A Willow Sex Chromosome Reveals Convergent Evolution of Complex Palindromic Repeats

Ran Zhou[1], David Macaya-Sanz[1], Craig H. Carlson[2], Jeremy Schmutz[3,4], Jerry W. Jenkins[3], David Kudrna[5], Aditi Sharma[4], Laura Sandor[4], Shengqiang Shu[4], Kerrie Barry[4], Gerald A. Tuskan[6,7], Tao Ma[8], Jianquan Liu[8,9], Matthew Olson[10], Lawrence B. Smart[2], Stephen P. DiFazio[1,11]

[1]Department of Biology, West Virginia University, Morgantown, WV 26506-6057

[2]Horticulture Section, School of Integrative Plant Science, Cornell University, New York State Agricultural Experiment Station, Geneva, NY 14456

[3]HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, USA

[4]Department of Energy Joint Genome Institute, Walnut Creek, California, USA

[5]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ, USA

[6] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[7]DOE-Center for Bioenergy Innovation (CBI), Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[8]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China

[9]State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology & College of Life Sciences, Lanzhou University, Lanzhou 730000, China

[10]Department of Biological Sciences, Texas Tech University, Box 43131, Lubbock, TX 79409-3131, USA

[11]Corresponding Author: spdifazio@mail.wvu.edu

R.Z.: razhou@mix.wvu.edu; D.M.-S.: david.macayasanz@mail.wvu.edu; C.H.C.: chc89@cornell.edu; J.S.: jschmutz@hudsonalpha.org; J.W.J.: jjenkins@hudsonalpha.org; D.K.: dkudrna@email.arizona.edu; A.S.: AditiSharma@lbl.gov; L.S.: LCSandor@lbl.gov; S.S.: SQShu@lbl.gov; K.B.: kwbarry@lbl.gov; G.A.T: tuskanga@ornl.gov; T.M.: matao.yz@gmail.com; J.L.: liujq@nwipb.cas.cn; M.O.: matt.olson@ttu.edu; L.B.S.: lbs33@cornell.edu

Abstract

Background

Sex chromosomes in a wide variety of species share common characteristics, including the presence of suppressed recombination surrounding sex determination loci. They have arisen independently in numerous lineages, providing a conclusive example of convergent evolution. Mammalian sex chromosomes contain multiple palindromic repeats across the non-recombining region that facilitate sequence conservation through gene conversion, and contain genes that are crucial for sexual reproduction. Plant sex chromosomes are less well understood, and in particular it is not clear how coding sequence conservation is maintained 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*. Two consecutive palindromes span over a region of 200 kb, with conspicuous 20 kb stretches of highly conserved sequences among the four arms. The closely-related species *S. suchowensis* also has two copies of a portion of the palindrome arm and provides strong evidence for gene conversion. Four genes in the palindrome are homologous to genes in the SDR 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 palindromic structures 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.

Keywords

Sex, gene conversion, W chromosome, palindrome, genome, *Salix*

Introduction

Sex chromosomes are fascinating not only because they contain the master regulatory genes that confer sex-specific traits, but also because of their unique evolutionary history [1]. 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 [1]. 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 [2,3]. Consequently, ancient sex chromosomes are difficult to sequence because they are often highly heterochromatic and have a large amount of repetitive and ampliconic DNA [1,4].

A striking characteristic of animal Y chromosomes is the presence of large palindromes in Y-ampliconic regions consisting of large inverted repeats with highly identical sequences that are undergoing gene conversion [5]. These palindromes are enriched for genes that are expressed in the testes and are important for male sexual development in mammals [4,6–8]. 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 [9]. However, such structures have not yet been described in plants.

Plant sex chromosomes have been identified in approximately 40 species but fewer than half of them are heteromorphic (readily discernable by cytology) [10,11]. In plants, male-specific Ys (MSYs) have been studied in papaya and persimmon. Both of these contain a female suppressor on the Y chromosome [12–14]. Recently, a female suppressing gene in asparagus has been identified on the Y chromosome using long-read sequencing technology with optical mapping [15]. Another study on octoploid strawberry found repeated translocations of a female-specific gene cassette [16]. RNA-seq was applied to study the switches in heterogamety in two *Silene* species as well [17]. These dynamic sexual systems provide an excellent opportunity to study the drivers of sex chromosome evolution.

Sex determination is similarly diverse within the Salicaceae family. SDRs have been consistently found on chromosome 15 with female heterogamety in multiple *Salix* species [18–20]. 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 [21,22]. Previously, we reported that the SDR occupies a large portion of the W chromosome in *S. purpurea*, because a large region on Chr15 shows lack of recombination and a large number of female-biased alleles [18,23]. This is different from the small SDR observed in *P. trichocarpa* and *P. balsamifera* [22,24]. 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 shared homologous

66 genes occur in the *Salix* and *Populus* SDRs, suggesting a potentially shared underlying
67 mechanism of sex determination in this family.

68 **Results**

69 **Genome assembly**

70 We present here highly contiguous genome assemblies of a female and a male *S.*
71 *purpurea*. The female assembly (94006 v4) consists of 452 contigs with an N50 of 5.1 Mb,
72 covering a cumulative total of 317.1 Mb. Similarly, the male assembly (Fish Creek v3), has 351
73 contigs and an N50 of 5.6 Mb, covering 312.9 Mb (Additional File 1: Table S1). Both assemblies
74 are partially phased in genomic regions where the two haplotypes are divergent. Alternative
75 haplotypes are represented by 421 contigs totaling 72.4 Mb in the female assembly, and 497
76 contigs totaling 149 Mb for the male. Using a genetic map from a large family derived from
77 progeny of the sequenced male genotype, we created assemblies representing the 19
78 chromosomes, containing 108 contigs totaling 288.3 Mb for the female, and 96 contigs totaling
79 288.5 Mb for the male. These represent over 90% of the assembled sequence in both cases,
80 though 344 and 255 contigs remained unplaced by the genetic map for the female and male,
81 respectively (Additional File 1:Table S2).

82 Because we expected the female-specific haplotype to be differentiated from the shared
83 haplotype in the SDR, we anticipated that much of this region would be assembled as separate
84 contigs, corresponding to W and Z haplotypes in the SDR. These can be readily differentiated by
85 examining the relative depth of coverage when aligning male versus female short read sequences
86 against these references. After identifying the location of the SDR based on the presence of sex-
87 biased markers [18], the initial Chromosome 15 assembly appeared to consist of a mix of Z and
88 W scaffolds in the SDR (Additional File 2: Figure S1a). We therefore sought to create a new

89 assembly with Z and W haplotypes assembled to separate chromosomes. To do this we first
90 identified the female-specific contigs (putative W contigs) using sex association and sex-specific
91 depth as criteria. This resulted in identifying 23 contigs that were putatively comprised primarily
92 of female-specific sequence (Additional File 1: Table S3). One scaffold was excluded because it
93 mostly consisted of an alternative haplotype of a longer contig in the PAR of Chr15W.

94 Many of these contigs lacked markers from our intercross map, particularly for those that
95 came from female-specific W (FSW) regions that were absent on the Z chromosome. We
96 therefore created new genetic maps that had a mix of SNP and indel markers that would be more
97 suited to capturing these hemizygous portions of the genome. The new genetic maps converged
98 to 19 major linkage groups representing the 19 chromosomes. The male backcross map
99 contained 8,715 markers, while the female backcross map contained 8,560 markers (Additional
100 File 1: Table S4). We used these to assemble a Z and a W version of Chr15 (Additional File 1:
101 Table S5). Thus, the current assembly (release ver5) contains 20 chromosomes, including
102 Chr15Z and Chr15W. A total of 6.56 Mb (95.7%) of the W-specific contig sequence, contained
103 in 17 contigs, was assembled to Chr15W using these maps. Four putative W scaffolds totaling
104 297 kb in length lacked mapped markers and could not be placed unambiguously.

105 **Location of the SDR**

106 We repeated sex association analysis with our new assembly with Chr15Z removed.
107 Among 54,959 tested SNPs, all 105 significantly sex-associated SNPs were present only on
108 Chr15W (Fig. 1a; Additional File 2: Figure S2a-c), and markers from PARs and other scaffolds
109 in the main genome did not show any sex association (Additional File 2: Figure S2a). The eight
110 top-ranking sex-associated markers were distributed from 7.66 Mb to 8.66 Mb. Sex-associated

111 markers were primarily heterozygous in females and homozygous in males, confirming our
112 previously-reported observation of ZW sex determination in *S. purpurea* [18].

113 **Composition of the sex chromosomes 15W and 15Z**

114 Chr15W is 15.7 Mb in length, composed of 22 contigs placed with the new genetic map.
115 For comparison, Chr15Z is only 13.3 Mb and is comprised of 16 contigs (Additional File 1:
116 Table S5; Figure 1). There are two pseudoautosomal regions (PARs) at each end of Chr15W that
117 are indistinguishable from the corresponding regions on Chr15Z. PAR1 is 2.7 Mb long and is
118 composed of two contigs, and PAR2 is 6.5 Mb and is comprised of three contigs (Fig. 1). These
119 regions are unphased and are therefore identical in the two assemblies.

120 The sex-determining region (SDR) spans from 2.7 Mb to 9.1 Mb on Chr15W and
121 occupies nearly 40% of the chromosome (hereafter referred to as the W-SDR). This region
122 undergoes minimal recombination in the mapping population (Additional File 2: Figure S3).
123 Reexamining sex-specific depth of the W-SDR, it is clear that this region of the genome is
124 mostly phased to separate the male and female haplotypes (Additional File 2: Figure S1b). One
125 contig (ws9) in the W-SDR is clearly chimeric based on the distribution of female-specific
126 sequence (Fig. 1c), and the depth profile (Additional File 2: Figure S1b). However, this contig is
127 only 300 kb in length, so it doesn't change the overall conclusion that the FSW is largely phased.
128 The region corresponding to the W-SDR on Chr15Z is about 4 Mb in length, and only occupies
129 28.2% of the chromosome (hereafter referred to as the Z-SDR). The size discrepancy is primarily
130 due to the FSW regions. Based on the depth ratios, the Z-degenerate regions are about 3.5 Mb
131 and W insertions are about 3.1 Mb in the W-SDR.

132 The W-SDR has lower gene density and higher repeat density than other portions of the
133 genome (Table 1). More specifically, both the W-SDR and the Z-SDR show lower gene density

134 on average than the PARs or other autosomes. Similarly, both the W-SDR and Z-SDR show
135 higher accumulation of Gypsy retrotransposons. Interestingly, Copia-LTRs occur at higher
136 density in the W-SDR region compared to the Z-SDR (15 per 100 kb vs 9 per 100 kb), (Kruskall-
137 Wallis test, P<2.2e-16) (Table 1), suggesting that these inserted following cessation of
138 recombination between these haplotypes.

139 **Gene content of the W chromosome**

140 There are 269 genes in PAR1, 778 genes in PAR2, and 488 genes in the W-SDR. In
141 contrast, the Z-SDR only contains 317 genes (Fig. 2; Additional File 1: Table S6-7). An
142 additional 29 genes are present on scaffold_844, which is likely derived from the Z haplotype,
143 but which lacked genetic markers to properly place it. There were 146 single copy mutual best
144 hits between the W-SDR and Z-SDR, referred to hereafter as Z degenerate genes (Fig. 2). The
145 SDR also contains 47 genes in tandem duplications, while the corresponding tandem repeats in
146 the Z-SDR contain 78 genes. Additionally, the Chr15W SDR contains 65 genes that have mutual
147 best hits on other autosomes, and 33 of these are tandemly duplicated in the SDR. In contrast, the
148 Z-SDR region contains only 14 such genes, only four of which are tandemly duplicated. These
149 putatively translocated genes comprise 13% of the W-SDR and only 4% of the Z-SDR, and
150 account for nearly half of the discrepancy in gene content between the haplotypes. An additional
151 52 genes in the W-SDR had a top hit to other genes in the genome, but the best hit was not
152 mutual, so these are lower confidence candidates for translocations or Z-degenerate genes. The
153 Z-SDR contained 20 such genes. The remaining genes had no significant hits to other genes in
154 the genome, presumably due to loss by deletion, or gaps in the sequence or annotation (125 in
155 the W-SDR and 58 in the Z-SDR).

156 **Z-degenerate Genes and Strata**

157 We used syntenic gene pairs identified through MCScanX between Chr15W and Chr15Z
158 to test if there are strata with different levels of synonymous substitutions (Ks), which would
159 indicate different phases of cessation of recombination [25]. There was little evidence to support
160 the presence of strata based on 151 pairs of Z-degenerate genes (Fig. 3 and Additional File 1:
161 Table S8). The average of Ks was 0.028. For comparison, the Ks between syntenic genes on
162 Chr01 for *S. purpurea* and *S. suchowensis* was 0.027, and the Ks between *S. purpurea* and *P.*
163 *trichocarpa* was 0.156 (Fig. 3).

164 **Translocations to the W-SDR and palindromic repeats**

165 The recently translocated genes are of particular interest because of the insights they can
166 provide into the evolutionary history of the sex chromosome. Among 65 genes putatively
167 translocated from autosomes to the W-SDR, 10 have best hits on Chr19 (manually annotated
168 genes excluded) (Additional File 1: Table S9). Contig ws19 is particularly enriched for
169 translocated genes, and merits a closer examination (Fig. 1). Contig ws19 contains 11
170 translocated genes, including four genes from Chr19 and four genes from Chr17 (Fig. 1). Many
171 of these translocated genes occur in two to four copies on ws19 in striking inverted repeat
172 configurations that are similar to the palindromic repeats that occur on mammalian Y
173 chromosomes (Fig. 4).

174 In *S. purpurea*, this region is female-specific and is composed of two palindromes.
175 Palindrome W.P1 spans about 42.7 Kb with a 2.6 kb spacer in the center, and Palindrome W.P.2
176 is immediately adjacent and spans over 165 kb (Table 2; Fig. 4a). A 20 kb sequence occurs in
177 inverted orientation and shows high sequence identity across the four arms of both palindromes
178 (Table 2; Fig. 5a). In palindrome W.P1 these are referred to as arm1 and arm2, and in
179 Palindrome W.P2 these are referred to as arm3a and arm4a (Table 2; Fig. 4a). Sequence identity

180 among these four arms is greater than 99% on average. The regions of high identity are disrupted
181 by a ~500 bp insertion in the center of arm4. Furthermore, arm3 has a 6.9 kb deletion at 11.7 kb,
182 followed by a stretch of 1.6 kb that can be aligned to the other arms in the same orientation (Fig.
183 5a). Additionally, there is a 12 kb stretch upstream of arm1 that shows high identity to portions
184 of arms 1 and 2. We call this the pre-arm for convenience (Table 2).

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

190 **Gene content of the palindromes**

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

197 The cytokinin response regulator is of particular interest because an ortholog of this gene
198 has also been found to be associated with sex in *Populus* [22], and is therefore an excellent
199 candidate as a master regulator of sex in the Salicaceae. The RR gene is highly conserved across
200 all four palindrome arms on the W-SDR (Fig. 5a,c). Interestingly, we also found a pseudogene
201 copy of the RR gene on the Z-SDR. This is the only one of the five genes that is present in some
202 form on the W-SDR, the Z-SDR, Chr19, and also in the SDR of *Populus*. There is a 2.6 kb

203 sequence inserted upstream of all RR copies in the palindrome, and not in the Z-SDR
204 pseudogene or on Chr19 (Additional File 2: Figure S4). This suggests that the W-SDR
205 palindrome formed after translocation from Chr19. Interestingly, the RR gene also occurs as
206 inverted repeats in all three locations in the genome (W-SDR, Z-SDR, and Chr19). However,
207 alignment of the W-SDR, Z-SDR and Chr19 versions demonstrates that the palindromes likely
208 formed independently, because the palindromic regions are different (Additional File 2: Figure
209 S4).

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

217 **Multiple LTR retrotransposons in the palindrome**

218 To gain further insight into the composition and history of the W-SDR, we used
219 LTRharvest and LTRdigest to annotate LTR retrotransposons in the palindromic region. We
220 identified one LTR retrotransposon in the pre-Arm region and 12 LTR retrotransposons in
221 palindrome W.P2 that have terminal repeats identified with coding regions (Fig.6a). These 13
222 retrotransposons are likely to be independent insertion events given that they have different long
223 terminal repeats as well as different target site duplications and do not occur in the same
224 position in the opposite arm of the palindrome (Additional file 1: Table S10). Given that there
225 are varying numbers of substitutions within the LTRs of the same retrotransposon, it appears that

226 these insertions have occurred repeatedly after establishment of the palindromes. Using a
227 previous estimation of the mutation rate in *P. tremula* (2.5×10^{-9} per year)[26], we estimate that
228 the oldest insertion occurred at least 8.6 ± 2.9 s.d. MYA from a nonautonomous LTR
229 retrotransposon, *Ltr-p2-a* (Fig. 6a and Additional file 1: Table S10). This is likely an
230 underestimate, since the *Salix* substitution rate is substantially higher than that of *Populus* [27].
231 Since the oldest substitutions occurred in Palindrome W.P2, we infer that this element became
232 established first (Fig. 6a). The LTRs of the nonautonomous elements *Ltr-p2-a* and *Ltr-p2-k* flank
233 the *SMR* and *RR* genes (Fig. 6c,d; Additional file 2: Figure S5), which raises the intriguing
234 possibility that these LTRs were involved in the translocation of these genes to this region. We
235 also found two highly identical LTRs from the same family in W.P1 (*Ltr-p2-b3* on arm3 and the
236 *Ltr-p2-b4* on arm4; Fig. 6a-c; Additional file 1: Table S10). There are truncated parts of this LTR
237 in the pre-arm and the spacer between arm1 and arm2 as well (Fig. 6b, c). These copies might be
238 a direct consequence of duplications and inversions that occurred during the formation of the
239 palindrome instead of independent insertions.

240 **Evidence for gene conversion in the palindromes**

241 We have shown that the palindromes are likely to be millions of years old, yet sequence
242 identity of portions of the palindrome arms remains high (Fig. 5a). A likely explanation for this
243 is gene conversion among the palindrome arms, as has been observed in the mammalian Y
244 chromosome palindromes [5]. To test for this, we searched for regions that had interspecific
245 polymorphisms relative to *S. suchowensis*, a closely-related species with ZW sex determination
246 [20]. If regions with interspecific polymorphisms lack paralogous sequence variation (PSV)
247 across the palindrome arms, then this would be excellent evidence of gene conversion. We
248 detected a 3 kb region within the palindromes where there are no PSVs in *S. purpurea* and only

249 one PSV in *S. suchowensis*, but substantial interspecific polymorphisms (Fig. 7). The depth of
250 this region is 4N as expected for the four copies of the palindrome arms in *S. purpurea*. In *S.*
251 *suchowensis*, the depth is between 2N and 3N, which indicates that there might be a palindrome
252 structure as well, though it might be incomplete. We also applied the same methods with
253 resequencing reads of two female and two male *S. viminalis* individuals (another *Salix* with ZW
254 sex determination) [19], but the palindromic region was not well covered by reads of either sex.
255 This may indicate that *S. viminalis* lacks the palindrome, though it is more distantly related to *S.*
256 *purpurea* than is *S. suchowensis*, so this may simply be due to excessive sequence divergence in
257 this region.

258 **Expression patterns of genes in the palindromes**

259 We examined expression profiles in multiple tissues of the two reference genomes to
260 validate the predicted transcripts and to determine how the expression patterns of genes in the
261 palindromes differ from their autosomal counterparts. As expected, most genes in the
262 palindromes show female-limited expression while the autosomal copies are generally not sex-
263 biased (Fig. 8a). The cytokinin response regulator (*RR*) (Sapur.15W073500) shows the highest
264 expression in catkin tissue, followed by expression in shoot tips and stems. On the contrary, two
265 autosomal copies on Chr19 show lower expression, limited to female catkins and male buds. The
266 four copies of the *SMR* gene show low expression in female catkins and other tissues, but the
267 autosomal copy on Chr19 (Sapur.019G001500) is expressed in all tissues (Fig. 8a). All five
268 copies of the *HCT* gene from the palindromes showed low expression in female catkins and roots
269 and higher expression in leaf tissues, shoot tips, and stems, all of which were female-biased. Two
270 copies of the DNA Primase gene from palindrome W.P2 also show high expression in leaf
271 tissues while the original copy on the autosome (Sapur.017G119600) was expressed across all

272 sampled tissues. Similarly, analysis of transcriptomic data of catkins from 10 females and 10
273 males in the F_2 family confirms that the genes in the palindromes are primarily expressed in
274 female tissue, in contrast to their autosomal paralogs (Fig. 8b).

275 **Discussion**

276 **The W chromosome in *S. purpurea***

277 Using male and female specific depth from a controlled cross pedigree, we have been
278 able to identify Z and W haplotypes from the SDR of a highly heterozygous species from a
279 standard PacBio assembly. We also show how presence-absence markers generated from
280 sequence depth in controlled cross progeny can be used to genetically map hemizygous portions
281 of the SDR. In a similar study of a young Y chromosome in asparagus, BioNano optical maps for
282 a YY individual were generated to improve genome contiguity, and sequence depth of coverage
283 was also treated as a QTL to aid the assembly because of the presence of large indels in the sex
284 chromosome [15]. Here, we showed that by combining long-read sequencing with GBS marker
285 data from a large F_2 family, we could efficiently identify the sex-specific regions of the sex
286 chromosome. However, unlike strategies like single-haplotype iterative mapping and sequencing
287 (SHIMS) that have been used in assemblies of mammalian Y chromosomes [4,6–8], our map-
288 based strategy could not provide a definitive order for the female-specific contigs due to lack of
289 recombination in the SDR.

290 Our W chromosome assembly has helped explain why sex association analyses
291 sometimes detect multiple unlinked loci on different chromosomes [18,22,28]. GWAS with a
292 nearly exhaustive collection of W contigs in the reference genome yields a single sex association
293 peak (Additional File 2: Figure S2). We have shown that the SDR contains many translocations
294 from autosomes. If those translocated segments are absent from the reference genome, W-

295 specific reads from a female individual align to the original source of those sequences, resulting
296 in false association peaks on the autosomes. This may also explain the multiple peaks detected in
297 *P. trichocarpa* and *P. balsamifera*, both of which have XY sex determination systems. In this
298 case, SNPs were identified based on alignments to a female reference genome, which lacks the
299 SDR [22].

300 Autosomal translocations are likely to be important contributors to the evolutionary
301 history of the sex chromosome in *S. purpurea*. The W chromosome is approximately 2.5 Mb
302 larger than the Z chromosome due to expansion of the SDR, including a substantial expansion of
303 gene content, driven in part by numerous translocations and subsequent expansion of autosomal
304 genes. Autosomal translocations have also been demonstrated in other sex chromosomes. For
305 example, ampliconic sequences on the human Y chromosome were acquired through
306 transpositions from diverse sources, and then amplified [4]. These ampliconic sequences account
307 for about 30% of the male-specific Y euchromatin [4]. These translocations and other structural
308 rearrangements may also suppress recombination and cause linkage between autosomal and sex-
309 linked genes [2]. In plants, the large sizes of young Y chromosomes have been attributed to the
310 accumulation of retrotransposons (reviewed in [29]).

311 Initial analyses in *P. trichocarpa* suggested that the SDR is much younger than the whole
312 genome duplication event that is shared by *Populus* and *Salix*, suggesting that the SDR became
313 established well after these genera diverged [22]. The SDR of *S. purpurea* also appears to be
314 established relatively recently. Furthermore, W and Z alleles in the nonrecombining SDR have
315 similar Ks values compared to those for orthologous autosomal genes from *S. purpurea* and *S.*
316 *suchowensis*. Given that the SDR is located in approximately the same portion of Chr15 in both
317 species, and both have ZW systems [18,20], it is reasonable to assume that the SDR became

318 established in this lineage prior to divergence of these two species, but well after divergence
319 from *Populus*.

320 Sex chromosomes commonly show evidence of "evolutionary strata" with markedly
321 different Ks values that represent different epochs of expansion of the SDR [30]. Under one
322 common model of sex chromosome evolution, these strata are the result of gradual SDR
323 expansion as sexually antagonistic alleles become incorporated into the SDR [25,31]. We
324 detected little evidence for the existence of such strata in *S. purpurea*. This corroborates a
325 previous analysis that failed to detect strata in *S. suchowensis* using an integrated segmentation
326 and clustering method, which was successful in finding strata in multiple other species, including
327 *P. trichocarpa* [32]. The failure of this method in *S. suchowensis* was hypothesized to be due to
328 the small size of the SDR [32]. However, the identified SDR in *S. purpurea* is about 6-7 Mb,
329 occupying more than one third of the W chromosome assembly. It appears instead that cessation
330 of recombination has not been a gradual long-term process in the *S. purpurea* SDR. One
331 explanation for the large size of this region is it partially overlaps with the centromere of Chr15,
332 as we previously reported [18]. It's possible that the repressed recombination in this region pre-
333 dated the translocation of a relatively small SDR cassette, as has been observed in octoploid
334 *Fragaria* [16]. This is consistent with the apparently small size of the region in *Populus* [22].
335 This is also consistent with the structure and composition of the palindromic repeats that we
336 discovered in *S. purpurea*, as detailed below.

337 **Sex chromosome palindrome repeats**

338 We have reported here the first observation of a large inverted repeat in a plant sex
339 chromosome, similar to the palindromic structures observed in mammalian sex chromosomes.
340 W.P1 and W.P2 of *S. purpurea* have a similar arrangement of arms as P1 and P3 in humans due

341 to the presence of highly homologous regions between the two palindromes. Similar palindromes
342 have been discovered on Y chromosomes of several mammals and avian W chromosomes
343 (reviewed by [5]). Large mammalian palindromes developed as a series of accumulations of
344 insertions from autosomes and maintained through arm-to-arm gene conversion. This
345 intrachromosomal gene conversion provides an essential function in maintaining coding
346 sequence integrity which otherwise would be compromised by the continuous accumulation of
347 deleterious mutations in the absence of homologous recombination (i.e., Muller's Ratchet)
348 [5,33,34]. The fact that these structures have independently evolved in non-recombining regions
349 of sex chromosomes is an intriguing case of convergent evolution of chromosome structure. It is
350 interesting to note that the chloroplast genome, another non-recombining chromosome in plants,
351 also contains a different large inverted repeat that undergoes gene conversion [35] and helps
352 maintain structural integrity of the genome, suggesting that this phenomenon may be common in
353 regions of the genome that lack recombination [36].

354 The *S. purpurea* palindromes are considerably smaller than mammalian palindromes, and
355 have only accumulated two major autosomal translocations, possibly reflecting their young age.
356 Another difference between the human palindrome and the one in *S. purpurea* is that the gene
357 conversion seems to be quite efficient across all the eight palindromes in humans, but the
358 observed regions under gene conversion in *S. purpurea* are much more limited. This is
359 particularly obvious in W.P2, compared to human P1, which has high sequence identity over
360 several Mb (Fig. 4).

361 LTR retrotransposons found in the palindrome are important at least for two reasons. The
362 observation of nonautonomous LTR retrotransposons suggests that translocations mediated
363 through transposable elements could provide a means of movement of genes (such as sex-

364 determining gene(s)), similar to what has been observed in octoploid *Fragaria* [16]. Secondly,
365 illegitimate intrachromosomal recombination events could be mediated by LTR retrotransposons
366 that have accumulated in the SDR [37], resulting in complex rearrangements like the
367 palindromic structures. Our observation of a truncated copy of a copia LTR located at the spacer
368 between arm1 and arm2 lends some support to this (Fig. 6). However, gene conversion prevents
369 us from properly inferring relationships between these LTRs and structural variation in the
370 palindrome by eliminating informative paralogous variations. Inferences about the order of
371 formation of palindromic arms will require analysis in a comparative phylogenetic framework.

372 **Type A Response Regulator is a sex determination candidate gene**

373 The Y chromosome palindromes in humans contain eight gene families that are expressed
374 predominantly in the testes and which are essential for spermatogenesis [5,38,39]. These genes
375 undergo extensive gene conversion and have high sequence conservation among the copies [5],
376 and they also show evidence of dosage sensitivity [40,41]. It is therefore reasonable to suspect
377 that the multi-copy genes in the *S. purpurea* palindrome might play important roles in sex
378 determination and/or sex dimorphism. Support for this hypothesis is provided by the type A
379 cytokinin response regulator homologs that occur in palindrome arms 1,2,3a, and 4a (Table 3).
380 The best ortholog of these genes in *P. trichocarpa* is Potri.019G133600. This gene grouped with
381 the Arabidopsis type A response regulators *ARR16* and *ARR17* in the original phylogenetic
382 analysis of this family in *Populus* [42]. Note that this gene was originally designated *PtRR11*, but
383 it is referred to as *RR9* in subsequent publications [43–45], so we will adopt that nomenclature
384 here to avoid confusion. *PtRR9* is expressed primarily in reproductive tissues in *Populus* [42,45],
385 and is also associated with sex in several *Populus* species [22,43,44]. Further supporting its
386 possible role in sex determination, it was the only gene in the *P. balsamifera* genome that

387 showed clear sex-specific differences in promoter and gene body methylation [43]. This raises
388 the intriguing possibility that this gene provides a shared mechanism of sex determination that
389 plays an essential role in the Chr19 SDR in *Populus* as well as the Chr15 SDR in *Salix*. This is
390 contrary to previous reports that detected no shared sequence composition of these two SDRs,
391 leading to a hypothesis of independent evolution of the sex chromosomes in these genera [20]

392 The cytokinin signaling pathway has emerged in recent years as a prominent candidate
393 for regulating floral development and sex expression in plants [46,47]. Cytokinin signaling in
394 Arabidopsis functions as a His-Asp phosphorelay system that transduces the signal from dual
395 function membrane-bound histidine kinases in the endoplasmic reticulum, through histidine
396 phosphotransferases to the type B response regulators (ARR). The type B ARR genes contain
397 both the Asp receiver domain that receives the phosphate from the AHP intermediates, as well as
398 a myb-like DNA binding domain that recognizes conserved motifs in the cytokinin response
399 pathways. The type A response regulators are activated by the Type B ARRs. Type A ARRs
400 contain the receiver domain but lack the DNA binding domain, and are generally negative
401 regulators of the cytokinin response (reviewed in [48]).

402 In Arabidopsis, disruption of this pathway at multiple stages has resulted in modified
403 reproductive development, with cytokinin typically being implicated in enhancing development
404 of female floral parts [46]. For example, it has recently been shown that the Type B response
405 regulators ARR1 and ARR10 bind to the promoter region of the floral homeotic gene
406 *AGAMOUS*, and down-regulation of these genes inhibits carpel regeneration in an in vitro assay
407 [49]. Similarly, triple mutants of *ARR1*, *ARR10*, and *ARR12* have severely compromised
408 gynoecial development [50]. Proper patterning of the gynoecium is hypothesized to depend upon
409 coordinated expression of genes that are positively regulated by cytokinin in the medial domain

410 of the developing gynoecium, and suppression of these genes in the lateral domains of the
411 gynoecium, potentially by *ARR16* [50]. Intriguingly, *ARR16* is one of the closest orthologs of
412 *PtRR9* [42].

413 The potential role of cytokinin signaling in dioecy has recently been highlighted by the
414 groundbreaking study by Akagi et al in kiwifruit (*Actinidia spp*) [47]. The authors identified a
415 Type C response regulator (*Shy Girl, SyGI*) on the Y chromosome that was associated with
416 maleness. Overexpression of this gene in Arabidopsis and *Nicotiana tabacum* caused suppression
417 of carpel development, supporting its potential role as a suppressor of female function [14]. This
418 work has some interesting parallels with the results reported here for *Salix* and *Populus*. First,
419 type C response regulators are essentially similar in structure to Type A response regulators, with
420 the main difference being that Type C is not induced by cytokinin. Interestingly, *PtRR9* also was
421 not induced by exogenous cytokinin application [42], though this has not yet been tested with
422 floral tissue. Second, *SyGI* was duplicated from an autosomal gene and subsequently gained a
423 new function on the Y chromosome, much like *SpRR9* has been duplicated from Chr19 in *S.*
424 *purpurea* and established a distinct pattern of expression, and presumably new functions.
425 However, *RR9* and *SyGI* are clearly not orthologous and likely perform different roles in
426 cytokinin signal transduction. This supports the view that there are numerous ways to achieve
427 separate sexes in plants, and it is likely that a myriad of mechanisms underlie the hundreds of
428 independent occurrences of dioecy in the angiosperms [51], even if a relatively small number of
429 pathways are involved [11,14].

430 **Conclusion**

431 We have shown that the SDR of *S. purpurea* contains large palindromic repeats that are
432 similar to the prominent palindromes on mammalian Y chromosomes. We further demonstrated

433 that the *Salix* palindrome is undergoing gene conversion, suggesting some functional similarities
434 to the mammalian sex chromosomes, a striking example of convergent evolution in chromosome
435 structure. We have also demonstrated that the coding sequence undergoing gene conversion in
436 the palindrome, *SpRR9*, is orthologous to a gene that is also associated with sex in *Populus*. This
437 gene is an excellent candidate for controlling sex determination through modulation of the
438 cytokinin signaling pathway. However, much remains to be determined about the underlying
439 mechanism of sex determination. Most importantly, it is currently unclear how the same gene is
440 functioning in an XY system in *Populus* and a ZW system in *Salix*. It is possible that the W
441 chromosome version acts as a dominant promoter of female function, while the Y version is a
442 dominant suppressor of female function, based on the putative roles of cytokinin and the type A
443 response regulators in female development in Arabidopsis. A detailed model should emerge
444 through comparative analysis of the W and Y chromosomes of multiple species in the
445 Salicaceae, which is currently underway.

446 **Methods**

447 ***Initial assembly of the genome***

448 Whole genome assemblies were produced for two *S. purpurea* clones: female clone
449 94006, and a male offspring of this clone, "Fish Creek" (clone 9882-34), which was derived from
450 a controlled cross between clone 94006 and male *S. purpurea* clone 94001. Clones 94001 and
451 94006 were collected from naturalized populations in upstate New York, USA. Sequencing reads
452 were collected using the Illumina and PACBIO platforms at the Department of Energy (DOE)
453 Joint Genome Institute (JGI) in Walnut Creek, California and the HudsonAlpha Institute in
454 Huntsville, Alabama. Illumina reads were sequenced using the Illumina HISeq platform, and the
455 PACBIO reads were sequenced using the RS platform. One 400bp insert 2x250 Illumina

456 fragment library was sequenced for total coverage of 183x in clone 94006 and 153x in Fish
457 Creek. Prior to use, Illumina reads were screened for mitochondria, chloroplast, and ΦX174
458 contamination. Reads composed of >95% simple sequence were removed. Illumina reads <50bp
459 after trimming for adapter and quality (q<20) were removed. For the PACBIO sequencing, a
460 total of 47 P6C4 chips (10 hour movie time) were sequenced for each genome with a p-read
461 yield of 39 Gb and a total coverage of ~110x per genome (Additional File 1: Table S11). The
462 assembly was performed using FALCON-UNZIP [52]and the resulting sequence was polished
463 using QUIVER [53]. Finally, homozygous SNPs and INDELs were corrected in the release
464 consensus sequence using ~80x of the 2x250 Illumina reads by aligning the reads using bwa
465 mem and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool
466 [54](Additional File 1: Table S12).

467 Chromosome-scale assemblies were created using a genetic map derived from 3,697 GBS
468 markers generated for a family of 497 F_2 progeny from a cross in which the male reference is the
469 father and the female reference is the grandmother. This map is described more completely in a
470 previous publication [18]. This intercross map was used to identify misjoins, characterized by an
471 abrupt change in the *S. purpurea* linkage group. Scaffolds were then oriented, ordered, joined,
472 and numbered using the intercross map and the existing 94006 v1 release assembly [18].
473 Adjacent alternative haplotypes were identified on the joined contigs, and these regions were
474 then collapsed using the longest common substring between the two haplotypes. Significant
475 telomeric sequence was identified using the $(TTTAGGG)_n$ repeat, and care was taken to make
476 sure that it was properly oriented in the production assembly. The remaining scaffolds were
477 screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be
478 a contaminant. Completeness of the euchromatic portion of the assembly was assessed by

479 aligning *S. purpurea* var 94006 v1 annotated genes to the assemblies. In both cases, 99.7% of the
480 genes were found.

481 ***Identification of W contigs***

482 Contigs derived from the W chromosome are expected to contain some large indels
483 compared to contigs from the Z chromosome due to the lack of recombination between W and Z.
484 These hemizygous regions should exclusively occur in the female specific SDR. To identify
485 these regions, we aligned 2x250 bp Illumina resequencing reads from female clone 94006 and
486 male clone Fish Creek to the new reference using Bowtie2 [55]. Depth of coverage was extracted
487 using samtools-1.2[56]. Median depth was calculated using a non-overlapping sliding window of
488 10 kb.

489 To verify if these hemizygous regions are strictly inherited in only female individuals, we
490 used the GBS data from the F_2 family. GBS reads of 195 offspring of each sex were aligned to
491 the v5 reference with Bowtie2. Due to low coverage and depth of the GBS markers per locus per
492 individual, bam files were merged according to sex in samtools-1.2. Depth was then called in
493 Samtools-1.2 with and max depth was limited to 80,000. Regions continuously covered by GBS
494 reads were defined as GBS intervals. Then, the median of each sex was calculated across all of
495 the intervals. We defined markers as female-specific by integrating the depth from both the F_2
496 GBS and 2x250 datasets (restricted to the GBS intervals) using two rules: 1) $\log_2(\frac{M_{195}+1}{F_{195}+1})<L$,
497 where L is the lower bounds of the distribution, defined by the fifth percentile divided by the
498 number of intervals tested (Additional File 2: Figure S6); and 2) $\log_2(\frac{94006_{2by250}+1}{Fish\ Creek_{2by250}+1})>5$. The
499 cutoff for the second criterion was based on the occurrence of a distinct peak in the distribution
500 of the ratios (Additional File 2: Figure S7). Scaffolds that contained at least three female-specific

501 markers were selected as candidate W-female-specific scaffolds (unmapped W pieces). Based on
502 these criteria, only two contigs from the original Chr15 assembly were from W contigs, and the
503 rest were from Z (Additional File 1: Table S5; Additional File 2: Figure S1a).

504 ***Assembly of the Z and W chromosomes***

505 Raw GBS reads used for the original map were demultiplexed and trimmed down to 64
506 bp for each read by process_radtags (in Stacks 1.44 [57]) with -c -q -r -t 64. Then, trimmed
507 reads of each sequenced individual from the F_2 family were aligned to the 19 chromosomes and
508 unmapped scaffolds from the main genome and alternative haplotypes from the v4 reference of
509 94006 using Bowtie 2 [55] with the --very-sensitive flag (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50) to
510 maintain a balance between sensitivity and accuracy. Upon examining the distribution of SNPs
511 in the genome, it became clear that the alternative haplotypes were preventing us from retrieving
512 markers in some regions in the genome, so we repeated the alignments using three different
513 reference sequences: 1) the 19 chromosomes, 2) unmapped scaffolds, and 3) alternative
514 haplotypes. Then, a wrapper script ref_map.pl in Stacks was used to call genotypes with -m 5
515 (minimum number of reads to create a tag for parents) and -P 3 (minimum number of reads to
516 create a tag for an offspring) on all progeny. Cross type "CP" was chosen since it was the one
517 closest to our cross. Offspring with poor coverage were removed from the downstream analysis.

518 Once all genotypes were retrieved through Stacks, markers from different loci showing
519 the exact same genotype/segregation across the progeny were binned and only markers from the
520 main genome were kept for mapping. Markers with severe segregation distortion or excessive
521 missing data were excluded, along with twelve offspring with very low call rate. Genotypes were
522 imputed and corrected based on inferring haplotypes in the two F1 parents from segregation of
523 the markers in the progeny.

524 The grandparents of the F_2 cross have extensive stretches of shared haplotypes, possibly
525 due to historic inbreeding in this naturalized population. This results in long runs of
526 heterozygosity and homozygosity in the F_1 progeny. This inhibits integration of backcross and
527 intercross markers by available mapping algorithms like those in the Onemap package [58]. To
528 circumvent this problem, all intercross markers were translated to female and male backcross
529 markers by identifying the parental origins of alleles based on parental phases and physical
530 position in the assembly. Also, putatively hemizygous markers were recoded as backcross
531 markers using sequence depth to infer genotypes. For example, markers with the segregation
532 pattern +/- x -/- were recoded as AB x BB. These genotypes were also imputed and corrected
533 based on the inferred haplotypes of the two F1 parents.

534 Onemap v2.1.1 was used to form initial linkage groups. For each chromosome, there are
535 two phased linkage groups from each backcross type. However, this phase information derived
536 from the F_2 family is only for the F1 parents, which cannot be directly used for phasing
537 haplotypes in the grandmother, clone 94006. By comparing parental genotypes from one LG to
538 those of the grandparents, we inferred which of the 94006 haplotypes were inherited by each F_1.
539 These results were used as a piece of evidence for identifying W-linked scaffolds/contigs, as well
540 as estimating the overall occurrence of chimeric contigs in the assembly. After building a
541 framework genetic map using markers from the main genome, non-distorted markers from
542 unmapped main scaffolds and alternative scaffolds were added.

543 All unmapped scaffolds were manually checked to see if they matched the phase
544 information or contained female-specific markers. Those that were identified as Z
545 scaffolds/contigs were excluded from the W map. The new W and Z were assembled using the
546 python package ALLMAPS [59] to order and orient scaffolds and reconstruct chromosomes

547 based on the genetic map. Only the order of the female backcross map was used to assemble the
548 W, and ALLMAPS was set not to break contigs. This new map-based assembly containing two
549 versions of chromosome 15 (Chr15Z and Chr15W) is version 5 of the *S. purpurea* var. 94006
550 genome.

551 To identify Z-degenerate regions and insertions in the female specific W (FSW), we
552 realigned the 2x250 reads of 94006 and Fish Creek to the 94006 v5 reference as described above,
553 except we removed Chr15Z out of the reference. Depth was calculated using samtools, and the
554 median depth of 50kb non-overlapping windows across was calculated with an in-house script.
555 Regions where medians of Fish Creek depth are no greater than 10 were considered as insertions
556 in the FSW, and regions with greater depth were considered Z-degenerate regions. This analysis
557 was repeated with a 10 kb window as well to enhance the resolution of the analysis.

558 ***Annotation of the genome***

559 Transcript assemblies were constructed from ~126M pairs of 2x76bp (94006) or 2x150bp
560 (Fish Creek) paired-end Illumina RNA-seq reads using PERTRAN. 188,628 transcript
561 assemblies were constructed using PASA from the RNA-seq transcript assemblies. Loci were
562 determined by transcript assembly alignments and/or EXONERATE alignments of proteins from
563 *Arabidopsis thaliana*, soybean, poplar, cassava, brachypodium, grape, and Swiss-Prot proteomes,
564 and high confidence *Salix purpurea* Fish Creek gene model peptides to repeat-soft-masked *S.*
565 *purpurea* 94006 genome using RepeatMasker with up to 2K BP extension on both ends unless
566 extending into another locus on the same strand. Gene models were predicted by homology-
567 based predictors, FGENESH+, FGENESH_EST (similar to FGENESH+, EST as splice site and
568 intron input instead of protein/translated ORF), and EXONERATE, PASA assembly ORFs (in-
569 house homology constrained ORF finder) and from AUGUSTUS via BRAKER1. The best

570 scored predictions for each locus were selected using multiple positive factors including EST and
571 protein support, and one negative factor: overlap with repeats. The selected gene predictions
572 were improved by PASA. Improvement includes adding UTRs, splicing correction, and adding
573 alternative transcripts. PASA-improved gene model proteins were subject to protein homology
574 analysis to the above mentioned proteomes to obtain Cscore and protein coverage. Cscore is a
575 protein BLASTP score ratio to MBH (mutual best hit) BLASTP score and protein coverage is
576 the highest percentage of protein aligned to the best homolog. PASA-improved transcripts were
577 selected based on Cscore, protein coverage, EST coverage, and its CDS overlapping with
578 repeats. The transcripts were selected if its Cscore is larger than or equal to 0.5 and protein
579 coverage larger than or equal to 0.5, or it has EST coverage, but its CDS overlapping with
580 repeats is less than 20%. For gene models whose CDS overlaps with repeats for more than 20%,
581 its Cscore must be at least 0.9 and homology coverage at least 70% to be selected. The selected
582 gene models were subject to Pfam analysis and gene models whose protein is more than 30% in
583 Pfam TE domains were removed. Incomplete gene models, low homology supported without
584 fully transcriptome supported gene models and short single exon (< 300 BP CDS) without
585 protein domain, or good expression gene models were manually filtered out.

586 To annotate potential genes or coding regions in the palindrome that were missed by the
587 automated annotation, the full nucleotide sequence of arm1 (about 20 kb) was submitted to the
588 Fgenesh online service (http://www.softberry.com/berry.phtml?topic=fgenesh) with specific
589 gene-finding parameters for *Populus trichocarpa*. The predicted peptide sequences were
590 searched against predicted proteins from *Populus trichocarpa* v3.0, and *Arabidopsis thaliana*
591 TAIR10 in Phytozome 12 (https://phytozome.jgi.doe.gov/) to find the closest homologous

592 annotation. The protein domains were identified using hmmscan in HMMER (v3.1b1,
593 http://hmmer.org/) against the Pfam-A domains (release 32, https://pfam.xfam.org).

594 ***Comparison of Z and W orthologous genes***

595 Allelic genes on the Z and W chromosomes were identified by performing a reciprocal
596 blastp of all primary annotated peptide sequences in the main genome with default parameters.
597 Mutual best hits were identified with over 95% identity over at least 80% of the transcript.
598 Tandem duplications were identified as genes with expectation values of $1x10^{-10}$ that occurred
599 within a 500 kb window. In these cases, one representative gene from each tandem array was
600 used as a representative sequence, and the mutual best hit outside the tandem array was identified
601 as above.

602 To identify allelic gene pairs for calculation of synonymous substitutions between the Z
603 and W alleles, a reciprocal blast of all primary annotated peptide sequences was run with
604 "blastall –p blastp -i -e 1e-20 -b 5 -v 5 -m 8", and MCscanX was run with default parameters
605 [60]. The synonymous and nonsynonymous substitution rate of each gene pair in each syntenic
606 block was estimated by aligning the sequences with CLUSTALW [61] and using the
607 calc_all_KaKs_pairs script in MCscanX. Only pairs between Chr15W and Chr15Z, and
608 scaffold_844 were used for estimating the divergence between Z and W.

609 ***Identification of sex-associated loci***

610 Loci associated with sex were identified using 60 non-clonal individuals from a
611 naturalized population of *S. purpurea* [18]. GBS reads from each individual were aligned to the
612 94006v5 genome without Chr15Z using Bowtie2. Genotypes were called in Stacks 1.14 using
613 the ref_map.pl wrapper and the populations module with a minimum minor allele frequency of

614 0.1 and a genotyping rate of 0.1. Loci with greater than 40% missing data were removed.
615 Association with sex was performed using emmax [62] as described previously [18].

616 ***Detection of palindromic repeats***

617 We conducted a systematic scan of large highly identical inverted regions. Each
618 chromosome and scaffold in the main genome was aligned to itself with LASTZ 1.03.66 with the
619 following flags: --gapped --exact=100 --step=20. An exact value of 100 means that 100-base
620 exact match is required to qualify as an High Scoring Pair in the output. Using a 1-kb sliding
621 window and 100-bp steps, an intrachromosomal hit was kept if it matched the following criteria:
622 1) the hit has the highest identity in the window, 2) greater than 2 kb in length, 3) aligned regions
623 are not further apart than 500 kb, and 4) there is a corresponding hit on the negative strand.
624 Candidate palindromes also satisfied the following criteria: 1) identity greater than 99.0%, 2) the
625 total number of hits from both strands should be less than 4, with at least one on the negative
626 strand, and 3) large inverted hits are at least 8 kb in length. Each region was manually checked
627 and assembly artifacts (e.g., corresponding to contig boundaries) and complex repetitive regions
628 were excluded.

629 Paralogous gene copies on autosomes were retrieved from the reciprocal blastp results
630 described above. Paralogous genes within the palindrome arms were aligned along with
631 paralogous copies from the autosomes using Muscle using default parameters provided in
632 MEGA 5. In a few cases, the resulting alignments were adjusted manually (Supplemental
633 Materials: AdditionalFile3). A Neighbor-Joining tree with default parameters was built using
634 MEGA 5 [63].

635 To identify recent insertions of transposable elements within the palindrome, LTRharvest
636 [64] was run with the sequence of the palindromic on Chr15W from 8,778 kb to 9,015 kb with

637 the target site duplication restricted to 5 bp to 20 bp. To find the protein domains in the coding
638 region, a protein domain search against Pfam-A domains (release 32) was performed using the
639 hidden Markov model methods implemented in LTRdigest (–hmms flag) [65]. Predicted LTR
640 retrotransposons were determined to be non-automonous when coding regions did not contain
641 any *gag* or *pol* related domains.

642 To estimate the substitution rate between the flanking LTR repeats, 5' and 3' repeats of
643 each LTR retrotransposon predicted from LTRharvest were aligned by MUSCLE using default
644 parameters provided in MEGA 5. After all gaps were removed, both number of differences and
645 substitution rate were estimated in MEGA5. For number of differences, transitions and
646 transversions were both included with a uniform rate. Substitution rate was modeled using the
647 Kimura 2-parameter model provided in MEGA5, and the rate variation among sites was modeled
648 with a gamma distribution (shape parameter = 1).

649 ***Detection of gene conversion***

650 As evidence of gene conversion, we searched for regions that were differentiated
651 between species but concordant among the palindrome arms. To accomplish this, we aligned
652 paired-end reads from a female clone of *S. suchowensis* (srx1561933) to the 94006 v5 female
653 reference, plus alternative haplotypes. This yielded an 82.9% overall alignment rate on average.
654 Mis-mapped reads originating from the autosomes were manually identified by scrutinizing the
655 alignments, and only reads that mapped exclusively to the palindromic regions were retained.
656 These reads were re-aligned to Arm1. SNPs and indels were called using mpileup with a
657 minimum mapping quality of 20 and depth less than 300.

658 ***Expression Profiling***

RNAseq data was obtained from catkins of 10 female and 10 male F_2 progeny. RNAseq data was 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 be highly identical 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 [66] 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 [67].

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).

Competing Interests

None.

Funding

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

682 Key Project for Basic Research (2012CB114504). The work conducted by the U.S. Department
683 of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department
684 of Energy under Contract No. DE-AC02-05CH11231.

685

686 **Author Contributions**

687 R.Z. and D.M.-S. assembled and analyzed the W chromosome; J.S. and J.W.J. assembled the
688 genome; D.K., A.S., L.S., G.A.T., C.C., and K.B. sequenced the genome and transcriptomes;
689 S.S. annotated the genome; S.D., L.B.S., T.M., J.L., and M.O. designed and led the study. S.D,
690 R.Z., and D.M.-S. wrote the manuscript.

691

692

693 **References**

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

696 2. Charlesworth D. Plant sex chromosome evolution. J. Exp. Bot. 2013;64:405–20.

697 3. Charlesworth B, Charlesworth D. A Model for the evolution of dioecy and gynodioecy. Am.
698 Nat. 1978;112:975–97.

699 4. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The
700 male-specific region of the human Y chromosome is a mosic of discrete sequence classes.
701 Nature. 2003;423:825–37.

702 5. Trombetta B, Cruciani F. Y chromosome palindromes and gene conversion. Hum. Genet.
703 2017;136:605–19.

704 6. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, Van Daalen SKM, Minx PJ, et al.
705 Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content.
706 Nature. 2010;463:536–9.

707 7. Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, et al. Strict
708 evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes.
709 Nature. 2012;483:82–6.

710 8. Soh YQS, Owens E, Brown LG, Alföldi J, Fulton RS, Wilson RK, et al. Sequencing the
711 Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex
712 Chromosomes. Cell. 2014;159:800–13.

713 9. Davis JK, Thomas PJ, Thomas JW. AW-linked palindrome and gene conversion in New
714 World sparrows and blackbirds. Chromosom. Res. 2010;18:543–53.

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

717 11. Renner SS. Pathways for making unisexual flowers and unisexual plants: Moving beyond the
718 "two mutations linked on one chromosome" model. Am. J. Bot. 2016;103:587–9.

719 12. Wang J, Na J, Yu Q, Gschwend AR, Han J, Zeng F, et al. Sequencing papaya X and Y^h
720 chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc. Natl. Acad.
721 Sci. 2012;109:13710–5.

722 13. Akagi T, Henry IM, Kawai T, Comai L, Tao R. Epigenetic Regulation of the Sex
723 Determination Gene MeGI in Polyploid Persimmon. Plant Cell. 2016;28:2905–15.

724 14. Henry IM, Akagi T, Tao R, Comai L. One Hundred Ways to Invent the Sexes: Theoretical
725 and Observed Paths to Dioecy in Plants. Annu. Rev. Plant Biol. 2018;69:553–75.

726 15. Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, Ayyampalayam S, et al. The
727 asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nat.
728 Commun. 2017;8:1279.

729 16. Tennessen JA, Wei N, Straub SCK, Govindarajulu R, Liston A, Ashman T-L. Repeated
730 translocation of a gene cassette drives sex-chromosome turnover in strawberries. PLOS Biol.
731 2018;16:e2006062.

732 17. Balounova V, Gogela R, Cegan R, Cangren P, Zluvova J, Safar J, et al. Evolution of sex
733 determination and heterogamety changes in section *Otites* of the genus *Silene*. Sci. Rep.
734 2019;9:1045.

735 18. Zhou R, Macaya-Sanz D, Rodgers-Melnick E, Carlson CH, Gouker FE, Evans LM, et al.
736 Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). Mol.

737 Genet. Genomics. 2018;293:1437–52.

738 19. Pucholt P, Rönnberg-Wästljung A-C, Berlin S. Single locus sex determination and female
739 heterogamety in the basket willow (*Salix viminalis* L.). Heredity. 2015;114:575–83.

740 20. Hou J, Ye N, Zhang D, Chen Y, Fang L, Dai X, et al. Different autosomes evolved into sex
741 chromosomes in the sister genera of *Salix* and *Populus*. Sci. Rep. 2015;5:9076.

742 21. Tuskan GA, DiFazio S, Faivre-Rampant P, Gaudet M, Harfouche A, Jorge V, et al. The
743 obscure events contributing to the evolution of an incipient sex chromosome in *Populus*: a
744 retrospective working hypothesis. Tree Genet. Genomes. 2012;8:559–71.

745 22. Geraldes A, Hefer CA, Capron A, Kolosova N, Martinez-Nuñez F, Soolanayakanahally RY,
746 et al. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*).
747 Mol. Ecol. 2015;

748 23. Carlson CH, Choi Y, Chan AP, Serapiglia MJ, Town CD, Smart LB. Dominance and sexual
749 dimorphism pervade the *Salix purpurea* L. transcriptome. Genome Biol. Evol. 2017;9:2377–94.

750 24. McKown AD, Klápště J, Guy RD, Soolanayakanahally RY, La Mantia J, Porth I, et al.
751 Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in
752 *Populus*. Sci. Rep. 2017;7:1831.

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

755 26. Ingvarsson PK. Multilocus patterns of nucleotide polymorphism and the demographic history
756 of *Populus tremula*. Genetics. 2008;180:329–40.

757 27. Hou J, Ye N, Dong Z, Lu M, Li L, Yin T. Major chromosomal rearrangements distinguish
758 willow and poplar after the ancestral “Salicoid” genome duplication. Genome Biol. Evol.

759 2016;8:1868–75.

760 28. Pucholt P, Hallingbäck HR, Berlin S. Allelic incompatibility can explain female biased sex
761 ratios in dioecious plants. BMC Genomics. 2017;18:251.

762 29. Hobza R, Cegan R, Jesionek W, Kejnovsky E, Vyskot B, Kubat Z. Impact of Repetitive
763 Elements on the Y Chromosome Formation in Plants. Genes. 2017;8:302.

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

765 31. Scotti I, Delph LF. Selective trade-offs and sex-chromosome evolution in *Silene latifolia*.
766 Evolution. 2006;60:1793–800.

767 32. Pandey RS, Azad RK. Deciphering evolutionary strata on plant sex chromosomes and fungal
768 mating-type chromosomes through compositional segmentation. Plant Mol. Biol. 2016;90:359–
769 73.

770 33. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, et al. Abundant
771 gene conversion between arms of palindromes in human and ape Y chromosomes. Nature.
772 2003;423:873–6.

773 34. Lange J, Skaletsky H, van Daalen SKM, Embry SL, Korver CM, Brown LG, et al.
774 Isodicentric Y Chromosomes and Sex Disorders as Byproducts of Homologous Recombination
775 that Maintains Palindromes. Cell. 2009;138:855–69.

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

778 36. Palmer JD, Thompson WF. Chloroplast DNA rearrangements are more frequent when a large
779 inverted repeat sequence is lost. Cell. 1982;29:537–50.

780 37. Devos KM, Brown JKM, Bennetzen JL. Genome Size Reduction through Illegitimate
781 Recombination Counteracts Genome Expansion in Arabidopsis. Genome Res. 2002;12:1075–9.

782 38. Navarro-Costa P, Plancha CE, Gonçalves J. Genetic Dissection of the AZF Regions of the
783 Human Y Chromosome: Thriller or Filler for Male (In)fertility? J. Biomed. Biotechnol.
784 2010;2010:1–18.

785 39. Krausz C, Casamonti E. Spermatogenic failure and the Y chromosome. Hum. Genet.
786 2017;136:637–55.

787 40. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, et al. Mammalian Y
788 chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014;508:494–9.

789 41. Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, et al. Avian W and
790 mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nat. Genet.
791 2017;49:387–94.

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

794 43. Bräutigam K, Soolanayakanahally R, Champigny M, Mansfield S, Douglas C, Campbell
795 MM, et al. Sexual epigenetics: gender-specific methylation of a gene in the sex determining
796 region of *Populus balsamifera*. Sci. Rep. 2017;7:45388.

797 44. Melnikova N V., Kudryavtseva A V., Borkhert E V., Pushkova EN, Fedorova MS,
798 Snezhkina A V., et al. Sex-specific polymorphism of MET1 and ARR17 genes in *Populus* ×
799 *sibirica*. Biochimie. 2019;162:26–32.

800 45. Chefdor F, Héricourt F, Koudounas K, Carqueijeiro I, Courdavault V, Mascagni F, et al.
801 Highlighting type A RRs as potential regulators of the dkHK1 multi-step phosphorelay pathway

802 in Populus. Plant Sci. 2018;277:68–78.

803 46. Wybouw B, De Rybel B. Cytokinin – A Developing Story. Trends Plant Sci. 2019;24:177–
804 85.

805 47. Akagi T, Henry IM, Ohtani H, Morimoto T, Beppu K, Kataoka I, et al. A Y-Encoded
806 Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response
807 Regulator in Kiwifruit. Plant Cell. 2018;30:780–95.

808 48. Kieber JJ, Schaller GE. Cytokinin signaling in plant development. Development.
809 2018;145:dev149344.

810 49. Rong XF, Sang YL, Wang L, Meng WJ, Zou CH, Dong YX, et al. Type-B ARRs Control
811 Carpel Regeneration Through Mediating AGAMOUS Expression in Arabidopsis. Plant Cell
812 Physiol. 2018;59:761–9.

813 50. Reyes-Olalde JI, Zúñiga-Mayo VM, Serwatowska J, Chavez Montes RA, Lozano-Sotomayor
814 P, Herrera-Ubaldo H, et al. The bHLH transcription factor SPATULA enables cytokinin
815 signaling, and both activate auxin biosynthesis and transport genes at the medial domain of the
816 gynoecium. PLOS Genet. 2017;13:e1006726.

817 51. Renner SS. The relative and absolute frequencies of angiosperm sexual systems: dioecy,
818 monoecy, gynodioecy, and an updated online database. Am. J. Bot. 2014;101:1588–96.

819 52. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased
820 diploid genome assembly with single-molecule real-time sequencing. Nat. Methods.
821 2016;13:1050–4.

822 53. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid,
823 finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods.

824 2013;10:563–9.

825 54. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
826 Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA
827 sequencing data. Genome Res. 2010;20:1297–303.

828 55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods.
829 2012;9:357–9.

830 56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
831 Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

832 57. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set
833 for population genomics. Mol. Ecol. 2013;22:3124–40.

834 58. Margarido GR, Souza AP, Garcia AA. OneMap: software for genetic mapping in outcrossing
835 species. Hereditas. 2007;144:78–9.

836 59. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust scaffold
837 ordering based on multiple maps. Genome Biol. 2015;16:3.

838 60. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection
839 and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.

840 61. Wilm A, Higgins DG, Valentin F, Blackshields G, McWilliam H, Wallace IM, et al. Clustal
841 W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8.

842 62. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance
843 component model to account for sample structure in genome-wide association studies. Nat
844 Genet. 2010;42:348–54.

845 63. Tamura K, Peterson D, Stecher G, Peterson N, Kumar S, Nei M. MEGA5: Molecular
846 Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and
847 Maximum Parsimony Methods. Mol. Biol. Evol. 2011;28:2731–9.

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

850 65. Steinbiss S, Willhoeft U, Gremme G, Kurtz S. Fine-grained annotation and classification of
851 de novo predicted LTR retrotransposons. Nucleic Acids Res. 2009;37:7002–13.

852 66. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware
853 quantification of transcript expression. Nat. Methods. 2017;14:417–9.

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

856

Table 1 Density of genes and LTR retrotransposons in different areas of the genome. Numbers in parentheses are standard deviations of frequency per 100 kb window.

Count per 100 kb	SDR	Z-SDR	PAR	Autosomes*
Genes	7.7(3.9)	7.3(3.4)	12.0(4.2)	11.4(4.2)
Gypsy-LTR	20.7(12.9)	21.8(13.4)	8.3(9.1)	10.0(11.0)
Copia-LTR	15.0(10.0)	9.0(6.7)	6.9(5.4)	7.9(7.0)

* All 18 chromosomes are included.

Table 2 Coordinates of palindromes in the female SDR.

	name	Start (bp)	End (bp)	Size (bp)	Gene families
	pre-arm	8,778,973	8,791,042	12,070	*R2,HCT*
Palindrome W.P1	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 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 $1x10^{-10}$.

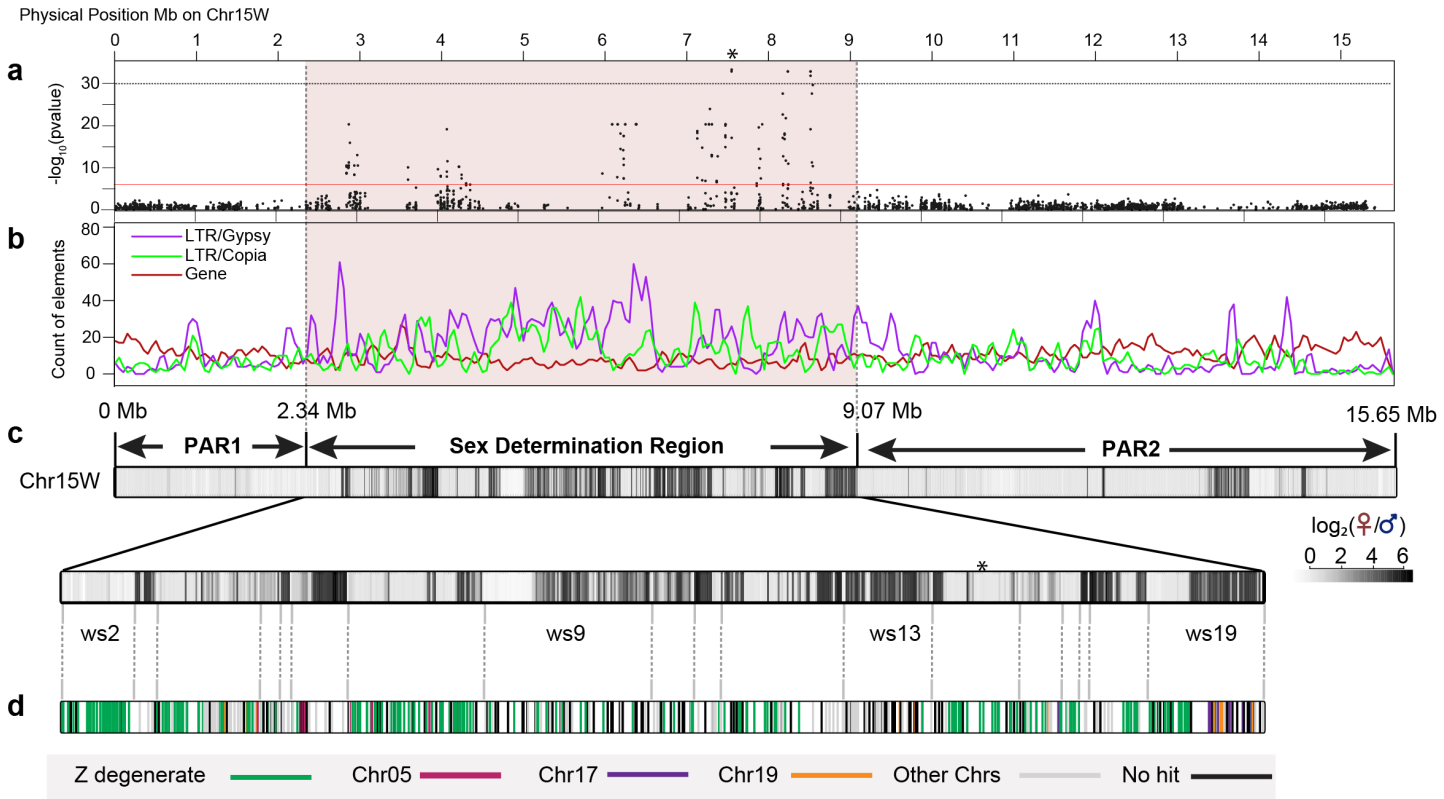

Figure 1. Genomic content of Chr15W and composition of the sex determination region (SDR). a. A Manhattan plot of Chr15W, based on GWAS using SNPs derived from aligning to a reference genome lacking Chr15Z. The Y axis is the negative logarithm of p values, and the red line indicates the Bonferroni cut 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.

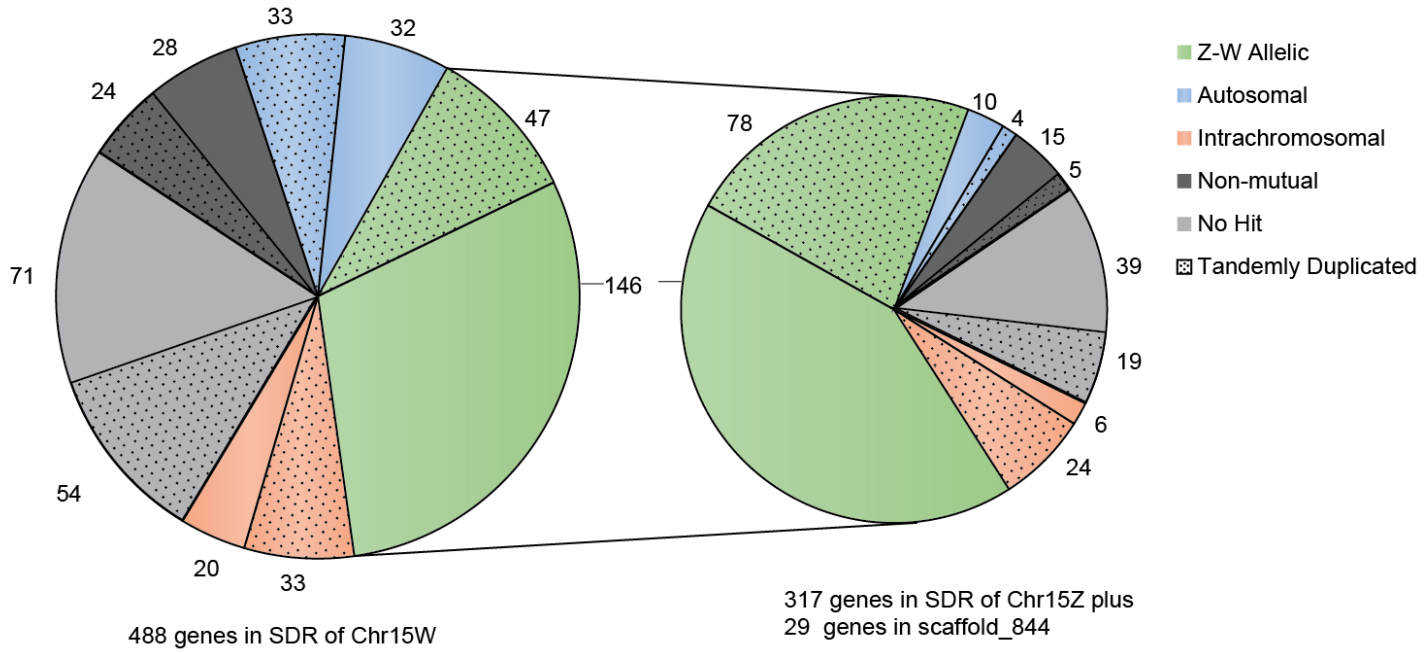

Figure 2. Annotated genes in Chr15W and Chr15Z. Genes are grouped according to the best non-self-hit in the annotated genome. Twenty-nine genes from an unmapped Z, scaffold_844 are also included. Stippled areas indicate genes of groups identified as tandem duplicates.

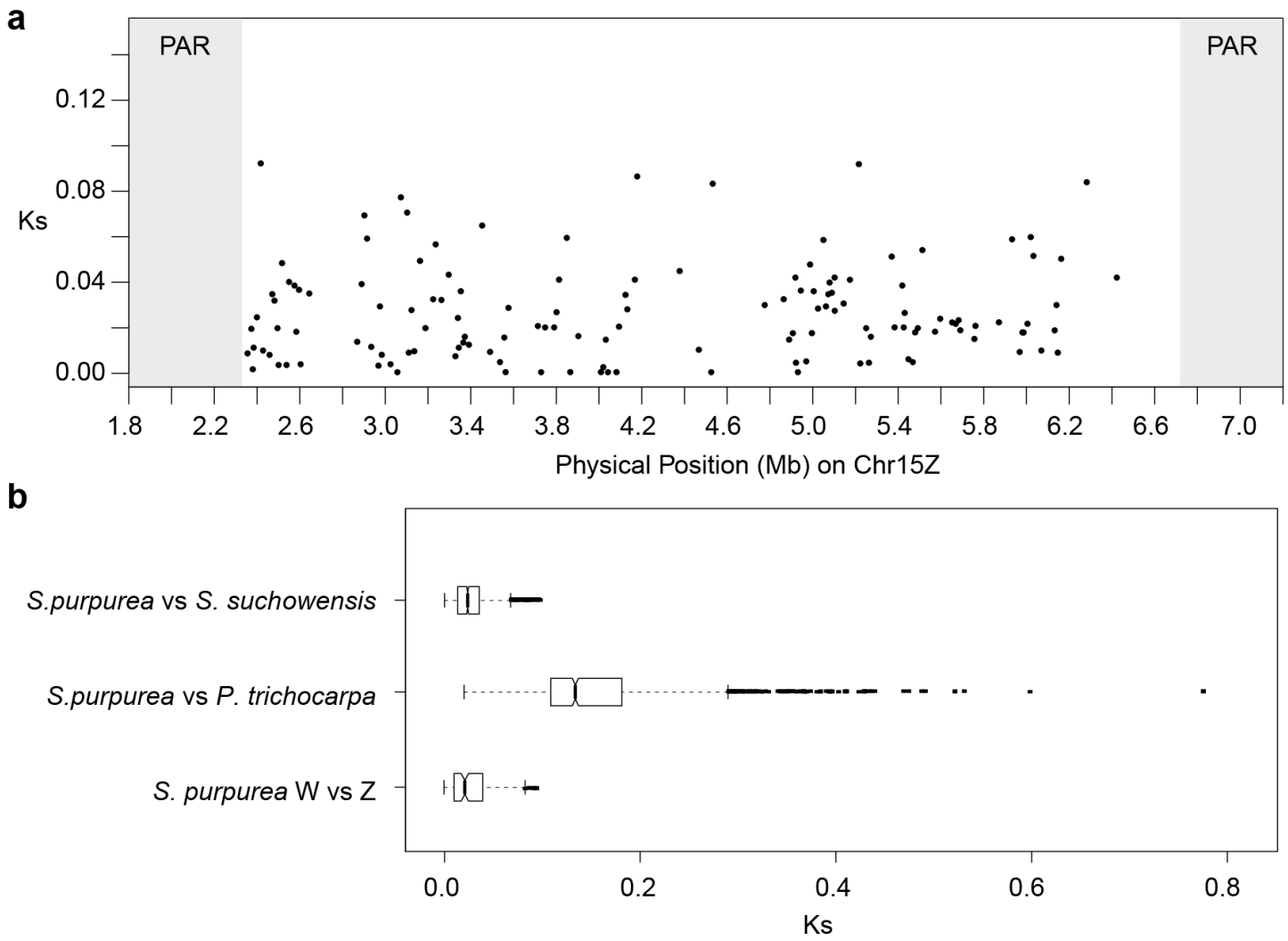

Figure 3. Synonymous substitution rates (Ks) for genes in the SDR. a. Comparison of syntenic genes in the W-SDR and Z-SDR. **b**. Boxplot showing distributions of interspecific synonymous substitutions for 1,203 syntenic genes on Chr01 for the closely-related species *S. purpurea* and *S. suchowensis* and for 1,377 genes on Chr01 in *S. purpurea* and *Populus trichocarpa*, compared to the distribution of substitutions between syntenic genes in the *S. purpurea* SDR.

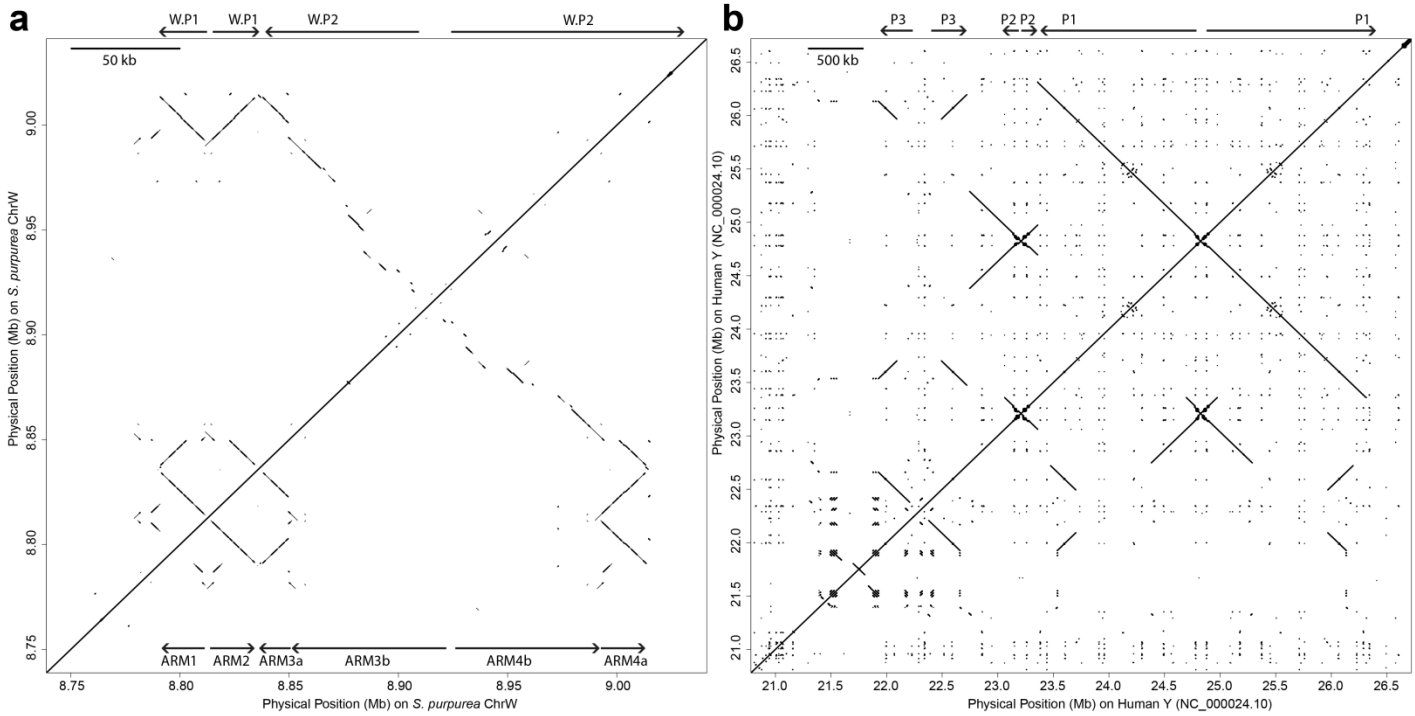

Figure 4. Palindromic repeats in the *S. purpurea* W chromsosme (a) and the *H. sapiens* Y chromosome (b). The dot plots were produced using LASTZ with identical settings. Note the different scales, indicated by the bar at the top right of each figure. *H. sapiens* palindromes are labeled following Skaletsky et al. [4].

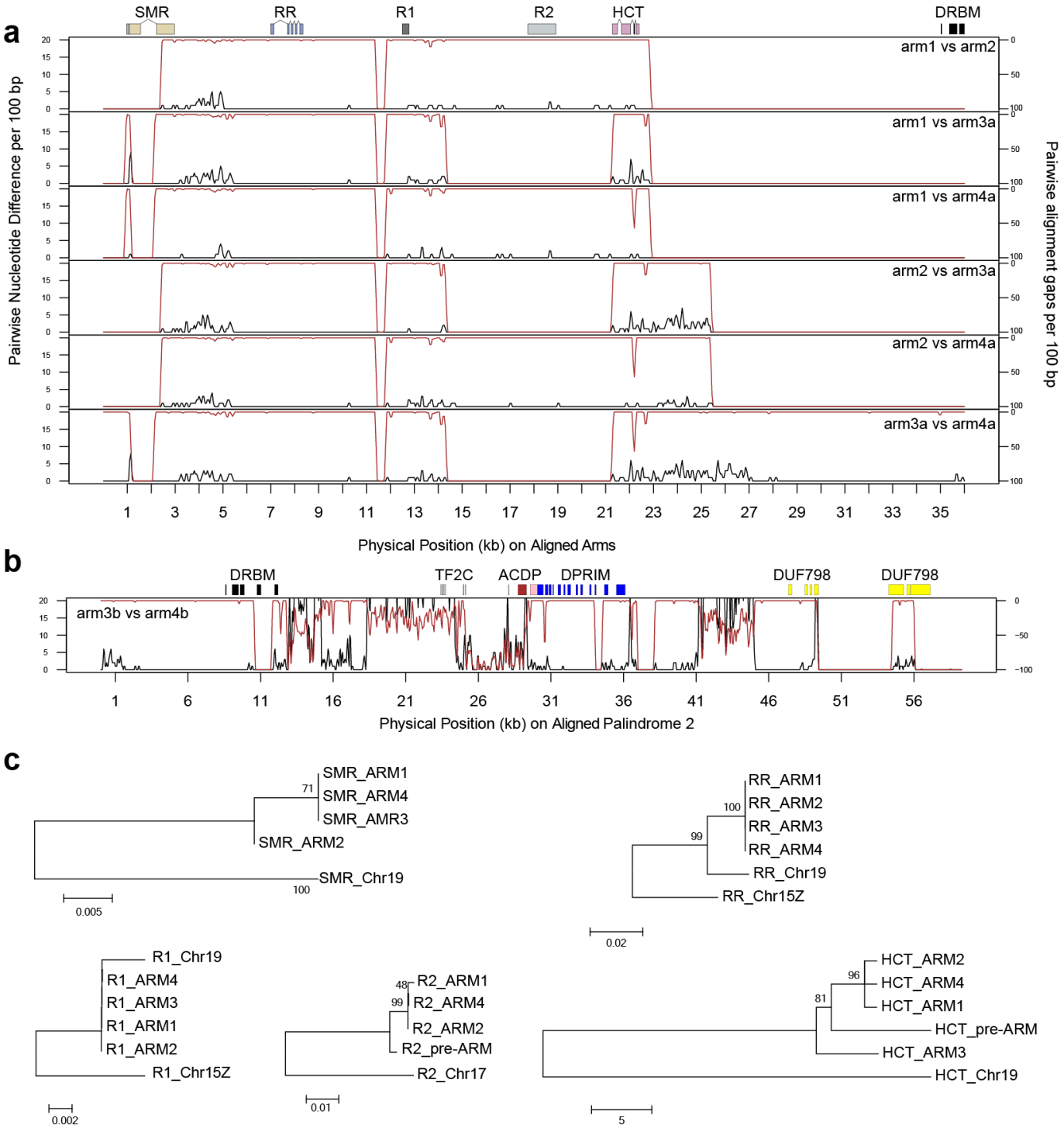

Figure 5. Sequence comparisons for the two palindromes. a. Comparison of the four arms that are shared among the two palindromes. The black line represents the number of nucleotide differences in 100 bp windows, while the red line indicates gaps in the alignment on an inverted scale. b. Comparison of the portions of palindrome 2 that are not shared with palindrome 1. c. Phylogenetic trees of five multi-copy genes in the palindromic region.

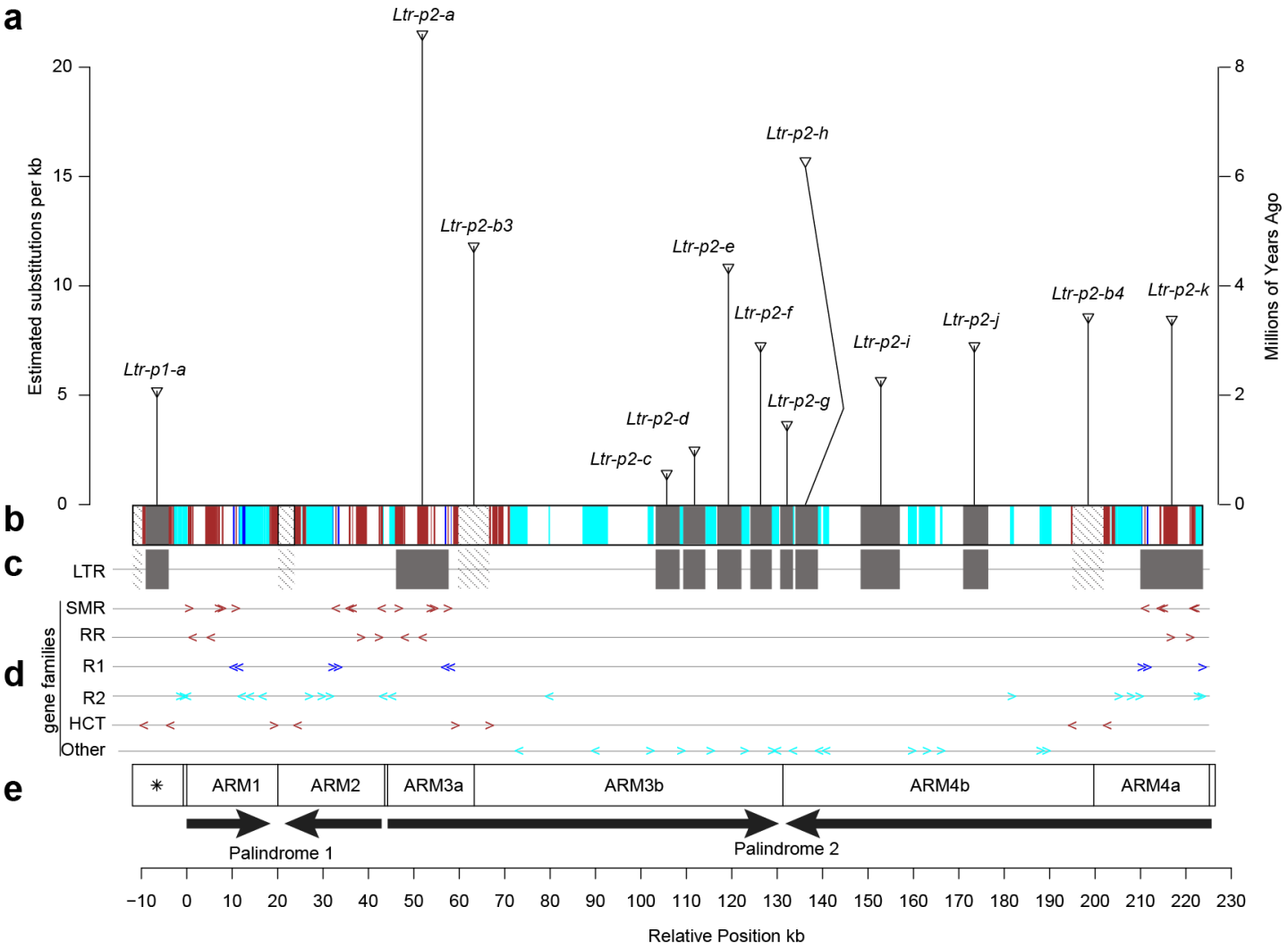

Figure 6. LTR retrotransposons, female specific genes, and palindromes. **a**. Each vertical line with a wedge on top represents each of the 13 TEs identified in the palindromic region by LTRharvest. The height of each line indicates the number of estimated nucleotide substitutions in the two LTRs (transposons 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.

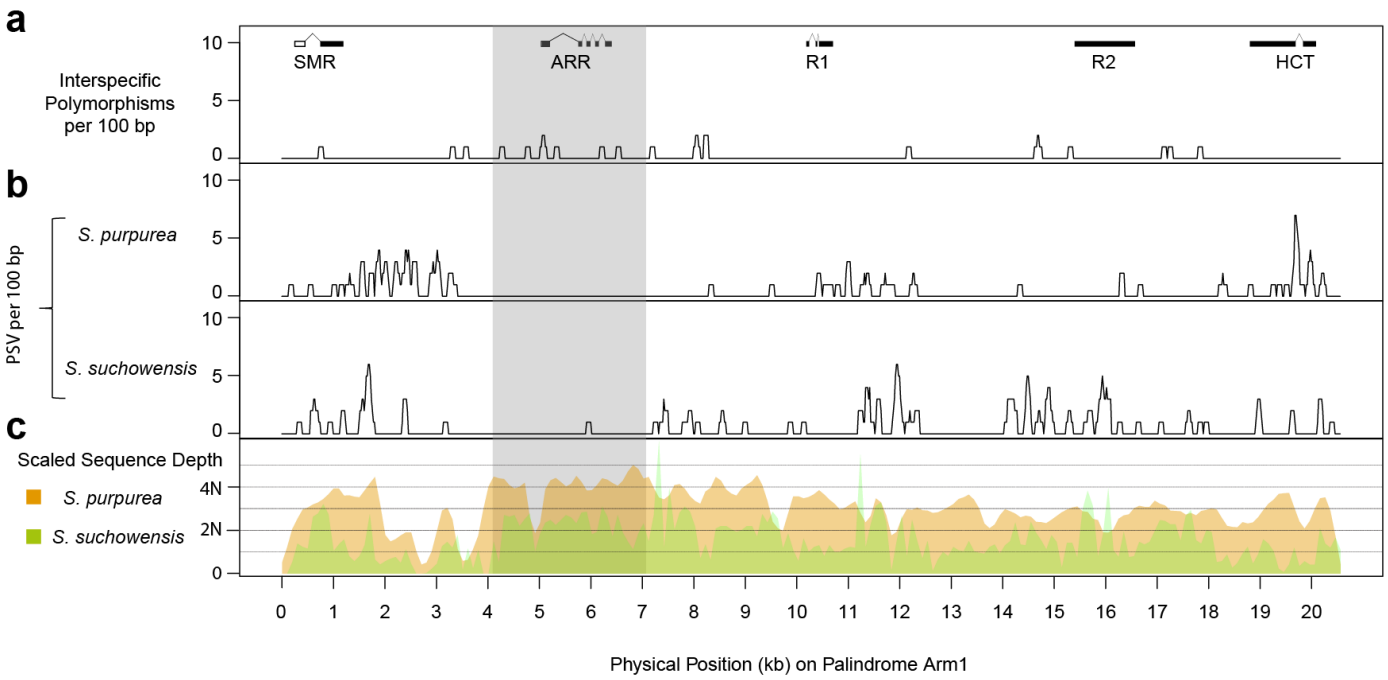

Figure 7. Sequence variation in the palindrome arms. a. Density of fixed differences between *S. purpurea* and *S. suchowensis* per 100 bp. **b**. Density of paralogous sequence variants (PSVs, differences among the 4 palindrome arms) in *S. purpurea* and *S. suchowensis*. **c**. Relative depth of Illumina 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.

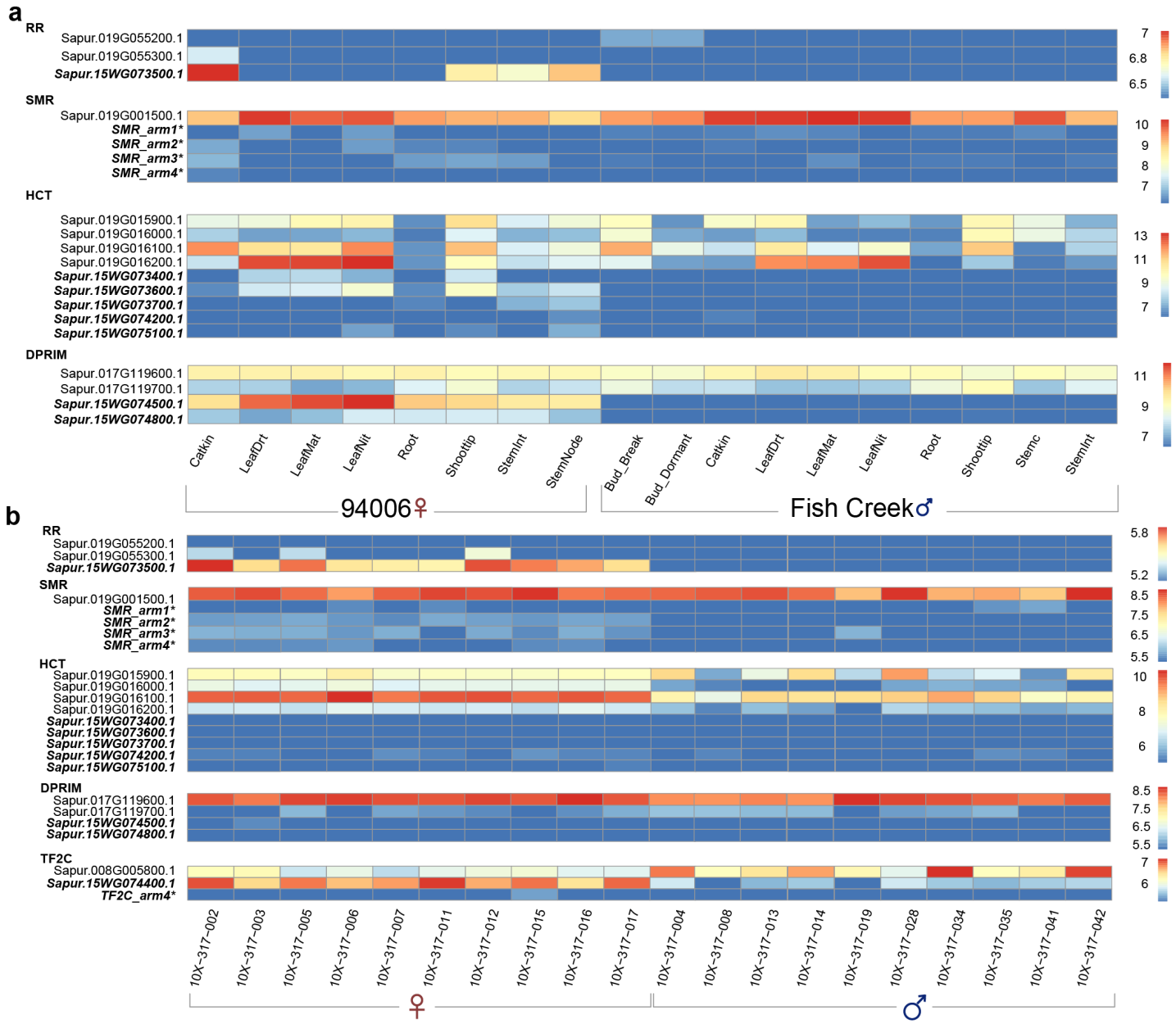

Figure 8. Expression profile of genes from the W palindromes and autosomal paralogs. **a**. Normalized read counts of genes in different tissues from clone 94006 (female) and Fish Creek (male). **b.** Normalized read counts of selected genes in catkins from 10 females and 10 males from an F_2 family. Gene labels in bold font are from the palindromes. Asterisks indicates manually annotated genes.