evolutionarily young pair of sex chromosomes.

Studies based on humans and other animals create a false impression of stability in sex determination systems, and their commonalities mask the diversity and turnover in sex determination mechanisms that are readily apparent when taking a broader taxonomic view (Moore et al. 2016). Dynamic SDRs and fast turnover of SDRs are likely to be quite common in plants, in which genetic control of sex appears to be poorly conserved (Charlesworth 2015; Moore et al. 2016). Studies focusing on the turnover of sex chromosomes are mostly from animals. The temporal order and directional trends of turnovers in sex-chromosomal rearrangement is not well understood due to this false impression (Bergero & Charlesworth 2009). Recently, a study on the SDRs of *Fragaria* octoploids provided the first case of transposition of a cassette of 14 kb of female-specific sequence among several chromosomes (Tennessen et al. 2018). However, this female-specific sequence does not answer questions about switches between male and female heterogamety, e.g. ZW and XY in Salicaceae. Determining the transition type at the chromosomal level, and the evolutionary forces responsible for transitions between male and female heterogamety are still not well understood (van Doorn & Kirkpatrick 2010). In plants, dioecy evolved independently in many clades allowing for a comparative approach that may reveal commonalities and peculiarities among independent origins of sex chromosomes (Ming et al. 2011).

Upon successful completion of my research project, we expect our contribution to be the first long-reads assembled Y chromosome with sex-specific contigs in *P. trichocarpa*, and as well as a W chromosome with sex-specific contigs in *S. purpurea*. We also expect to provide a detailed characterization of W and Y chromosomes, delineate important differences between X and Y chromosomes, and present evidence of shared mechanisms between *S. purpurea* and *P. trichocarpa*. A series of important questions in terms of the size of the SDR, its gene content, and other genomic features are inconclusive because of limited knowledge about the Y and W chromosomes. The completion of my research will be innovative because it will establish the first genomic assembly of the sex chromosomes and the SDRs therein by using novel methods to detect both male-specific sequences in the Y and female-specific sequences in the W. The products of the proposed work provide an essential genomic resource for both breeding projects and studies of the evolution of sex in plants. This will immediately solve the demand for genomic resources about sex chromosome in breeding and reaches new horizons in the evolution of sex chromosomes in plants.

References

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

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. 14:113–24.

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

Beukeboom LW, Perrin N (2014) The Evolution of Sex Determination. Oxford University Press, New York, NY

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. Philos. Trans. R. Soc. London. Ser. B Biol. Sci. 355:1563–1572.

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

Charlesworth D. 2016. Plant Sex Chromosomes. Annu. Rev. Plant Biol. 67:397–420.

Chin C-S et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10:563–569.

Cronk QCB, Needham I, Rudall PJ. 2015. Evolution of Catkins: Inflorescence Morphology of Selected Salicaceae in an Evolutionary and Developmental Context. Front. Plant Sci. 6:1030.

Delph LF, Arntz AM, Scotti-Saintagne C, Scotti I. 2010. THE GENOMIC ARCHITECTURE OF SEXUAL DIMORPHISM IN THE DIOECIOUS PLANT SILENE LATIFOLIA. Evolution (N. Y). 64:2873–2886.

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

Geraldes A et al. 2015. Recent y chromosome divergence despite ancient origin of dioecy in poplars (Populus). Mol. Ecol. 24:3243–3256.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17:333–351.

Gunter LE, Roberts GT, Lee K, Larimer FW, Tuskan G a. 2003. The development of two flanking SCAR markers linked to a sex determination locus in Salix viminalis L. J. Hered. 94:185–189.

Harkess A et al. 2017. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nat. Commun. 8:1279.

Henry IM, Akagi T, Tao R, Comai L. 2018. One Hundred Ways to Invent the Sexes: Theoretical and Observed Paths to Dioecy in Plants. Annu. Rev. Plant Biol. 69:553–575.

Hobza R et al. 2017. Impact of repetitive elements on the Y chromosome formation in plants. Genes (Basel). 8.

Hou J et al. 2015. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci. Rep. 5:e9076.

Jayakumar V, Sakakibara Y. 2017. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. Brief. Bioinform. 20:866–876.

Kersten B, Pakull B, Groppe K, Lueneburg J, Fladung M. 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. (Stuttg). 16:411–418.

Koren S et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736.

Li C, Lin F, An D, Wang W, Huang R. 2018. Genome sequencing and assembly by long reads in plants. Genes (Basel). 9.

Lin J, Gibbs JP, Smart LB. 2009. Population genetic structure of native versus naturalized sympatric shrub willows (Salix; Salicaceae). Am. J. Bot. 96:771–785.

Ma T et al. 2013. Genomic insights into salt adaptation in a desert poplar. Nat. Commun. 4:2797. https://doi.org/10.1038/ncomms3797.

McKown AD et al. 2017. Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in Populus. Sci. Rep. 7:1831.

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

Moore RC, Harkess AE, Weingartner LA. 2016. How to be a seXY plant model: A holistic view of sex-chromosome research. Am. J. Bot. 103:1379–1382.

Pakull B et al. 2011. Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a Populus tremula × Populus tremuloides cross. Can. J. For. Res. 41:245–253.

Pakull B, Kersten B, Lüneburg J, Fladung M. 2014. A simple PCR-based marker to determine sex in aspen. Plant Biol. 047300:256–261.

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

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.

Sterck L et al. 2005. EST data suggest that poplar is an ancient polyploid. New Phytol. 167:165–170.

Tennessen JA et al. 2018. Repeated translocation of a gene cassette drives sex-chromosome turnover in strawberries. PLOS Biol. 16:e2006062. https://doi.org/10.1371/journal.pbio.2006062.

Tuskan GA et al. 2006. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science (80-.). 313:1596–1604.

Tuskan GA 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.

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

Zhou R et al. 2018. Characterization of a large sex determination region in Salix purpurea L. (Salicaceae). Mol. Genet. Genomics 293:1437–1452.

Zsuffa L. 1990. Genetic improvement of willows for energy plantations. Biomass 22:35–47.

CHAPTER II

CHARACTERIZATION OF A LARGE SEX DETERMINATION REGION IN *SALIX PURPUREA* L. (SALICACEAE)

Abstract

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.

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 & 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*.

Methods

Genome Assembly

This work is based on v1.0 of the *S. pupurea* genome (available at http://phytozome.jgi.doe.gov). Briefly, 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 ~140X 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% to 85% of the willow genome is captured in the current assembly. Gene annotations were accomplished using the Phytozome pipeline (Goodstein et al. 2012). The RepeatModeler (v1.0.8) package (http://www.repeatmasker.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 8,531 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.mstmap.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 135X and 117X, 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.uniprot.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 cut-off (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://phytozome.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). Briefly, total RNA was extracted using the Spectrum^{TM} Total Plant RNA Kit. Libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit. Libraries were sequenced on the Illumina HiSeq platform (1x100 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 cM 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 cM 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 Mbp to 1.9 Mbp and from ~10.9 Mbp 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 ×

10^{-7}, Fig. 2; Fig. S3). Among these markers, 41 were located on chr 15, from 10.7 Mb 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. Additionally, 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*.

S. 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 ($P < 2.2\times10^{-16}$). 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 nonrecombining 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 Mb to 15.3 Mb) overlaps with a large region (9.8 Mb to 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 and 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 *P*5948, <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 4 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.

Table 2.1 Bayesian credible intervals for sex QTL on chromosome 15.

	Physical Map		Genetic Map	
	Start (bp)	End (bp)	Start (cM)	End (cM)
Female Map	10,939,613	11,569,298	225.42	240.17
Male Map	372,445	1,881,243	326.48	347.17
Intercross	11,401,384	15,091,498	55.69	58.22

Table 2.2 Length of intervals that lacked coverage in alignments of 2x250 bp reads against the reference genome assembly (also derived from female clone 94006). Number in the parentheses is the percentage of the total genome composition in that category that lacked coverage.

	Whole Genome	Fish Creek (♂)	94006 (♀)
Total Length	348,745,509	14,564,089 (4.18)	562,813 (0.16)
Chromosomes	251,661,964	7,693,428 (3.06)	303,356 (0.12)
Scaffolds	97,083,545	6,870,661 (7.08)	259,457 (0.27)
Repeats	98,506,863	5,328,429 (5.41)	260,598 (0.26)
Genes	120,852,638	2,654,305 (2.20)	78,325 (0.06)
SDR	3,073,122	480,360 (15.63)	4,814 (0.16)

Table 2.3 Significantly overrepresented GO terms of candidate genes from SDR.

Description	GO term	Number of genes in SDR	Number of genes outside SDR	P value
Microtubule motor activity	GO:0003777	7	91	4.73×10^{-6}
Kinesin complex	GO:0005871	7	92	5.07×10^{-6}
Microtubule-based movement	GO:0007018	7	92	5.07×10^{-6}
Microtubule binding	GO:0008017	7	133	4.84×10^{-5}

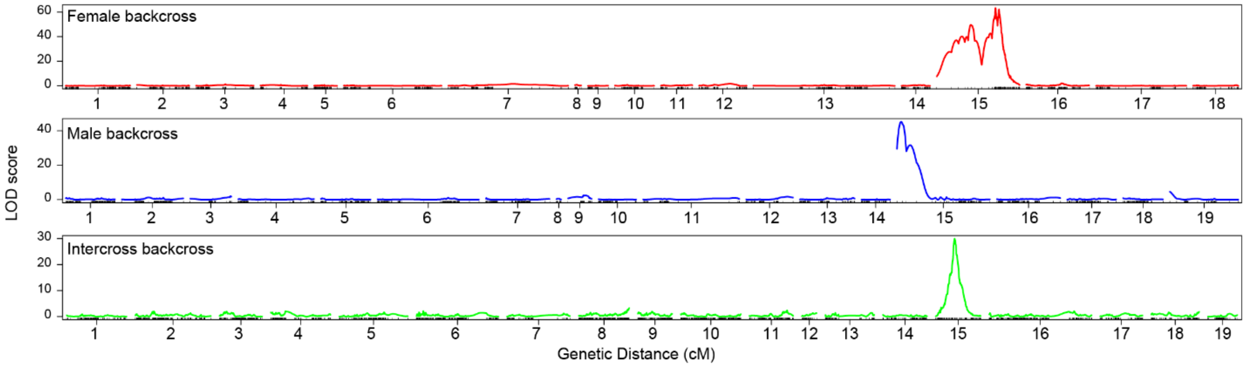


Figure 2.1 QTL for sex in an F_2 *S. purpurea* cross. From top to bottom are LOD scans for female backcross (red), male backcross (blue) and intercross (green) markers across the 19 major *S. purpurea* linkage groups. Chromosome 15 has a very strong QTL sex in all three maps, and the male backcross also shows a weak peak on chromosome 19 (LOD=4.68; table 1)**.**

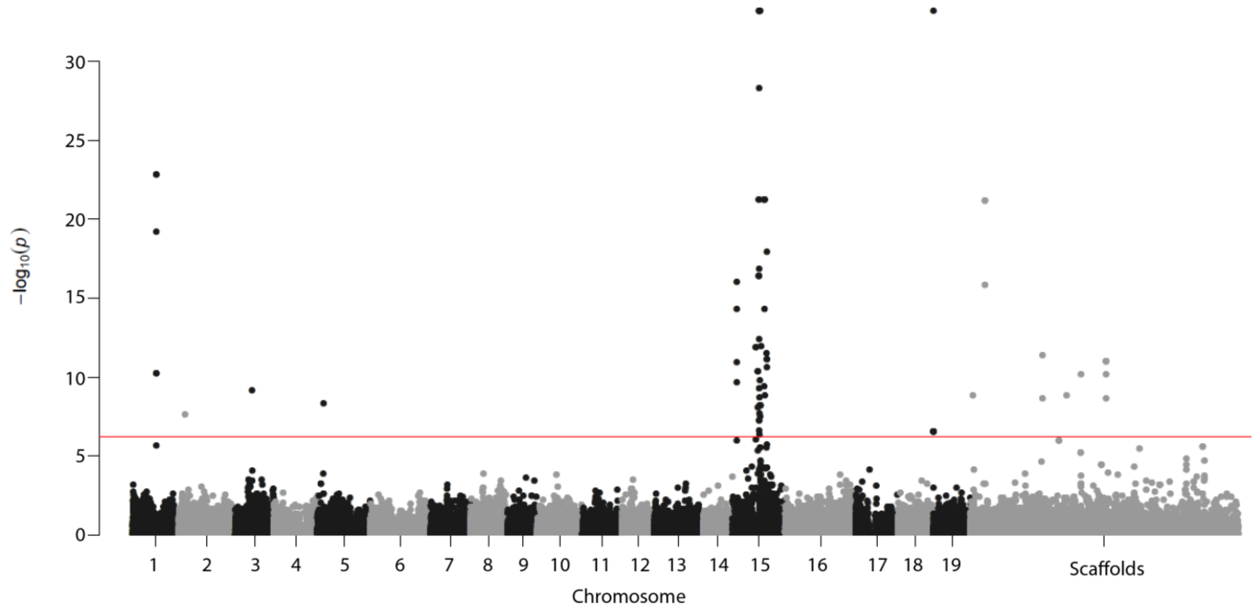


Figure 2.2 Manhattan plot derived from genome-wide association analysis for sex determination. The Y-axis shows the strength of association ($-\log_{10}$(P value)) for each SNP ordered by chromosome and SNP position (x axis). The horizontal line indicates significance after a Bonferroni correction for multiple testing.

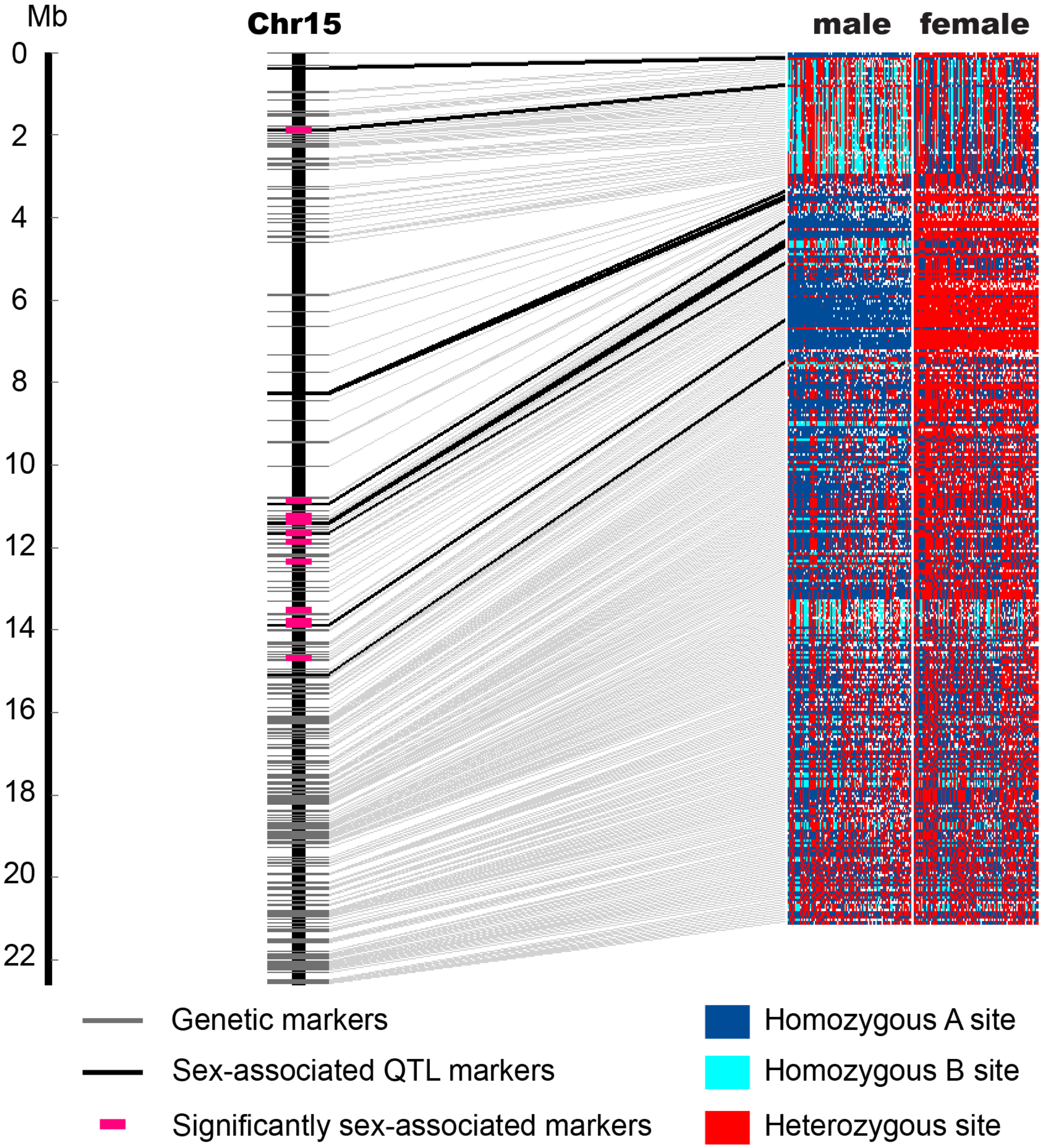


Figure 2.3 Genotype configuration of chromosome 15 in males and females from the F_2 family. Markers from all three genetic maps are shown as horizontal lines corresponding to their physical positions on the chromosome 15 physical assembly. Markers with top LOD scores in each map are colored as black. Significantly associated markers from the GWAS analysis with P < $1x10^{-7}$ are indicated by fuschia marks on the physical map. Each marker is connected between

physical map and its genotype configurations with 100 selected progenies of each sex. Genotypes of QTL markers are colored according to their homozygosity or heterozygosity.

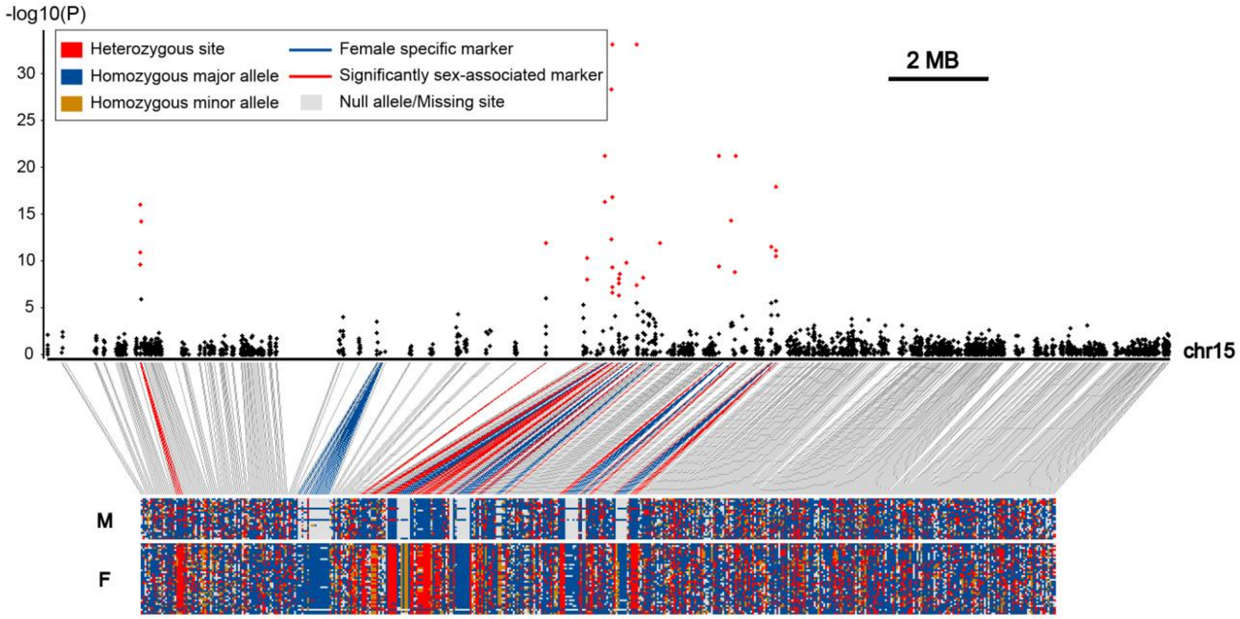


Figure 2.4 Genotype configurations of markers on chromosome 15 from the *S. purpurea* association population. The top is a blowup of chromosome 15 from the Manhattan plot in Fig. 2, with significantly sex-associated markers colored red. The bottom shows the genotype configurations in the association 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).

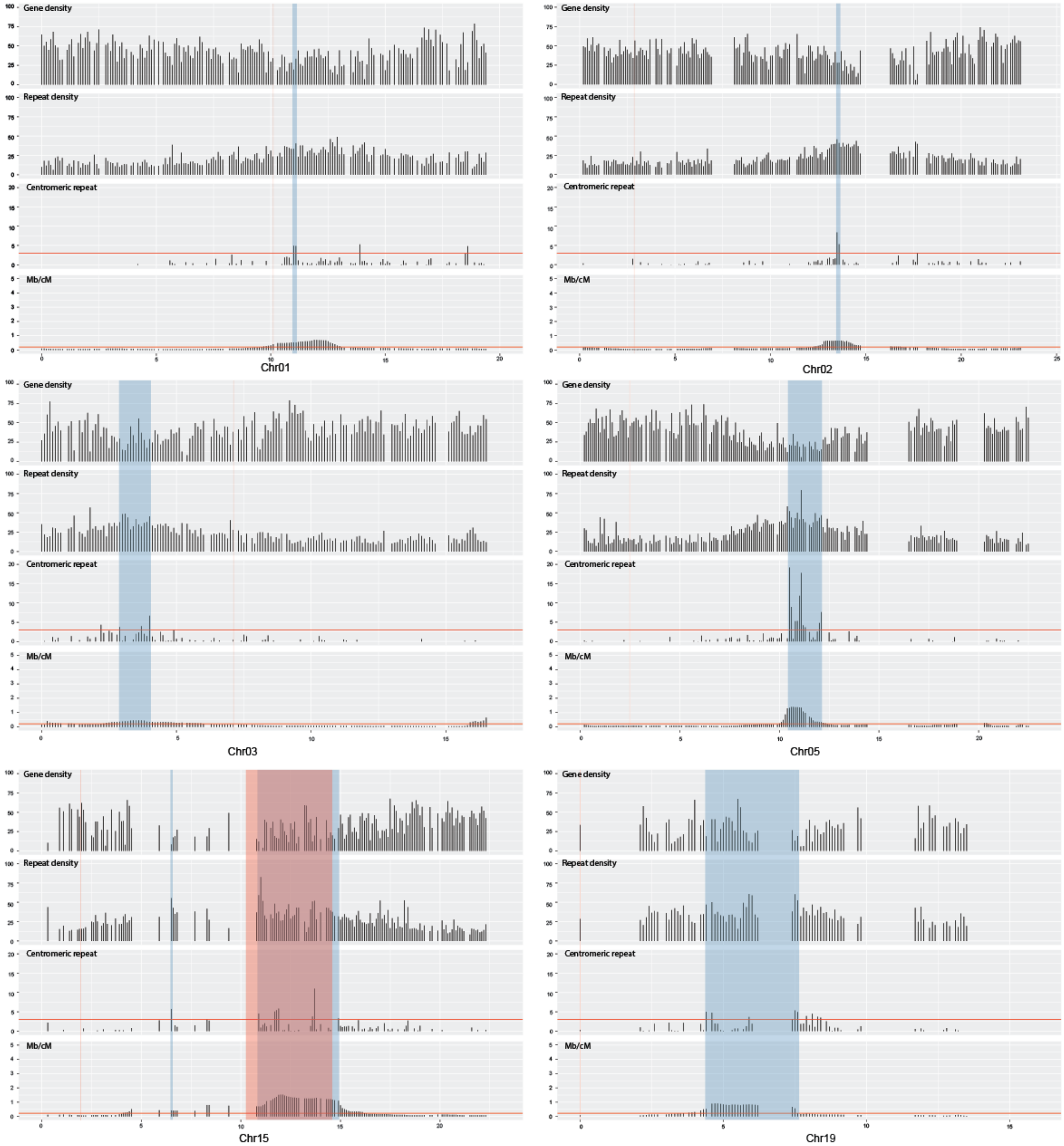


Figure 2.5 Delineation of putative centromeres relative to the SDRs. Bar plots represent, from the top, gene density, repeat density, density of centromeric repeats, and physical:genetic distance ratio (Mb/cM) in 100 kb windows. Blue shading shows positions of putative centromeres, as defined by empirical thresholds represented by horizontal red lines, and red shading represents the SDR.

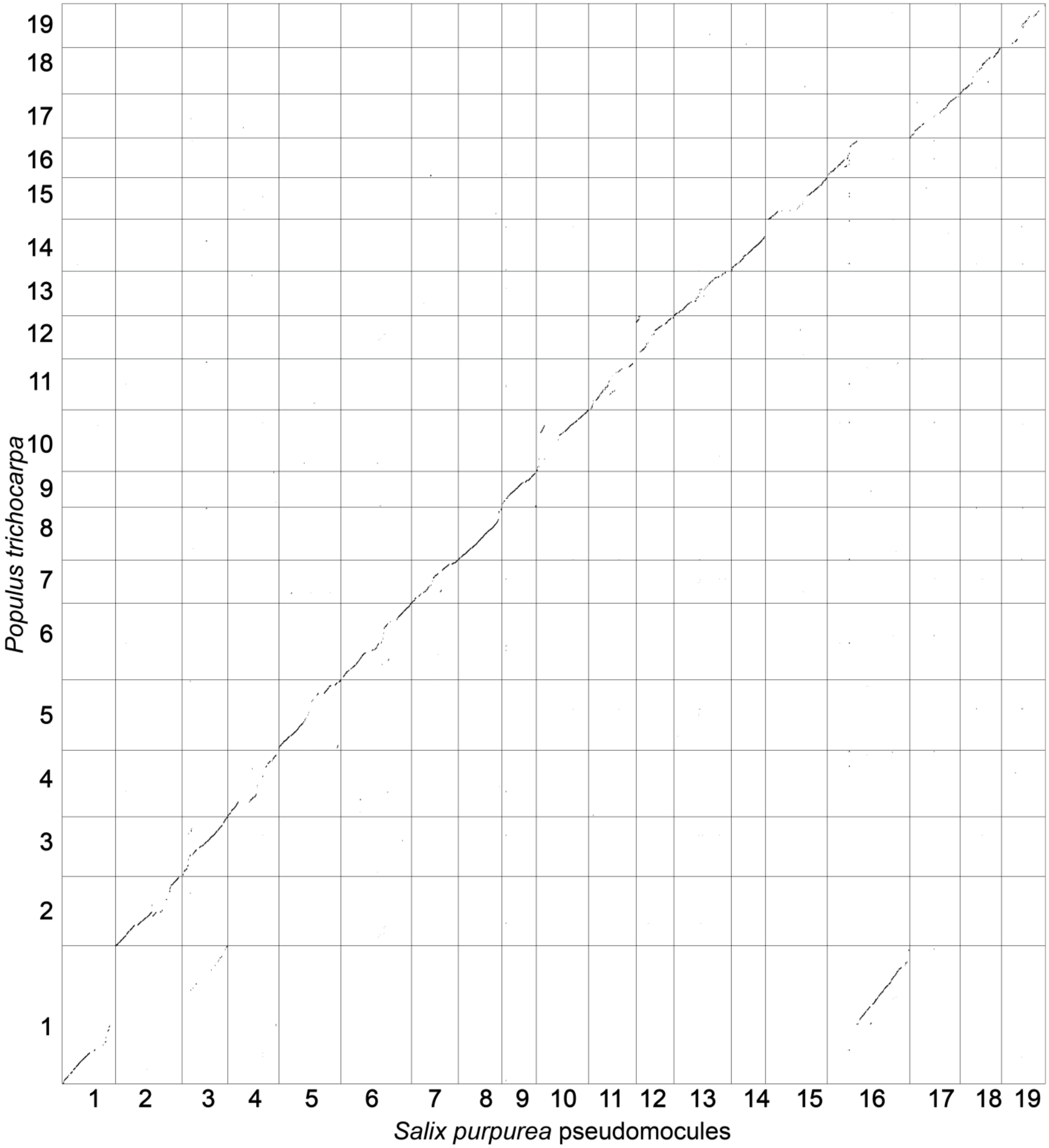


Figure 2.6 Comparison between the *S. purpurea* (x-axis) and *P. trichocarpa* (y-axis) genomes, with parameters set to exclude paralogous segments derived from the most recent whole genome duplication.

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 like 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.

Conclusions

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. 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.

Statement of Original Publication

This work was originally published in the Dec, 2018 edition of *Molecular Genetics and Genomics* (Zhou et al. 2018) https://doi.org/10.1007/s00438-018-1473-y

References

Ainsworth C (2000) Boys and girls come out to play: The molecular biology of dioecious plants. Ann Bot 86:211–221. doi: 10.1006/anbo.2000.1201

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. doi: 10.2307/2260821

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. doi: 10.1093/jhered/89.1.44

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. doi: 10.1038/nrg.2015.28

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

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

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

Ashman T-L, Kwok A, Husband BC (2013) Revisiting the Dioecy-Polyploidy Association: Alternate Pathways and Research Opportunities. Cytogenet Genome Res 140:241–255. doi: 10.1159/000353306

Bachtrog D, Mank J, Peichel CL, et al (2014) Sex Determination: Why So Many Ways of Doing

It? PLoS Biol 12:e1001899. doi: 10.1371/journal.pbio.1001899

Barrett SCH, Hough J (2013) Sexual dimorphism in flowering plants. J Exp Bot 64:67–82. doi: 10.1093/jxb/err313

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

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. doi: 10.1534/genetics.106.070110

Beukeboom LW, Perrin N (2014) The Evolution of Sex Determination. Oxford University Press, New York, NY

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. doi: 10.3732/ajb.90.9.1389

Bull JJ, Charnov EL (1977) Changes in the heterogametic mechanism of sex determination. Heredity 39:1–14. doi: 10.1038/hdy.1977.38

Burby PE, Simmons LA (2017) MutS2 promotes homologous recombination in *Bacillus subtilis*. J Bacteriol 199:e00682-16. doi: 10.1128/JB.00682-16

Cai G, Cresti M (2009) Organelle motility in the pollen tube: a tale of 20 years. J Exp Bot 60:495–508. doi: 10.1093/jxb/ern321

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. doi:

10.1093/gbe/evx174

Charlesworth D (2006) Evolution of Plant Breeding Systems. Curr Biol 16:726–735. doi: 10.1016/j.cub.2006.07.068

Charlesworth D (2016) Plant Sex Chromosomes. Annu Rev Plant Biol 67:397–420. doi: 10.1146/annurev-arplant-043015-111911

Charlesworth D (2015) Plant contributions to our understanding of sex chromosome evolution. New Phytol 208:52–65. doi: 10.1111/nph.13497

Charlesworth D, Charlesworth B (1978) Population genetics of partial male-sterility and the evolution of monoecy and dioecy. Heredity 41:137–153. doi: 10.1038/hdy.1978.83

Charnov EL (1982) The theory of sex allocation. Monogr Popul Biol 18:1–355. doi: 10.1017/CBO9781107415324.004

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. doi: 10.1371/journal.pone.0147671

DePristo M a, 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. doi: 10.1038/ng.806

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

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. doi: 10.1371/journal.pone.0019379

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. doi: 10.1186/1471-2105-13-S4-S14

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. doi: 10.1093/nar/gkl735

Fukui K, Kuramitsu S (2011) Structure and Function of the Small MutS-Related Domain. Mol Biol Int 2011:1–9. doi: 10.4061/2011/691735

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. doi: 10.1007/s10886-007-9257-6

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. doi: 10.1007/s11295-007-0085-1

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. doi: 10.1111/mec.13126

Glaubitz JC, Casstevens TM, Lu F, et al (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e0090346. doi: 10.1371/journal.pone.0090346

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. doi: 10.3732/ajb.1500424

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 U S A 108:1513–1518. doi: 10.1073/pnas.1017351108

Goodstein DM, Shu S, Howson R, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186. doi: 10.1093/nar/gkr944

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. doi: 10.1093/gbe/evw127

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. doi: 10.1038/srep09076

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. doi: 10.1038/ng.548

Karp A, Hanley SJ, Trybush SO, et al (2011) Genetic Improvement of Willow for Bioenergy and Biofuels. J Integr Plant Biol 53:151–165. doi: 10.1111/j.1744-7909.2010.01015.x

Karrenberg S, Kollmann J, Edwards PJ (2002) Pollen vectors and inflorescence morphology in four species of *Salix*. Plant Syst Evol 235:181–188. doi: 10.1007/s00606-002-0231-z

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. doi: 10.1111/plb.12048

Kunkel T a, Erie D a (2005) DNA mismatch repair. Annu Rev Biochem 74:681–710. doi: 10.1146/annurev.biochem.74.082803.133243

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. doi: 10.4236/jbm.2016.412018

Lin H, Niu L, McHale N a, et al (2013) Evolutionarily conserved repressive activity of WOX proteins mediates leaf blade outgrowth and floral organ development in plants. Proc Natl Acad Sci U S A 110:366–371. doi: 10.1073/pnas.1215376110

Lloyd DG (1979) Evolution towards dioecy in heterostylous populations. Plant Syst Evol 131:71–80. doi: 10.1007/BF00984123

Mank JE (2009) Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. Am Nat 173:141–150. doi: 10.1086/595754

Melters DP, Bradnam KR, Young H a, et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10. doi: 10.1186/gb-2013-14-1-r10

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327. doi: 10.1016/j.ygeno.2010.03.001

Ming R, Bendahmane A, Renner SS (2011) Sex chromosomes in land plants. Annu Rev Plant Biol 62:485–514. doi: 10.1146/annurev-arplant-042110-103914

Ming R, Moore PH (2007) Genomics of sex chromosomes. Curr Opin Plant Biol 10:123–130. doi: 10.1016/j.pbi.2007.01.013

Mock KE, Callahan CM, Islam-Faridi MN, et al (2012) Widespread triploidy in western North American aspen (*Populus tremuloides*). PLoS One 7:e48406. doi: 10.1371/journal.pone.0048406

Moore EC, Roberts RB (2013) Polygenic sex determination. Curr Biol 23:R510-2. doi: 10.1016/j.cub.2013.04.004

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. doi: 10.1371/journal.pbio.0030004

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. doi: 10.1016/S0092-8674(02)00691-8

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

Otto SP, Pannell JR, Peichel CL, et al (2011) About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet 27:358–367. doi: 10.1016/j.tig.2011.05.001

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. doi: 10.1007/s11295-009-0204-2

Pakull B, Kersten B, Lüneburg J, Fladung M (2014) A simple PCR-based marker to determine sex in aspen. Plant Biol 17:256–261. doi: 10.1111/plb.12217

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. doi: 10.1007/s11103-015-0422-y

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. doi: 10.1007/s11295-010-0297-7

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

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. doi: 10.1038/ng1847

Pucholt P, Hallingbäck HR, Berlin S (2017a) Allelic incompatibility can explain female biased sex ratios in dioecious plants. BMC Genomics 18:251. doi: 10.1186/s12864-017-3634-5

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. doi: 10.1038/hdy.2014.125

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. doi: 10.1093/molbev/msx144

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. doi: 10.1073/pnas.1321897111

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. doi: 10.3732/ajb.1400196

Rice WWR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution

38:1416–1424. doi: 10.2307/2408385

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. doi: 10.1007/s12155-014-9521-x

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, New York, NY, pp 15–38

Sterck L, Rombauts S, Jansson S, et al (2005) EST data suggest that poplar is an ancient polyploid. New Phytol 167:165–170. doi: 10.1111/j.1469-8137.2005.01378.x

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. doi: 10.1139/B07-058

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. doi: 10.1007/s11295-012-0495-6

Tuskan GA, DiFazio S, Jansson S, et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604. doi: 10.1126/science.1128691

Ueno N, Suyama Y, Seiwa K (2007) What makes the sex ratio female-biased in the dioecious tree *Salix sachalinensis*? J Ecol 95:951–959. doi: 10.1111/j.1365-2745.2007.01269.x

van Doorn GS, Kirkpatrick M (2007) Turnover of sex chromosomes induced by sexual conflict. Nature 449:909–912. doi: 10.1038/nature06178 van Doorn GS, Kirkpatrick M (2010) Transitions between male and female heterogamety caused by sex-antagonistic selection. Genetics 186:629–645

Vyskot B, Hobza R (2015) The genomics of plant sex chromosomes. Plant Sci 236:126–135. doi: 10.1016/j.plantsci.2015.03.019

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. doi: 10.1073/pnas.1207833109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1207833109

Westergaard M (1958) The Mechanism of Sex Determination in Dioecious Flowering Plants. Adv Genet 9:217–281. doi: 10.1016/S0065-2660(08)60163-7

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. doi: 10.1093/molbev/msm088

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. doi: 10.1101/gr.7076308

Zhou W, Lu Q, Li Q, et al (2017) PPR-SMR protein SOT1 has RNA endonuclease activity. Proc Natl Acad Sci U S A 114:E1554–E1563. doi: 10.1073/pnas.1612460114

Supplementary Materials

Link to supplemental tables https://static-content.springer.com/esm/art%3A10.1007%2Fs00438-018-1473-y/MediaObjects/438_2018_1473_MOESM1_ESM.xlsx

Supplemental Table 2.1 Significant markers (LOD>3.5) from QTL mapping of sex. The table includes linkage group (LG), map positions (in centimorgans), map type (female backcross, F, male backcross, M, and intercross, IC), the physical scaffold from the genome assembly, the physical position of the marker in the genome assembly, and the frequency of different genotype configurations in the progeny.

Supplemental Table 2.2 Number of unfiltered GBS markers produced by the Tassel pipeline for the F_2 family 317. Markers/100kb is the average number of markers per 100 kb interval. F:M Backcross is the ratio of markers in a Female Backcross configuration (heterozygous in the female parent, homozygous in the male parent) to markers in the Male Backcross configuration (homozygous in female parent, heterozygous in male parent).

Supplemental Table 2.3 Results of GWAS for sex. The table includes all significant markers ($p<1x10^{-7}$).

Supplemental Table 2.4 Best matches for secondary *S. purpurea* SDRs to the *S. purpurea* and *P. trichocarpa* genomes. “Secondary Blast Hit” is the best blastn hit to the *S. purpurea* genome, after excluding self hits.

Supplemental Table 2.5 Markers showing a female-specific genotype configuration (one allele observed in females, none in males). These are presumably derived from W segments included in the genome assembly.

Supplemental Table 2.6 Scaffolds with >30% female-specific sequence. "Proportion W" is a calculation based on the proportion of the scaffold, after excluding gaps, that is present in the female sequence but absent in the male sequence (Female-Specific).

Supplemental Table 2.7 Repeat composition of the *S. purpurea* chromosomes.

Supplemental Table 2.8 Predicted genes found within the SDR of *S. purpurea*. "W Overlap" and "W proportion" represent the intersection of the location of the gene with female-specific genome segments. Omega values are the ratio of nonsynonymous (dN) to synonymous (dS) substitutions between the *S. purpurea* and *P. trichocarpa* orthologs. Multiple values are provided in cases with multiple *Populus* orthologs, presumably due to lineage-specific expansion.

Link to supplemental figures: https://static-content.springer.com/esm/art%3A10.1007%2Fs00438-018-1473-y/MediaObjects/438_2018_1473_MOESM2_ESM.pdf

Supplemental Figure 2.1 Pairwise Scaled Identity by State (IBS) for the (a) complete association population (N=112), (b) the complete F_2 full sib Family (N=497), and (c) the association population with clones removed (N=75). The IBS cutoff used for identifying clonal pairs was 0.9.

Supplemental Figure 2.2 Frequency of mapped markers with and without segregation distortion in family 317 for males and females. A. Markers in female-backcross configuration. B. Markers in male-backcross configuration. Notice the lack of undistorted (normal) markers on chr 19 in female backcross configuration.

Supplemental Figure 2.3 Quantile–Quantile (Q–Q) plots of observed and expected P-values for the GWAS for sex. Red line indicates X = Y.

Supplemental Figure 2.4 Box plots of average observed heterozygosity for males and females for sex-associated loci in the *S. purpurea* association population.

Supplemental Figure 2.5 Distribution of differences in null allele frequency between females and males in the association population. Extreme values are shaded in red.

Supplemental Figure 2.6 Proportion of reference sequence gaps ("assembly Ns") in regions that showed no coverage in the female (a) or male (b) reference-based alignments. The male had 0 coverage primarily in regions with minimal reference gaps, suggesting that these are regions that are present in the female sequence and absent in the male.

Supplemental Figure 2.7 Box plot showing that the proportion of repeat elements is elevated in the SDR.

Supplemental Figure 2.8 Delineation of putative centromeres relative to the SDRs, for chromosomes not shown in the main text. 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. The position of the SDRs are indicated by vertical red shading.

Supplemental Figure 2.9 Box plots comparing the composition of putative centromeric intervals to the rest of the genome, including (from top to bottom) gene content, total repeat content, presence of putative centromere-associated repeat elements, and physical:genetic distance ratio (Mb/cM).

Supplemental Figure 2.10 Dot plot derived from aligning the *S. suchowensis* SDR (primarily located on scaffold64) to *S. purpurea* chr 15 using lastz.

Supplemental Figure 2.11 Alignment of Kinesin genes from the SDR of *S. purpurea* and their closest ortholog in *P. trichocarpa.* SapurV1A.1267s0010 is artificially truncated due to an assembly gap overlapping with the gene. Conserved domains are highlighted and labeled. Tandem duplicate pairs are 1.) SapurV1A.0719s0080 and SapurV1A.0719s0090; and 2.) SapurV1A.1267s0010 and SapurV1A.1267s0020.

CHAPTER III

A WILLOW SEX CHROMOSOME REVEALS CONVERGENT EVOLUTION OF COMPLEX PALINDROMIC REPEATS

Abstract

Background: Sex chromosomes have arisen independently in a wide variety of species, yet they share common characteristics, including the presence of suppressed recombination surrounding sex determination loci. Mammalian sex chromosomes contain multiple palindromic repeats across the non-recombining region that show sequence conservation through gene conversion, and contain genes that are crucial for sexual reproduction. In plants, it is not clear if palindromic repeats play a role in maintaining sequence conservation in the absence of homologous recombination.

Results: Here we present the first evidence of large palindromic structures in a plant sex chromosome, based on a highly contiguous assembly of the W chromosome of the dioecious shrub *Salix purpurea*. The W chromosome has an expanded number of genes due to transpositions from autosomes. It also contains two consecutive palindromes that span a region of 200 kb, with conspicuous 20 kb stretches of highly conserved sequences among the four arms that show evidence of gene conversion. Four genes in the palindrome are homologous to genes in the sex determination regions of the closely related genus *Populus*, which is located on a different chromosome. These genes show distinct, floral-biased expression patterns compared to paralogous copies on autosomes.

Conclusion: The presence of palindromes in sex chromosomes of mammals and plants highlights the intrinsic importance of these features in adaptive evolution in the absence of recombination. Convergent evolution is driving both the independent establishment of sex chromosomes as well as their fine-scale sequence structure.

Introduction

Sex chromosomes carry genes that confer or control sex-specific traits (Bachtrog 2013). In theory, the heterogametic (sex-specific) sex chromosome evolved from an autosome. There are two important features in sex determination regions (SDRs): suppressed recombination and the presence of sequences that only occur in one sex (Bachtrog 2013). Furthermore, many sex chromosomes have lost most of their original genes over evolutionary time, and accumulated repetitive sequences such as transposable elements and tandem gene duplications (Charlesworth 2013; Charlesworth & Charlesworth 1978). Consequently, sex chromosomes can be difficult to sequence because they are often highly heterochromatic and have a large amount of repetitive and ampliconic DNA (Skaletsky et al. 2003; Bachtrog 2013).

A striking characteristic of mammalian sex chromosomes is the presence of large palindromes in ampliconic regions of the X and Y chromosomes that consist of large inverted repeats with highly identical sequences that are undergoing gene conversion (Betrán et al. 2012a; Trombetta & Cruciani 2017). Ampliconic sequences on the human Y chromosome were acquired through transpositions from diverse sources, and then amplified (Skaletsky et al. 2003). These ampliconic sequences account for about 30% of the Y euchromatin (Skaletsky et al. 2003). The human Y chromosome palindromes contain eight gene families that are expressed predominantly in the testes and which are essential for spermatogenesis (Navarro-Costa et al. 2010; Trombetta & Cruciani 2017; Krausz & Casamonti 2017). These genes undergo extensive gene conversion and have high sequence identity among the copies (Trombetta & Cruciani 2017). Other palindromes occur in the genome, but those on the sex chromosomes are by far the largest and have the highest rates of gene conversion (Trombetta & Cruciani 2017; Warburton et al. 2004). Palindromes have also been found on the W chromosomes of New World sparrows and blackbirds, suggesting that this may be a widespread feature of sex chromosomes (Davis et al. 2010). However, such structures have not yet been described in plants.

Unlike in most animals, there is a lack of obvious sex chromosome heteromorphism in most dioecious plant species (i.e., differences are not readily discernable by cytology) (Ming et al. 2011; Bachtrog et al. 2014). Sex determination systems are quite diverse in plants, and the mechanisms of sex determination have been identified for an increasing number of species in recent years (Henry et al. 2018). For example, Y chromosomes have been intensively studied in papaya and persimmon. Both of these contain a female suppressor on the Y chromosome (Jianping Wang et al. 2012; Akagi et al. 2016; Henry et al. 2018). Recently, a female suppressing gene in asparagus has been identified on the Y chromosome using long-read sequencing technology with optical mapping (Harkess et al. 2017). Another study on octoploid strawberry found repeated transpositions of a female-specific gene cassette (Tennessen et al. 2018). The genus *Silene* does have clearly heteromorphic sex chromosomes, and has been a long-standing model for sex determination in XY plants. Female-suppressing and male-promoting factors were identified in *Silene* the 1950's using genetic approaches (Westergaard 1958). More recently it has been shown that some species of *Silene* have ZW sex determination systems, though it remains unclear if there are commonalities in the underlying mechanisms of sex determination in XY and ZW species (Balounova et al. 2019).

Sex determination is similarly diverse within the Salicaceae family. SDRs have been consistently found on chromosome 15 with female heterogamety in multiple *Salix* species (Zhou et al. 2018; Pucholt et al. 2015; Hou et al. 2015). This is quite different from the closely-related genus *Populus* where sex determining regions consistently occur on chromosome 19, with most species showing male heterogamety (Tuskan et al. 2012; Geraldes et al. 2015). Previously, we reported that the SDR occupies a large portion of the W chromosome in *S. purpurea* with suppressed recombination extending over ~5 Mb (Zhou et al. 2018; Carlson et al. 2017). This is substantially larger than the SDR in *P. trichocarpa* and *P. balsamifera*, which appears to be approximately 100 kb in size (Geraldes et al. 2015; McKown et al. 2017). However, due to the structural complexity of the SDRs, none of these studies have thus far included an in-depth analysis of the sequence composition and structure of the SDRs, and it is unclear whether there is a common underlying mechanism of sex determination. Here we present a much more complete assembly of the *S. purpurea* W chromosome and report for the first time in plants a palindromic repeat structure that is similar to the one found on mammalian Y chromosomes. We also demonstrate that gene content is expanded on the W chromosome, and homologous genes occur in the *Salix* and *Populus* SDRs, suggesting that there may be some overlap in the underlying mechanisms of sex determination in this family.

Methods

Initial assembly of the genome

Whole genome assemblies were produced for two *S. purpurea* clones: female clone 94006, and a male offspring of this clone, "Fish Creek" (clone 9882-34), which was derived from a controlled cross between clone 94006 and male *S. purpurea* clone 94001. Clones 94001 and 94006 were collected from naturalized populations in upstate New York, USA. Sequencing reads were collected using the Illumina and PACBIO platforms at the Department of Energy (DOE) Joint Genome Institute (JGI) in Walnut Creek, California and the HudsonAlpha Institute in Huntsville, Alabama. Illumina reads were sequenced using the Illumina HISeq platform, and the PACBIO reads were sequenced using the RS platform. One 400bp insert 2x250 Illumina fragment library was sequenced for total coverage of 183x in clone 94006 and 153x in Fish Creek. Prior to use, Illumina reads were screened for mitochondria, chloroplast, and ΦX174 contamination. Reads composed of >95% simple sequence were removed. Illumina reads <50bp after trimming for adapter and base quality (q<20) were removed. For the PACBIO sequencing, a total of 47 P6C4 chips (10 hour movie time) were sequenced for each genome with a p-read yield of 39 Gb and a total coverage of ~110x per genome (Additional file 1: Table S11). The assembly was performed using FALCON-UNZIP (Chin et al. 2016) and the resulting sequence was polished using QUIVER (Chin et al. 2013). Finally, to correct false polymorphisms resulting from errors in PacBio reads, homozygous SNPs and INDELs were corrected in the release consensus sequence using ~80x of the 2x250 Illumina reads from the reference individual. This was accomplished by aligning the reads using bwa mem and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al. 2010)(Additional file 1: Table S12).

Chromosome-scale assemblies were created using a genetic map derived from 3,697 GBS markers generated for a family of 497 F_2 progeny from a cross in which the male reference is the father and the female reference is the grandmother. This map is described more completely in a previous publication (Carlson et al. 2019). This intercross map was used to identify misjoins, characterized by an abrupt change in the *S. purpurea* linkage group. Scaffolds were then oriented, ordered, joined, and numbered using the intercross map and the existing 94006 v1 release assembly (Zhou et al. 2018). Adjacent alternative haplotypes were identified on the joined contigs, and these regions were then collapsed using the longest common substring between the two haplotypes. Significant telomeric sequence was identified using the $(TTTAGGG)_n$ repeat, and care was taken to make sure that it was properly oriented in the production assembly. The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Completeness of the euchromatic portion of the assembly was assessed by aligning *S. purpurea* var 94006 v1 annotated genes to the assemblies. In both cases, 99.7% of the genes were found.

Identification of W contigs

Contigs derived from the W chromosome are expected to contain some large indels compared to contigs from the Z chromosome due to the lack of recombination between W and Z. These hemizygous regions should exclusively occur in the W haplotype of SDR. To identify these regions, we aligned 2x250 bp Illumina resequencing reads from female clone 94006 and male clone Fish Creek to the new reference using Bowtie2 (Langmead & Salzberg 2012a). Depth of coverage was extracted using samtools-1.2 (Li et al. 2009). Median depth was calculated using a non-overlapping sliding window of 10 kb.

To verify if these hemizygous regions are strictly inherited in only female individuals, we used the GBS data from the F_2 family. GBS reads of 195 offspring of each sex were aligned to the v5 reference with Bowtie2. Due to low coverage and depth of the GBS markers per locus per individual, bam files were merged according to sex in samtools-1.2. Depth was then called in Samtools-1.2 with and max depth was limited to 80,000. Regions continuously covered by GBS reads were defined as GBS intervals. Then, the median of each sex was calculated across all of the intervals. We defined markers as female-specific by integrating the depth from both the F_2 GBS and 2x250 datasets (restricted to the GBS intervals) using two rules: 1) $\log_2(\frac{M_{195}+1}{F_{195}+1})<L$, where L is the lower bounds of the distribution, defined by the fifth percentile divided by the number of intervals tested (Additional file 2: Fig. S7); and 2) $\log_2(\frac{94006_{2by250}+1}{Fish\ Creek_{2by250}+1})>5$. The cutoff for the second criterion was based on the occurrence of a distinct peak in the distribution of the ratios (Additional file 2: Fig. S8). Scaffolds that contained at least three sex-linked markers were selected as candidate W scaffolds. Based on these criteria, only two contigs from the original Chr15 assembly were from W contigs, and the rest were from Z (Additional file 1: Table S5; Additional file 2: Fig. S1a).

Assembly of the Z and W chromosomes

Raw GBS reads used for the original map were demultiplexed and trimmed down to 64 bp for each read by process_radtags (in Stacks 1.44 (Catchen et al. 2013)) with -c -q -r -t 64. Then, trimmed reads of each sequenced individual from the F_2 family were aligned to the 19 chromosomes and unmapped scaffolds from the main genome and alternative haplotypes from the v4 reference of 94006 using Bowtie 2 (Langmead & Salzberg 2012b) with the --very-sensitive flag (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50) to maintain a balance between sensitivity and accuracy. Upon examining the distribution of SNPs in the genome, it became clear that the alternative haplotypes were preventing us from retrieving markers in some regions in the genome, so we repeated the alignments using three different reference sequences: 1) the 19 chromosomes, 2) unmapped scaffolds, and 3) alternative haplotypes. Then, a wrapper script ref_map.pl in Stacks was used to call genotypes with -m 5 (minimum number of reads to create a tag for parents) and -P 3 (minimum number of reads to create a tag for an offspring) on all progeny. Cross type “CP” was chosen since it was the one closest to our cross. Offspring with poor coverage were removed from the downstream analysis.

Once all genotypes were retrieved through Stacks, markers from different loci showing the exact same genotype/segregation across the progeny were binned and only markers from the main genome were kept for mapping. Markers with severe segregation distortion or excessive missing data were excluded, along with twelve offspring with very low call rate. Genotypes were imputed and corrected based on inferring haplotypes in the two F_1 parents from segregation of the markers in the progeny.

The grandparents of the F_2 cross have extensive stretches of shared haplotypes, possibly due to historic inbreeding in this naturalized population. This results in long runs of heterozygosity and homozygosity in the F_1 progeny. This inhibits integration of backcross and intercross markers by available mapping algorithms like those in the Onemap package (Margarido et al. 2007). To circumvent this problem, all intercross markers were translated to female and male backcross markers by identifying the parental origins of alleles based on parental phases and physical position in the assembly. Also, putatively hemizygous markers were recoded as backcross markers using sequence depth to infer genotypes. For example, markers with the segregation pattern +/- x -/- were recoded as AB x BB. These genotypes were also imputed and corrected based on the inferred haplotypes of the two F1 parents.

Onemap v2.1.1 was used to form initial linkage groups. For each chromosome, there are two phased linkage groups from each backcross type. However, this phase information derived from the F_2 family is only for the F_1 parents, which cannot be directly used for phasing haplotypes in the grandmother, clone 94006. By comparing parental genotypes from one LG to those of the grandparents, we inferred which of the 94006 haplotypes were inherited by each F_1. These results were used as a piece of evidence for identifying W-linked scaffolds/contigs, as well as estimating the overall occurrence of chimeric contigs in the assembly. After building a framework genetic map using markers from the main genome, non-distorted markers from unmapped main scaffolds and alternative scaffolds were added.

All unmapped scaffolds were manually checked to see if they matched the phase information or contained sex-linked markers. Those that were identified as Z scaffolds/contigs were excluded from the W map. The new W and Z were assembled using the python package ALLMAPS (Tang et al. 2015) to order and orient scaffolds and reconstruct chromosomes based on the genetic map. Only the order of the female backcross map was used to assemble the W, and ALLMAPS was set not to break contigs. This new map-based assembly containing two versions of chromosome 15 (Chr15Z and Chr15W) is version 5 of the *S. purpurea* var. 94006 genome.

To identify Z-W homologous regions (analogous to X-degenerate regions in mammalian sex chromosomes) and insertions in the W haplotype, we realigned the 2x250 reads of 94006 and Fish Creek to the 94006 v5 reference using Bowtie2 as described above, except we removed Chr15Z from the reference. Depth was calculated using samtools, and the median depth of 50kb non-overlapping windows was calculated with an in-house perl script. Regions where medians of Fish Creek depth are no greater than 10 were considered as insertions in the FSW, and regions with greater depth were considered Z-W homologous regions. This analysis was repeated with a 10 kb window as well to enhance the resolution.

Annotation of the genome

Transcript assemblies were constructed from ~126M pairs of 2x76bp (94006) or 2x150bp (Fish Creek) paired-end Illumina RNA-seq reads using PERTRAN. 188,628 transcript assemblies were constructed using PASA from the RNA-seq transcript assemblies. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from *Arabidopsis thaliana*, soybean, poplar, cassava, brachypodium, grape, and Swiss-Prot proteomes, and high confidence *Salix purpurea* Fish Creek gene model peptides, with up to 2 kb extension on both ends unless extending into another locus on the same strand. The reference genome was soft-masked using RepeatMasker. Gene models were predicted by the homology-based predictors. FGENESH+, FGENESH_EST, and EXONERATE, by PASA assembly of ORFs, and from AUGUSTUS via BRAKER1. The best scored predictions for each locus were selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement included adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subjected to protein homology analysis to the above mentioned proteomes to obtain Cscore (the ratio of mutual best hit BLASTP scores) and percentage of protein aligned to the best homolog. The transcripts were selected if its Cscore was greater than or equal to 0.5 and protein coverage greater than or equal to 0.5. Alternatively, proteins with EST coverage were accepted if overlap with repeats was less than 20%. For gene models with greater than 20% CDS overlap with repeats, the Cscore cutoff was 0.9 and homology coverage was at least 70%. The selected gene models were subjected to Pfam analysis and gene models with more than 30% in Pfam TE domains were removed. Incomplete gene models with low homology and transcriptome support and short single exon proteins (< 300 BP CDS) lacking conserved domains or transcriptome support were manually filtered out.

To annotate potential genes or coding regions in the palindrome that were missed by the automated annotation, the full nucleotide sequence of arm1 (about 20 kb) was submitted to the Fgenesh online service (http://www.softberry.com/berry.phtml?topic=fgenesh) with specific gene-finding parameters for *Populus trichocarpa*. The predicted peptide sequences were searched against predicted proteins from *Populus trichocarpa* v3.0, and *Arabidopsis thaliana* TAIR10 in Phytozome 12 (https://phytozome.jgi.doe.gov/) to find the closest homologous annotation. The protein domains were identified using hmmscan in HMMER (v3.1b1, http://hmmer.org/) against the Pfam-A domains (release 32, https://pfam.xfam.org).

Comparison of Z and W orthologous genes

Homologous genes on the Z and W chromosomes (Z-W homologs) were identified by performing a reciprocal blastp of all primary annotated peptide sequences in the main genome with default parameters. Mutual best hits were identified with over 90% identity over at least 70% of the transcript. Tandem duplications were identified as genes with expectation values of $1x10^{-10}$ that occurred within a 500 kb window. In these cases, one representative gene from each tandem array was used as a representative sequence, and the mutual best hit outside the tandem array was identified as above. Genes that lacked hits in the Z-SDR were searched against the Populus trichocarpa v3.0 reference genome. Those with hits to Chr15 in Populus were designated as "Ancestral" under the assumption that the homolog was present prior to the establishment of the SDR in *S. purpurea*, but was subsequently lost from the Z-SDR. Those genes that lacked hits to Chr15 in either species but which had a mutual best hit meeting the above criteria to an autosomal gene were designated as autosomal transpositions. Genes that could not be readily categorized due to a lack of mutual best hits satisfying the above criteria were designated as "Non-mutual" or "No Hit" as appropriate.

To identify homologous gene pairs for calculation of synonymous substitutions between the Z and W alleles, a reciprocal blast of all primary annotated peptide sequences was run with "blastall –p blastp -i -e 1e-20 -b 5 -v 5 -m 8", and MCscanX was run with default parameters (Y. Wang et al. 2012). The synonymous and nonsynonymous substitution rate of each gene pair in each syntenic block (d_S and d_N, respectively) was estimated by aligning the sequences with CLUSTALW (Wilm et al. 2007) and using the yn00 function in PAML (Yang 2007). Only pairs between the W-SDR and Z-SDR (including the unmapped scaffold_844) were used for estimating the divergence between Z and W haplotypes. It is important to note that this analysis does not control for polymorphism within populations, so it may be an overestimate of divergence.

Identification of sex-associated loci

Loci associated with sex were identified using 60 non-clonal individuals from a naturalized population of *S. purpurea* (Gouker et al. 2019). GBS reads from each individual were aligned to the 94006v5 genome without Chr15Z using Bowtie2. Genotypes were called in Stacks 1.14 using the ref_map.pl wrapper and the populations module with a minimum minor allele frequency of 0.1 and a genotyping rate of 0.1. Loci with greater than 40% missing data were removed. Association with sex was performed using emmax (Kang et al. 2010) as described previously (Zhou et al. 2018).

Detection of palindromic repeats

We detected the palindromic repeats by aligning the SDR region to itself with LASTZ 1.03.66 with the following flags: --gapped --exact=100 --step=20. Paralogous gene copies on autosomes were retrieved from the reciprocal blastp results described above. Paralogous genes within the palindrome arms were aligned along with paralogous copies from the autosomes using Muscle using default parameters provided in MEGA 5. In a few cases, the resulting alignments were adjusted manually (Supplemental Materials: AdditionalFile3). A Neighbor-Joining tree with default parameters was built using MEGA 5 (Tamura et al. 2011).

To identify recent insertions of transposable elements within the palindrome, LTRharvest (Ellinghaus et al. 2008) was run with the sequence of the palindromic portion of the W-SDR from 8,778 kb to 9,015 kb with the target site duplication restricted to 5 bp to 20 bp. To find the protein domains in the coding region, a protein domain search against Pfam-A domains (release 32) was performed using the hidden Markov model methods implemented in LTRdigest (–hmms flag) (Steinbiss et al. 2009). Predicted LTR retrotransposons were determined to be non-automonous when coding regions did not contain any *gag* or *pol* related domains.

We estimated time since transposition based on the number of substitutions between the two LTR arms (SanMiguel et al. 1998). To estimate the substitution rate between the flanking LTR repeats, 5’ and 3’ repeats of each LTR retrotransposon predicted from LTRharvest were aligned by MUSCLE using default parameters provided in MEGA 5. After all gaps were removed, both number of differences and substitution rate were estimated in MEGA5. For number of differences, transitions and transversions were both included with a uniform rate. Substitution rate was modeled using the Kimura 2-parameter model provided in MEGA5, and the rate variation among sites was modeled with a gamma distribution (shape parameter = 1).

The time since transposition was estimated based on the mutation rate previously reported for *Populus tremula* ($2.5x10^{-9}$ per year)(Ingvarsson 2008).

Detection of gene conversion

As evidence of gene conversion, we searched for regions that were differentiated between species but concordant among the palindrome arms (Rozen et al. 2003). To accomplish this, we first aligned paired-end reads from a female clone of *S. suchowensis* (srx1561933) to the 94006 v5 female reference, plus alternative haplotypes, using Bowtie2 with the --local flag. This yielded an 82.9% overall alignment rate on average. The Illumina reads described above for clone 94006 were mapped using identical parameters. All reads aligning to the palindromes were extracted and compared to the whole genome using blastn. Mis-mapped reads originating from the autosomes were manually identified by scrutinizing the alignments, and only reads that mapped exclusively to the palindromic regions were retained. These reads were then re-aligned to a new reference consisting exclusively of arm 1 of the *S. purpurea* palindrome. SNPs and indels were called using mpileup and filtered to exclude loci with a minimum site quality <Q20 or depth >300.

Expression Profiling

RNAseq data was obtained from catkins of 10 female and 10 male F_2 progeny. RNAseq data were also obtained from multiple tissues of clones 94006 and Fish Creek. All sequences were Illumina 2x150 bp reads, except for 94006, which were 2x76 bp reads. Transcripts from the palindrome can have high sequence identity among arms and with other paralogous sequences on the autosomes, which can complicate estimation of gene expression. Thus, all predicted coding sequences from the same gene family in the palindrome were aligned to the autosomal paralogs, and conserved sequences were masked in the reference genome. Salmon-0.11.3 (Patro et al.

2017) was used to quantify (salmon quant) the raw read count for each sample mentioned above with the gcBias flag as suggested by the developers. Heatmaps were generated separately for each group of palindrome genes, using $\log_2$ transformed data normalized with respect to library size or by variance stabilizing transformations (VST) using the R packages pheatmap and Deseq2 (Love et al. 2014).

Results

Genome assembly

We present here highly contiguous genome assemblies of a female and a male *S. purpurea*. The female assembly (94006 v4) consists of 452 contigs with an N50 of 5.1 Mb, covering a cumulative total of 317.1 Mb. Similarly, the male assembly (Fish Creek v3), has 351 contigs and an N50 of 5.6 Mb, covering 312.9 Mb (Additional file 1: Table S1). Both assemblies are partially phased in genomic regions where the two haplotypes are divergent. Alternative haplotypes are represented by 421 contigs totaling 72.4 Mb in the female assembly, and 497 contigs totaling 149 Mb for the male. Using a genetic map from a large intercross family derived from progeny of the sequenced male genotype, we created assemblies representing the 19 chromosomes, containing 108 contigs totaling 288.3 Mb for the female, and 96 contigs totaling 288.5 Mb for the male. These represent over 90% of the assembled sequence in both cases, though 344 and 255 contigs remained unplaced by the genetic map for the female and male, respectively (Additional file 1:Table S2). The mapped and unplaced contigs are hereafter collectively referred to as the main genome, which excludes the alternative haplotypes.

Because we expected the W haplotype to be differentiated from the Z haplotype in the SDR, we anticipated that much of this region would be assembled as separate contigs. These can be readily differentiated by examining the relative depth of coverage when aligning male versus female short read sequences against these references. After identifying the location of the SDR based on the presence of sex-linked markers (Zhou et al. 2018), the initial Chromosome 15 assembly appeared to consist of a mix of Z and W scaffolds in a region we infer to be within the SDR (Additional file 2: Fig. S1a). We therefore sought to create a new assembly with Z and W haplotypes assembled to separate chromosomes. To do this we first identified the putative W contigs using sex association in a population of 60 unrelated individuals and differential depth of coverage in males and females from an F_2 pedigree as criteria (Zhou et al. 2018). This resulted in identifying 23 contigs that were putatively comprised primarily of sequence derived from the W haplotype (Additional file 1: Table S3). One scaffold was excluded because it mostly consisted of an alternative haplotype of a longer contig of Chr15W.

Many of these contigs lacked markers from the intercross map that was used in the original genome assembly (Zhou et al. 2018), particularly for those that came from portions of the W haplotype that were absent from the Z chromosome. We therefore created new genetic maps that had a mix of SNP and indel markers that would be more suited to capturing these hemizygous portions of the genome. The new genetic maps converged to 19 major linkage groups representing the 19 chromosomes. The male backcross map contained 8,715 markers, while the female backcross map contained 8,560 markers (Additional file 1: Table S4). We used these to assemble a Z and a W version of Chr15 (Additional file 1: Table S5). Thus, the current assembly (release ver5) contains 20 chromosomes, including Chr15Z and Chr15W. A total of 6.56 Mb (95.7%) of the W-specific contig sequence, contained in 17 contigs, was assembled to Chr15W using these maps. Four putative W scaffolds totaling 297 kb in length lacked mapped markers and could not be placed unambiguously.

Location of the SDR

We repeated sex association analysis for the 60 unrelated individuals using our new assembly with Chr15Z removed. Among 54,959 tested Genotyping by Sequencing (GBS) SNPs, all 105 significantly sex-linked SNPs were present only on Chr15W (Fig. 1a; Additional file 2: Fig. S2a-c), and markers from PARs and other scaffolds in the main genome did not show any sex association (Additional file 2: Fig. S2a). The eight top-ranking sex-associated markers were distributed from 7.66 Mb to 8.66 Mb. Sex-associated markers were primarily heterozygous in females and homozygous in males, confirming our previously-reported observation of ZW sex determination in *S. purpurea* (Zhou et al. 2018).

Composition of chromosomes 15W and 15Z

Chr15W is 15.7 Mb in length, composed of 22 contigs placed with the new genetic map. For comparison, Chr15Z is only 13.3 Mb and is comprised of 16 contigs (Additional file 1: Table S5; Fig. 1). There are two pseudoautosomal regions (PARs), one at each end of Chr15W, that are indistinguishable from the corresponding regions on Chr15Z. PAR1 is 2.3 Mb long and is composed of one contig, and PAR2 is 6.5 Mb and is comprised of three contigs (Fig. 1). These regions are unphased and are therefore identical in the two assemblies.

The W-linked sex-determining region (SDR) is 6.8 Mb in length, and occupies nearly 40% of the chromosome (hereafter referred to as the W-SDR). This region undergoes minimal recombination in the mapping population (Additional file 2: Fig. S4). Reexamining male and female depth of coverage of the W-SDR, it is clear that this region of the genome is mostly phased to separate the male and female haplotypes (Additional file 2: Fig. S1b). The region corresponding to the W-SDR on Chr15Z is only about 4 Mb in length, and only occupies 28.2% of the chromosome (hereafter referred to as the Z-SDR) (Additional file 2: Fig. S3). Based on the ratio of male and female depth of coverage, the Z-W homologous regions that are present on both the Z and W chromosome are about 3.5 Mb and insertions that are unique to the W are about 3.1 Mb in the W-SDR (Fig. 1c).

The W-SDR has lower gene density and higher repeat density than other portions of the genome, suggesting that repetitive elements have accumulated in this region (Table 1). More specifically, both the W-SDR and the Z-SDR show lower gene density on average than the PARs or other autosomes. Similarly, both the W-SDR and Z-SDR show higher accumulation of Gypsy retrotransposons. Interestingly, Copia-LTRs occur at higher density in the W-SDR region compared to the Z-SDR (10.9% of W-SDR vs 5.9% of Z-SDR), (Kruskall-Wallis test, $P<2.2e-16$) (Table 1), suggesting that these inserted following cessation of recombination between these haplotypes.

Gene content of the W chromosome

There are 269 genes in PAR1, 778 genes in PAR2, and 488 genes in the W-SDR. In contrast, the Z-SDR only contains 317 genes (Fig. 2; Additional file 1: Table S6-7). An additional 29 genes are present on scaffold_844, which is likely derived from the Z haplotype, but which lacked genetic markers to properly place it. To evaluate the completeness of the Z chromosome, we compared the gene content of this region to that from the Fish Creek male reference genome. The Z-SDR region was comprised of four contigs spanning from 2.86 to 7.10 Mb in Fish Creek, containing a total of 333 genes. Since the size and gene content were very similar between the Z chromosomes of the male and female references, we are restricting our analysis to the female to simplify the comparison.

There were 156 single copy mutual best hits between the W-SDR and Z-SDR, referred to hereafter as Z-W homologs (analogous to X-degenerate genes on mammalian sex chromosomes) (Fig. 2). The W-SDR also contains 32 genes in tandem duplications, while the corresponding tandem repeats in the Z-SDR contain 56 genes. Additionally, the W-SDR contains 40 genes that have mutual best hits on other autosomes, and 33 of these are tandemly duplicated in the SDR. In contrast, the Z-SDR region contains only 11 such genes, only six of which are tandemly duplicated. These putatively transposed genes comprise 8% of the W-SDR and only 3% of the Z-SDR. Another 54 genes in the W-SDR resulted from intrachromosomal transpositions and subsequent tandem duplication, while only 7 genes in this category are found on the Z-SDR. In total, these transposed and ampliconic genes account for more than half of the discrepancy in gene content between the haplotypes. An additional 103 genes in the W-SDR had a top hit to other genes in the genome, but the best hit was not mutual, so these are lower confidence candidates for transpositions or Z-W homologs. The Z-SDR contained 54 such genes. The remaining genes had no significant hits to other genes in the genome, presumably due to loss by deletion, or gaps in the sequence or annotation (85 in the W-SDR and 42 in the Z-SDR).

Z-W Homologs and Strata

We used syntenic gene pairs identified through MCScanX between the W-SDR and Z-SDR to test if there are strata with different degrees of divergence based on synonymous substitutions (d_S), which would indicate different phases of cessation of recombination (Bergero & Charlesworth 2009). There was little evidence to support the presence of strata based on 156 pairs of Z-W homologs (Fig. 3 and Additional file 1: Table S8). The average d_S was 0.027± 0.020 SE. For comparison, the d_S between syntenic genes on Chr01 for *S. purpurea* and *S. suchowensis* was 0.045±0.0022 SE, and the d_S between *S. purpurea* and *P. trichocarpa* was 0.146±0.0022 SE for syntenic genes on Chr01 (Fig. 3).

Transpositions to the W-SDR and palindromic repeats

The recently transposed genes are of particular interest because they could provide a potential mechanism for establishment of the SDR, and could highlight genes that are potential candidates for sex determination and/or sex antagonism (van Doorn & Kirkpatrick 2007). Among 40 genes putatively transposed from autosomes to the W-SDR, 7 have best hits on Chr19 (manually annotated genes excluded) (Additional file 1: Table S9). Contig ws19 is particularly enriched for transposed genes, and merits a closer examination (Fig. 1). Contig ws19 contains 11 transposed genes, including four genes from Chr19 and four genes from Chr17 (Fig. 1). Many of these transposed genes occur in two to four copies on ws19 in striking inverted repeat configurations that are similar to the palindromic repeats that occur on mammalian Y chromosomes (Fig. 4).

In *S. purpurea*, this region is female-specific (i.e., it occurs in all females but in no males) and is composed of two palindromes. Palindrome W.P1 spans about 42.7 Kb with a 2.6 kb spacer in the center, and Palindrome W.P.2 is immediately adjacent and spans over 165 kb (Table 2; Fig. 4a). A 20 kb sequence occurs in inverted orientation and shows high sequence identity across the four arms of both palindromes (Table 2; Fig. 5a). In palindrome W.P1 these are referred to as arm1 and arm2, and in Palindrome W.P2 these are referred to as arm3a and arm4a (Table 2; Fig. 4a). Sequence identity among these four arms is greater than 99% on average. The regions of high sequence identity are disrupted by a ~500 bp insertion in the center of arm4. Furthermore, arm3 has a 6.9 kb deletion at 11.7 kb, followed by a stretch of 1.6 kb that can be aligned to the other arms in the same orientation (Fig. 5a). Additionally, there is a 12 kb stretch upstream of arm1 that shows high identity to portions of arms 1 and 2. We call this the pre-arm for convenience (Table 2).

Palindrome W.P2 contains an additional inverted repeat that is missing from W.P1. We refer to this as arm3b and arm4b (Table 2; Fig. 4a). Sequence identity is somewhat lower between these two arms compared to the other four, ranging from 96% to 99% over most of their length. Furthermore, the regions of high identity are disrupted by numerous insertions and deletions (Fig. 5b).

Gene content of the palindromes

There are five genes duplicated across arms 1, 2, 3a and 4a of both palindromes. These are the Small Muts-Related protein (SMR), a Type-A cytokinin response regulator (RR), two genes that contain an NB-ARC domain (R1 and R2), and a Hydroxycinnamoyl-CoA shikimate/ hydroxycinnamoyl transferase (HCT) (Table 3). All of these genes except R2 have clear paralogous copies on Chr19. There is very little sequence divergence among most of these paralogs in the palindromes (Fig. 5).

The cytokinin response regulator is of particular interest because an ortholog of this gene has also been found to be associated with sex in *Populus* (Geraldes et al. 2015), and is therefore an excellent candidate as a sex determination gene in the Salicaceae. The RR gene is highly conserved across all four palindrome arms on the W-SDR (Fig. 5a,c). Interestingly, we also found a pseudogene copy of the RR gene on the Z-SDR. This is the only one of the five genes that is present in some form on the W-SDR, the Z-SDR, Chr19, and also in the SDR of *Populus*. There is a 2.6 kb sequence inserted upstream of all RR copies in the palindrome, and not in the Z-SDR pseudogene or on Chr19 (Additional file 2: Fig. S5). This suggests that the W-SDR palindrome formed after transposition from Chr19. Interestingly, the RR gene also occurs as inverted repeats in all three locations in the genome (W-SDR, Z-SDR, and Chr19). However, alignment of the W-SDR, Z-SDR and Chr19 versions demonstrates that the palindromes likely formed independently, because the palindromic regions are different (Additional file 2: Fig. S5).

There are an additional five genes in the W.P2 palindrome. Three of these genes occur as inverted repeats: a DNA-directed primase/polymerase protein (*DRBM*), a DNA primase (*DPRIM*), and a protein containing Domain of Unknown Function 789 (*DUF789*). In addition, there is a homolog of ARGONAUTE 4 (*TF2C*) and a CBS domain protein (*ACDP*) in single copy. Four of these genes were apparently transposed from Chr17 (Table 3). This leads us to the hypothesis that after these genes were transposed to the W-SDR they underwent several rounds of structural rearrangements, including duplications, inversions, and deletions.

Multiple LTR retrotransposons in the palindrome

To gain further insight into the composition and history of the W-SDR, we used LTRharvest and LTRdigest to annotate LTR retrotransposons in the palindromic region. We identified one LTR retrotransposon in the pre-Arm region and 12 LTR retrotransposons in palindrome W.P2 that have terminal repeats identified with coding regions (Fig.6a). These 13 retrotransposons are likely to be independent insertion events given that they have different long terminal repeats as well as different target site duplications and do not occur in the same position in the opposite arm of the palindrome (Additional file 1: Table S10). Given that there are varying numbers of substitutions within the LTRs of the same retrotransposon, it appears that these insertions have occurred repeatedly after establishment of the palindromes. Using a previous estimation of the mutation rate in *P. tremula* (2.5x10^{-9} per year)(Ingvarsson 2008), we estimate that the oldest insertion occurred at least 8.6 ± 2.9 s.d. MYA from a nonautonomous LTR retrotransposon, *Ltr-p2-a* (Fig. 6a and Additional file 1: Table S10). This is likely an underestimate, since the *Salix* substitution rate is substantially higher than that of *Populus* (Hou et al. 2016). Since the oldest substitutions occurred in Palindrome W.P2, we infer that this element became established first (Fig. 6a). The LTRs of the nonautonomous elements *Ltr-p2-a* and *Ltr-p2-k* flank the *SMR* and *RR* genes (Fig. 6c,d; Additional file 2: Fig. S6), which raises the intriguing possibility that these LTRs were involved in the transposition of these genes to this region. However, the target site duplications for these copies are identical across the palindrome arms, suggesting that the duplications and rearrangements of these genes in the W-SDR did not involve these elements (Fig. S6). We also found two highly similar LTRs from the same family in W.P1 (*Ltr-p2-b3* on arm3 and the *Ltr-p2-b4* on arm4; Fig. 6a-c; Additional file 1: Table S10). There are truncated parts of this LTR in the pre-arm and the spacer between arm1 and arm2 as well (Fig. 6b, c). These copies might be a direct consequence of duplications and inversions that occurred during the formation of the palindrome instead of independent insertions.

Evidence for gene conversion in the palindromes

We have shown that the palindromes are likely to be millions of years old based on the retrotransposon analysis, yet sequence identity of portions of the palindrome arms remains high (Fig. 5a). The most parsimonious explanation for this is gene conversion among the palindrome arms, as has been observed in the mammalian Y chromosome palindromes (Trombetta & Cruciani 2017; Rozen et al. 2003). To test for this, we searched for regions that had interspecific base substitutions relative to *Salix suchowensis*, a closely-related species with ZW sex determination (Hou et al. 2015). If regions with interspecific substitutions lack paralogous sequence variation (PSV) across the palindrome arms, then this would be excellent evidence of gene conversion (Rozen et al. 2003). We detected a 3 kb region within the palindromes where there are no PSVs in *S. purpurea* and only one PSV in *S. suchowensis*, but substantial interspecific polymorphisms (Fig. 7). The depth of this region is 4N as expected for the four copies of the palindrome arms in *S. purpurea*. In *S. suchowensis*, the depth is between 2N and 3N, which indicates that there might be a palindrome structure as well, though it might be incomplete. We also applied the same methods with resequencing reads of two female and two male *S. viminalis* individuals (another *Salix* with ZW sex determination) (Pucholt et al. 2015), but the palindromic region was not well covered by reads of either sex. This may indicate that *S. viminalis* lacks the palindrome, though it is more distantly related to *S. purpurea* than is *S. suchowensis*, so this may simply be due to excessive sequence divergence in this region.

Expression patterns of genes in the palindromes

We examined expression profiles in multiple tissues of the two reference genomes to validate the predicted transcripts and to determine how the expression patterns of genes in the palindromes differ from their autosomal counterparts. Most genes in the palindromes show female-limited expression while the autosomal copies are generally not sex-biased (Fig. 8a). The cytokinin response regulator (*RR*) (Sapur.15W073500) shows the highest expression in catkin tissue, followed by expression in shoot tips and stems. On the contrary, two autosomal copies on Chr19 show lower expression, limited to female catkins and male buds. The four copies of the *SMR* gene show low expression in female catkins and other tissues, but the autosomal copy on Chr19 (Sapur.019G001500) is expressed in all tissues (Fig. 8a). All five copies of the *HCT* gene from the palindromes showed low expression in female catkins and roots and higher expression in leaf tissues, shoot tips, and stems, all of which were female-biased. Two copies of the DNA Primase gene from palindrome W.P2 also show high expression in leaf tissues while the original copy on the autosome (Sapur.017G119600) was expressed across all sampled tissues. Similarly, analysis of transcriptomic data of catkins from 10 females and 10 males in the F_2 family confirms that the genes in the palindromes are primarily expressed in female tissue, in contrast to their autosomal paralogs (Fig. 8b).

Table 3.1 Cumulative size in Mb of genes and LTR retrotransposons in different areas of the genome. Numbers in parentheses are percentages of the proportion of the specific type of regions.

Category	**W-SDR**	**Z-SDR**	**PAR**	**Autosomes***
Genes	1.56 (23.8)	1.14 (26.8)	3.72 (41.9)	104.31 (38.1)
Total Repeats	3.16 (48.1)	1.81 (42.4)	2.58 (29.0)	89.17 (32.6)
Gypsy-LTR	0.86 (13.2)	0.55 (12.8)	0.38 (4.3)	15.45 (5.6)
Copia-LTR	0.72 (10.9)	0.25 (5.9)	0.37 (4.1)	13.87 (5.1)

* All 18 chromosomes are included.

Table 3.2 Coordinates of palindromes in the female SDR.

	Name	Start (bp)	End (bp)	Size (bp)	Gene families
Palindrome W.P1	pre-arm	8,778,973	8,791,042	12,070	*R2,HCT*
	arm1	8,790,932	8,811,002	20,071	*SMR,RR,R1,R2,HCT*
	Spacer1	8,811,003	8,814,588	3,586	
	arm2	8,814,589	8,834,138	19,550	*SMR,RR,R1,R2,HCT*
Palindrome W.P2	arm3a	8,836,813	8,850,772	13,960	*SMR,RR,R1,HCT*
	arm3b	8,850,773	8,920,527	69,755	*DRBM,TF2C,DPRIM,DUF789*
	Spacer2	Unidentified			
	arm4b	8,920,528	8,993,098	72,571	*DRBM,ACDP,DPRIM,DUF789*
	arm4a	8,993,099	9,013,390	20,292	*SMR,RR,R1,R2,HCT*

Table 3.3 Genes present in Palindromes 1 and 2.

	Gene Symbol	**Number of Copies**	**GeneID**	**Chromosome of the non-W best hit**	**Best Hit in *A. thaliana***	**Arabidopsis name or description (function)**	**Best Hit in *P. trichocarpa* v3**	**Identity *of P. trichocarpa* Best Hit**
Palindromes W.P1 and W.P2	*SMR*	4[a]	Manually annotated	Chr19	AT5G23520	*SMR* (Small MutS Related) domain-containing protein)	Potri.T013000	90.70
	RR	4	Sapur.15WG073500 Sapur.15WG073900 Sapur.15WG074000 Sapur.15WG075200	Chr19	AT3G56380	*ARR17* (type A cytokinin response regulator)	Potri.019G133600	92.81
	R1	4[a]	Sapur.15WG073800 Sapur.15WG074100	Chr15Z	AT4G27220	NB-ARC domain-containing disease resistance protein	Potri.T012900	81.00
	R2	3(1)[a]	Manually annotated	Chr17	AT4G27220	NB-ARC domain-containing disease resistance protein	Potri.T013300	61.23
	HCT	4(1)	Sapur.15WG073400 Sapur.15WG073600 Sapur.15WG073700 Sapur.15WG074200 Sapur.15WG075100	Chr19[b]	AT5G48930	*HCT* (hydroxycinnamoyl-coa shikimate/quinate hydroxycinnamoyl transferase)	Potri.018G104700	58.02
Palindrome W.P2 only	*DRBM*	2	Sapur.15WG074300 Sapur.15WG075000	Chr17	AT1G09700	*ATDRB1* (dsRNA binding protein)	Potri.017G126700	61.95
	TF2C	1	Sapur.15WG074400	Chr08[c]	AT2G27040	*AGO4* (*ARGONAUTE 4*, siRNA mediated gene silencing)	NA	NA
	ACDP	1	Sapur.15WG074900	Chr17	AT5G52790[d]	CBS domain protein with DUF21 (transmembrane transporter)	Potri.017G147900	83.33
	DPRIM	2	Sapur.15WG074500 Sapur.15WG074800	Chr17	AT5G52800	DNA primase	Potri.017G148000	92.52
	DUF789	2	Sapur.15WG074600 Sapur.15WG074700	Chr17	AT1G03610	*DUF789* (protein of unknown function)	Potri.017G152600	86.03

[a] Manually annotated transcripts were included in the count. Numbers in the parenthesis are from a fragment in the upstream portion of W.P1 that is homologous to part of W.P1. [b] This cluster of tandem duplications on Chr19 in *S. purpurea* is not present on Chr19 in *P. trichocarpa*. [c] The palindrome gene contains only a truncated blast hit to Sapur.008G005800 on Chr08. [d] This best hit with an expected value of 8x10^{-3} due to a sequence length of 84 aa. Expected values of the remaining *A. thaliana* were less than 1x10^{-10}.

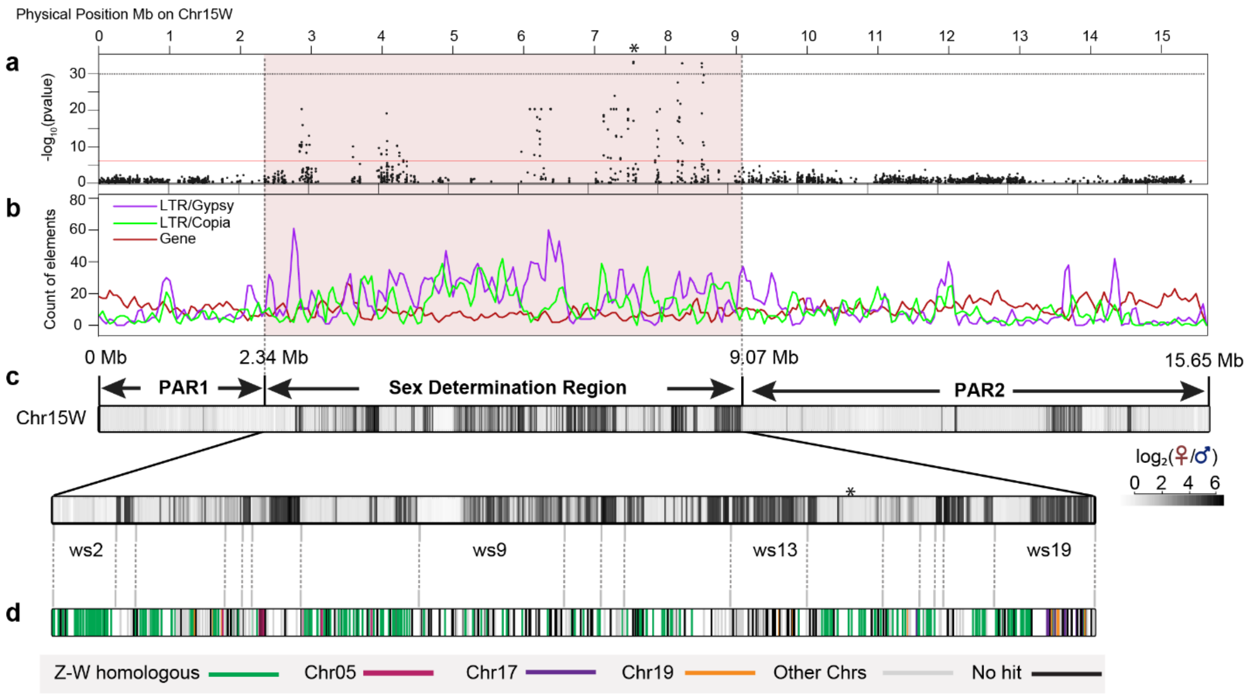


Figure 3.1 Genomic content of Chr15W and composition of the sex determination region (SDR). a. A Manhattan plot of Chr15W, based on GWAS using SNPs derived from aligning to a reference genome lacking Chr15Z. The Y axis is the negative logarithm of p values, and the red line indicates the Bonferroni cut off. **b**. Count of LTR elements including Gypsy and Copia, as well as genes in 100 kb windows with a 50 kb step size. **c**. Distribution of female-biased sequence on Chr15W, along with a more detailed view of the SDR below. Each colored block shows the log$_2$ of the ratio of female and male depth in 10 kb windows. Vertical gray lines below the figure show the boundaries of the contigs in the SDR. **d**. Each tick represents a gene in the SDR. Colors indicate putative origins of the genes based on blastp versus the rest of the genome.

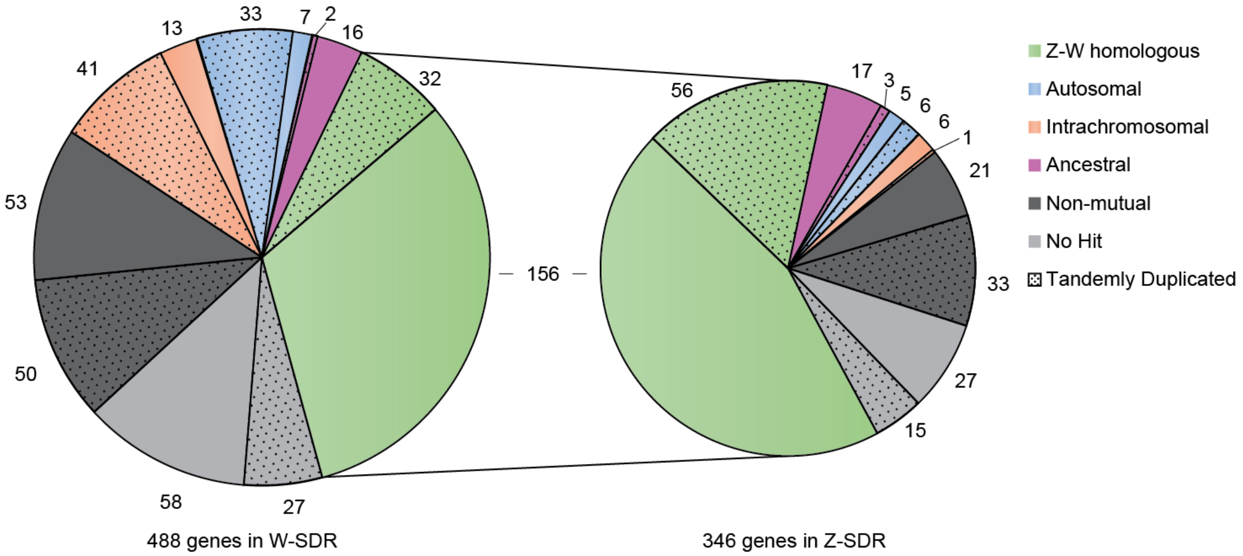


Figure 3.2 Annotated genes in Chr15W and Chr15Z. Genes are grouped according to the best non-self-hit in the annotated genome. Twenty-nine genes from an unmapped Z, scaffold_844 are also included. Stippled areas indicate genes of groups identified as tandem duplicates.

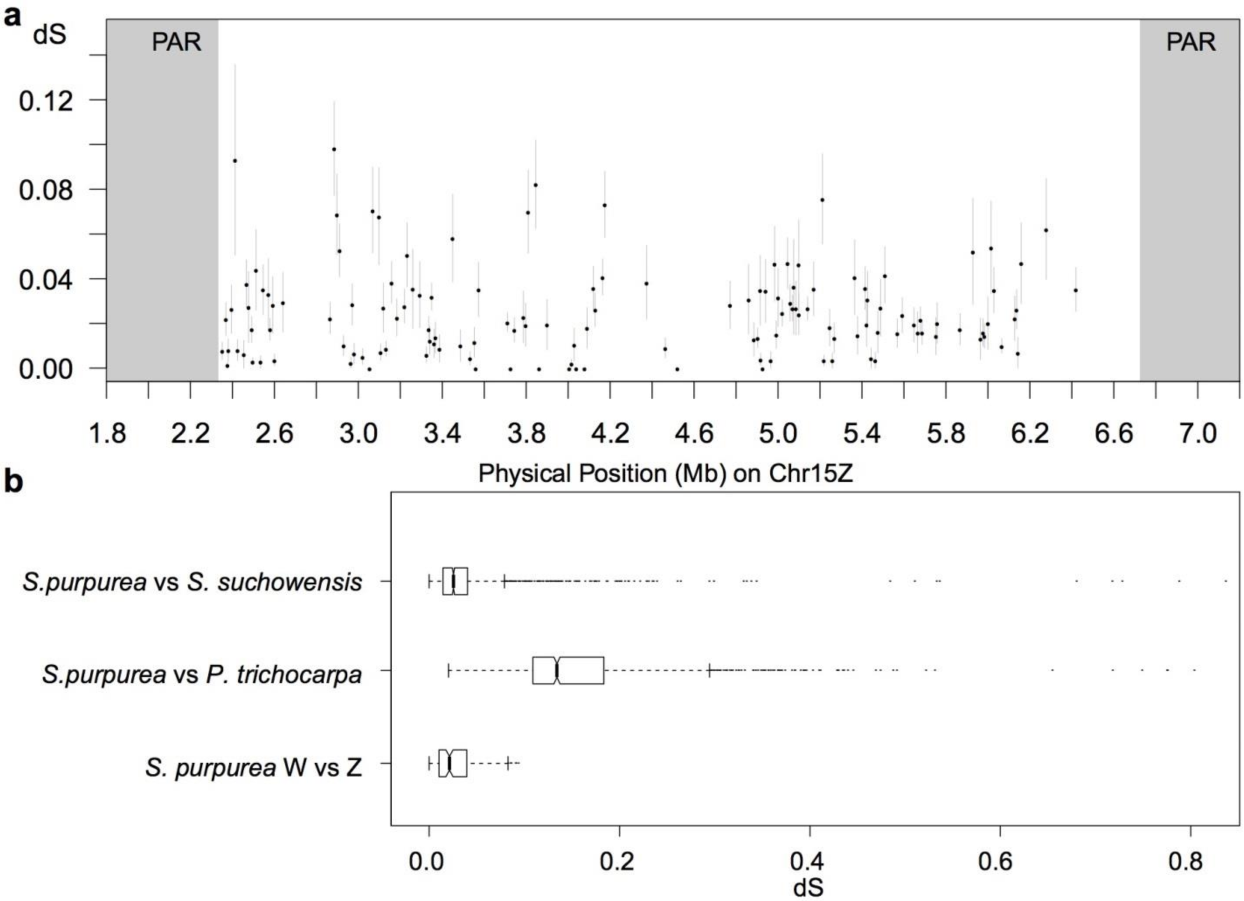


Figure 3.3 Synonymous substitution rates (*d*s) for genes in the SDR. a. Comparison of syntenic genes in the W-SDR and Z-SDR. Bars represent standard errors. **b**. Boxplot showing distributions of interspecific synonymous substitutions for 1,365 syntenic genes on Chr01 for the closely-related species *S. purpurea* and *S. suchowensis* and for 1,363 genes on Chr01 in *S. purpurea* and *Populus trichocarpa*, compared to the distribution of substitutions between syntenic genes in the *S. purpurea* SDR.

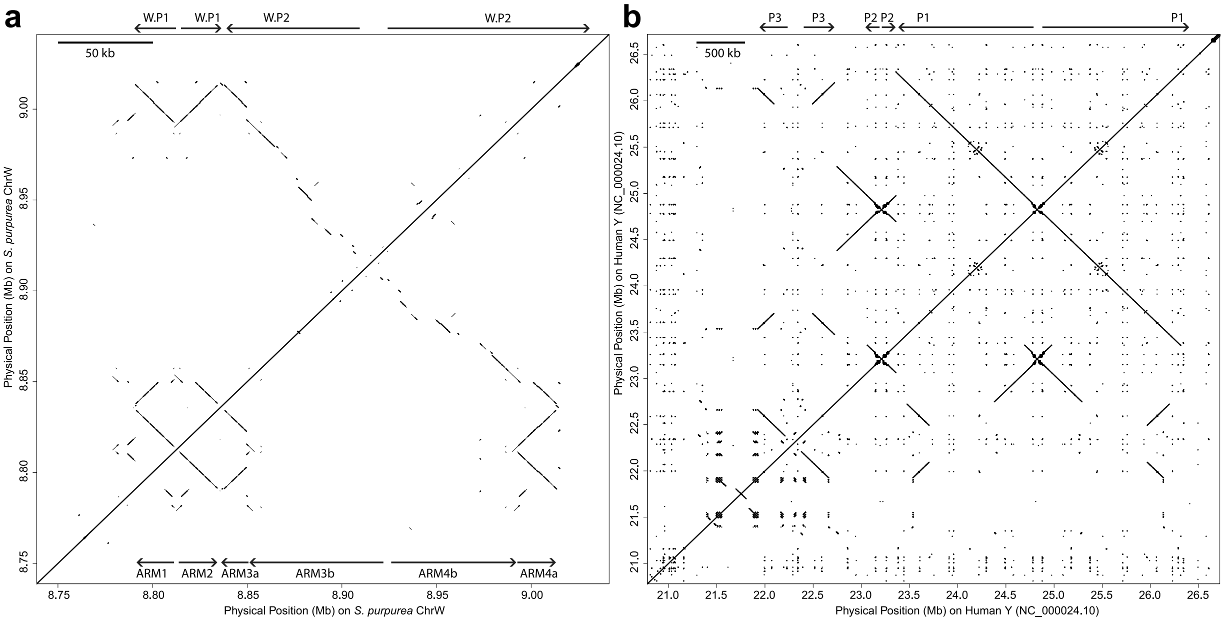


Figure 3.4 Palindromic repeats in the *S. purpurea* W chromsosme (a) and the *H. sapiens* Y chromosome (b). The dot plots were produced using LASTZ with identical settings. Note the different scales, indicated by the bar at the top right of each figure. *H. sapiens* palindromes are labeled following Skaletsky et al. (Skaletsky et al. 2003).

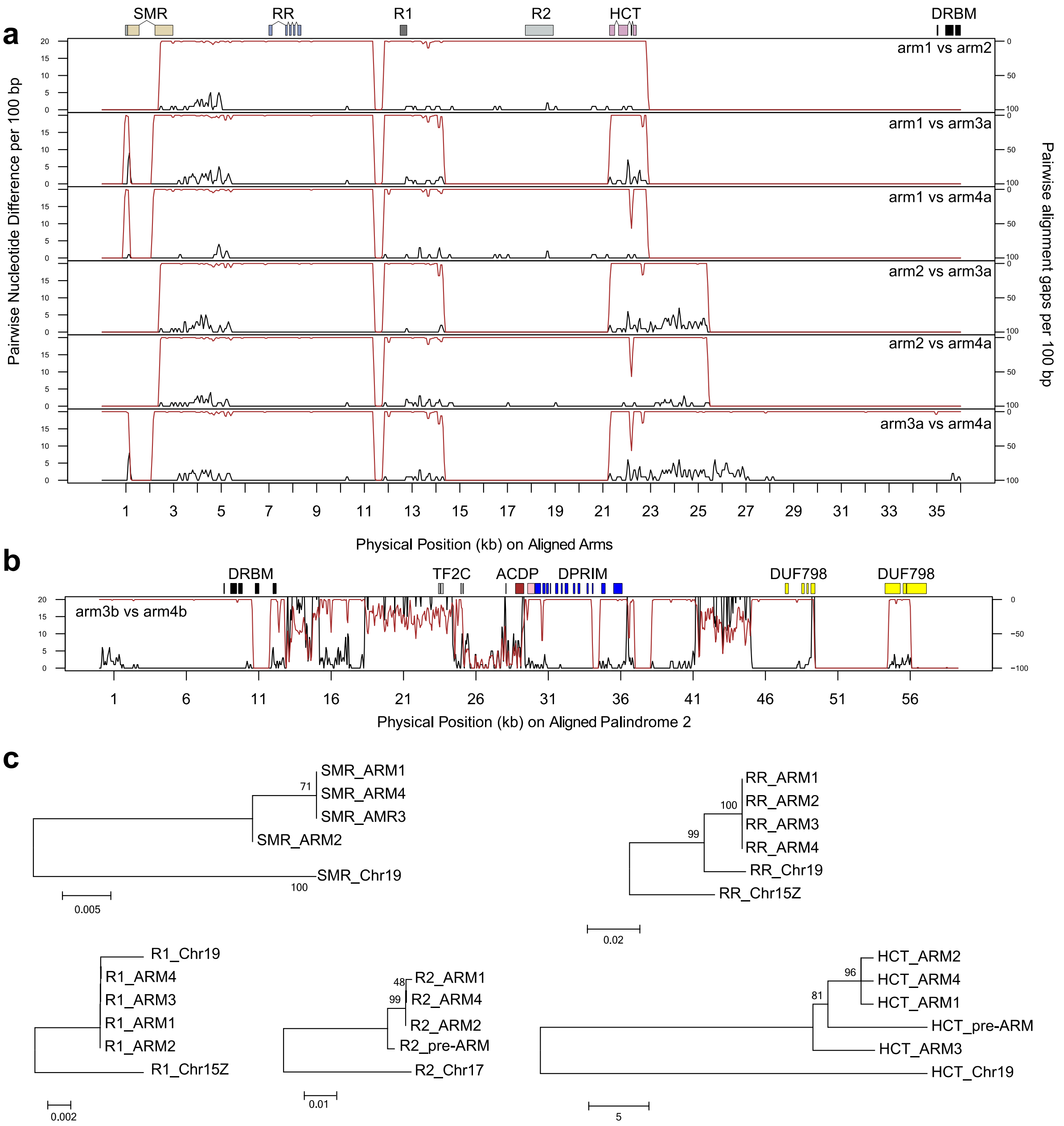


Figure 3.5 Sequence comparisons for the two palindromes. a. Comparison of the four arms that are shared among the two palindromes. The black line represents the number of nucleotide differences in 100 bp windows, while the red line indicates gaps in the alignment on an inverted scale. **b**. Comparison of the portions of palindrome 2 that are not shared with palindrome 1. **c**. Phylogenetic trees of five multi-copy genes in the palindromic region.

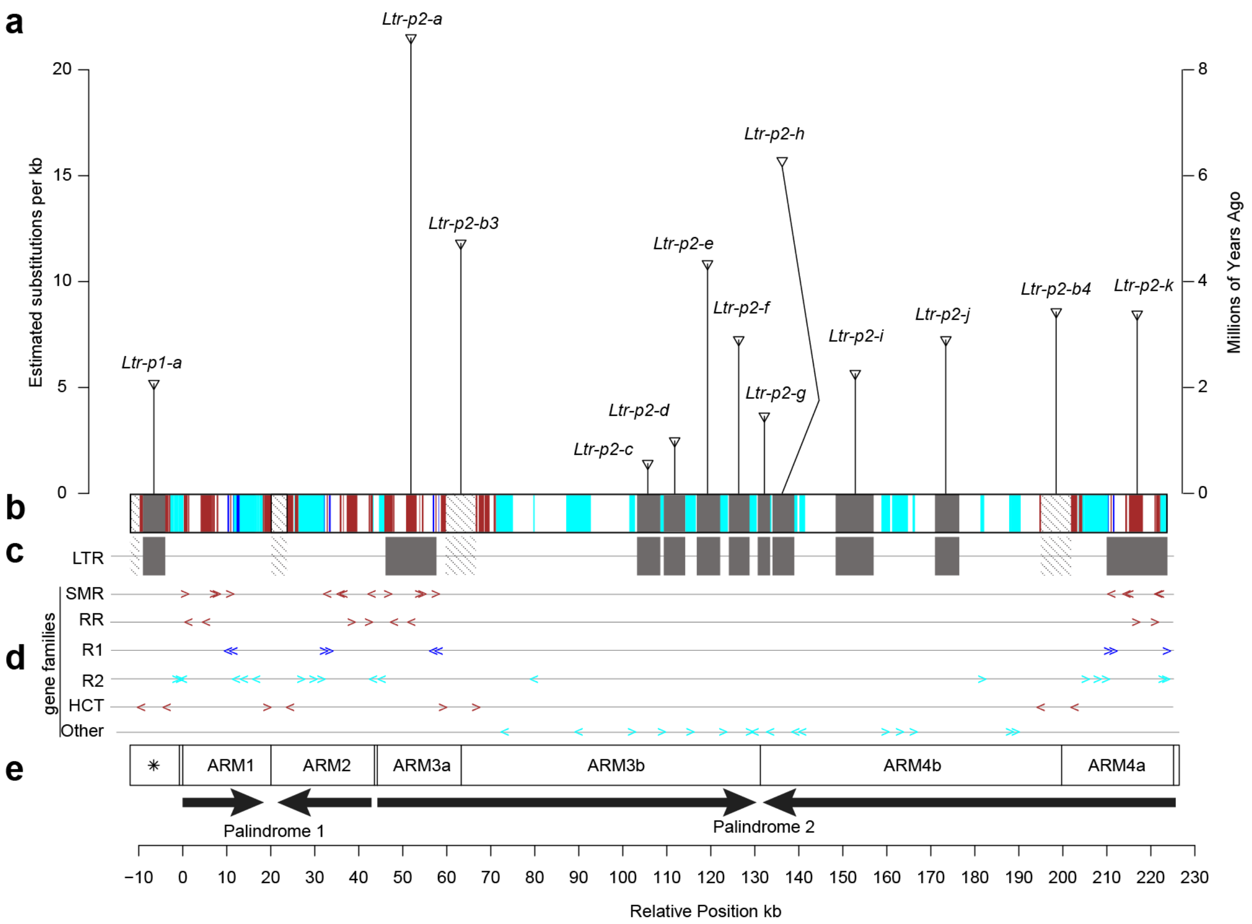


Figure 3.6 LTR retrotransposons, female specific genes, and palindromes. a. Each vertical line with a wedge on top represents each of the 13 TEs identified in the palindromic region by LTRharvest. The height of each line indicates the number of estimated nucleotide substitutions in the two LTRs (transposons a-h), and an approximation of the insertion time based on the mutation rate in *P. tremula* (Ingvarsson 2008). **b**. Colored boxes represent putative chromosomal origins of genes in the palindrome. Dark red, Chr19, cyan, Chr17. Blue boxes represent genes with paralogs on the Z chromosome. **c**. The positions of 13 LTRs (shaded boxes). Hatched boxes represent incomplete duplications derived from *Ltr-p2-b3/b4*. **d**. Exon positions and orientations, represented by colored arrows. **e**. Schematic representation of female-specific palindromes. The

box with a star represents a homologous region derived from part of one of the arms (preARM). Directions of arrows indicate the relative orientations of the four arms.

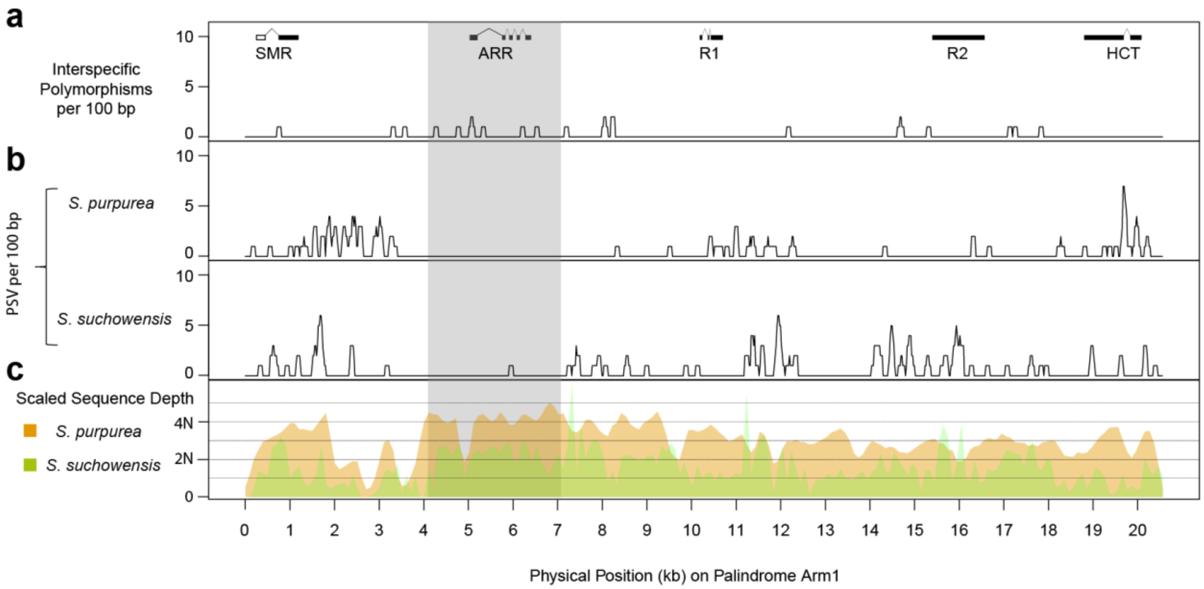


Figure 3.7 Sequence variation in the palindrome arms. a. Density of fixed differences between *S. purpurea* and *S. suchowensis* per 100 bp. **b**. Density of paralogous sequence variants (PSVs, differences among the 4 palindrome arms) in *S. purpurea* and *S. suchowensis*. **c**. Relative depth of Illumina sequence reads aligned to a reference sequence of one arm of the *S. purpurea* palindrome, where 2N represents the expected depth of read alignment across the whole genome. The grey shaded area represents a segment of the palindrome that is enriched for interspecific fixed variants, but depleted in PSVs, providing strong evidence for differential gene conversion in the two lineages.

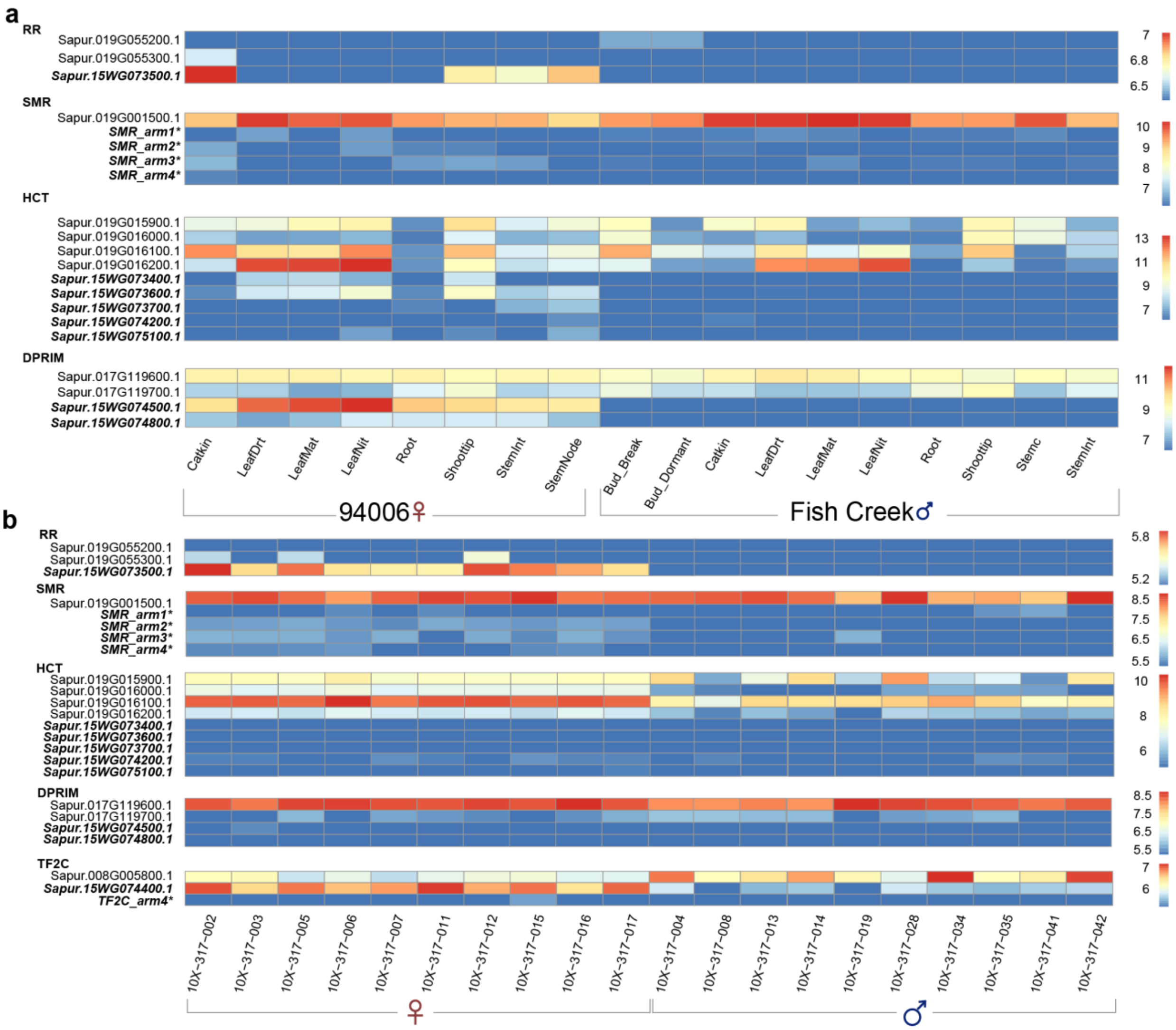


Figure 3.8 Expression profile of genes from the W palindromes and autosomal paralogs. a. Normalized read counts of genes in different tissues from clone 94006 (female) and Fish Creek (male). **b.** Normalized read counts of selected genes in catkins from 10 females and 10 males from an F_2 family. Gene labels in bold font are from the palindromes. Asterisks indicates manually annotated genes.

Discussion

The W chromosome in *S. purpurea*

Using depth of coverage for males and females from a controlled cross pedigree, we have been able to identify Z and W haplotypes from the SDR of a highly heterozygous species from a standard PacBio assembly. We also show how presence-absence markers generated from sequence depth in controlled cross progeny can be used to genetically map hemizygous portions of the SDR. In a similar study of a young Y chromosome in asparagus, BioNano optical maps for a YY individual were generated to improve genome contiguity, and sequence depth of coverage was also treated as a QTL to aid the assembly because of the presence of large indels in the sex chromosome (Harkess et al. 2017). Here, we showed that by combining long-read sequencing with GBS marker data from a large F_2 family, we could efficiently identify the male and female haplotypes in the SDR. However, unlike strategies like single-haplotype iterative mapping and sequencing (SHIMS) that have been used in assemblies of mammalian Y chromosomes (Skaletsky et al. 2003; Hughes et al. 2010, 2012; Soh et al. 2014), our map-based strategy could not provide a definitive order for the W contigs due to lack of recombination in the SDR.

The W-SDR is approximately 2.5 Mb larger than the Z-SDR. This is due in part to a greater accumulation of transposable elements, which account for approximately 1.35 Mb of this difference. This is consistent with expectations for sex chromosome evolution where transposable elements are expected to accumulate in regions with suppressed recombination (Charlesworth 2016; Ming & Moore 2007; Bachtrog 2013). However, gene content of the sex chromosome is expected to decrease due to the absence of recombination and reduced efficiency of purifying selection (Bachtrog 2013; Bergero & Charlesworth 2009). Instead, we observed that gene content is expanded in the W-SDR, driven in part by numerous transpositions and subsequent expansion of autosomal genes. Autosomal transpositions have also been demonstrated in other sex chromosomes, including mammalian Y chromosomes (Trombetta & Cruciani 2017). The recently-formed neo-Y chromosome of *Drosophila miranda* also shows massive expansion of genes that have been translocated from autosomes, and these are enriched for genes contributing to sex-specific functions (Bachtrog et al. 2019).

Sex chromosomes commonly show evidence of "evolutionary strata" with markedly different levels of sequence divergence that represent different epochs of expansion of the SDR (Charlesworth 2016). Under one common model of sex chromosome evolution, these strata are the result of multiple periods of SDR expansion as sexually antagonistic polymorphisms become incorporated into the SDR (Bergero & Charlesworth 2009; Scotti & Delph 2006). Although the identified SDR in *S. purpurea* is about 6-7 Mb, occupying more than one third of the W chromosome assembly, we detected little evidence for the existence of such strata. This corroborates a previous analysis that failed to detect strata in *S. suchowensis* using an integrated segmentation and clustering method (Pandey & Azad 2016). It appears that cessation of recombination has not been a gradual long-term process in the *S. purpurea* SDR, although it is certainly possible that the oldest strata have decayed to the point where they cannot be meaningfully aligned. An explanation for the large size of this region is that it partially overlaps with the centromere of Chr15, as we previously reported (Zhou et al. 2018). It is possible that the repressed recombination in this region pre-dated the transposition of a relatively small SDR cassette, as has been observed in octoploid *Fragaria* (Tennessen et al. 2018). This is consistent with the apparently small size of the region in *Populus* (~100 kb), which is located on a different chromosome (Geraldes et al. 2015). This is also consistent with the structure and composition of the palindromic repeats that we discovered in *S. purpurea*, which are excellent candidates as sex determination loci, as detailed below.

Sex chromosome palindrome repeats

We have reported here the first observation of a large inverted repeat in a plant sex chromosome, similar to the palindromic structures observed in mammalian sex chromosomes. We have further demonstrated that these palindromes are undergoing gene conversion, suggesting functional similarities to mammalian sex chromosome palindromes. W.P1 and W.P2 of *S. purpurea* have a similar arrangement of arms as P1 and P3 in humans due to the presence of highly homologous regions between the two palindromes. Similar palindromes have been also been discovered on Y chromosomes of other mammals, as well as avian W chromosomes (reviewed by (Trombetta & Cruciani 2017; Betrán et al. 2012b)). Large mammalian palindromes developed as a series of accumulations of insertions from autosomes and maintained through arm-to-arm gene conversion. This intrachromosomal gene conversion can maintain coding sequence integrity which otherwise would be compromised by the continuous accumulation of deleterious mutations in the absence of homologous recombination (i.e., Muller's Ratchet) (Trombetta & Cruciani 2017; Rozen et al. 2003; Lange et al. 2009; Betrán et al. 2012b). The fact that these structures have independently evolved in non-recombining regions of sex chromosomes is an intriguing case of convergent evolution of chromosome structure. Interestingly, the chloroplast genome, another non-recombining chromosome in plants, also contains a different large inverted repeat that undergoes gene conversion (Goulding et al. 1996) and helps maintain structural integrity of the genome, suggesting that this phenomenon may be common in regions of the genome that lack recombination (Palmer & Thompson 1982). However, it is also important to note that not all palindromic repeats occur in regions of the genome with suppressed recombination, most notably the large palindromes on the mammalian X chromosome. Palindromes may therefore play another role beyond maintenance of sequence integrity, such as mitigating expression of sexually antagonistic genes (Warburton et al. 2004) or in gene dosage compensation in the heterogametic sex (Bellott et al. 2014, 2017).

The *S. purpurea* palindromes are considerably smaller than mammalian palindromes, and have only accumulated two major autosomal transpositions (from Chr17 and Chr19), possibly reflecting their young age. Another difference between the human palindrome and the one in *S. purpurea* is that the gene conversion seems to be quite efficient across all the eight palindromes in humans, but the observed regions under gene conversion in *S. purpurea* are much more limited. This is particularly obvious in W.P2, compared to human P1, which has high sequence identity over several Mb (Fig. 4). Nevertheless, we found strong evidence for gene conversion in the cytokinin response regulator gene, based on an absence of PSVs. The ortholog of this gene in *S. suchowensis* has accumulated divergent nucleotide substitutions, which also seem to be homogenized among copies. This is a clear signature of gene conversion, and is unlikely to result from purifying selection or very recent independent duplication events (Rozen et al. 2003).

Evidence for a possible shared evolutionary history for the *Populus* and *Salix* SDRs

Initial analyses in *P. trichocarpa* suggested that the SDR is much younger than the whole genome duplication event that is shared by *Populus* and *Salix*, suggesting that the SDR became established well after these genera diverged (Geraldes et al. 2015). The low divergence between homologs in the fully sex-linked region (i.e., between Chr15W and Chr15Z homologs) shows that the SDR *of S. purpurea* evolved recently. Furthermore, given that the SDR is located in approximately the same portion of Chr15 in both *S. purpurea* and *S. suchowensis,* and both have ZW systems (Hou et al. 2015; Zhou et al. 2018), it is reasonable to assume that the SDR became established in this lineage prior to divergence of these two species, but well after divergence from *Populus*, which has an XY SDR on Chr19. On this basis, it has been hypothesized that these SDRs have independent evolutionary origins (Hou et al. 2015). We believe that our results point toward a single origin of dioecy in these genera, as well as shared components of an underlying sex determination system focused on cytokinin-mediated regulation.

Support for this hypothesis is provided by the type A cytokinin response regulator homologs that occur in palindrome arms 1,2,3a, and 4a (Table 3), which show strong evidence of ongoing gene conversion and female-specific expression in *S. purpurea*. The best ortholog of these genes in *P. trichocarpa* is Potri.019G133600 (this gene was originally designated *PtRR11*, but it is referred to as *RR9* in subsequent publications (Bräutigam et al. 2017; Melnikova et al. 2019; Chefdor et al. 2018), so we will adopt that nomenclature here to avoid confusion). *PtRR9* grouped with the *Arabidopsis thaliana* type A response regulators *ARR16* and *ARR17* in the original phylogenetic analysis of this family in *Populus* (Ramírez-Carvajal et al. 2008). The *ARR16* gene has been implicated in gynoecial development in Arabidopsis (Reyes-Olalde et al. 2017). *PtRR9* is expressed primarily in reproductive tissues in *Populus* (Chefdor et al. 2018; Ramírez-Carvajal et al. 2008), and is also associated with sex in several *Populus* species (Geraldes et al. 2015; Bräutigam et al. 2017; Melnikova et al. 2019). Further supporting its possible role in sex determination, it was the only gene in the *P. balsamifera* genome that showed clear sex-specific differences in promoter and gene body methylation (Bräutigam et al. 2017). This raises the intriguing possibility the mechanisms of sex determination in ZW *Salix* and XY *Populus* share common regulatory elements and a shared evolutionary origin.

The cytokinin signaling pathway has emerged in recent years as a prominent candidate for regulating floral development and sex expression in plants (Wybouw & De Rybel 2019;

Akagi et al. 2018). The potential role of cytokinin signaling in dioecy has recently been highlighted by the groundbreaking study by Akagi et al in kiwifruit (*Actinidia spp*) (Akagi et al. 2018). The authors identified a Type C response regulator (*Shy Girl, SyGI*) on the Y chromosome that was associated with maleness. Overexpression of this gene in Arabidopsis and *Nicotiana tabacum* caused suppression of carpel development, supporting its potential role as a suppressor of female function (Henry et al. 2018). This work has some interesting parallels with the results reported here for *Salix* and *Populus*. First, type C response regulators are essentially similar in structure to Type A response regulators, with the main difference being that Type C is not induced by cytokinin. Interestingly, *PtRR9* also was not induced by exogenous cytokinin application (Ramírez-Carvajal et al. 2008), though this has not yet been tested with floral tissue. Second, *SyGI* was duplicated from an autosomal gene and subsequently gained a new function on the Y chromosome, much like *SpRR9* has been duplicated from Chr19 in *S. purpurea* and established a distinct pattern of expression, and presumably new functions. However, *RR9* and *SyGI* are clearly not orthologous and likely perform different roles in cytokinin signal transduction. This supports the view that there are numerous ways to achieve separate sexes in plants, and it is likely that a myriad of mechanisms underlie the hundreds of independent occurrences of dioecy in the angiosperms (Renner 2014), even if a relatively small number of pathways are involved (Henry et al. 2018; Renner 2016).

Conclusion

We have shown that the SDR on the W chromosome of *S. purpurea* has expanded gene content compared to the corresponding region on the Z chromosome, due in part to autosomal genes that have been transposed and expanded in the region of suppressed recombination. We further demonstrated that some of these transposed genes are arranged as palindromic repeats that are undergoing gene conversion, suggesting some functional similarities to the mammalian sex chromosomes. This is a striking example of convergent evolution in chromosome structure. We have also demonstrated that the coding sequence undergoing gene conversion in the palindrome, *SpRR9,* is orthologous to a gene that is also associated with sex in *Populus*. This gene is an excellent candidate for controlling sex determination through modulation of the cytokinin signaling pathway. However, much remains to be determined about the underlying mechanism of sex determination. Most importantly, it is currently unclear how the same gene is functioning in an XY system in *Populus* and a ZW system in *Salix*. It is possible that the W chromosome version acts as a dominant promoter of female function, while the Y version is a dominant suppressor of female function, based on the putative roles of cytokinin and the type A response regulators in female development in Arabidopsis. A detailed model should emerge through comparative analysis of the W and Y chromosomes of multiple species in the Salicaceae, which is currently underway. If the underlying mechanism shares common regulatory elements, this will be the first case demonstrating XY and ZW systems that are controlled by the same pathway in plants.

Acknowledgements

We thank Fred Gouker and the teams at the New York State Agricultural Experiment Station and West Virginia University for help with management and phenotyping of the field trials used in this study. We also thank the teams at Phytozome and JGI for facilitating public distribution of the data. We also thank Jennifer Hawkins, Niels Müller and Mathias Fladung for helpful discussions. This work was supported by the NSF Dimensions of Biodiversity Program (DEB-1542509 to S.D. and DEB-1542599 to M.O.). Support was also provided by the National Natural Science Foundation of China (31590821, 31561123001, 31500502, 41871044), National Key Research and Development Program of China (2017YFC0505203, 2016YFD0600101), and the National Key Project for Basic Research (2012CB114504). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Statement of Original Publication

This work has been accepted for publication in *Genome Biology* in February 2020. https://doi.org/10.1186/s13059-020-1952-4

References

Akagi T et al. 2018. A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. Plant Cell 30:780–795.

Akagi T, Henry IM, Kawai T, Comai L, Tao R. 2016. Epigenetic Regulation of the Sex Determination Gene MeGI in Polyploid Persimmon. Plant Cell 28:2905–2915.

Bachtrog D et al. 2014. Sex Determination: Why So Many Ways of Doing It? PLoS Biol. 12:e1001899.

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

Bachtrog D, Mahajan S, Bracewell R. 2019. Massive gene amplification on a recently formed *Drosophila* Y chromosome. Nat. Ecol. Evol. 3:1587–1597.

Balounova V et al. 2019. Evolution of sex determination and heterogamety changes in section *Otites* of the genus *Silene*. Sci. Rep. 9:1045.

Bellott DW et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nat. Genet. 49:387–394.

Bellott DW et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508:494–499.

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

Betrán E, Demuth JP, Williford A. 2012a. Why Chromosome Palindromes? Int. J. Evol. Biol. 2012:207958.

Betrán E, Demuth JP, Williford A. 2012b. Why Chromosome Palindromes? Int. J. Evol. Biol. 2012:1–14.

Bräutigam K et al. 2017. Sexual epigenetics: gender-specific methylation of a gene in the sex determining region of *Populus balsamifera*. Sci. Rep. 7:45388.

Carlson CH et al. 2017. Dominance and sexual dimorphism pervade the *Salix purpurea* L. transcriptome. Genome Biol. Evol. 9:2377–2394.

Carlson CH et al. 2019. Joint linkage and association mapping of complex traits in shrub willow (*Salix purpurea* L.). Ann. Bot.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22:3124–3140.

Charlesworth B, Charlesworth D. 1978. A Model for the evolution of dioecy and gynodioecy. Am. Nat. 112:975–997.

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

Charlesworth D. 2016. Plant Sex Chromosomes. Annu. Rev. Plant Biol. 67:397–420.

Chefdor F et al. 2018. Highlighting type A RRs as potential regulators of the dkHK1 multi-step phosphorelay pathway in *Populus*. Plant Sci. 277:68–78.

Chin C-S et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10:563–569.

Chin C-S et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13:1050–1054.

Davis JK, Thomas PJ, Thomas JW. 2010. AW-linked palindrome and gene conversion in New

World sparrows and blackbirds. Chromosom. Res. 18:543–553.

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

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.

Geraldes A et al. 2015. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). Mol. Ecol.

Gouker FE, DiFazio SP, Bubner B, Zander M, Smart LB. 2019. Genetic diversity and population structure of native, naturalized, and cultivated *Salix purpurea*. Tree Genet. Genomes 15:47.

Goulding SE, Wolfe KH, Olmstead RG, Morden CW. 1996. Ebb and flow of the chloroplast inverted repeat. Mol. Gen. Genet. 252:195–206.

Harkess A et al. 2017. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nat. Commun. 8:1279.

Henry IM, Akagi T, Tao R, Comai L. 2018. One Hundred Ways to Invent the Sexes: Theoretical and Observed Paths to Dioecy in Plants. Annu. Rev. Plant Biol. 69:553–575.

Hou J et al. 2015. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci. Rep. 5:9076.

Hou J et al. 2016. Major chromosomal rearrangements distinguish willow and poplar after the ancestral ‘Salicoid’ genome duplication. Genome Biol. Evol. 8:1868–1875.

Hughes JF et al. 2010. Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. Nature 463:536–539.

Hughes JF et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature 483:82–86.

Ingvarsson PK. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics 180:329–340.

Kang HM et al. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354.

Krausz C, Casamonti E. 2017. Spermatogenic failure and the Y chromosome. Hum. Genet. 136:637–655.

Lange J et al. 2009. Isodicentric Y Chromosomes and Sex Disorders as Byproducts of Homologous Recombination that Maintains Palindromes. Cell 138:855–869.

Langmead B, Salzberg SL. 2012a. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Langmead B, Salzberg SL. 2012b. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357. https://doi.org/10.1038/nmeth.1923.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Margarido GR, Souza AP, Garcia AA. 2007. OneMap: software for genetic mapping in outcrossing species. Hereditas 144:78–79.

McKenna A et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McKown AD et al. 2017. Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in *Populus*. Sci. Rep. 7:1831.

Melnikova N V. et al. 2019. Sex-specific polymorphism of MET1 and ARR17 genes in *Populus* × *sibirica*. Biochimie 162:26–32.

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

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

Navarro-Costa P, Plancha CE, Gonçalves J. 2010. Genetic Dissection of the AZF Regions of the Human Y Chromosome: Thriller or Filler for Male (In)fertility? J. Biomed. Biotechnol. 2010:1–18.

Palmer JD, Thompson WF. 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell 29:537–550.

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.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14:417–419.

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

Ramírez-Carvajal G a, Morse AM, Davis JM. 2008. Transcript profiles of the cytokinin response regulator gene family in *Populus*. New Phytol. 177:77–89.

Renner SS. 2016. Pathways for making unisexual flowers and unisexual plants: Moving beyond the ‘two mutations linked on one chromosome’ model. Am. J. Bot. 103:587–589.

Renner 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.

Reyes-Olalde JI et al. 2017. The bHLH transcription factor SPATULA enables cytokinin signaling, and both activate auxin biosynthesis and transport genes at the medial domain of the gynoecium. PLOS Genet. 13:e1006726.

Rozen S et al. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. 20:43–45.

Scotti I, Delph LF. 2006. Selective trade-offs and sex-chromosome evolution in *Silene latifolia*. Evolution (N. Y). 60:1793–1800.

Skaletsky H et al. 2003. The male-specific region of the human Y chromosome is a mosic of discrete sequence classes. Nature 423:825–837.

Soh YQS et al. 2014. Sequencing the Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex Chromosomes. Cell 159:800–813.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 37:7002–7013.

Tamura K et al. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum

Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28:2731–2739.

Tang H et al. 2015. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16:3.

Tennessen JA et al. 2018. Repeated translocation of a gene cassette drives sex-chromosome turnover in strawberries. PLOS Biol. 16:e2006062.

Trombetta B, Cruciani F. 2017. Y chromosome palindromes and gene conversion. Hum. Genet. 136:605–619.

Tuskan GA 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.

Wang Jianping 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.

Wang Y. et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40:e49.

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes. Genome Res. 14:1861–1869.

Westergaard M. 1958. The Mechanism of Sex Determination in Dioecious Flowering Plants. Adv. Genet. 9:217–281.

Wilm A et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Wybouw B, De Rybel B. 2019. Cytokinin – A Developing Story. Trends Plant Sci. 24:177–185.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zhou R et al. 2018. Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). Mol. Genet. Genomics 293:1437–1452.

Availability of Data and Materials

All sequence data used in this manuscript have been deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra). Accession numbers are available in Additional file 1: Table S13. The genome assemblies and annotations are available through Phytozome (https://phytozome-next.jgi.doe.gov).

Supplementary Materials

Additional File 1: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-020-1952-4/MediaObjects/13059_2020_1952_MOESM1_ESM.xlsx

Supplemental Table 3.1 contains assembly statistics for the *S. purpurea* reference genomes.

Supplemental Table 3.2 describes caffold and contig size for the map-based assemblies of *S. purpurea.*

Supplemental Table 3.3 provides information about contigs identified as sex-specific based on depth and the presence of sex-associated markers.

Supplemental Table 3.4 provides numbers of markers in genetic maps derived from GBS data from an F_2 cross of *S. purpurea.*

Supplemental Table 3.5 is a summary of reassembly of Chr15W and Chr15Z.

Supplemental Table 3.6 summarizes results of all-vs-all blastp of W-SDR genes in the S. purpurea v5.0 genome.

Supplemental Table 3.7 summarizes results of all-vs-all blastp of Z-SDR genes in the S. purpurea v5.0 genome.

Supplemental Table 3.8 provides d_N and d_S values calculated on W and Z allele pairs from syntenic blocks identified by MCScanX.

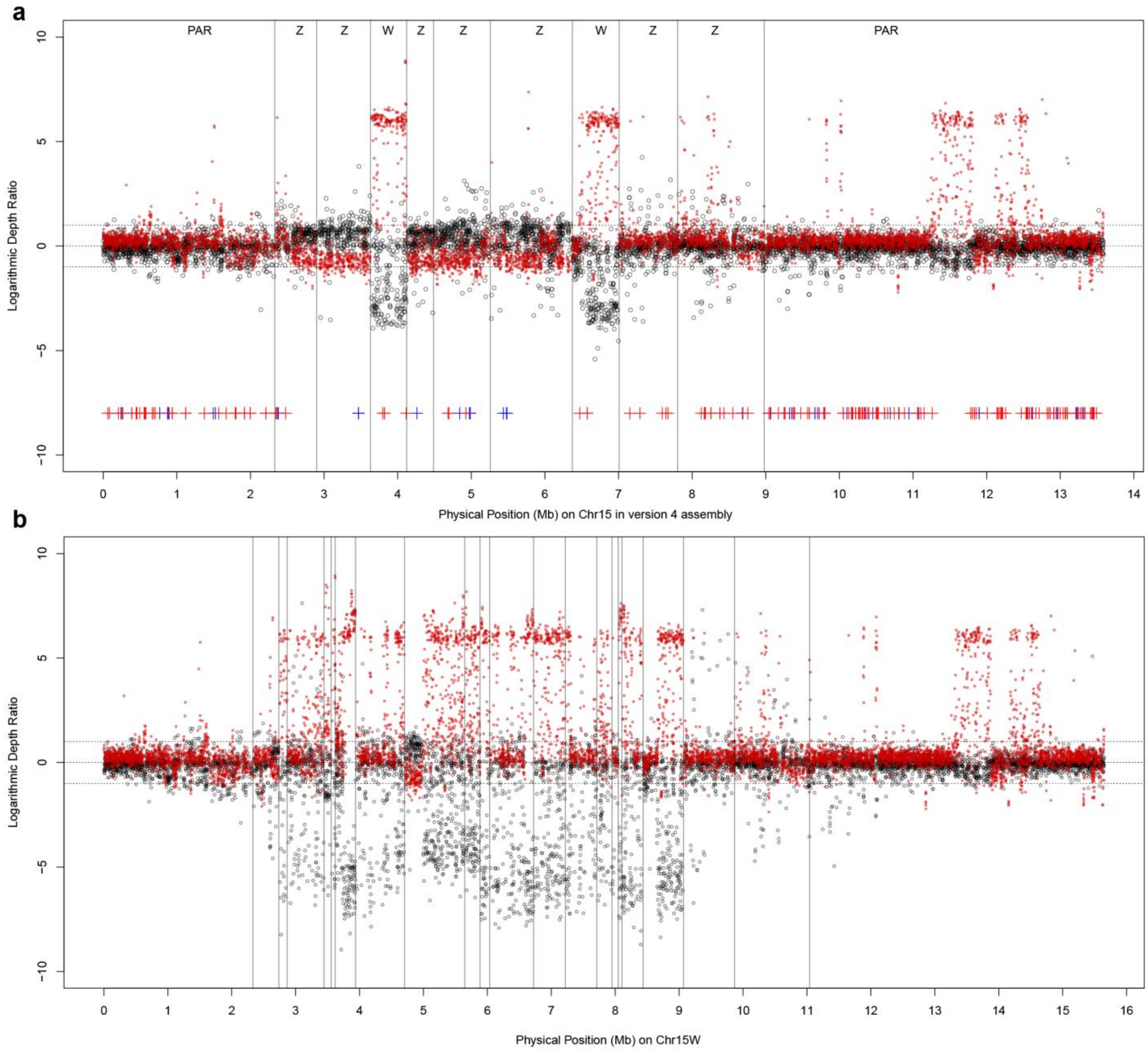
Supplemental Table 3.9 summarizes the number and origins of translocations to the W chromosome, based on mutual best blast hits.

Supplemental Table 3.10 LTR retrotransposons identified through the LTRharvest.

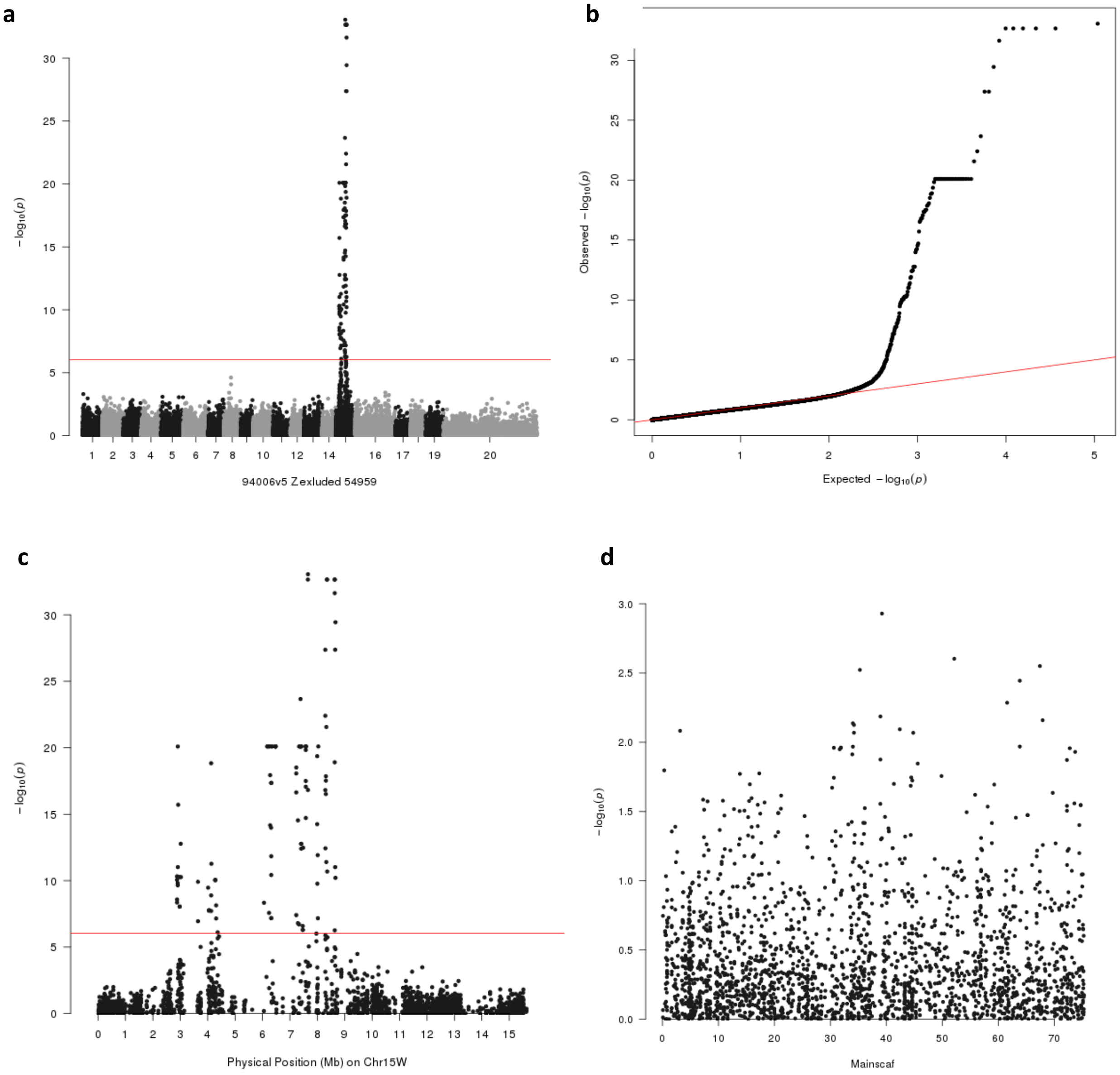
Supplemental Table 3.11 describes coverage of PacBio sequence reads for different size classes for the two *S. purpurea* genome assemblies.

Supplemental Table 3.12 provides the results of SNP and INDEL corrections of genome assemblies of *S. purpurea.*

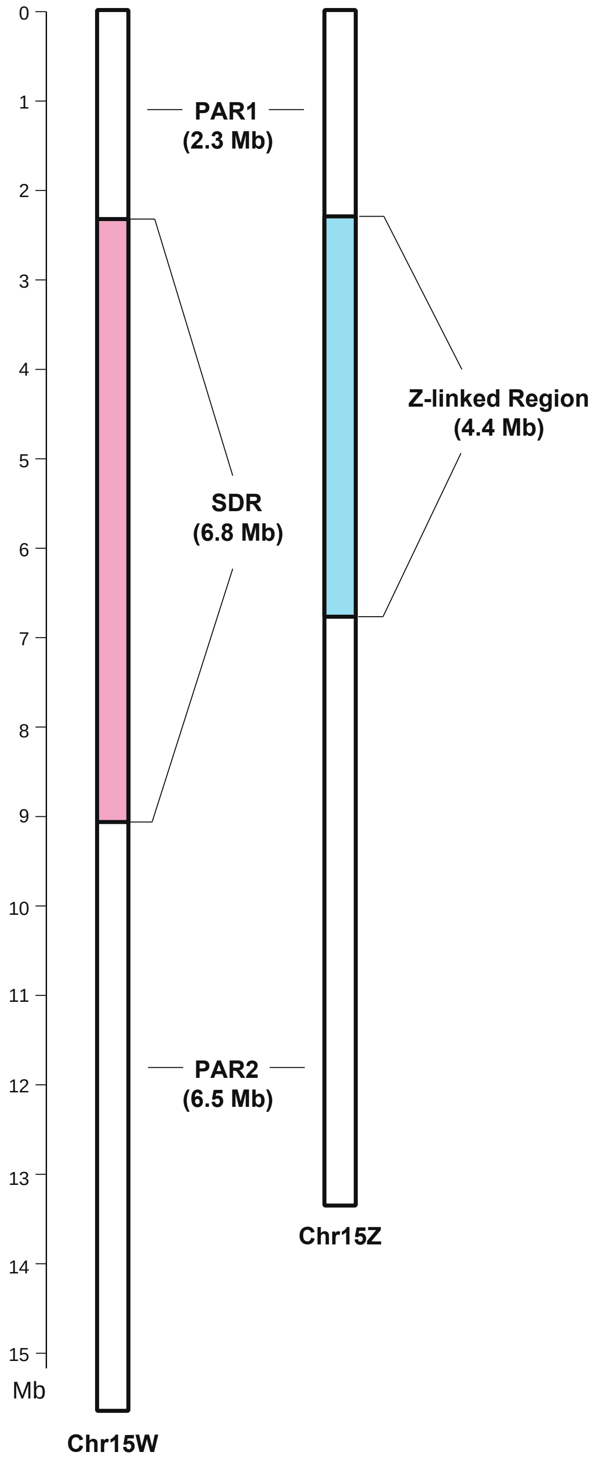
Supplemental Table 3.13 lists NCBI accession numbers of sequences used in this study.



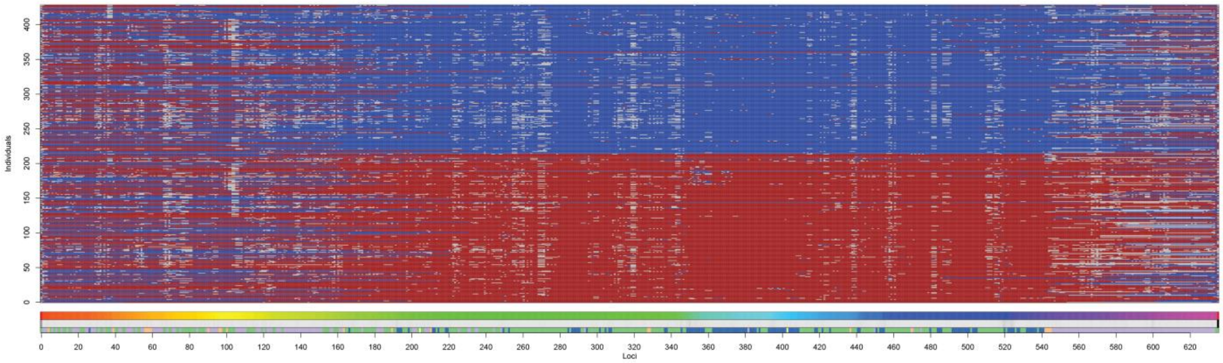
Supplemental Figure 3.1 Sex-specific depth of Chr15 for two assemblies of female clone 94006. a. Initial assembly guided by mapped SNP markers only. Boundaries of contigs are represented with vertical lines. Each contig is categorized as pseudo-autosomal region (PAR), Z, or W according to the logarithmic depth ratio between female and male sequence alignments. Ratios of GBS marker depth ($\log_2$(M/F)) for 200 progeny from an F2 pedigree (family 317) are shown by black dots, and ratios of the two reference individuals from Illumina 2x250 resequencing reads $\log_2$(F/M) are shown by red dots. Near the bottom, red-crosses represent markers that are inherited from the female parent, and blue crosses are markers inherited from the male parent. **b**. Chromosome 15W, following reassembly using scaffolds with female-specific alleles and/or female-biased depth ratios.



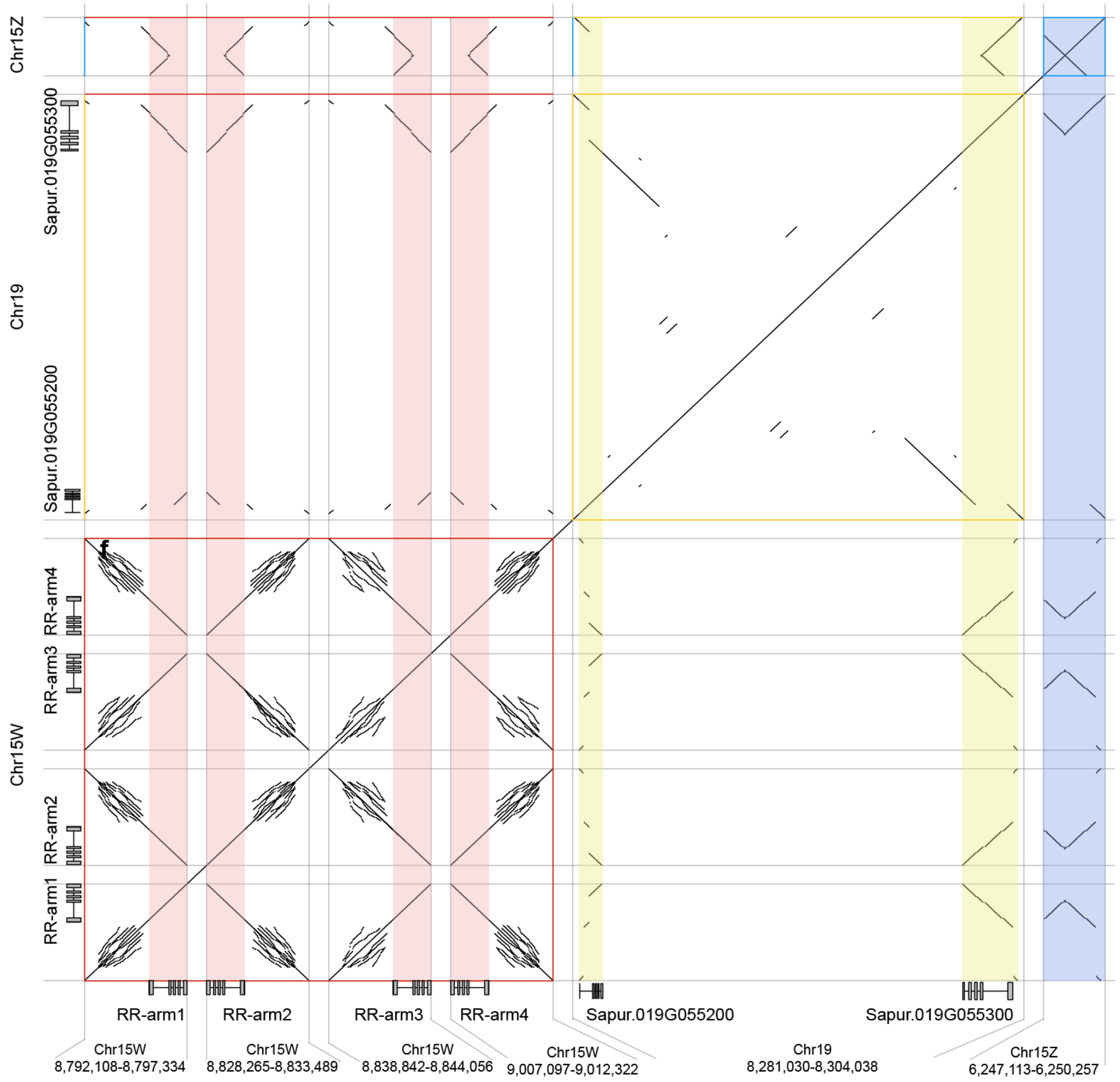
Supplemental Figure 3.2 Association of sex with new 94006v5 assembly. a. Manhattan plots showing association of sex repeated with the v5 genome assembly, with Chr15Z removed. The analysis was performed with a natural population of 60 non-clonal individuals. The red line indicates a Bonferroni cutoff 9.10 x 10^{-7} with 54,959 tested SNPs. **b**. QQ-plot for the association analysis. **c**. Manhattan plot for Chromosome 15W. **d**. Manhattan plot for unplaced scaffolds from the main genome. None of these showed significant association with sex.



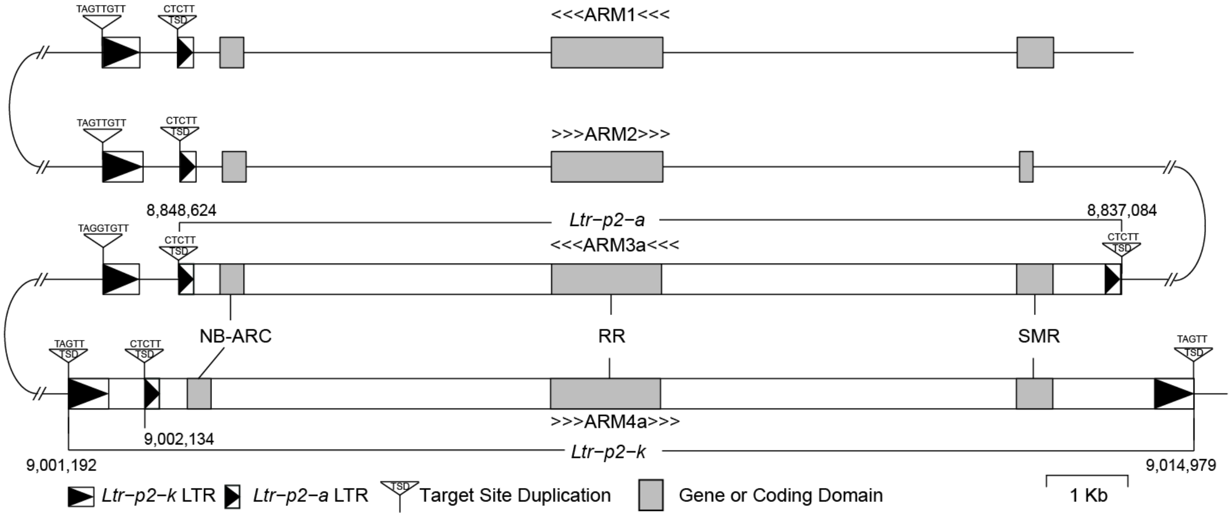
Supplemental Figure 3.3 Relative sizes and composition of Chr15Z and Chr15W. The PAR regions are unphased, and identical between the two chromosomes, while the W-SDR and the Z-linked region are mostly phased, and contain different sequence contigs**.**



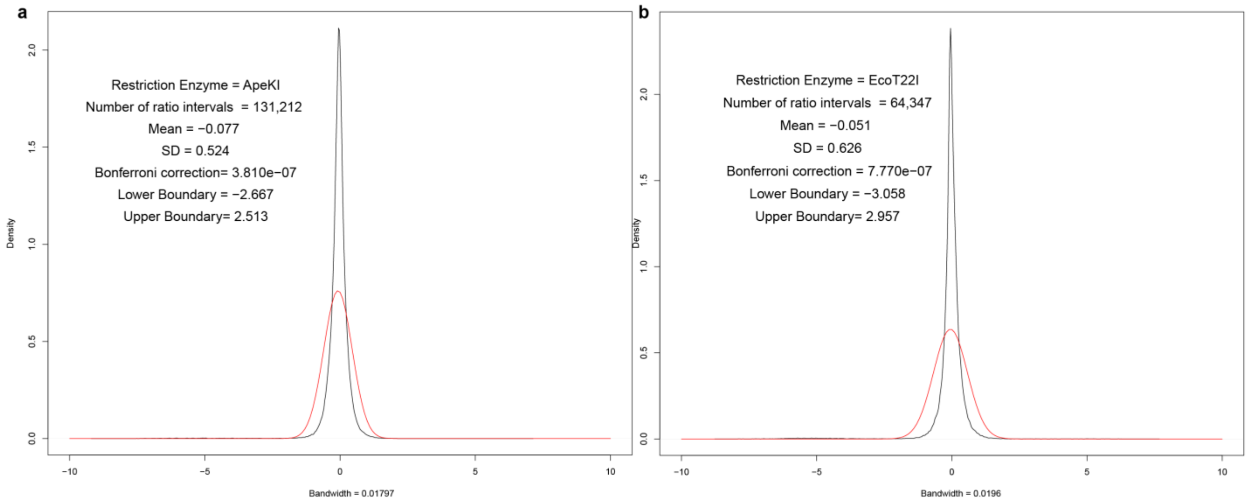
Supplemental Figure 3.4 Recombination among parental haplotypes in F2 progeny. GBS markers are ordered along the genetic map. 214 F_2 progeny from each sex are in the rows, with males at the top of the figure. Red cells represent alleles derived from the W haplotype, and blue cells represent the Z haplotype according to the maternal (Wolcott) genetic map. Gray cells represent missing data that could not be imputed.



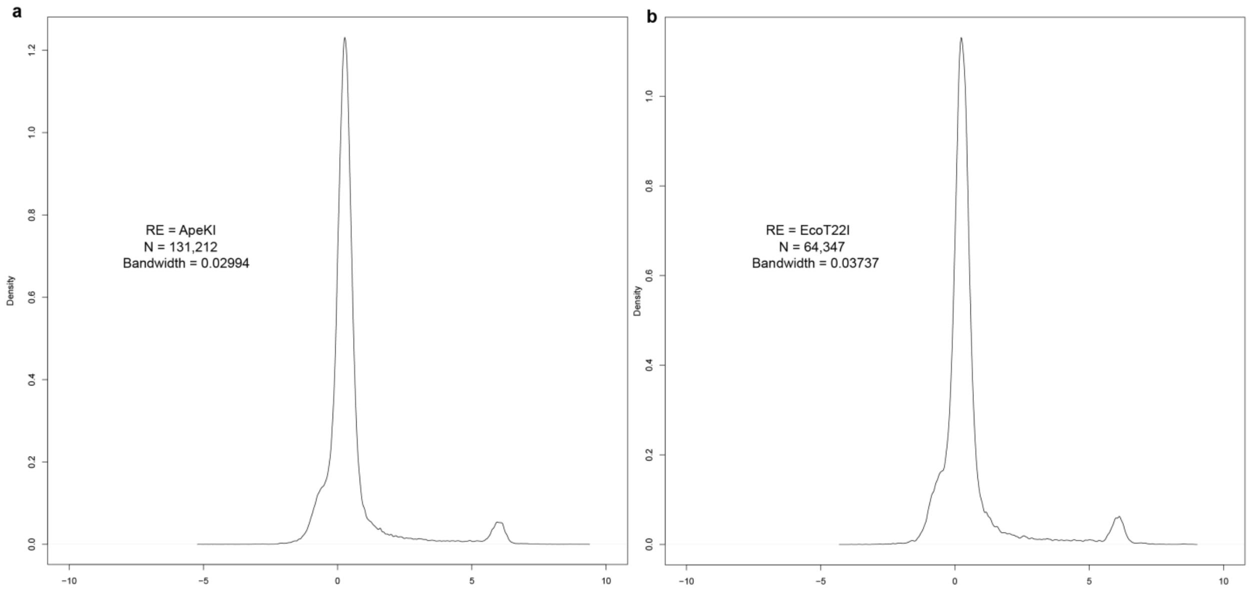
Supplemental Figure 3.5 Dotplots of regions containing portions of the RR gene. This dot plot was generated from data produced by aligning sequences from identified regions containing the RR genes (complete or partial) on Chr15W palindrome (red), Chr15Z (blue), and Chr19 (yellow) using LASTZ. Colored shading indicates the X axis location of genes and genes models, which are also displayed on both axes. Notice that the Chr15Z block (blue) contains a truncated portion of the gene, and was not annotated. Three colored squares along the diagonal line show the palindromic structures. Horizontal and vertical lines with different colors indicate the area of pairwise alignments between RR genes from different chromosomes.



Supplemental Figure 3.6 Arrangement of *Ltr-p2-a* and *Ltr-p2-k* in the palindrome. Two non-autonomous LTR retrotransposons on arm3 and arm4 are shown with their target site duplication (TSD) sequences, long terminal repeats (LTRs), and genes or domains highlighted. Duplicated sequence features are also labeled on arm1 and arm2. Numbers indicate the coordinates of these transposable elements on the W chromosome.



Supplemental Figure 3.7 Distribution of depth ratios (log$_2$($\frac{M_{195}+1}{F_{195}+1}$)) **of GBS reads aligned to the female 94006 v4 genome**. The distribution of depth ratios of GBS markers of *Ape*KI is indicated by the black line, and normal distribution with the same mean and standard deviation (SD) is indicated with a red line. To detect outliers, such as intervals only covered in one sex, lower and upper boundaries were determined according to the Bonferroni corrected percentile (0.05/number of intervals) of this normal distribution. **b**. The same process was applied with the GBS markers that were generated from *Eco*T22I.



Supplemental Figure 3.8 Distribution of depth ratios $\log_2\left(\frac{94006_{2by250}+1}{Fish\ Creek_{2by250}+1}\right)$ **of Ilumina 2x250 reads aligned to the female 94006 v4 genome**. **a.** Counts are only from intervals defined by GBS markers from the F_2 family to facilitate comparisons. The peak around 6 putatively represents sequences derived from the W chromosome, as well as deletions in Fish Creek relative to 94006. **b**. The same process was applied with the GBS markers that were generated from *Eco*T22I.

CHAPTER IV

HOMOLOGOUS INVERTED REPEATS PRESENT IN THE SEX DETERMINATION REGION OF *POPULUS TRICHOCARPA*

Abstract

The ages and sizes of a sex-determination region (SDR) are difficult to determine in non-model species. Due to the lack of recombination and enrichment of repetitive elements in SDRs, the quality of assembly with short sequencing reads is low. Unique features present in the sequence of SDRs help provide clues about how SDRs are established and how they evolve in the absence of recombination. Several *Populus* species have been reported with a male heterogametic configuration of sex (XX/XY system) mapped on chromosome 19, but the exact location of the SDR has been inconsistent among species. Lack of resolution in the size and location of the SDR in *Populus trichocarpa* exacerbates the situation further when the SDR is compared across other species. Here we present the first complete assembly of the SDR on the Y chromosome of a male individual of *P. trichocarpa*. We identified homologous gene sequences in the SDR of *P. trichocarpa* and the SDR of the W chromosome in *S. purpurea*. We show that the inverted repeats (IRs) found in the Y-SDR and the W-SDR are lineage-specific. We hypothesize that although the two IRs are derived from the same orthologous gene within each species, the newly-increased copy could maintain the original function through gene conversion, as is the case of the palindromic repeats in *S. purpurea*. Alternatively, the truncated inverted repeats in *P. trichocarpa* could function as a template for regulatory elements by being transcribed into regulatory RNAs that target the homologous gene. These findings highlight the idea that diverse sex-determining systems may be achieved through a similar evolutionary pathway, thereby providing a possible mechanism to explain the lability of sex-determination systems in plants.

Introduction

The evolution of sex is a fundamental yet complex mystery to biologists. Phenomena like sexual selection were used by Darwin as an example of natural selection for explaining the phenotypic differences between sexes in many species (Darwin 1859). Furthermore, the genetic mechanisms of sex determination have long fascinated molecular biologists, and the remarkable diversity of mechanisms in plants is just starting to be understood (Henry et al. 2018). Unlike gonochorous animals, angiosperm plants are largely cosexual, meaning that each individual has both sex functions. Some cosexual species have hermaphroditic flowers, and some are monoecious where pistils and stamens are present on different flowers within the same individual. Dioecious species represent about 5% of plants (Renner 2014). This does not mean that dioecy is rare. Instead, it occurs across many angiosperm phyla (Renner 2014; Henry et al. 2018), which indicates that the evolution of dioecy has occurred many times in plants. Another difference between animals and plants is that a range of reproduction modes can be found in just one genus, such as the genus *Silene*, which contains hermaphroditism, dioecy, and several intermediate modes as well (Balounova et al. 2019).

Sex chromosomes are generally considered to have evolved from a pair of autosomes with arrested recombination around the sex-determining loci (Charlesworth 2013). The cessation of recombination along with chromosomal rearrangements contributes to the further divergence of the proto sex chromosomes, which eventually leads to fully established sex chromosomes (Bachtrog 2013; Charlesworth 2013). Two main sex determination systems are commonly seen in animals and plants. One is female heterogamety, or ZW/ZZ, where females carry a pair of different sex chromosomes, such as in birds. On the contrary, in male heterogamety, or XX/XY, males carry a pair of different sex chromosomes, such as is seen in mammals. With modern sequencing techniques, it became possible to ask questions related to the characteristic features of the structure and evolution of sex chromosomes, which is often found to be one important feature in dioecious species. Several reported sex chromosomes in plants are homomorphic, in contrast to the strongly heteromorphic sex chromosomes in mammals. This indicates young ages of sex chromosomes in plants. It is likely that most plants have dynamic sex-determination regions (SDRs) which show rapid turnover resulting in poor conservation of the genetic mechanisms controlling of sex (Charlesworth 2015; Moore et al. 2016). Studies focusing on the turnover of sex chromosomes are mostly from animals. The temporal order and directional trends of turnovers in sex-chromosomal rearrangement are not well understood due to this false impression (Bergero & Charlesworth 2009). Recently, a study on the SDRs of *Fragaria* octoploids provided the first case of translocation of a cassette of 14 kb of female-specific sequence among several chromosomes (Tennessen et al. 2018). In *Silene*, section Otites has both female and male heterogamety systems and a possible change from female to male heterogamety within this section might have occurred (Balounova et al. 2019). Almost all species in the Salicaceae are dioecious (Cronk et al. 2015). However, both female and male heterogamety systems are reported to be found in this family (Geraldes et al. 2015; Pucholt et al. 2015; Hou et al. 2015; Zhou et al. 2018).

The sex of many species in *Populus* was reported as female heterogametic (Westergaard 1958). With more advanced molecular techniques, chromosome 19 has been shown to be male heterogametic in several *Populus* studies (Paolucci et al. 2010; Pakull et al. 2011, 2014; Geraldes et al. 2015). Although sex determination has been mapped to Chr19 in *Populus*, Chr19 is not the only chromosome containing sex-specific markers in sex association analysis (Geraldes et al. 2015). The inconsistent location of the SDR on multiple chromosomes in *Populus* is conspicuous compared to the consistent identification of SDRs around the center of Chr15 in several *Salix* species (Pucholt et al. 2015; Hou et al. 2015; Zhou et al. 2018). Multiple locations of sex-specific markers in *Populus* were proposed to be associated with the erroneous assembly of portions of the SDR in the reference genome (Geraldes et al. 2015). Furthermore, the SDR in *P. trichocarpa* was inferred to be small and compact with less than 20 genes spanning ~100 kbp on chromosome 19 (Geraldes et al. 2015), in contrast to the SDR of *Salix purpurea*, which contains 488 genes and spans over nearly 7 Mb (Chapter3). However, the previous results in *Populus* are based mainly on alignment of short read sequences to a reference genome derived from a female individual, which would lack the SDR in this XY species. More definitive conclusions can be drawn from assembly and analysis of a male reference genome.

In the study of this chapter, we established a new assembly derived from a male *P. trichocarpa* clone. By identifying sex-linked genetic markers in this new assembly, we identified the sex-determination region in the Y chromosome and described the genomic composition of this Y-SDR in detail. We also inferred the age of the SDR from the substitution rates estimated from the terminal repeats of autonomous LTR transposons. Finally, we tested if a shared sex-determining element was present in both genera. With these findings, we provide a possible interpretation of the relationship between two different sex-determining systems in *S. purpurea* and *P. trichocarpa*.

Methods

Initial genome assembly

Clone Stettler-14 is a male *P. trichocarpa* tree growing near Mt. Hood, Oregon. The tree was originally collected as part of a study to determine the rates of somatic mutation and variation in methylation status (Hofmeister et al. 2019). The genome was sequenced to 118 .58x depth using PacBio technology, with an average read length of 10,477 bp. The genome was assembled using CANU v1.4 and polished using QUIVER. The assembled genome contained 392.3 Mb of sequence and the contig N50 was 7.5 Mb. The genome also contained ~232.2 Mb of alternative haplotypes. Full details of the assembly and annotation can be found in Hofmeister et al. (2020).

Variants calling of individuals of natural population

100 unrelated individuals of each sex were selected to perform the sex association. The 2x100 bp resequencing reads of each individual were aligned to sequences in the main genome from the male reference genome through Bwa mem 0.7.17 (Li & Durbin 2009) with flags -M -t 8 -R. Duplicated reads were marked with MarkDuplicates from Picard (http://broadinstitute.github.io/picard/). These alignments were used to retrieve variants through the HaplotypeCaller of GATK (Van der Auwera et al. 2013). VariantFiltration of GATK was applied to filter variants with "AF < 0.01 || AF > 0.99 || QD < 10.0 || ExcessHet > 20.0 || FS > 10.0 || MQ < 58.0" in the -filter-expression flag as: 1) if allele frequency is lower than 0.01 or above 0.99; 2) the QUAL score normalized by allele depth is smaller than 10; 3) Phred-scaled p-value for exact test of excess heterozygosity is over 20; 4) Phred-scaled p-value using Fisher's exact test to detect strand bias is over 10; 5) RMS Mapping Quality is smaller than 58. The same

steps were applied when the alignments were generated with reference sequences of alternative haplotypes from the male reference genome.

Sex-association analysis

All SNP variants generated from previous steps were further selected with a minor allele frequency above 0.05 for sex-association analysis. The sex-association was performed with the same 100 females and 100 males by using the Fisher's exact test provided in plink v1.07 (Purcell et al. 2007). If the *P*-value of a tested marker was lower than the Bonferroni correction (with α=0.05), then it was considered to be significantly sex-associated. In the analysis of Stettler-14 V1 main genome, 4,586,112 SNPs were tested with a Bonferroni correction at 1.09 x 10^{-8}. In the analysis of Stettler-14 V2 main genome, 5,302,648 SNPs were tested with a Bonferroni correction at 9.43 x 10^{-9}.

Identifying the sex-specific covered region

To find alternative haplotypes derived from sex chromosomes (either X or Y), we aligned the same reads from 100 unrelated individuals of each sex with Bwa mem 0.7.17 (Li & Durbin 2009) to a reference that contain sequences from both the main genome and alternative haplotypes. Depth was calculated on the merged bam file from individuals of same sex using Samtools-1.2 (Li et al. 2009) and max depth was limited to 80,000. The median depth of 1kb non-overlapping windows was calculated with an in-house python script. These 1kb intervals were retained if the total median depth was no less than 400 to avoid inaccurate estimation on the depth ratio. If the depth ratio $\log_2(\frac{F_{100}+1}{M_{100}+1})$ of the interval was smaller than -1, then the interval was considered as male-biased. If the log ratio was greater than 1, then it was considered as female-biased.

Genetic linkage mapping

Three half-sib families of male parents from a half-diallel designed cross (7 × 7) were used to generate three genetic maps. Similar protocol as described above was used to call variants. For each half-sib cross, only markers in backcross configuration were used. Onemap (Margarido et al. 2007) was used to cluster markers into linkage groups and estimate the genetic distances. For computational reasons, markers of each cross were divided into two sets (even vs odd indexes), so two maps were created for each cross, totalling six maps. In addition, a map generated from the interspecific cross 52124 (*P. deltoides* × *P. trichocarpa*) was used to increase the accuracy. These seven maps were combined using allmaps (Tang et al. 2015) to recreate the chromosomes (details below).

Identification of contigs from SDR and reconstructing Y chromosome

After taking sex-association SNPs and male-biased intervals into account, we identified one Y-linked contig that was originally placed on Chr18 in the v1 genome. We also identified three alternative haplotypes of this contig, presumably derived from the X chromosome. To evaluate the placement of this Y-linked contig, we compared the order of markers in a genetic map derived from a controlled cross to the order in the physical assembly (Figure 1). The Chr18 placement was clearly incorrect based on this analysis, which indicated that the contig containing the SDR should be placed on Chr19 (Figure 1), as was previously shown (Geraldes et al. 2015). We therefore broke the chromosomal scaffolds into contigs at 10 kb gap intervals. The genetic map was used to produce a new assembly with allmaps (Tang et al. 2015). The orientation of each chromosome was determined by comparison to scaffolds from the corresponding region of the Nisqually-1 v4 assembly. For chromosome 19, corresponding region of the Nisqually-1 v4 assembly (scaffold N.Chr19:1-1,632,082 bp and scaffold_25, which contained a large number of sex-associated SNPs (N.scaffold_25: 1-640,640 bp) was also used for adjusting the order and orientations as well. With the adjusted order and orientation, we manually built chromosome 19Y with the contig carrying the SDR and rest contigs in chromosome 19 with 10,000 bp gap insertion between those contigs. Given a finding of small SDR size in chromosome 19Y, instead of constructing the whole X chromosome, we only built an X-linked scaffold for the SDR by concatenating those three X-linked contigs according to the order and orientation of the SDR. Alignments of these contigs and the Y-SDR were accomplished using lastz-1.04 (Harris 2007).

Gene annotation on the SDR and X-linked scaffold

To annotate potential coding genes that were missed by the automated annotation in the SDR and the X scaffold, the new Y-SDR contig and the X scaffold were submitted to the Fgenesh (Solovyev et al. 2006) online service (http://www.softberry.com/berry.phtml?topic=fgenesh) with specific gene-finding parameters for *Populus trichocarpa*. The predicted peptide sequences were searched against predicted proteins from *Populus trichocarpa* v3.0, and *Arabidopsis thaliana* TAIR10 in Phytozome 12 (https://phytozome.jgi.doe.gov/) to find the closest homologous annotation. Only predicted genes that have at least one hit in either species were retained as valid predictions.

Estimation of the divergence of the SDR

To identify allelic gene pairs for calculation of synonymous substitutions between the Z and W alleles, a reciprocal blast of all annotated peptide sequences was performed by blastp with a limit of a maximum number of hits at 5, and MCscanX (Wang et al. 2012) was run with default parameters. The synonymous and nonsynonymous substitution rate of each gene pair in each syntenic block (dS and dN, respectively) was estimated by aligning the sequences with

CLUSTALW (Wilm et al. 2007) and using the yn00 function in PAML (Yang 2007). Gene pairs with dS values smaller than 0.5 were kept for estimating the divergence between X and Y.

Identification of recently inserted LTR retrotransposable elements and repetitive elements

To identify recent insertions of transposable elements in the SDR and corresponding X interval, LTRharvest (Ellinghaus et al. 2008) was run with the sequence of the SDR (Chr19Y: 1-120,000 bp) and the X scaffold with the target site duplication restricted to 4 bp to 20 bp. To find the protein domains in the coding region, a protein domain search against Pfam-A domains (release 32) was performed using the hidden Markov model methods implemented in LTRdigest (–hmms flag) (Steinbiss et al. 2009). The same methods described in chapter 3 were used to estimate the substitution rates between the LTR repeats.

Short tandem duplications were initially identified through TRF 4.09 (Benson 1999) with 2 5 7 80 10 50 2000 -l 2 -d. Then, regions that contain no less than 1000 bp with a typical telomeric repeat motif (TTTAGGG)$_n$-3’ or (CCCTAAA)$_n$-3’ were designated as telomeric repeats (Richards & Ausubel 1988). For centromeres, we decided to use the assembled sequence. We set the filter to search for a region that contains a periodical length between 150 bp and 400 bp with a number of copies greater than 50 for candidates of centromeres. The RepeatModeler (v1.0.8) package (http://www.repeatmasker.org) was used to identify and mask repetitive elements in the genome.

Expression of the inverted repeats

RNA-seq reads from flower tissues of three females (BESC423, 443, 842) and three males (GW9592, 9840, 9911) were retrieved from the JGI portal (https://genome.jgi.doe.gov/portal/). Each set of RNA-seq reads were aligned to the Stettler-14

V2 reference genome with HISAT2 (Kim et al. 2019). The alignments from the inverted repeats were visually checked for accuracy in the Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv/). All replicates of the same stage of the same individual were merged with samtools-1.2 (Li et al. 2009). The number of reads per site was retrieved with the depth flag by samtools. Depth was calculated from the median of coverage in each 100 bp window for visualization.

Inference of phylogenetic relationship of the homologous sequences in the SDRs

The shared sequence between the Y-SDR in *P. trichocarpa* and W-SDR in *S. purpurea* was identified using reciprocal blastp searches using the predicted proteins from each interval. Only best mutual hits were taken as shared genes. Due to the incompleteness of the response regulator fragments in the inverted repeats in the Y-SDR, the coding sequence of the complete homologous gene *Po14v11g057342m* was used to annotate those fragments. Given poor bootstrap values when short fragments of truncated response regulator were used in the alignment, we decided to use only the longest fragment on ARM-4a as the representative sequence of the response regulator fragments in the inverted repeats. Homologous sequences identified between the two SDRs were aligned by MUSCLE using default parameters provided in MEGA 5 (Tamura et al. 2011) and alignment was manually checked and adjusted if any alignment error was found. The neighbor-joining method was used for building the phylogenetic tree with the substitution rate modeled by Kimura 2-parameter model provided in MEGA5, and the rate variation among sites was modeled with a gamma distribution (shape parameter = 1).

Results

Assembly of the new version of Stettler-14 with the Y chromosome

The new version (V2) of the Stettler-14 contains 19 chromosomes, 7 contigs plus mitochondrion and chloroplast genomes in the main genome, which spans 391 Mb in total (Table 1). Among 122 contigs assembled in 19 chromosomes in the V1 genome, 62 remained in the same order and 50 of them were adjusted with our new map in the V2 assembly (Table 2). Thus, a new V2 assembly contains 112 contigs mapped in 19 chromosomes. Two contigs from the V1 chromosome 7 (scaffold_4005 and scaffold_4006 in V2) could not be mapped with our new map, thus they were kept in the new main genome with other unmapped scaffolds from V1. Additionally, the remaining eight contigs in the V1 chromosomes were identified as alternative haplotypes of the assembly, so they were kept with other alternative haplotypes from V1. Changes of contig positions between the two assemblies can be found in the Table 2. The new Y chromosome contains 8 contigs in the new assembly with a total length of 16.5 Mb (Table 3).

Genomic composition of the new assembly

The total size of annotated repetitive elements is 159.8 Mb taking up about 41.1% of main genome (Table 4). LTR-Copia elements occupy about 3.8% of the assembly and LTR-Gypsy occupies about 11.6%. Over the 19 chromosomes, 32.2% of the genomic regions are annotated with genes. Interestingly, nearly half (49%) of the chromosome 19 consists of repetitive elements and one third of them are LTR-Gypsy (Table 4). Due to the large number of repetitive elements on chromosome 19, the gene space only comprises 28.3% of the chromosome, the lowest among all chromosomes. On the contrary, chromosome 9 contains the lowest amount of repetitive elements at 30.1% of its size, and has the highest gene content (40.0%).

Identification of sex-associated scaffolds based on SNP associations

In the Stettler-14 V1 main genome, 4,586,112 SNP variants called from GATK were tested with the association of the sex by the Fisher's exact test. This yields 119 sex-associated SNPs (*P*-value < 1.09 x 10^{-8}) and all of them were found within a 300 kb stretch on Chr18 ranging from 15,993,536 bp to 16,289,766 bp in the V1 assembly (Figure 2). In alternative haplotypes, 91 SNP variants (*P*-value < 1.66 x 10^{-8}) were identified to be sex-associated from 3,017,607 tested SNP variants. These sex-associated SNPs helped us identify scaffold_43 and scaffold_1208 to be sex-associated. Scaffold_43 contains 33 sex-associated markers and scaffold_1208 contains 56 sex-associated markers. Further alignment of scaffold_43 and scaffold_1208 also confirmed that they were alternative haplotypes of chromosome 19 (Table 5). Scaffold_71 and scaffold_1121 are not considered to be sex-linked because there is only one sex-associated SNP in each of them.

Using the new assembled main genome as a reference, we re-called the genotypes from the same set of individuals. 5,302,648 SNPs in the main genome called from GATK were tested for association with sex. This yields 200 sex-associated SNPs (*P*-value < 9.43 x 10^{-9}) and all of them were found within a 300 kb stretch from the beginning of chromosome 19Y (Figure 3). A majority of sex-associated SNPs are found within the first 120 kb of the Y chromosome, with the remaining marginally significant sex-associated SNPs scattered around two regions at 160 kb and 300 kb (Figure 3d).

The distribution of genotype configurations of the 200 sex-associated markers matches a male heterogametic system (XX/XY system) (Figure 4). About 146 markers are configured as homozygous XX in females, while 138 markers are configured as heterozygous XY in males (Figure 4c). This confirms the Y haplotypes are present in the main reference genome, while alternative alleles are from X haplotypes. Additionally, the preponderance of female null alleles distributed from 10 kb to 50 kb shows the reference contains at least 40 kb of male-specific Y regions that are not covered in females (Figure 4b). The majority of sex-associated markers occur within 115 kb, suggesting that the SDR is confined to this region (Figure 4c).

Male-specific regions

To identify potential male-specific sequences in the assembly, we also performed depth analysis as described in chapter 3. In the main genome, the depth analysis discloses that same contig with sex-associated SNPs also contains 107 male-biased markers. The average of these 107 male-biased markers shows an extremely biased depth toward males with M:F depth about 9:1. This means these markers are from a male-specific region with male coverage only. Further examination of the coordinates of these male-biased markers confirms that they are from the same contig where 119 sex-associated SNPs were found (Figure 4b). Among the analyzed alternative haplotype scaffolds, scaffold_43 and scaffold_1534 were found to contain 10 (out of 310) and 5 (out of 31) male-biased depth markers. However, for these male-biased markers, the depth of males is only about twice that of the females in both scaffolds, which is not expected to be present in an XX/XY system. Since the reference used for depth analysis contains sequences from both the main genome and alternative haplotypes, we suspect this could be an artifact due to the extra copy in the reference. Further alignment of scaffold_1534 confirms that this scaffold is an alternative haplotype of Chr19Y with high sequence similarity (>99%).

Genomic composition of the Y-SDR

Approximately 7,800 bp at the end of the SDR was comprised of short tandem repeats of telomere repeat motif (TTTAGGG)n-3' (Figure 5a). Similarly, one of its alternative haplotypes, scaffold_1208 contains about 4,000 bp telomere at its end. The Y-SDR is about 120 kb at the beginning of chromosome 19 assembly and it contains about 50 kb of sequence that is only present in male haplotypes (Figure 4b). The rest of the X-degenerate regions contain the majority of sex-associated markers identified above (Figure 4c). The male-specific regions consist primarily of fragments from Gypsy-LTR elements according to our analysis while the X-degenerate region does not show discrimination between the types of repetitive elements (Figure 5b). Additional identification of four autonomous LTRs allows us to roughly estimate the minimal age of the SDR (Figure 5d). These Y-linked autonomous LTRs inserted into the Y chromosome after the cessation of recombination. No autonomous LTR was found in the male-limited regions. All four LTRs are found to be inserted around the X-degenerate region but absent from X alternative haplotypes. Among these four autonomous LTRs, a Gypsy type LTR, *Ltr-y-a* shows the highest substitution rates of 33.95 substitutions per 1 kb, which means that this oldest insertion occurred no later than around 13.6±3.7 SE million years ago. The remaining four LTRs have lower substitution rates (Table 6).

Five genes are annotated in the X-degenerate region of the SDR (Figure 5c, Table 7). including several sex candidates reported in a previous study of the SDR in *P. trichocarpa* (Geraldes et al. 2015). The corresponding X alleles of these genes are also identified in the previous Nisqually version 3 genome and current Nisqually version 4 genome (Table 7). To estimate the divergence after the arrest of recombination in the SDR, we compared the annotations from two X-haplotypes (a misplaced contig and scaffold_25) in the Nisqually version 4. The estimated synonymous substitutions rate or d_S values between X and Y alleles are different among different genes. The Po14v11g055355m (function unknown) does not contain any synonymous substitutions but only nonsynonymous substitutions. Estimated d_S values of the other three genes are 0.0176, 0.0224, and 0.0669, where *MET1* (Po14v11g055360m) furthest from the male-specific region has the lowest substitution rate (Table 7). Interestingly, *TCP-1*

(Po14v11g055363m) has the highest substitution rate, which is also the gene closest to the male-specific region. All of these d_S values were smaller than the previous estimates of average d_S 0.146±0.0022 SE between *S. purpurea* and *P. trichocarpa* (Zhou et al. 2019). A further search of the orthologous genes in a female reference (94006) of *S. purpurea* by using these Y-SDR genes showed that Po14v11g055355m was the only ortholog containing a hit on chromosome 19 in *S. purpurea*. The rest genes do not have hits on the chromosome 19 in *S. purpurea*. Both *MET1* and *TCP-1* have hits to Sapur.004G100800 and Sapur.004G101000 on chromosome 4 in *S. purpurea*, which matches what we observed in the translocation analysis (Figure 5e). The R-gene, Po14v11g055357m was excluded from the divergence analysis due to an excessive number of hits in the genome. When these genes were searched against a male *S. purpurea* reference, Po14v11g055355m and the *MET1* gene have hits to SpFC.19G000200 and SpFC.19G000100 from chromosome 19 in the male *S. purpurea* reference.

The inverted repeats (IRs) in the Y-SDR

In the Y-SDR, one of the features in the male-specific region is a cluster of five homologous arms arranged as inverted repeats (IRs) that might be derived from duplications and structural variations (Figure 5a). By aligning the sequence from 20 kb to 45 kb of the Y chromosome, five arms were identified based on their sequence identity (Table 8 and Figure 6). The longest IR is formed between ARM-2 and ARM-3, and two arms have a similar length of about 3.8 kb with an identity of 93.3%. The Spacer-1 is around 2 kb between the two arms, which are not homologous to these arms. ARM-4a and partial sequence of the ARM-3 can also form an IR structure with a 2.7 kb spacer sequence, Spacer-2 between the two arms (Table 8). ARM-1 and ARM-4b are shorter than the other arms but both contain homologous sequences of other arms (Figure 6). According to the sequence analysis of recent transpositions into the SDR, all five arms have a high sequence identity (>90%) to the pseudoautosomal region (PAR) at the other end of the Y chromosome.

These IRs were found to share sequence identity (>90%) with a response regulator gene (*PtRR11/9, Potri.019G133600* in *P. trichocarpa* V3), *Po14v11g057342m* (Chr19: 16,454,242-16,457,207) at the other end of the Y chromosome. All five arms contain the first exon of this gene model but none of them contains the full length of the gene model (Table 9). Both the last two exons (exon 5 and exon 6) are absent from these arms. ARM-1 only contains the first exon which does not contain any coding sequence. The only copy of exon 4 in the SDR is in the spacer between ARM-3 and ARM-4a with transcript-order along with exon 1-3 on ARM-4a (Figure 6c). All of the introns between exons in this region are also present in order based on the alignments to the gene model of *PtRR17*. The Spacer-2 between ARM-3 and ARM-4a also contains a fragment from chromosome 9 (Figure 4e), which includes upstream sequence and the first exon of a Glutamyl-tRNA reductase gene (*Po14v11g032403m*, Chr09: 7,655,369-7,659,100), an orthologous gene of atHEMA in *Arabidopsis thaliana*. By comparing these IRs to the coding sequence of *PtRR11/9*, we noticed that all IR arms have lost the ability to encode a complete protein due to the frameshift caused by deletions or insertions, or even loss of start codon in ARM-3.

The expression of these IRs was detected by using RNA-seq of flower tissues from three males (Figure 7). We found male-specific expression in the region from 20 to 40 kb on chromosome 19. The fragments derived from the first exon of *Po14v11g032403m,* a homolog of atHEMA in the Spacer-2 between ARM-3 and ARM-4a showed expression in both the middle and late flower stages. The fragments of exon1, exon2, and exon 3 from ARM-2 and ARM-3 were expressed in all three samples (Figure 7). Thus, these IRs are able to be transcribed into

RNAs. However, based on the alignment of coding regions, they are unlikely to code for a protein.

The origin of IRs

Given the presence of homologous response regulator gene or fragments in inverted repeats of both SDRs in two species, we decided to test if the translocated duplication events to these inverted repeats are independent lineage-specific events. In the constructed phylogenetic tree, each translocation appears as a lineage-specific duplication after the split of two species instead of one shared translocation (Figure 8). This suggests that translocations from the autosomes are lineage-specific and independent events after the split of the two species (Figure 8).

Table 4.1 The statistics of the version 2 assembly of Stettler-14.

CHR	Gap Size (bp)	Gap Insertion count	SEQ(bp)	SEQ contig count	TotalSize (bp)
Chr01	130,000	13	49,548,573	14	49,678,573
Chr02	40,000	4	25,269,097	5	25,309,097
Chr03	40,000	4	23,594,591	5	23,634,591
Chr04	60,000	6	23,120,857	7	23,180,857
Chr05	30,000	3	23,900,588	4	23,930,588
Chr06	90,000	9	26,750,282	10	26,840,282
Chr07	10,000	1	14,820,512	2	14,830,512
Chr08	30,000	3	20,221,161	4	20,251,161
Chr09	10,000	1	12,976,220	2	12,986,220
Chr10	40,000	4	22,571,514	5	22,611,514
Chr11	50,000	5	18,712,988	6	18,762,988
Chr12	60,000	6	14,978,488	7	15,038,488
Chr13	40,000	4	15,789,923	5	15,829,923
Chr14	70,000	7	18,810,830	8	18,880,830
Chr15	20,000	2	15,039,279	3	15,059,279
Chr16	40,000	4	14,606,863	5	14,646,863
Chr17	50,000	5	15,968,856	6	16,018,856
Chr18	50,000	5	15,901,547	6	15,951,547
Chr19	70,000	7	16,457,936	8	16,527,936
Grand Total (Chr)	930,000	93	389,040,105	112	389,970,105
scaffold_758			105,391		105,391
scaffold_2190			60,967		60,967
scaffold_2269			59,578		59,578
scaffold_3504			33,156		33,156
scaffold_3526			31,461		31,461
scaffold_4005			111,088		111,088
scaffold_4006			154,143		154,143
Chloroplast			157,033		157,033
Mitochondrion			803,750		803,750
Grand Total (main genome)			390,556,672		391,486,672

Table 4.2 The comparison of contigs between version1 and version2 assembly of Stettler-14

ContigID	V1.CHR	V1.start	V1.end	V2.CHR	V2.start	V2.end	length	ori	Adjusted?N:same as versio1
Chr01_25740966_33211283	Chr01	25740966	33211283	Chr01	24507806	31978123	7470318	+	Adjusted
Chr01_33221284_41876894	Chr01	33221284	41876894	Chr01	31988124	40643734	8655611	+	Adjusted
Chr03_1_184912	Chr03	1	184912	Chr01	40653735	40838646	184912	+	Adjusted
Chr01_41886895_42008194	Chr01	41886895	42008194	Chr01	40848647	40969946	121300	+	Adjusted
Chr01_42018195_47333803	Chr01	42018195	47333803	Chr01	40979947	46295555	5315609	+	Adjusted
Chr01_47343804_48286028	Chr01	47343804	48286028	Chr01	46305556	47247780	942225	+	Adjusted
Chr01_48296029_50716821	Chr01	48296029	50716821	Chr01	47257781	49678573	2420793	+	Adjusted
Chr03_194913_3350160	Chr03	194913	3350160	Chr03	1	3155248	3155248	+	Adjusted
Chr16_2482109_2847456	Chr16	2482109	2847456	Chr03	3165249	3530596	365348	+	Adjusted
Chr03_3360161_3797499	Chr03	3360161	3797499	Chr03	3540597	3977935	437339	+	Adjusted
Chr03_3807500_7425933	Chr03	3807500	7425933	Chr03	3987936	7606369	3618434	+	Adjusted
Chr03_7435934_23454155	Chr03	7435934	23454155	Chr03	7616370	23634591	16018222	+	Adjusted
Chr07_7438823_15504207	Chr07	7438823	15504207	Chr07	6765128	14830512	8065385	+	Adjusted
Chr10_947427_5449970	Chr10	947427	5449970	Chr10	865268	5367811	4502544	+	Adjusted
Chr10_5459971_10934104	Chr10	5459971	10934104	Chr10	5377812	10851945	5474134	+	Adjusted
Chr10_10944105_22693673	Chr10	10944105	22693673	Chr10	10861946	22611514	11749569	+	Adjusted
Chr11_93173_7401520	Chr11	93173	7401520	Chr11	1	7308348	7308348	+	Adjusted
Chr11_7637837_9580910	Chr11	7637837	9580910	Chr11	7318349	9261422	1943074	+	Adjusted
Chr11_9590911_10101600	Chr11	9590911	10101600	Chr11	9271423	9782112	510690	+	Adjusted
Chr11_10111601_14177274	Chr11	10111601	14177274	Chr11	9792113	13857786	4065674	+	Adjusted
Chr11_14187275_18804531	Chr11	14187275	18804531	Chr11	13867787	18485043	4617257	+	Adjusted
Chr11_18814532_19082476	Chr11	18814532	19082476	Chr11	18495044	18762988	267945	+	Adjusted
Chr11_1_83172	Chr11	1	83172	Chr12	813653	896824	83172	+	Adjusted
Chr10_865268_937426	Chr10	865268	937426	Chr12	906825	978983	72159	+	Adjusted
Chr12_813653_2393975	Chr12	813653	2393975	Chr12	988984	2569306	1580323	-	Adjusted
Chr12_2403976_6200597	Chr12	2403976	6200597	Chr12	2579307	6375928	3796622	+	Adjusted
Chr12_6210598_7294464	Chr12	6210598	7294464	Chr12	6385929	7469795	1083867	+	Adjusted
Chr12_7304465_14863157	Chr12	7304465	14863157	Chr12	7479796	15038488	7558693	+	Adjusted
Chr13_5036445_8641180	Chr13	5036445	8641180	Chr13	4709215	8313950	3604736	+	Adjusted
Chr13_8651181_9002748	Chr13	8651181	9002748	Chr13	8323951	8675518	351568	-	Adjusted
Chr13_9012749_9393014	Chr13	9012749	9393014	Chr13	8685519	9065784	380266	+	Adjusted
Chr13_9403015_16157153	Chr13	9403015	16157153	Chr13	9075785	15829923	6754139	+	Adjusted
Chr14_15835894_17647670	Chr14	15835894	17647670	Chr14	15520489	17332265	1811777	-	Adjusted
Chr14_15520489_15825893	Chr14	15520489	15825893	Chr14	17342266	17647670	305405	-	Adjusted
Chr01_24507806_25730965	Chr01	24507806	25730965	Chr14	17657671	18880830	1223160	+	Adjusted
Chr16_2857457_8747733	Chr16	2857457	8747733	Chr16	2482109	8372385	5890277	+	Adjusted
Chr16_8757734_9428119	Chr16	8757734	9428119	Chr16	8382386	9052771	670386	+	Adjusted
Chr16_9438120_9574562	Chr16	9438120	9574562	Chr16	9062772	9199214	136443	-	Adjusted

Chr16_9584563_15022211	Chr16	9584563	15022211	Chr16	9209215	14646863	5437649	+	Adjusted
Chr18_8942904_10910179	Chr18	8942904	10910179	Chr18	8765020	10732295	1967276	+	Adjusted
Chr18_10920180_15785410	Chr18	10920180	15785410	Chr18	10742296	15607526	4865231	+	Adjusted
Chr18_16342221_16676241	Chr18	16342221	16676241	Chr18	15617527	15951547	334021	+	Adjusted
Chr18_15795411_16332220	Chr18	15795411	16332220	Chr19	1	536810	536810	-	Adjusted
Chr19_1_454219	Chr19	1	454219	Chr19	546811	1001029	454219	-	Adjusted
Chr19_1728559_2300400	Chr19	1728559	2300400	Chr19	1011030	1582871	571842	+	Adjusted
Chr19_464220_1067824	Chr19	464220	1067824	Chr19	1592872	2196476	603605	-	Adjusted
Chr19_1077825_1718558	Chr19	1077825	1718558	Chr19	2206477	2847211	640734	+	Adjusted
Chr19_2310401_6813220	Chr19	2310401	6813220	Chr19	2857211	7360030	4502820	+	Adjusted
Chr19_6823221_14488160	Chr19	6823221	14488160	Chr19	7370031	15034970	7664940	+	Adjusted
Chr19_14498161_15981126	Chr19	14498161	15981126	Chr19	15044971	16527936	1482966	+	Adjusted
Chr01_1_9888560	Chr01	1	9888560	Chr01	1	9888560	9888560	+	N
Chr01_9898561_18250635	Chr01	9898561	18250635	Chr01	9898561	18250635	8352075	+	N
Chr01_18260636_20397240	Chr01	18260636	20397240	Chr01	18260636	20397240	2136605	+	N
Chr01_20407241_22037535	Chr01	20407241	22037535	Chr01	20407241	22037535	1630295	+	N
Chr01_22047536_23134204	Chr01	22047536	23134204	Chr01	22047536	23134204	1086669	+	N
Chr01_23144205_23907092	Chr01	23144205	23907092	Chr01	23144205	23907092	762888	+	N
Chr01_23917093_24497805	Chr01	23917093	24497805	Chr01	23917093	24497805	580713	+	N
Chr02_1_4663646	Chr02	1	4663646	Chr02	1	4663646	4663646	+	N
Chr02_4673647_18230933	Chr02	4673647	18230933	Chr02	4673647	18230933	13557287	+	N
Chr02_18240934_18519700	Chr02	18240934	18519700	Chr02	18240934	18519700	278767	+	N
Chr02_18529701_19304897	Chr02	18529701	19304897	Chr02	18529701	19304897	775197	+	N
Chr02_19314898_25309097	Chr02	19314898	25309097	Chr02	19314898	25309097	5994200	+	N
Chr04_1_9886930	Chr04	1	9886930	Chr04	1	9886930	9886930	+	N
Chr04_9896931_10422575	Chr04	9896931	10422575	Chr04	9896931	10422575	525645	+	N
Chr04_10432576_12261701	Chr04	10432576	12261701	Chr04	10432576	12261701	1829126	+	N
Chr04_12271702_13333728	Chr04	12271702	13333728	Chr04	12271702	13333728	1062027	+	N
Chr04_13343729_13822371	Chr04	13343729	13822371	Chr04	13343729	13822371	478643	+	N
Chr04_13832372_17934164	Chr04	13832372	17934164	Chr04	13832372	17934164	4101793	+	N
Chr04_17944165_23180857	Chr04	17944165	23180857	Chr04	17944165	23180857	5236693	+	N
Chr05_1_12930902	Chr05	1	12930902	Chr05	1	12930902	12930902	+	N
Chr05_12940903_13058191	Chr05	12940903	13058191	Chr05	12940903	13058191	117289	+	N
Chr05_13068192_13583035	Chr05	13068192	13583035	Chr05	13068192	13583035	514844	-	N
Chr05_13593036_23930588	Chr05	13593036	23930588	Chr05	13593036	23930588	10337553	+	N
Chr06_1_716779	Chr06	1	716779	Chr06	1	716779	716779	+	N
Chr06_726780_9565911	Chr06	726780	9565911	Chr06	726780	9565911	8839132	+	N
Chr06_9575912_13853036	Chr06	9575912	13853036	Chr06	9575912	13853036	4277125	+	N
Chr06_13863037_14947758	Chr06	13863037	14947758	Chr06	13863037	14947758	1084722	+	N
Chr06_14957759_15148964	Chr06	14957759	15148964	Chr06	14957759	15148964	191206	+	N
Chr06_15158965_15401667	Chr06	15158965	15401667	Chr06	15158965	15401667	242703	+	N
Chr06_15411668_15599556	Chr06	15411668	15599556	Chr06	15411668	15599556	187889	+	N

Chr06_15609557_16037684	Chr06	15609557	16037684	Chr06	15609557	16037684	428128	+	N
Chr06_16047685_26378503	Chr06	16047685	26378503	Chr06	16047685	26378503	10330819	+	N
Chr06_26388504_26840282	Chr06	26388504	26840282	Chr06	26388504	26840282	451779	+	N
Chr07_1_6755127	Chr07	1	6755127	Chr07	1	6755127	6755127	+	N
Chr08_1_15252432	Chr08	1	15252432	Chr08	1	15252432	15252432	+	N
Chr08_15262433_15828470	Chr08	15262433	15828470	Chr08	15262433	15828470	566038	-	N
Chr08_15838471_19958690	Chr08	15838471	19958690	Chr08	15838471	19958690	4120220	+	N
Chr08_19968691_20251161	Chr08	19968691	20251161	Chr08	19968691	20251161	282471	+	N
Chr09_1_1219669	Chr09	1	1219669	Chr09	1	1219669	1219669	+	N
Chr09_1229670_12986220	Chr09	1229670	12986220	Chr09	1229670	12986220	11756551	+	N
Chr10_1_290450	Chr10	1	290450	Chr10	1	290450	290450	+	N
Chr10_300451_855267	Chr10	300451	855267	Chr10	300451	855267	554817	+	N
Chr12_1_803652	Chr12	1	803652	Chr12	1	803652	803652	-	N
Chr13_1_4699214	Chr13	1	4699214	Chr13	1	4699214	4699214	+	N
Chr14_1_2391150	Chr14	1	2391150	Chr14	1	2391150	2391150	+	N
Chr14_2401151_14660941	Chr14	2401151	14660941	Chr14	2401151	14660941	12259791	+	N
Chr14_14670942_15098390	Chr14	14670942	15098390	Chr14	14670942	15098390	427449	+	N
Chr14_15108391_15380144	Chr14	15108391	15380144	Chr14	15108391	15380144	271754	+	N
Chr14_15390145_15510488	Chr14	15390145	15510488	Chr14	15390145	15510488	120344	+	N
Chr15_1_5988787	Chr15	1	5988787	Chr15	1	5988787	5988787	+	N
Chr15_5998788_6388890	Chr15	5998788	6388890	Chr15	5998788	6388890	390103	+	N
Chr15_6398891_15059279	Chr15	6398891	15059279	Chr15	6398891	15059279	8660389	+	N
Chr16_1_2472108	Chr16	1	2472108	Chr16	1	2472108	2472108	+	N
Chr17_1_1439050	Chr17	1	1439050	Chr17	1	1439050	1439050	+	N
Chr17_1449051_3538016	Chr17	1449051	3538016	Chr17	1449051	3538016	2088966	+	N
Chr17_3548017_6416780	Chr17	3548017	6416780	Chr17	3548017	6416780	2868764	+	N
Chr17_6426781_8272838	Chr17	6426781	8272838	Chr17	6426781	8272838	1846058	+	N
Chr17_8282839_9088194	Chr17	8282839	9088194	Chr17	8282839	9088194	805356	+	N
Chr17_9098195_16018856	Chr17	9098195	16018856	Chr17	9098195	16018856	6920662	+	N
Chr18_1_3825649	Chr18	1	3825649	Chr18	1	3825649	3825649	+	N
Chr18_3835650_6140443	Chr18	3835650	6140443	Chr18	3835650	6140443	2304794	+	N
Chr18_6150444_8755019	Chr18	6150444	8755019	Chr18	6150444	8755019	2604576	+	N
Chr01_50726822_50762295	Chr01	50726822	50762295	scaffold_4001	1	35474	35474	+	ToAltHap
Chr04_23190858_23275917	Chr04	23190858	23275917	scaffold_4002	1	85060	85060	+	ToAltHap
Chr06_26850283_26963253	Chr06	26850283	26963253	scaffold_4003	1	112971	112971	+	ToAltHap
Chr07_6765128_7143591	Chr07	6765128	7143591	scaffold_4004	1	378464	378464	+	ToAltHap
Chr07_7153592_7264679	Chr07	7153592	7264679	scaffold_4005	1	111088	111088	+	unmapped
Chr07_7274680_7428822	Chr07	7274680	7428822	scaffold_4006	1	154143	154143	+	unmapped
Chr11_7411521_7627836	Chr11	7411521	7627836	scaffold_4007	1	216316	216316	+	ToAltHap
Chr13_4709215_5026444	Chr13	4709215	5026444	scaffold_4008	1	317230	317230	+	ToAltHap
Chr13_16167154_16265704	Chr13	16167154	16265704	scaffold_4009	1	98551	98551	+	ToAltHap
Chr18_8765020_8932903	Chr18	8765020	8932903	scaffold_4010	1	167884	167884	+	ToAltHap

Table 4.3 The contigs used in new Y chromosome assembly.

ChrID	recode	Start	End	Scaffold/ContigID in v1	Note	Length (bp)	Orientation
Chr19Y	yc1	1	536,810	Chr18_15795411_16332220	SDR*	536,810	-
Chr19Y	yc2	546,811	1,001,029	Chr19_1_454219/Contig1	PAR	454,219	-
Chr19Y	yc3	1,011,030	1,582,871	Chr19_1728559_2300400/Contig4	PAR	571,842	-
Chr19Y	yc4	1,592,872	2,196,476	Chr19_464220_1067824/Contig2	PAR	603,605	+
Chr19Y	yc5	2,206,477	2,847,210	Chr19_1077825_1718558/Contig3	PAR	640,734	+
Chr19Y	yc6	2,857,211	7,360,030	Chr19_2310401_6813220/Contig5	PAR	4,502,820	+
Chr19Y	yc7	7,370,031	15,034,970	Chr19_6823221_14488160/Contig6	PAR	7,664,940	+
Chr19Y	yc8	15,044,971	16,527,936	Chr19_14498161_15981126/Contig7	PAR	1,482,966	+

* This contig contains both SDR and PAR.

Table 4.4 Cumulative size in Mb of genes and LTR retrotransposons across 19 chromosomes in the genome. Four columns on the right are percentages of the proportion of the specific type of content.

CHR	**Total Repeat**	**LTR-Copia**	**LTR-Gypsy**	**Gene**	**TotalRepeat%**	**LTR/Copia%**	**LTR/Gypsy%**	**Gene%**
Chr01	21.6	2.2	5.7	15.3	43.6	4.4	11.5	30.9
Chr02	9.8	0.7	2.5	8.4	38.8	2.9	9.7	33.4
Chr03	9.2	0.9	2.6	7.8	39.1	3.7	11.0	33.0
Chr04	10.0	1.0	2.9	7.1	43.1	4.2	12.7	30.8
Chr05	9.4	0.8	2.6	7.9	39.2	3.2	11.0	32.9
Chr06	10.1	0.9	2.4	9.1	37.7	3.2	8.9	34.2
Chr07	6.0	0.5	1.7	4.6	40.4	3.5	11.2	31.3
Chr08	7.2	0.6	2.1	7.2	35.8	3.1	10.5	35.6
Chr09	3.9	0.2	0.9	5.2	30.1	1.8	7.2	40.0
Chr10	8.1	0.7	2.3	8.2	35.9	2.9	10.0	36.4
Chr11	8.9	1.0	2.6	5.4	47.6	5.2	13.8	29.0
Chr12	6.4	0.7	1.8	4.5	42.6	4.9	12.2	29.8
Chr13	6.7	0.7	2.0	5.2	42.7	4.4	12.6	32.8
Chr14	7.8	0.7	2.1	5.9	41.4	3.8	11.3	31.6
Chr15	6.2	0.6	1.9	4.9	41.5	3.9	12.7	32.7
Chr16	6.4	0.6	1.9	4.3	43.6	3.9	12.7	29.1
Chr17	7.3	0.7	2.4	4.8	45.4	4.3	15.1	30.0
Chr18	6.8	0.6	2.1	4.9	42.6	3.7	13.1	31.1
Chr19	8.1	0.9	2.7	4.7	49.0	5.7	16.5	28.3
Total	159.8	14.9	45.2	125.4	41.1	3.8	11.6	32.2

Table 4.5 The contigs used in the assembly of the corresponded alternative haplotype on X.

seqID	ChrID.v2	start.v2	end.v2	SDR	ori	Size (bp)	ChrID.v1	start.v1	end.v1
scaffold_1208	Chr19X	1	85,028	altHap	-	85,028	scaffold_1208	1	85,028
scaffold_1534	Chr19X	95,029	169,954	altHap	-	74,926	scaffold_1534	1	74,926
scaffold_43	Chr19X	179,955	543,096	altHap	+	363,142	scaffold_43	1	363,142

Table 4.6 Four automonous LTR identified in the Y-SDR

CHR	**LTR-ID**	**superFamily**	**SITE count**	**Substitution Rate (SE)**	**element start**	**element end**	**l/rLTR length**	**TSD motif**	**Pfam**
Chr19Y	*Ltr-y-a*	Gypsy	162	0.078(0.025)	64,125	69,169	175/162	aaat	Retrotrans_gag-223..315;RVP_2-479..564;RVT_1-724..861
Chr19Y	*Ltr-y-b*	Gypsy	409	0.007(0.004)	90,484	95,727	409/409	tattt	Retrotrans_gag-262..351;RVP_2-512..601;RVT_1-746..906;rve-1255..1363;Chromo-1552..1599
Chr19Y	*Ltr-y-c*	Copia	166	0.045(0.017)	96,610	98,445	166/166	tttc	UBN2_3-153..247;RVT_2-244..308
Chr19Y	*Ltr-y-d*	Copia	294	0.065(0.016)	99,790	104,250	303/295	ttca	DUF4219-126..152;UBN2-204..281;gag_pre-integrs-514..572;rve-587..665;RVT_2-992..1139

Standard errors were obtained by a bootstrap procedure (1000 replicates). The Super family for each LTR retrotransposon was classied based on an online LTR classifier(http://ltrclassifier.ird.fr/LTRclassifier/form.html).

TSD: target site duplication; LTR: long terminal repeat.

Table 4.7 Annotated genes in the SDR on chromosome 19 of current Stettler-14 with their homologous genes in other *P. trichocarpa* genomes.

geneID	V2.chr	V2.start	V2.end	size	V2.ori	Description	annotation in V3 genome	Nisqually V4	dS(S.E.)	dN(S.E.)
Po14v11g055363m	Chr19	52,354	56,656	4,303	+	T-complex protein 1 subunit gamma (TCP-1,CCT3, TRIC5)	Potri.018G138200 ; Potri.T046300	Potriv41g055126m;	0.0737(0.0136)	0.0016(0.0012)
								Potriv41g057391m	0.0600(0.012)	0.0008(0.0008)
Po14v11g055362m	Chr19	59,327	69,212	9,886	+	Chloride channel protein CLC-C	Potri.018G138100 ; Potri.T046200	Potriv41g055125m;	0.0117(0.0044)	0.0206(0.0078)
								Potriv41g057390m	0.033(0.0075)	0.0105(0.0025)
Po14v11g055360m	Chr19	73,031	82,422	9,392	+	similar to DNA (cytosine-5)-methyltransferase AthI (EC 2.1.1.37)	Potri.018G138000 ; Potri.T046100	Potriv41g055122m;	0.0194(0.0041)	0.006(0.0013)
								Potriv41g057386m	0.0158(0.0036)	0.0057(0.0013)
Po14v11g055357m	Chr19	96,006	105,907	9,902	+	Archaeal ATPase (Arch_ATPase) // Leucine rich repeat (LRR_8)	Potri.018G137900 (*)	NA	NA	NA
Po14v11g055355m	Chr19	116,621	118,021	1,401	-	hypothetical protein	Potri.018G137700	Potriv41g055119m;	0(0)	0.0206(0.0078)
								Potriv41g057380m	0(0)	0.0236(0.0084)

Table 4.8 The physical positions of inverted repeats in Chr19Y.

	start	**end**	**size (bp)**	**ARMs**
Chr19Y	23,726	25,349	1,624	ARM-1
Chr19Y	25,381	29,199	3,819	ARM-2
	29,200	31,389	2,190	Spacer-1
Chr19Y	31,390	35,225	3,836	ARM-3
	35,226	37,885	2,660	Spacer-2
Chr19Y	37,886	40,646	2,761	ARM-4a
Chr19Y	40,531	41,958	1,428	ARM-4b

Table 4.9 Fragments of Response regulator genes in inverted arms compared to the complete paralogous gene Po14v11g057342m.

Po14v11g057342m (Chr19)	**size (bp)**	**ARM-1**	**ARM-2**	**ARM-3**	**ARM-4a**	**ARM-4b**
exon1 (3'-UTR)	76	+	-	+	-	+
exon2	139	Absent	-	+	-	Absent
exon3	74	Absent	-	+(trancated)	-	+
exon4	78	Absent	Absent	Absent	-(Spacer)	Absent
exon5	71	Absent	Absent	Absent	Absent	Absent
exon6 (5'-UTR)	373	Absent	Absent	Absent	Absent	Absent

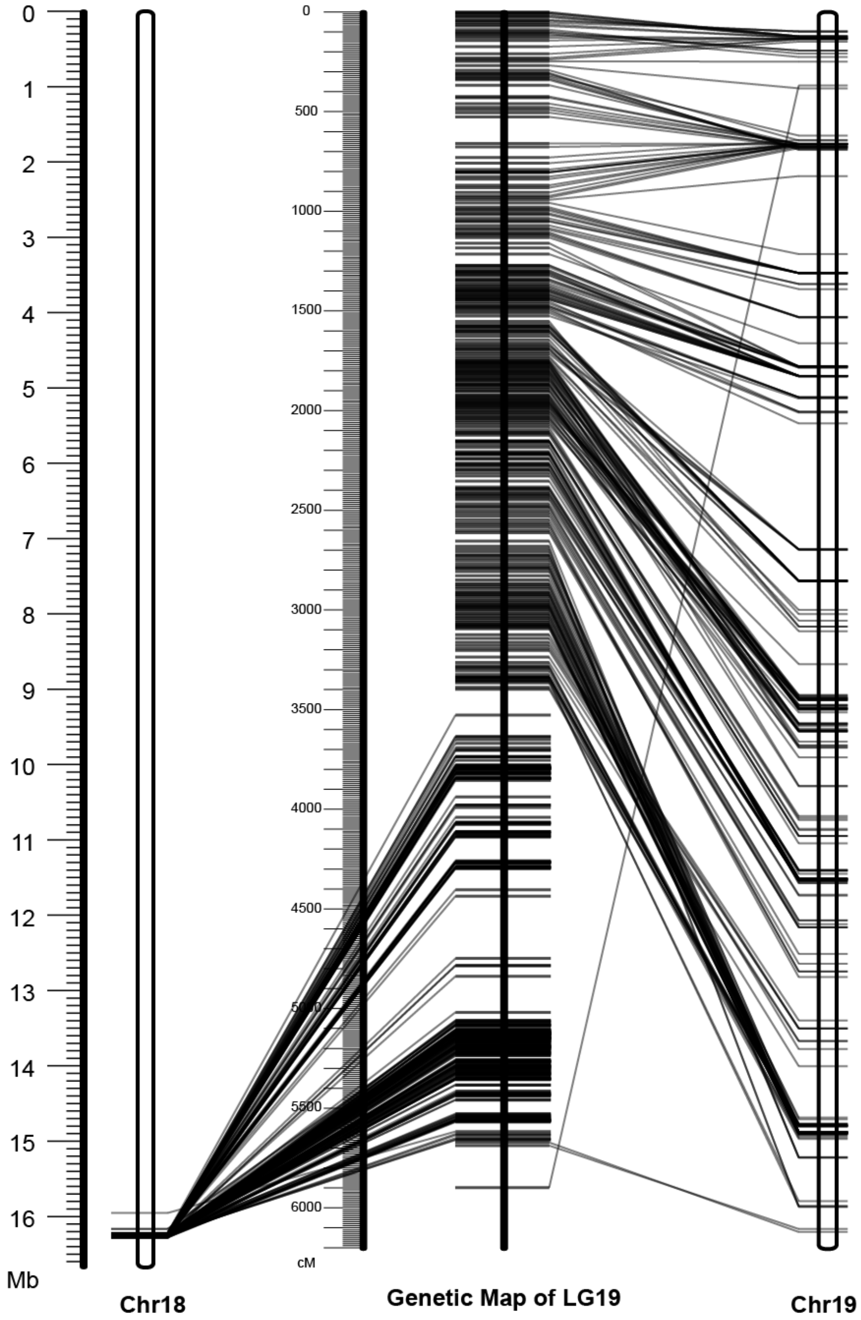


Figure 4.1 The comparisons between the genetic map of chromosome 19 and the physical assemblies of chromosome 18 and 19 in Stettler-14 V1. The physical assemblies of chromosome 18 and 19 are on each side with unfilled rectangles, and the built genetic map of chromosome 19 was shown in the middle. Each horizontal tick represents a genetic marker and its physical position and genetic position is connected with a line.

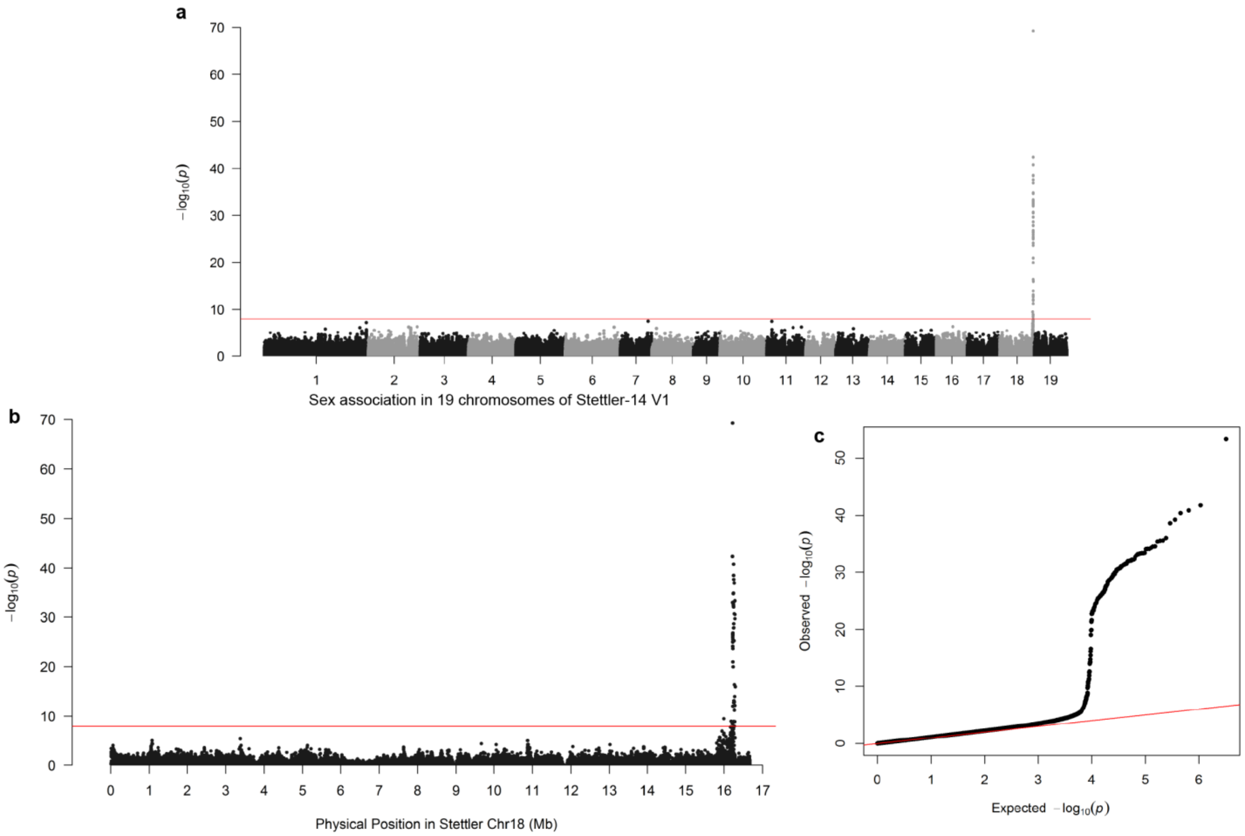


Figure 4.2 Sex association analysis with markers from the V1 main genome of Stettler-14. a. Manhattan plot of *P*-values from sex-association analysis with 200 individuals in 19 chromosomes. **b.** A close look at the sex-associated markers on chromosome 18. The red line indicates the Bonferroni cutoff (1.09×10^{-8}) in a and c. **c.** A quantile-quantile plot of the *P*-values from the association analysis. For convenience, plotted markers are a subset of the original dataset.

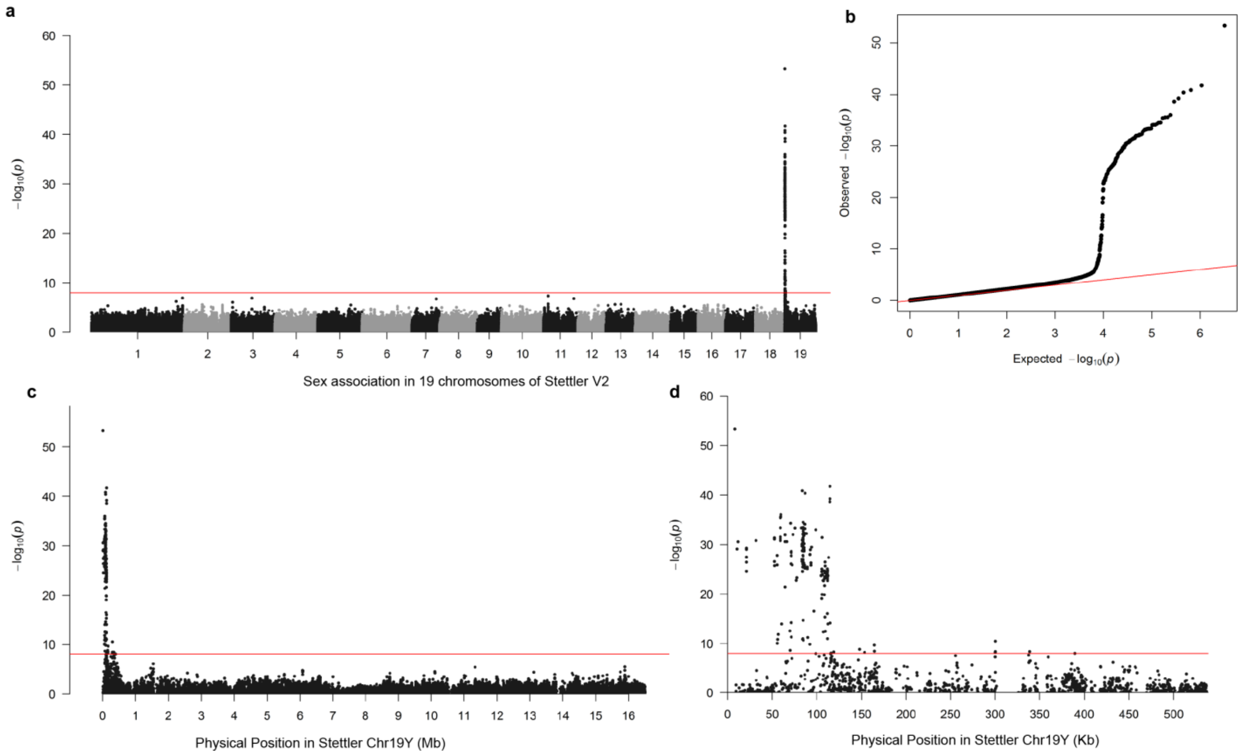


Figure 4.3 Sex association analysis with markers from the main genome. a. Manhattan plot of *P*-values from sex-association analysis with 200 individuals in 19 chromosomes, which shows a single clear peak at the end of the new Chr19Y assembly. The red line indicates the Bonferroni cutoff (9.43×10^{-9}) in a, c, and d. **b.** A quantile-quantile plot of the *P*-values the association analysis was displayed. For displaying convenience, plotted markers are a subset of the original dataset. **c.** A close look at the sex-associated markers on chromosome 19. **d.** A further zoom-in at the sex-associated markers on the first contig (Y-linked haplotype) of chromosome 19.

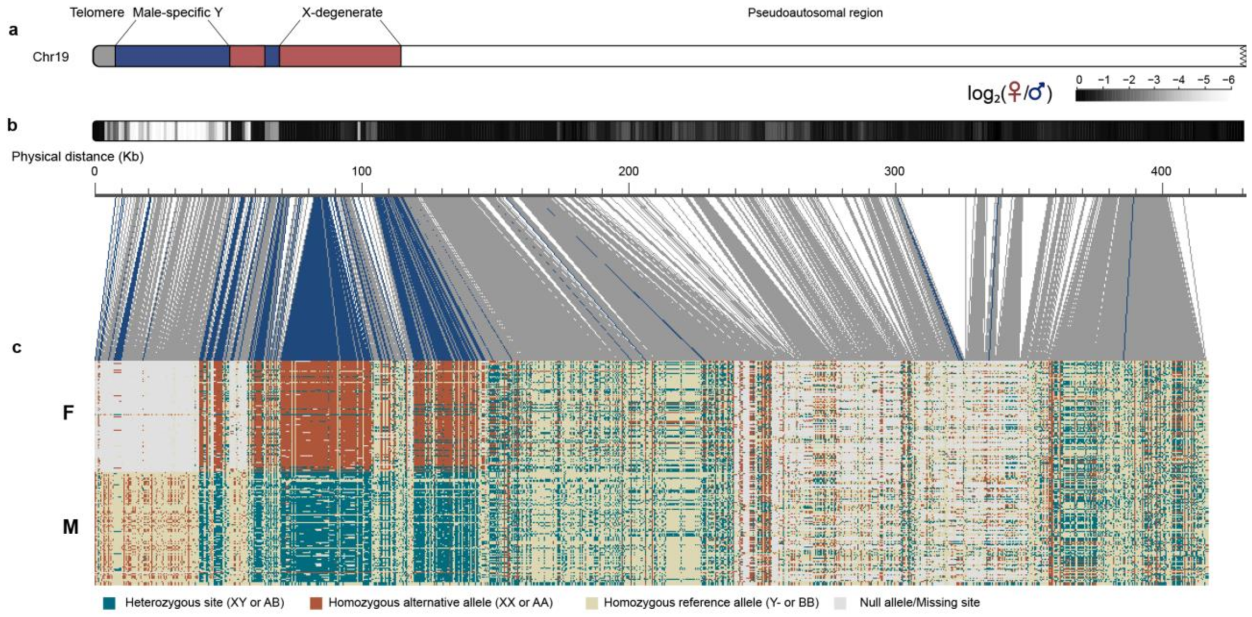


Figure 4.4 Genotype configurations of 200 individuals in the identified sex-linked region. a. The schedule of the identified sex-linked region in chromosome 19 in *P. trichocarpa* Stettler-14. **b.** Distribution of male-biased sequence in the SDR. Each colored block shows the log$_2$ of the ratio of female and male depth in 1 kb windows with X haplotype excluded from the reference. **c.** The genotype configuration of the SDR and a physically linked 300 kb pseudoautosomal region in chromosome 19. The links between **a** and **b** show the physical positions of SNP sites and are highlighted with blue color when the site is sex-associated.

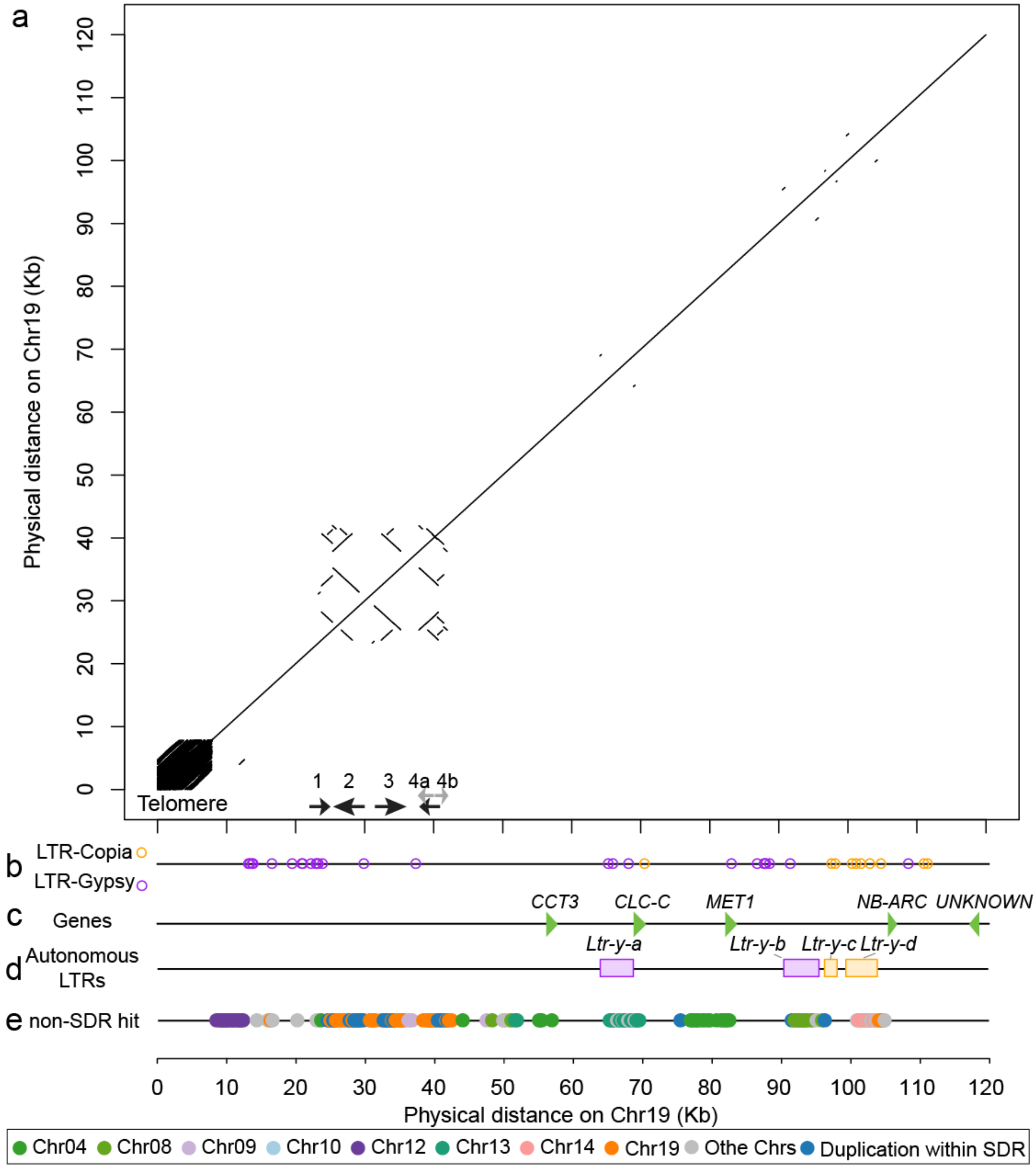


Figure 4.5 Dotplot and landscape of genomic contents of the Y-SDR in *P. trichocarpa*. a. A dotplot of the alignment between the genomic sequence in the SDR in *P. trichocarpa* Stettler-14 to itself. **b.** LTR-Copia and LTR-Gypsy elements identified from RepeatMasker were plotted as circles. **c.** Five genes in the SDR are shown with green triangles. **d**. autonomous LTRs identified

from LTRdigest/LTRharvest are shown with colored rectangles. Purple ones are from LTR-Gypsy superfamily and orange ones are from LTR-Copia superfamily. **e**. Translocations identified in the SDR. Non-SDR hits of 200 bp sequence chunks from SDR. Only single hit was kept.

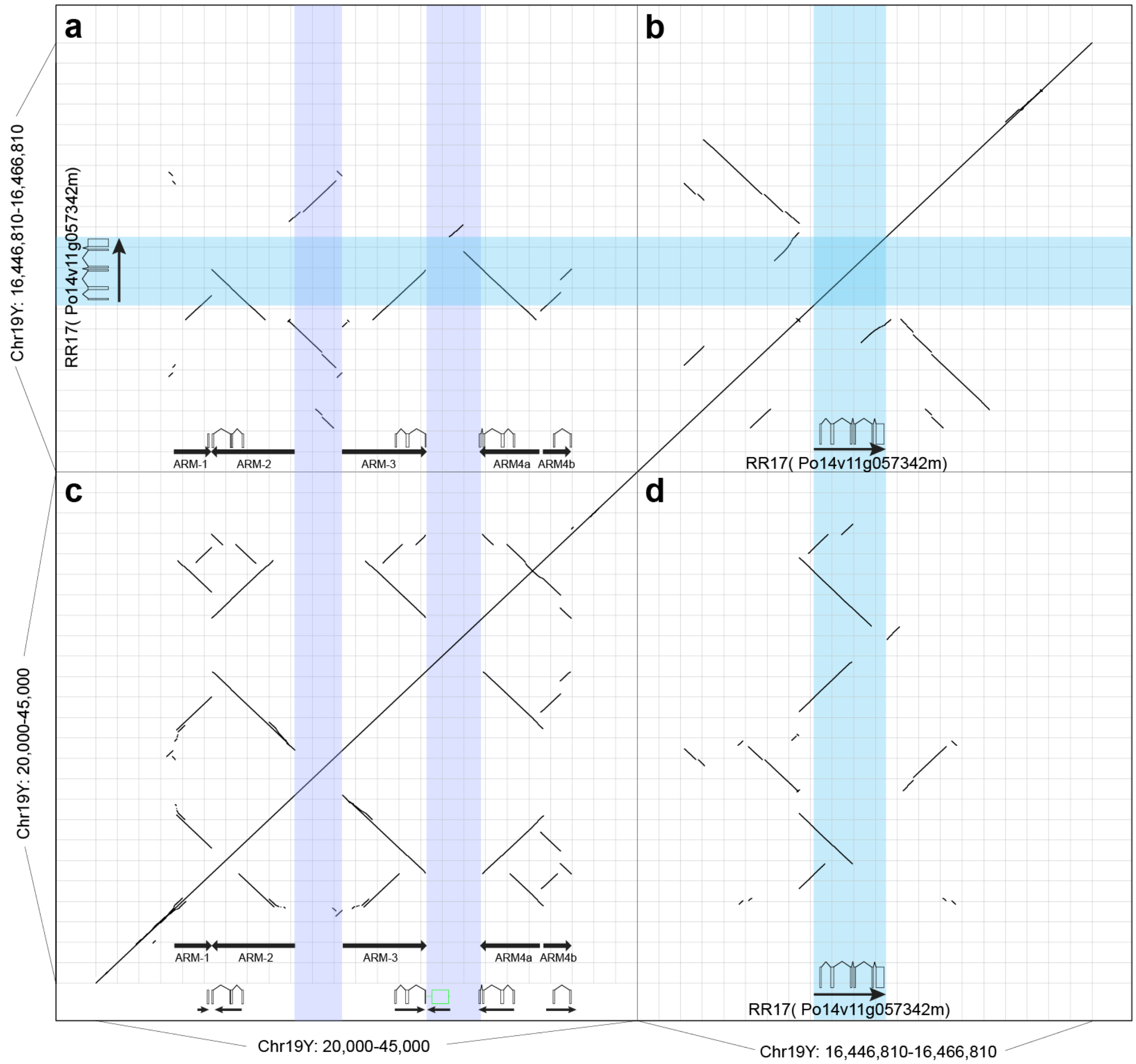


Figure 4.6 Dotplots of the male-specific inverted repeats in the SDR on the Y chromosome and comparison to the closest paralogous gene Po14v11g057342m. a. A dotplot of the alignment between male-specific inverted repeats in the SDR in *P. trichocarpa* Stettler-14 to the region of RR17 Po14v11g057342m on chromosome 19. **b.** a dotplot of the self-alignment from the genomic sequence from 16,446,810 to 16,466,810 on chromosome 19 where RR17 is. **c.** a dotplot of the self-alignment between the genomic sequence from 20 kb to 45 kb on chromosome 19 where male-specific inverted repeats are. **d**. a mirror image of the dotplot of **a**.

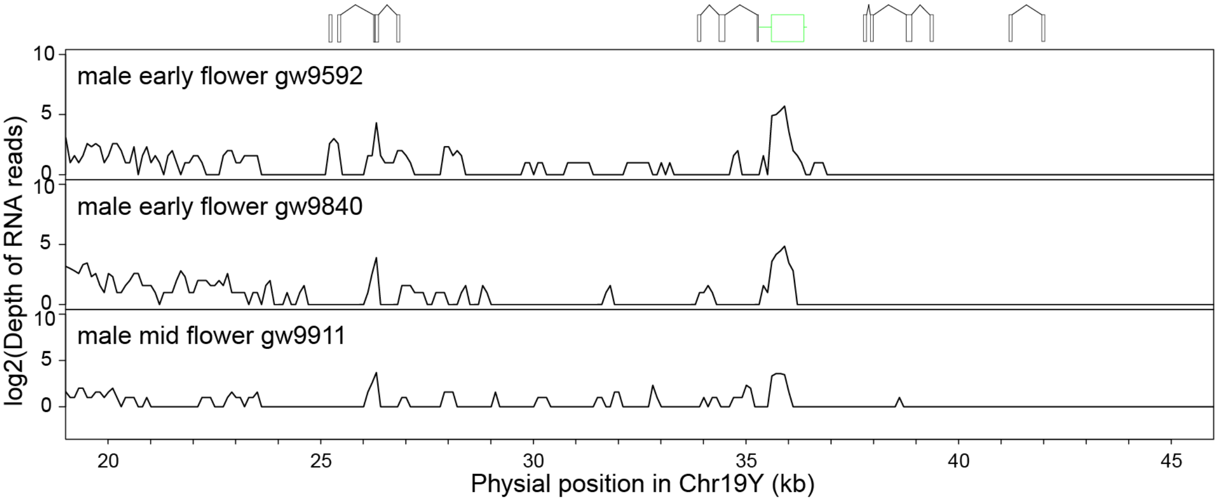


Figure 4.7 RNA-seq reads depth in the inverted repeats. The expression of fragments of the response regulator gene in the male-specific invert repeats was quantified by logarithmic of counts of RNA-seq reads in three male individuals sampled from different flowering stages. Fragments of gene models were displayed to help visualization, where the green box is a fragment of ATHEMA.

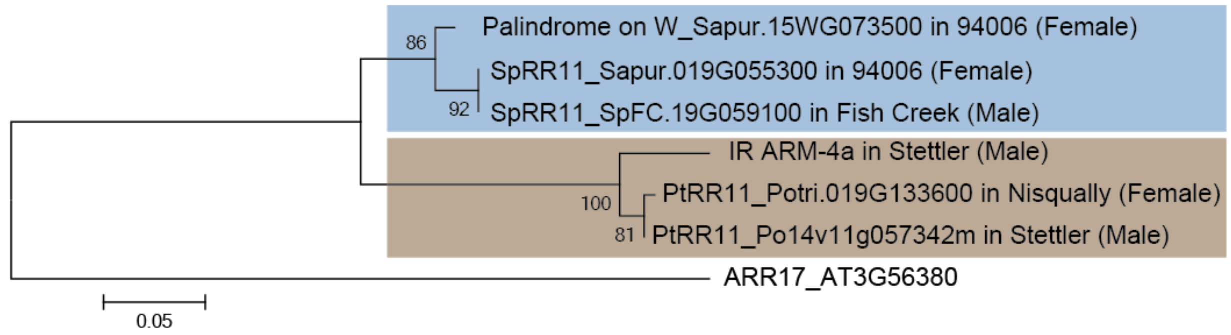

86
Palindrome on W_Sapur.15WG073500 in 94006 (Female)
SpRR11_Sapur.019G055300 in 94006 (Female)
92
SpRR11_SpFC.19G059100 in Fish Creek (Male)
IR ARM-4a in Stettler (Male)
100
PtRR11_Potri.019G133600 in Nisqually (Female)
81
PtRR11_Po14v11g057342m in Stettler (Male)
ARR17_AT3G56380
0.05

Figure 4.8 Phylogenetic relationship of homologous *PtRR11*. A neighbor-joining tree was constructed based on the aligned coding sequences of homologous PtRR11 in *P. trichocarpa* and *S. purpurea*. The coding sequence of *At3g56380* in *Arabidopsis thaliana* was used as an outgroup. Branch length represents the substitution rates and bootstrap values were estimated with 1000 replicates in MEGA.

Discussion

Determining the ages and sizes of the SDR in non-model species is difficult, even with genome sequencing (Charlesworth 2016). Here, we showed that the SDR in *P. trichocarpa* is quite small at approximately 115 kb, as previously claimed by Geraldes et al. 2015. Our improved assembly coupled with estimation of depth of coverage across the genome shows that the male-specific region is at least 40 kb, which is longer than four small male-specific contigs with an average length 1,877 bp from the previous study (Geraldes et al. 2015). Such as a small size of the SDR may simply reflect a recent origin of the SDR in *P. trichocarpa*: insufficient time has elapsed to allow for the expansion in this region (Charlesworth 2013). Nevertheless, despite the small size of the SDR, if the male-female sequence differences in the SDR were shared across several related species, the age of the genetic sex determination might be old (Charlesworth 2013). Chromosome 19 shows an overall higher proportion of repetitive elements than other chromosomes (Table 4), which indicates its unusual genome dynamics compared to the other chromosomes. Thus far, the exact sizes of the SDRs in other *Populus* species are still unknown due to lack of a reliable Y chromosome sequence. Instead, the size of the SDR has thus far been estimated using map-based methods (Pakull et al. 2011, 2014; Kersten et al. 2014). Thus, more data in other *Populus* species is required to confirm the age of the sex-determination loci.

In the previous analysis of sex association in *P. trichocarpa*, *POPTR_0009s08410* (*AtHEMA1*) on Chr09, and *POPTR_0019s15410* (*ARR17*) were found in the regions significantly associated with sex (Geraldes et al. 2015). The authors suspected the assembly of the genome could be erroneous given the inconsistent locations of the association signals. In contrast with previous sex-linked signals over multiple chromosomes in the genome, the signals of sex-linked markers in our studies are well clustered within a 115 kb region. Using the complete assembly and annotation of the SDR region of Stettler-14, we showed that neither of these genes is actually associated with sex. Instead, transposed fragments of these genes are located in the SDR, thereby causing a false signal when the X chromosome is used as a reference genome. This is a common problem for SDRs that contain sex-specific sequence, when the homogametic sex is used as a reference genome (chapter 3) . We have previously shown that the SDR of *S. purpurea* also contains abundant sequences transposed from autosomes (Zhou et al. 2019). Unfortunately, we could not identify reliable recent insertions of non-autonomous LTRs into the male-specific region in *P. trichocarpa* as we did for the female-specific region in *S. purpurea*, thus we could not evaluate if these transpositions are related to LTR movements.

The Y-SDR in *P. trichocarpa* is different from the W-SDR in *S. purpurea* from several perspectives. The large size of the W-SDR was shown to be related to the accumulation of repetitive elements (chapter 3). Also, the number of genes in the X-degenerate regions is different in the two species due to their dramatically different sizes. There are 156 Z-W homologous genes in the W-SDR of *S. purpurea* but only 5 X-Y homologous genes in the *P. trichocarpa* Y-SDR. None of these genes were orthologous. By estimating the synonymous substitution rates of four gene pairs between the Y-SDR and the X-haplotype, we showed that the divergence after the arrest of recombination between X and Y haplotypes was likely to have begun after the split of *S. purpurea* and *P. trichocarpa* (Zhou et al. 2019). This again indicates that the age of the SDR might be young in both species, but further evidence from related species is needed to confirm this. Despite the differences between the Y-SDR and W-SDR, we discovered that a very similar sequence feature is present in the Y-SDR, which is the cluster of inverted repeats (IRs).

Indeed, the male-specific region is mostly composed of a cluster of homologous IR that could be a result of transposition to the SDR followed by several duplications. Similar genomic structures, large identical IRs, (palindromes) also have been observed in the female-specific region of the SDR in *S. purpurea* (chapter 3). Both homologous IRs in the SDRs of the Y or W chromosome in *P. trichocarpa* and *S. purpurea* are essentially derived from genomic duplication. The differences between them in the two species are striking. The four homologous arms that form palindromes in *S. purpurea* are mostly identical due to gene conversion within sequence identity above 99.5%. In contrast, the IRs found in the Y-SDR in *P. trichocarpa* show markedly lower sequence identity ranging from 90% to 95% between arms. The size of the homologous arms in *S. purpurea* is about 20 kb, with only a large (~ 7kb) deletion on one of the arms (Zhou et al. 2019). In contrast, the size of the IR arms in *P. trichocarpa* is no more than 3.8 kb. These homologous IR arms also contain incomplete fragments from only one gene family, while homologous arms of the palindrome in *S. purpurea* contain four copies from five gene families, and additional copies of other genes in the degenerated palindrome arms (Zhou et al. 2019). These differences indicate that the evolution and functions of the SDRs in the two species might be different. These IRs in *P. trichocarpa* are unlikely to play the same function as the ones from palindromes in *S. purpurea*.

Coincidentally enough, a set of IRs that are homologous to a response regulator gene, a possible female-promoting gene (Zhou et al. 2019), is present in the male-specific region in the SDR of *P. trichocarpa*. Homologous arms in the palindrome of *S. purpurea* also contain four nearly identical copies of this cytokinin response regulator gene. Recently, a Type-C response regulator, SyGI was shown to acquire a gynoecium-specific expression after the *Actinidia*-specific duplication event (Akagi et al. 2018). Previous analysis of the methylation in the female reference showed that a response regulator gene (*PtRR11/9*, *Potri.019G058900*) was the only gene in the *P. balsamifera* genome that showed clear sex-specific methylation differences through its promoter and gene body (Bräutigam et al. 2017). This gene is also associated with sex in other *Populus* species (Chefdor et al. 2018; Melnikova et al. 2019). So how is a female-promoting gene turned into a gynoecium-suppressor that is present in the male-specific region of Y-SDR?

Given the loss of the ability to encode a complete protein of these IRs and independent translocation to the Y-SDR, we suspect that the function of these IRs might be different from those highly identical arms in the *S. purpurea* palindrome. Gene silencing induced by a dsRNA species in posttranscriptional gene silencing (PTGS) is observed more frequently in inverted repeat transgenes than in direct repeats in several transgenic experiments in plants (Tijsterman et al. 2002). In *A. thaliana*, an inverted duplication of a target gene can create microRNAs that facilitate site-specific cleavage or translational repression of the targets (Allen et al. 2004). The known methods for achieving PTGS could be through either RNA-directed DNA methylation (RdDM) or through processed antisense siRNAs guiding sequence-specific degradation of complementary mRNAs (Aufsatz et al. 2002). Meanwhile, a methyltransferase gene (Potriv41g057386m) is present in the X-degenerate region. Furthermore, exon1 of PtRR11/9 which does not contain any coding sequences were retained in all IRs arms. This would not simply be a coincidence of finding the IRs that are targeting the homologous PtRR11/9 and a methyltransferase gene in the SDR of *P. trichocarpa*. These findings suggest that these IRs might function as a template for regulatory RNAs along with *MET1,* through either the DNA methylation or short interfering RNAs, leading to the silencing of the female-promoting gene *PtRR11/9* outside the SDR and suppressing the development of female reproductive organs. Here, we hypothesize that the dosage of the *PtRR11/9, Potri.019G058900* is crucial to the female function, which might be the same case for the orthologous gene *Sapur.019G055300* in *S. purpurea*. But the distinctive fate of the duplicates after the translocation into the Y-SDR in *P. trichocarpa* is that these inverted repeats might have become templates for regulatory RNAs that could reduce the dosage of the homologous gene, instead of maintaining the function of the original copy. On the contrary, the palindrome arms in *S. purpurea* might still maintain the original copy function but with a selection favored dosage effects in females. Large and nearly identical (>99%) IRs have been found in both X and Y chromosomes in humans, mouse and several other mammals (Hobza et al. 2017; Trombetta & Cruciani 2017). The recent finding of large homologous IRs in the W chromosome and short IRs in this study shed light on their important role in the sex determination in plants as well. Additional data and analysis from other species in Salicaceae are required to test our hypothesis about these inverted repeats in plant SDRs.

References

Akagi T et al. 2018. A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. Plant Cell 30:780–795.

Allen E et al. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat. Genet. 36:1282–1290.

Aufsatz W, Mette MF, van der Winden J, Matzke AJM, Matzke M. 2002. RNA-directed DNA methylation in &lt;em&gt;Arabidopsis&lt;/em&gt; Proc. Natl. Acad. Sci. 99:16499 LP – 16506.

Van der Auwera GA et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinforma. 43:11.10.1-11.10.33.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. 14:113–24.

Balounova V et al. 2019. Evolution of sex determination and heterogamety changes in section *Otites* of the genus *Silene*. Sci. Rep. 9:1045.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

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

Bräutigam K et al. 2017. Sexual epigenetics: gender-specific methylation of a gene in the sex determining region of *Populus balsamifera*. Sci. Rep. 7:45388.

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

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

Charlesworth D. 2016. Plant Sex Chromosomes. Annu. Rev. Plant Biol. 67:397–420.

Chefdor F et al. 2018. Highlighting type A RRs as potential regulators of the dkHK1 multi-step phosphorelay pathway in *Populus*. Plant Sci. 277:68–78.

Cronk QCB, Needham I, Rudall PJ. 2015. Evolution of Catkins: Inflorescence Morphology of Selected Salicaceae in an Evolutionary and Developmental Context. Front. Plant Sci. 6:1030.

Darwin C. 1859. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. London : John Murray, 1859 https://search.library.wisc.edu/catalog/9934839413602122.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.

Geraldes A et al. 2015. Recent y chromosome divergence despite ancient origin of dioecy in poplars (Populus). Mol. Ecol. 24:3243–3256.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University.

Henry IM, Akagi T, Tao R, Comai L. 2018. One Hundred Ways to Invent the Sexes: Theoretical and Observed Paths to Dioecy in Plants. Annu. Rev. Plant Biol. 69:553–575.

Hobza R et al. 2017. Impact of repetitive elements on the Y chromosome formation in plants. Genes (Basel). 8.

Hofmeister BT et al. 2019. The somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. bioRxiv 862623.

Hou J et al. 2015. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci. Rep. 5:e9076.

Kersten B, Pakull B, Groppe K, Lueneburg J, Fladung M. 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. (Stuttg). 16:411–418.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37:907–915.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Margarido GR, Souza AP, Garcia AA. 2007. OneMap: software for genetic mapping in outcrossing species. Hereditas 144:78–79.

Melnikova N V. et al. 2019. Sex-specific polymorphism of MET1 and ARR17 genes in *Populus × sibirica*. Biochimie 162:26–32.

Moore RC, Harkess AE, Weingartner LA. 2016. How to be a seXY plant model: A holistic view of sex-chromosome research. Am. J. Bot. 103:1379–1382.

Pakull B et al. 2011. Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a Populus tremula × Populus tremuloides cross. Can. J. For. Res. 41:245–253.

Pakull B, Kersten B, Lüneburg J, Fladung M. 2014. A simple PCR-based marker to determine sex in aspen. Plant Biol. 047300:256–261.

Paolucci I 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.

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

Purcell S et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575.

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.

Richards EJ, Ausubel FM. 1988. Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. Cell 53:127–136.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7:S10.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 37:7002–7013.

Tamura K et al. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28:2731–2739.

Tang H et al. 2015. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16:3.

Tennessen JA et al. 2018. Repeated translocation of a gene cassette drives sex-chromosome turnover in strawberries. PLOS Biol. 16:e2006062. https://doi.org/10.1371/journal.pbio.2006062.

Tijsterman M, Ketting RF, Plasterk RHA. 2002. The Genetics of RNA Silencing. Annu. Rev. Genet. 36:489–519.

Trombetta B, Cruciani F. 2017. Y chromosome palindromes and gene conversion. Hum. Genet. 136:605–619.

Wang Y et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40:e49–e49.

Westergaard M. 1958. The Mechanism of Sex Determination in Dioecious Flowering Plants. Adv. Genet. 9:217–281.

Wilm A et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zhou R et al. 2019. A Willow Sex Chromosome Reveals Convergent Evolution of Complex Palindromic Repeats. bioRxiv 710046.

Zhou R et al. 2018. Characterization of a large sex determination region in Salix purpurea L. (Salicaceae). Mol. Genet. Genomics 293:1437–1452.

CHAPTER V

GENERAL CONCLUSIONS

Over the last decade, the application of sequencing technology has been very informative in the fields of evolution and genetics. As one of the mysterious, interesting, yet fundamental questions to us, how sex evolves is always one of the hardest questions. Several models have been developed to explain the evolution of separate sexes and sex chromosomes, including models that invoke one locus, as well as models with two loci, models that pass through a gynodioecious intermediate, and models that pass through a monoecious intermediate (Charlesworth 2013; Olson et al. 2017; Henry et al. 2018). The sex-determining region (SDR) is like a “black hole” of genomics, due to its resistance to genetic mapping due to lack of recombination (making mapping impossible) and recalcitrance to short-reads sequence assembly, due to its highly repetitive nature. Although young sex chromosomes are often found in plants, this does not necessarily indicate that the molecular differences are small between X and Y chromosomes. In fact, recent studies in several plant sex chromosomes show that turnovers or transitions could have happened. The important indication from my studies is the lability of the sex determination systems found in the Salicaceae family.

In *S. purpurea*, I used both quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) to investigate the sex-linked or sex-associated regions in the genome assembly (chapter 2). I determined that the SDR is located on chromosome 15 with female heterogamety (ZW/ZZ). This indicates that the location of SDRs in the genome and sex configurations might be well conserved across *S. purpurea*, *S. suchowensis,* and *S. viminalis* (Hou et al. 2015; Pucholt et al. 2015). Female-specific allele drop-out along with analysis of the female-specific coverage in the SDR on the W chromosome suggests that the SDR is large, which had not been confirmed in other reported *Salix* species. This is important because the SDR in *P. trichocarpa* was inferred to be around 100 kb (Geraldes et al. 2015). The result of the identification of the centromere inside the SDR also suggests that the suppressed recombination might have already been in place before the occurrence of the sex-determining gene. This raises the possibility that accepted models of sex chromosome evolution may need to be modified to allow for the pre-existence of suppressed recombination in the region where sex determination genes transpose or evolve (Bachtrog 2013; Charlesworth 2013).

We also showed that the incompleteness of the assembly of the SDR might be the cause for scattered association peaks observed from several chromosomes and scaffolds. In particular, with greatly improved SDRs in chapters 3 and 4, I showed that the single association peak in the sex-association analysis only appears when the reference contains well-assembled Y (or W) haplotypes. When the reference is XX or ZZ, such as Nisqually-1 in *P. trichocarpa*, the reads from XY (or ZW) individuals will be aligned the locus where they match best in the reference genome. If the SDR is missing from the reference, the reads are usually aligned to the best paralogous regions in the autosomes, which results in the association peak being distributed across the genome with a main peak in the SDR. This might become severe when the translocation from autosomes to the SDR occurred recently because this isn't enough time for degenerations of the Y (or W) to proceed for them to be distinguished from the autosomal paralogs.

Beyond improving the completeness and contiguity of the W and Z chromosome assembly in *S. purpurea*, the finding of a large palindromic structure in a plant sex chromosome for the first time and related genetic analysis in chapter 3 added to our knowledge of the complex but interesting features in the sex chromosomes. In accordance with the original finding of large palindromes in human Y chromosomes (Skaletsky et al. 2003), I found that the palindrome in *S. purpurea* shares a very similar structure as the one in the human Y chromosome. Additionally, I presented evidence to show that the palindrome in *S. purpurea* could still be under gene conversion, which was also shown among palindromic arms in the human Y by Rozen et al. (2003) via comparing the human and chimpanzee Y. These results raise intriguing questions about why large palindromes are maintained in sex chromosomes. One reason is that palindromes allow intrachromosomal gene conversion that can eliminate deleterious mutations, or propagate beneficial mutations as mechanisms to protect against Y degeneration (Betrán et al. 2012; Hobza et al. 2017). Insertions of LTR retrotransposons in the degenerated arms of palindrome W.P2 also contrast to the insertions found in arms undergoing gene conversion. This reveals the highly active transposable elements were accumulated in the non-recombining regions, which is different from the almost homogenous components in arms undergoing gene conversion. This suggests that palindromes might be favored as mechanisms to remedy Y degeneration.

In contrast to the substantial proportion of mammalian sex chromosomes occupied by palindromes, palindromes occupy only a small portion of the SDR on the W chromosome of *S. purpurea*. Gene content in the palindrome also reveals that female-specific regions are impacted not only by the degeneration of the proto sex chromosomes but also by translocations from the autosomes. Genes in the palindromes mostly have autosomal paralogs in the genome (chapter 3). Given the dynamic genomic environment modulated by the transposable elements in plants, the lability of the sex determination systems might be a synergistic outcome of the combined effects of transposable elements and selection.

With several intriguing results in chapter 3, it is natural to investigate if these features are shared in a closely related sister species, *P. trichocarpa*. Although a large SDR is found in *S. purpurea*, the size of the SDR on Y in *P. trichocarpa* is as small as about 115 kb. The analysis of the synonymous substitution rate (d_S) between the X and Y alleles suggests that the age of the SDR is younger than the divergence of two genera. Neither the same sex chromosome nor the same heterogamety is shared between these two species. In *S. purpurea*, we confirmed the chromosome 15 is the sex chromosome with a female heterogametic system (ZW/ZZ), whereas chromosome 19 is the sex chromosome with a male heterogametic system (XX/XY) in *P. trichocarpa*. Given highly syntenic genomes in these two species (chapter 2), this prompted us to look for the labile sex determination systems for the family Salicaceae. With an improved assembly of chromosome 19, we found a cluster of homologous inverted repeats (IRs) in the SDR of the Y chromosome in *P. trichocarpa*. In contrast to the one we found in the *S. purpurea*, both the spanning size of these IRs and the lengths of arms are smaller in *P. trichocarpa*. The identities among the arms from the IRs are also lower than the ones from *S. purpurea*, which suggests a lack of gene conversion between those IRs in the SDR of *P. trichocarpa*. However, the finding of a shared homologous gene family, Arabidopsis *response regulator 17* (*ARR17*) in both IRs and palindromes raises the intriguing possibility about shared sex-determining mechanisms and questions related to transitions between female and male heterogamety in the family Salicaceae.

The *ARR17* gene plays a key role in the cytokinin signaling pathway, which is crucial for the development of the reproductive organs in plants (To & Kieber 2008; Hwang et al. 2012; Kieber & Schaller 2018). The effect of cytokinin on sex seems to be labile among species (Louis et al. 1990; Bracale et al. 1991). Recently, a type-C cytokinin response regulator was identified as a potential sex-determining gene in the Y-specific region in the genus *Actinidia* (kiwifruits) (Akagi et al. 2018). Four nearly identical copies of *ARR17* orthologs have been found in the W palindromes of *S. purpurea*. On the contrary, degenerated fragments of incomplete *ARR17* orthologs occur as inverted repeats in the Y-SDR of *P. trichocarpa*. Further phylogenetic analysis showed that these additional copies in both SDRs are independent or lineage-specific duplications. Several results from previous studies showed that *ARR17* homolog (*PtRR11/9*) in *Populus* had sex-specific patterns in expression and methylation (Ramírez-Carvajal et al. 2008; Bräutigam et al. 2017; Melnikova et al. 2019), we propose a model to explain the lability of the sex-determining mechanisms observed: 1) a gene from a pathway of producing or sensing one of the cytokinin that could regulate the development of reproductive organs could be a sex-determining gene; 2) the development of reproductive organs is strongly associated with either the dosage of this gene; 3) in *S. purpurea*, four complete genes in the W-SDR might be selectively advantageous in females, which might have retained the same function as the homologous *PtRR11/9*. Instead, the incomplete fragments in the IRs can be transcribed into regulatory RNAs that target the homologous *PtRR11/9* in *P. trichocarpa*. This means that regulation on the expression of *ARR17* homolog in both species might be the key to the sex-determining mechanism. When the expression of *ARR17* homolog is high, the pathway suppresses the male development, leading to female flowers. When the expression of *ARR17* homolog is silenced, the pathway promotes the male development without the suppression from *ARR17* homolog. This requires us to assume that *ARR17* homolog is necessary for the development of normal female organs in these two species. Additionally, the pathway with *ARR17* homolog silenced is supposed to carry female-suppressing function based on this model. The finding of unusual gene duplications in the SDRs suggests that gene duplications in plant SDRs are worth careful examination in future studies. The fate of these duplicates is particularly interesting under a scenario without homologous recombination between the X- and Y-chromosomes. For example, a retrotransposed copy could form a chimeric gene with a new function when it is integrated with an existing gene (Innan & Kondrashov 2010).

With the results from chapters 2 through 4, more important questions emerge to be answered in the future. These questions include: what are the main reasons for the different sizes of the SDRs in *S. purpurea* and *P. trichocarpa* if the SDRs are young in both genera? Additional data from other related species will be required to confirm the age of this small SDR in *P. trichocarpa*. With more SDRs being described in other related species, the evolution of the sex-determining system in the family should be revealed in a higher resolution. With these data, I expect that the transition between female and male heterogamety can be tested with our hypothesis about the female- and male- promoting genes. Questions like if those transitions are homologous or non-homologous will be addressed from the identification of SDRs in each species. The identification of inverted repeats in *S. purpurea* and *P. trichocarpa* raises the requirement of thoroughly searching these genomic structures in other species, which might help us understand their importance in the SDR. Thus, our results provide the first description of the SDR directly from a genome assembly. In my studies, we identified important sex-linked markers and described the genomic features of SDRs in detail. We believe that these findings will benefit future studies on the evolution of sex chromosomes as well as the understanding of the sex-determination systems in plants.

References

Akagi T et al. 2018. A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. Plant Cell 30:780–795.

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

Betrán E, Demuth JP, Williford A. 2012. Why Chromosome Palindromes? Int. J. Evol. Biol. 2012:207958.

Bracale M et al. 1991. Sex determination and differentiation in Asparagus officinalis L. Plant Sci. 80:67–77.

Bräutigam K et al. 2017. Sexual epigenetics: gender-specific methylation of a gene in the sex determining region of *Populus balsamifera*. Sci. Rep. 7:45388.

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

Geraldes A et al. 2015. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). Mol. Ecol.

Henry IM, Akagi T, Tao R, Comai L. 2018. One Hundred Ways to Invent the Sexes: Theoretical and Observed Paths to Dioecy in Plants. Annu. Rev. Plant Biol. 69:553–575.

Hobza R et al. 2017. Impact of repetitive elements on the Y chromosome formation in plants. Genes (Basel). 8.

Hou J et al. 2015. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci. Rep. 5:e9076.

Hwang I, Sheen J, Müller B. 2012. Cytokinin Signaling Networks. Annu. Rev. Plant Biol.

63:353–380.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11:97–108.

Kieber JJ, Schaller GE. 2018. Cytokinin signaling in plant development. Development 145:dev149344.

Louis J-P, Augur C, Teller G. 1990. Cytokinins and Differentiation Processes in &lt;em&gt;Mercurialis annua&lt;/em&gt; Plant Physiol. 94:1535 LP – 1541.

Melnikova N V. et al. 2019. Sex-specific polymorphism of MET1 and ARR17 genes in *Populus × sibirica*. Biochimie 162:26–32.

Olson MS, Hamrick JL, Moore R. 2017. Breeding Systems, Mating Systems, and Genomics of Gender Determination in Angiosperm Trees. In: Groover, A & Cronk, Q, editors. Springer International Publishing, Switzerland, pp. 139–158.

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

Ramírez-Carvajal G a, Morse AM, Davis JM. 2008. Transcript profiles of the cytokinin response regulator gene family in *Populus*. New Phytol. 177:77–89.

Rozen S et al. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876.

Skaletsky H et al. 2003. The male-specific region of the human Y chromosome is a mosic of discrete sequence classes. Nature 423:825–837.

To JPC, Kieber JJ. 2008. Cytokinin signaling: two-components and more. Trends Plant Sci.

13:85–92.