Article

De Novo Assembly and Annotation of 11 Diverse Shrub Willow (*Salix*) Genomes Reveals Novel Gene Organization in Sex-Linked Regions

Brennan Hyden [1,2,*], Kai Feng [2,*], Timothy B. Yates [2], Sara Jawdy [2], Chelsea Cereghino [2], Lawrence B. Smart [1,] and Wellington Muchero [2,**]**

[1] Horticulture Section, School of Integrative Plant Science, Cornell University, Geneva, NY
[2] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN
* These authors contributed equally to the work.
* Correspondence: lbs33@cornell.edu, mucherow@ornl.gov; Tel.: (L.B.S. +1 315-787-2490; W.M. +1 865-576-0223)

Abstract: Poplar and willow species in the Salicaceae are dioecious, yet have been shown to use different sex determination systems located on different chromosomes. Willows in the section Vetrix are interesting for comparative studies of sex determination systems, yet genomic resources for these species are still quite limited. Only a few annotated reference genome assemblies are available, despite many species in use in breeding programs. Here we present *de novo* assemblies and annotations of 11 shrub willow genomes from six species. Copy number variation of candidate sex determination genes within each genome was characterized and revealed remarkable differences in putative master regulator gene duplication and deletion. We also analyzed copy number and expression of candidate genes involved in floral secondary metabolism, and identified substantial variation across genotypes, which can be used for parental selection in breeding programs. Lastly, we report on a genotype that produces only female descendants and identified gene presence/absence variation in the mitochondrial genome that may be responsible for this unusual inheritance.

Keywords: Salix; shrub willow; genome assembly; sex determination

1. Introduction

The genus *Salix* and the Salicaceae family are of growing scientific interest for their use as model systems to understand sex determination and sex chromosome dynamics. Salicaceae is almost entirely dioecious, and contains approximately 30 species of *Populus* and over 300 species of *Salix* [1] yet, both the location of the sex determination system (SDR) as well as the sex inheritance mechanism (ZW vs XY) differ across clades within this family. In the shrub willows (section Vetrix), the sex determination region has been localized in to Chr15, with a ZW system of inheritance [2-5], while Chr15XY is predominant in the section Salix, including *S. arbutifolia [6]*, and a Chr07 XY sex determination system has been identified in the tree willows (section Protitea) *S. nigra [7], S. chaenomeloides [6],* and *S. dunnii [8]*. In *Populus*, Chr19 XY, Chr19 ZW, and Chr14 XY sex-determination systems have all been reported [9]. The precise genes responsible for sex determination in Salicaceae are still being studied but are thought to involve a presence/absence of expression of *ARR17* in *Populus* [10] and tree willows [6] (sections Salix and Protitea). Shrub willows in the section Vetrix, on the other hand, may possess a two-gene system of sex determination; dosage level of *ARR17* in combination with *GATA15* has been suggested as the mechanism of sex determination in the shrub willow *S. purpurea*, based on expression and resequencing evidence from a set of monoecious families in this species [11] *AGO4*, *DRB1*, and three hypothetical proteins have also been proposed as potential master regulators of sex in *S. purpurea* [12].

 doi:10.20944/preprints202212.0185.v1

Shrub willows in the section Vetrix are a dioecious crop grown widely across the northern hemisphere for a variety of horticultural uses, including for bioenergy, as ornamentals, and for ecological restoration purposes [1,13]. Commonly cultivated shrub willow species include European natives *S. purpurea* and *S. viminalis*, the Chinese species *S. suchowensis*, Japanese natives *S. integra* and *S. udensis*, and *S. koriyanagi* from Korea [1,14,15]. Together, these six aforementioned species represent a broad range of genetic diversity across the section Vetrix, including two subsections: Helix (*S. purpurea, S. suchowensis, S. integra, S. koriyanagi*) and Vimen (*S. viminalis, S. udensis*) [2,14]. Due to both the dynamic nature of the SDR within this genus and the unique mechanism of sex determination in *S. purpurea* [12], there is an interest in comparing the gene content of the sex determination regions across shrub willow species, in order confirm the two-gene model from *S. purpurea* and to identify any additional shifts in sex determination genes during the evolution of this clade.

Salix are primarily insect pollinated and as such must produce a suite of secondary metabolites to attract pollinators [16-18]. Previous studies that characterized terpenoid and flavonoid profiles in *Salix* catkins have shown substantial differential expression of these compounds based on sex, which influences pollinator attraction [16,19]. Secondary metabolites also play a known role in defense against herbivory across plant species [20,21]. QTL mapping of floral terpenoid, flavonoid, and phenolic glucoside production and identification of candidate genes has been conducted in *S. purpurea* and candidate genes for many specific compounds have been identified [19]. However, as of yet there has been little effort to compare these candidate genes between related species. Characterizing the presence, copy number, and expression of secondary metabolite genes across *Salix* species is therefore useful for understanding biological differences in floral secondary metabolite production, and their effects on pollinator attraction an herbivory.

Genomic resources for the genus *Salix* are still under development, with the shrub willows being the most well-studied group with several assembled genomes and recent advances in QTL mapping of various traits, including yield, insect resistance, and rust susceptibility [22,23]. Within the section Vetrix, reference genomes are currently available for a female *S. viminalis* [24], a female *S. suchowensis* [25], a male *S. purpurea* ('Fish Creek'), a female *S. purpurea* (94006) [26], and a monoecious *S. purpurea* [11], the latter two of which have fully assembled Chr15Z and Chr15W sex chromosomes. Here we present *de novo* assembly and annotation of 11 *Salix* genomes across six shrub willow species, including three newly sequenced and assembled species. Among these 11 genomes is a reassembly of 94006, a *S. purpurea* female that was used for the *S. purpurea* 94006 v5.1 reference genome (https://phytozome-next.jgi.doe.gov/info/Spurpurea_v5_1, last accessed 2 December 2022) and is the mother of the male 'Fish Creek' used for the *S. purpurea* v3.1 reference (https://phytozome-next.jgi.doe.gov/info/SpurpureaFishCreek_v3_1, last accessed 2 December 2022) produced by the US Department of Energy Joint Genome Institute (JGI) [26,27]. A male *S. purpurea*, (94001, the father of 'Fish Creek'), two female (P294, P295) and one male (P63) *S. suchowensis*, one female *S. integra* (P336), one male (04-FF-016) and one female (SH3) *S. koriyanagi*, one female (07-MBG-5027) and one male ('Jorr') *S. viminalis*, and a male *S. udensis* (04-BN-051) were also sequenced. These particular genotypes are of interest since previous research has reported F_1 crosses to *S. purpurea* 94001 and 94006 for each of these genotypes along with linkage maps, phenotypic analysis, and QTL mapping in the progeny [2].

For each assembly and annotation, gene content across the Chr15W SDR regions was characterized. Notably, nearly all previously identified candidate sex determination genes are missing from *S. koriyanagi, S. viminalis*, and *S. udensis* which suggests a unique sex determination mechanism in these species that may not involve *ARR17* as shown in *Populus* and *S. purpurea* [12,28]. Furthermore, the expression and copy number variation of various secondary metabolite genes was assessed, including candidates for known dimorphic floral volatile and phenolic glycoside compounds [19]. Finally, we present data that

 doi:10.20944/preprints202212.0185.v1

supports an all-female inheritance in the *Salix integra* P336 descendants and identify a missing mitochondrial *RPL10* gene as a candidate mechanism for this inheritance.

2. Results

2.1. Assembly and Annotation

Oxford Nanopore read length and quality distributions for each assembly are shown in Fig. S1. Mean genome coverage ranged from 45x to 103x. Contig N50 values ranged from 300.36 Kb in 04-FF-016, to 804.25 kb in P336. Assembly lengths were relatively consistent within species and subsections. *S. suchowensis* had the largest genome size, with a mean of 375 Mb, while the mean size of the *S. viminalis* genome was only 288 Mb. All assemblies had a Eudicot core gene BUSCO score of above 95%. Assembly statistics are shown in Table 1.

Table 1. Assembly statistics of 11 genomes, with *S. purpurea* 94006 v5.1 and 'Fish Creek' v3.1 assemblies for comparison. *scaffold number of 04-FF-016 prior to manual cutting of chimeric scaffolds.

Genome	Species	Subsection	Sex	Total Assembly Length	Number of Scaffolds	Number of Contigs	Contig N50 (KB)	Largest Contig (MB)	Mean Coverage	Assembly BUSCO Score
JGI 94006 v5.1	*S. pupurea*	Helix	F	328,137,719	348	NA	NA	NA	NA	97.0%
JGI 'Fish Creek' v3.1	*S. purpurea*	Helix	M	312,123,941	274	NA	NA	NA	NA	97.2%
94006	*S. purpurea*	Helix	F	338,238,421	179	2,675	319.30	4.75	72	95.8%
94001	*S. purpurea*	Helix	M	332,407,318	136	2,696	232.30	3.67	55	95.8%
P63	*S. suchowensis*	Helix	M	369,253,841	135	2,243	383.13	3.78	58	96.2%
P294	*S. suchowensis*	Helix	F	375,803,650	173	2,589	325.52	2.46	57	95.8%
P295	*S. suchowensis*	Helix	F	382,054,263	135	1,982	435.71	2.16	62	96.3%
P336	*S. integra*	Helix	F	312,752,820	111	1,246	804.25	5.99	60	96.7%
SH3	*S. koriyanagi*	Helix	F	339,158,221	147	2,922	335.52	2.19	45	95.5%
04-FF-016	*S. koriyanagi*	Helix	M	349,107,755	152*	2,983	300.36	2.27	75	95.1%
07-MBG-5027	*S. viminalis*	Vimen	F	293,303,539	171	1,716	532.84	4.16	103	95.7%
'Jorr'	*S. viminalis*	Vimen	M	282,587,186	197	2,136	442.89	3.81	51	96.1%
04-BN-051	*S. udensis*	Vimen	M	315,877,065	140	2,087	396.09	4.45	51	95.5%

Annotation BUSCO scores ranged from 77.9% in P336 (*S. integra*) to 92.9% in 'Jorr' (*S. viminalis*). The mean number of annotated genes across all genomes was 32,166, while the mean number of annotated transcripts was 40,679. The estimated number of missing genes, relative to 94006 v5.1, ranged from 3,706 in the 94006 reassembly to 4,973 in SH3. Genes in genome-specific orthogroups ranged from 331 in 'Jorr' to 1026 in 94006. Annotation statistics are shown in Table 2.

Table 2. Summary statistics from 11 genome annotations, with *S. purpurea* 94006 v5.1 and 'Fish Creek' v3.1 assemblies for comparison.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0185.v1

Genome	Species	Annotation BUSCO Score	Genes	Transcripts	Genes Missing	Genome-Specific Orthogroups	Genes in Specific Orthogroups
JGI 94006 v5.1	*S. pupurea*	97.0%	35125	57462	NA	NA	NA
JGI 'Fish Creek' v3.1	*S. purpurea*	97.2%	34464	46943	NA	NA	NA
94006	*S. purpurea*	82.2%	31938	36199	3706	379	1026
94001	*S. purpurea*	91.1%	31470	39196	4164	336	770
P63	*S. suchowensis*	84.9%	30530	37310	4663	229	534
P294	*S. suchowensis*	89.7%	34681	38788	4002	298	730
P295	*S. suchowensis*	87.2%	30719	36507	4532	217	574
P336	*S. integra*	77.9%	29907	34327	4733	225	574
SH3	*S. koriyanagi*	86.1%	30539	36436	4973	181	442
04-FF-016	*S. koriyanagi*	87.0%	30478	36226	4856	229	543
07-MBG-5027	*S. viminalis*	89.0%	31708	37991	3732	267	706
'Jorr'	*S. viminalis*	92.9%	30524	34112	4420	138	331
04-BN-051	*S. udensis*	86.5%	30382	36483	4902	270	609

Comparative genomics analysis with Orthofinder assigned 391,057 out of 407,955 transcripts across all 11 annotations to 49,209 orthogroups (95.2%). 2,769 orthogroups were genome-specific and accounted for 1.7% of all transcripts. Orthogroup assignment had a G50 of 11 and an O50 of 11,958 among assigned genes. 10,799 orthogroups had all 11 genomes represented, 3,401 of which were single-copy orthogroups. Phylogenetic analysis of the annotated gene sets using Orthofinder grouped genomes of the same species together (Fig. S2).

2.2. Sex determination gene analysis

Copy numbers of each candidate sex-determination gene in the Chr15W locus, as well as the Chr15Z exon 1 and Chr19 full-length copies of *ARR17* varied among genotypes (Table 3). Candidate genes were present in expected numbers in *S. purpurea* 94006, consistent with the JGI *S. purpurea* 94006 v5.1 reference genome, with the exception of one fewer *ARR17* copy as well as *GATA15* assembled on Chr17 instead of Chr15W, which are likely due to errors during assembly or scaffolding. P294, P295, and P336 have two Chr15 *ARR17* copies and one Chr15 *AGO4* copy. Most candidate sex determination genes, including *ARR17*, *AGO4*, *GATA15*, were missing from SH3 and 07-MBG-5097 (Table 3).

Table 3. Copy number of candidate sex determination genes across the 11 annotated genomes with *S. purpurea* 94006 v5.1 and 'Fish Creek' v3.1 assemblies for comparison. **GATA15* was assembled to Chr17, but this is likely an assembly error. **a fourth *ARR17* was identified on a purged haplotig.

Gene ID	Function	9400 6 JGI (F)	'Fish Creek' JGI (M)	9400 6 (F)	9400 1 (M)	P29 5 (F)	P29 4 (F)	P63 (M)	P33 6 (F)	SH3 (F)	04-FF-016 (M)	07-MBG-5097 (F)	'Jorr' (M)	04-BN-051 (M)
Sapur.15WG073500	ARR17	4	0	3**	0	2	2	0	2	0	0	0	0	0
Sapur.019G053300	ARR17	2	2	2	2	1	1	1	1	1	1	1	1	1
ARR17 15Z exon 1	ARR17	1	1	1	2	1	1	1	1	1	1	0	0	0

 doi:10.20944/preprints202212.0185.v1

Sapur.15WG062800	GATA15	1	0	1*	0	1	1	0	0	0	0	0	0
Sapur.15WG074400	AGO4	3	0	3	0	1	1	0	1	0	0	0	0
Sapur.15WG074300	DRB1	2	0	5	1	1	2	1	1	5	2	2	1	2
Sapur.15WG074900	hypothetical	1	0	1	0	1	1	0	0	0	0	0	0	0
Sapur.15WG075300	hypothetical	1	0	0	0	2	2	0	2	0	0	0	0	0
Sapur.15WG075700	hypothetical	2	0	3	1	0	0	0	1	0	9	0	0	0

2.3. Secondary metabolism gene analysis

BLASTN analysis of secondary metabolism genes revealed variation in copy number between genomes for many genes (Table 4 and Supplementary Table S1). Total combined expression across all eight tissue types for secondary metabolism genes also showed substantial variation between genotypes (Table 5).

Table 4. Flavonoid, terpenoid, and phenolic glucoside genes with annotated copy number from each genome.

Gene family description	Gene copy number 94006 v5.1	Associated compound/family	94006	94001	P63	P294	P295	P336	SH3	04-FF-016	07-MBG-5027	'Jorr'	04-BN-051
4-coumarate:CoA ligase	1	unknown glucoside	1	1	1	1	1	1	1	1	1	1	1
chalcone-flavonone isomerase	1	chalcone; isosalipurposide; chalconaringenin 4'-glucoside; catechin; naringenin; prunin; salipurposide; flavenoid	1	1	2	1	1	1	1	1	1	1	1
arogenate/prephenate dehydratase	2	prunin; isosalicin	2	1	2	1	1	1	1	1	2	1	1
cinnamyl alcohol dehydrogenase-like protein	1	benzeneacetaldehyde	1	1	1	1	0	0	1	2	2	0	0
Chalcone synthase	10	chalcone	12	12	12	11	12	11	11	10	9	11	13

 doi:10.20944/preprints202212.0185.v1

Flavenol synthase 1	3	flavenoid	2	2	1	4	1	2	2	1	1	1	1
farnesyltransferase A	1	terpene; farnesene	1	1	1	1	1	1	1	1	1	1	1
geranylgeranyltransferase type I beta subunit	1	terpene	0	0	0	0	0	0	0	0	0	0	0
total-pinoresinol reductase 1	2	isosalicin; tremuloidin; phenolic glycoside	2	3	3	3	3	3	3	2	2	2	2
geranylgeranyl transferase alpha subunit	9	terpene	0	1	0	0	0	1	2	1	0	1	0
RAB geranylgeranyl transferase beta subunit	1	terpene	0	0	0	0	0	0	0	0	0	0	0
flavonol synthase/flavanone 3-hydroxylase	4	flavenoid; tremulacin	6	3	4	7	6	7	7	5	4	6	6
terpene synthase 03	9	terpenoid; beta-ocimene; beta-pinene; farnesene; isoprene	11	5	10	11	7	11	5	7	7	3	3
terpene synthase 14	2	terpene; linalool	2	3	4	6	3	5	3	3	3	7	3
terpene synthase 21	15	terpene; sesquiterpene	6	9	3	17	3	9	7	2	7	5	4
pinene synthase	2	terpene; pinene	2	0	1	1	2	1	1	1	0	0	0
oxidosqualene cyclase	1	squalene	4	4	2	2	2	1	1	2	2	4	7
coniferyl aldehyde 5-hydroxylase	2	kaempferol-3-O-glucoside; prunin	2	3	3	4	2	2	2	3	2	2	2
dihydroflavonol 4-reductase-like1	1	isosalicin	1	1	1	2	2	1	2	2	1	1	1
squalene synthase	2	squalene	4	3	2	3	3	4	3	2	4	3	4
terpene synthase 04	2	terpene; linalool	1	2	2	2	2	2	1	2	2	2	2
geranylgeranyl diphosphate synthase	1	terpene	1	1	1	1	1	1	1	1	1	1	1
geranylgeranyl pyrophosphate synthase	2	terpene	2	2	2	2	3	2	3	2	2	2	2
geranylgeranyl reductase	3	terpene	6	5	5	4	5	5	5	4	5	4	5

 doi:10.20944/preprints202212.0185.v1

geranyl diphosphate synthase	1	terpene	0	0	0	0	0	0	0	0	0	0	0
phytoene desaturation 1	1	phytoene	1	1	2	2	2	1	1	1	1	1	1
UDP-glucose flavonoid 3-O-glucosyltransferase	6	phenolic glycoside	1	5	0	6	6	1	0	0	0	0	5
total-UDP-glucose:flavonoid 7-O-glucosyltransferase	2	phenolic glycoside	2	2	2	2	2	2	2	2	2	2	2
phytoene desaturase 3	1	phytoene	0	1	0	2	0	0	0	0	1	0	1
phytoene synthase	4	phytoene	3	3	2	2	2	2	3	2	2	2	5
squalene monooxygenase	8	squalene	12	9	20	0	0	0	15	12	7	7	10
squalene epoxidase 3	1	squalene	0	1	0	0	0	0	1	0	0	0	1
terpene synthase 02	2	terpene	5	1	1	1	1	1	1	1	2	3	1
terpene synthase 10	1	terpene	0	0	0	0	0	0	0	0	0	0	0
total-UDP-glucose flavonoid 3-O-glucosyltransferase	20	phenolic glycoside; prunin	9	31	8	24	23	11	10	8	10	8	29
UDP-sugar flavonoid 7-O-glycosyltransferase	1	phenolic glycoside	1	3	1	3	2	1	0	1	2	2	0
Chalcone-flavanone isomerase	10	chalcone; flavenoid; unknown phenolic glucoside	2	3	5	2	3	2	2	1	2	2	1
flavonol synthase/flavanone 3-hydroxylase	6	flavenoid	7	5	12	6	11	10	7	7	5	8	4
dihydroflavonol 4-reductase	1	flavenoid; benzeneacetaldehyde	1	1	3	1	1	1	1	2	1	1	1
UDP-N-acetylglucosamine transferase subunit alg13	1	kaempferol-3-O-glucoside	0	0	0	1	0	1	1	1	0	1	1
geranylgeranyl diphosphate reductase	2	terpene	3	3	3	3	1	2	3	2	4	3	4
beta-pinene synthase	1	terpene; beta-pinene	3	0	2	4	3	4	2	1	1	0	0

 doi:10.20944/preprints202212.0185.v1

benzoyl coenzyme A: benzyl alcohol benzoyl transferase	1	unknown glucoside	2	2	1	1	1	1	1	2	0	0	1

Table 5. Flavonoid, terpenoid, and phenolic glucoside normalized expression levels.

Gene family Description	Associated compound/family	94006	94001	P63	P294	P295	P336	SH3	04-FF-016	07-MBG-5027	'Jorr'	04-BN-051
4-coumarate:CoA ligase	unknown glucoside	311.1	358.2	679.2	844.7	1056.4	1117.1	1419.8	1530.6	631.1	495.5	1116.4
chalcone-flavonone isomerase	chalcone; isosalipurposide; chalconaringenin 4'-glucoside; catechin; naringenin; prunin; salipurposide; flavenoid	1040.5	1103.9	1219.4	907.4	2684.7	2282.2	1145.6	1690.8	2299.2	3322.4	3920.7
arogenate/prephenate dehydratase	prunin; isosalicin	415.6	388.1	427.9	425.5	432.9	281.1	362.1	754.1	266.5	333.3	615.3
cinnamyl alcohol dehydrogenase-like protein	benzeneacetaldehyde	759.9	1039.2	932.9	345.0	549.7	1093.4	1654.5	1122.3	474.9	444.2	203.3
Chalcone synthase	chalcone	4546.4	4439.1	6871.8	7672.3	11059.4	7784.6	4565.6	5304.4	9541.2	7902.9	8363.3
Flavenol synthase 1	flavenoid	67.0	22.0	194.2	232.3	93.9	90.8	735.4	266.2	149.2	64.9	278.0
farnesyltransferase A	terpene; farnesene	311.9	240.3	310.9	276.8	260.4	349.8	399.0	298.4	312.6	260.2	309.3
geranylgeranyltransferase type I beta subunit	terpene	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
total-pinoresinol reductase 1	isosalicin; tremuloidin; phenolic glycoside	96.3	80.7	35.6	52.0	63.8	31.0	50.9	52.6	51.3	97.1	122.3
geranylgeranyl transferase alpha subunit	terpene	920.1	1208.0	179.4	151.3	182.3	158.1	160.2	146.9	221.4	191.4	190.8
RAB geranylgeranyl transferase beta subunit	terpene	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
flavonol synthase/flavanone 3-hydroxylase	flavenoid; tremulacin	504.9	890.9	1047.8	1177.2	953.6	919.8	726.0	792.9	902.5	735.4	583.2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0185.v1

terpene synthase 03	terpenoid; beta-ocimene; beta-pinene; farnesene; isoprene	1585.0	1174.6	3570.8	674.7	2517.7	240.8	1689.0	1800.2	1677.8	995.6	2131.5
terpene synthase 14	terpene; linalool	3.5	6.6	11.0	29.6	8.2	104.1	58.8	100.1	23.1	52.9	8.1
terpene synthase 21	terpene; sesquiterpene	152.8	151.7	6.2	134.0	110.7	257.6	525.9	486.8	4.9	244.9	223.5
pinene synthase	terpene; pinene	0.2	6.4	0.2	5.1	3.8	0.2	0.7	20.8	8.4	142.3	17.3
oxidosqualene cyclase	squalene	12.9	23.6	26.9	55.6	24.7	39.3	43.4	30.2	5.2	8.4	35.7
coniferyl aldehyde 5-hydroxylase	kaempferol-3-O-glucoside; prunin	623.8	426.8	818.1	1206.8	698.6	932.8	408.0	907.4	513.5	782.5	756.9
dihydroflavonol 4-reductase-like1	isosalicin	1.2	0.2	1.7	6.0	2.0	7.8	0.3	0.3	3.0	7.3	48.3
squalene synthase	squalene	757.6	551.0	538.0	593.1	530.4	484.4	401.5	460.9	561.6	640.5	647.7
terpene synthase 04	terpene; linalool	0.8	1.0	0.3	2.4	1.7	6.2	32.6	31.7	0.1	0.2	0.4
geranylgeranyl diphosphate synthase	terpene	0.7	0.1	2.8	2.4	1.8	0.6	2.5	2.4	12.4	17.0	1.7
geranylgeranyl pyrophosphate synthase	terpene	278.6	231.0	330.4	374.7	344.5	456.6	173.4	318.9	249.9	386.1	210.4
geranylgeranyl reductase	terpene	587.9	669.0	727.6	674.7	623.1	642.5	552.4	625.8	637.2	668.0	612.4
geranyl diphosphate synthase	terpene	0.0	0.1	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.1	0.2
phytoene desaturation 1	phytoene	824.3	821.2	867.2	1074.0	802.3	747.3	975.1	1323.1	730.1	693.0	589.3
UDP-glucose flavonoid 3-O-glucosyltransferase	phenolic glycoside	476.7	1083.2	476.2	962.7	1836.6	251.0	278.6	286.7	304.7	218.7	150.3
total-UDP-glucose:flavonoid 7-O-glucosyltransferase	phenolic glycoside	84.9	146.1	88.9	148.9	38.6	61.9	74.7	275.2	23.8	7.8	23.8
phytoene desaturase 3	phytoene	1.9	2.9	4.7	1.5	4.8	1.2	2.2	2.0	1.5	0.9	2.3
phytoene synthase	phytoene	397.7	286.4	377.8	244.5	348.0	165.6	260.6	311.1	551.2	207.9	349.0
squalene monooxygenase	squalene	662.1	593.1	744.8	1432.8	895.5	1001.7	877.1	957.9	595.6	749.2	927.6
squalene epoxidase 3	squalene	0.7	2.4	6.3	1.5	7.9	3.9	6.2	4.6	0.4	1.4	4.3
terpene synthase 02	terpene	6.2	20.6	1.1	1.2	1.5	35.1	15.3	36.9	1.3	10.4	0.5

 doi:10.20944/preprints202212.0185.v1

terpene synthase 10	terpene	0.0	0.7	0.2	0.1	0.4	3.9	2.3	5.2	0.7	1.4	0.0
total-UDP-glucose flavonoid 3-O-glucosyltransferase	phenolic glycoside; prunin	5934.47	4744.3	3697.9	2115.6	3210.1	2691.9	2452.7	2662.8	2215.2	1833.0	3496.3
UDP-sugar flavonoid 7-O-glycosyltransferase	phenolic glycoside	1157.9	810.6	163.3	136.6	137.7	0.1	0.1	0.1	689.1	523.7	0.1
Chalcone-flavanone isomerase	chalcone; flavenoid; unknown phenolic glucoside	2245.2	2000.8	1896.4	2788.5	3661.5	2921.6	2528.5	2955.5	2661.6	3001.0	2918.6
flavonol synthase/flavanone 3-hydroxylase	flavenoid	306.3	387.1	287.7	539.4	367.0	766.6	352.8	391.4	624.6	872.0	1087.9
dihydroflavonol 4-reductase	flavenoid; benzeneacetaldehyde	568.3	587.6	781.3	662.9	820.7	774.9	591.7	557.7	190.9	244.3	186.7
UDP-N-acetylglucosamine transferase subunit alg13	kaempferol-3-O-glucoside	51.8	100.9	53.3	79.7	60.9	77.5	62.5	104.0	84.7	94.3	75.9
geranylgeranyl diphosphate reductase	terpene	3681.37	3066.8	3119.7	2853.1	3958.5	2283.5	1871.1	3363.7	2337.3	3370.2	2694.4
beta-pinene synthase	terpene; beta-pinene	0.8	8.2	97.5	21.1	98.7	230.0	94.0	381.2	41.8	187.2	75.1
benzoyl coenzyme A: benzyl alcohol benzoyl transferase	unknown glucoside	1894.17	1181.8	739.3	390.2	392.4	422.6	323.3	957.0	250.6	185.5	101.8

2.4. P336 crosses and progeny

All progeny in the eight families generated with P336 as the female parent, including F_1 progeny and second-generation progeny, were female, with over 75% of plants flowering in each family at the time of data collection (Table 6).

Table 6. Summary of families generated with P336 as the mother and maternal grandmother and resulting scores of sex on the progeny.

Family ID	Mother	Maternal Species	Father	Paternal Species	Progeny	Percent Flowering	Percent female
13X-426	P336	*S. integra*	94001	*S. purpurea*	284	98%	100%
20X-565	P336	*S. integra*	Fish Creek	*S. purpurea*	210	75%	100%
20X-564	P336	*S. integra*	94003	*S. purpurea*	252	77%	100%
20X-278	P336	*S. integra*	P63	*S. suchowensis*	212	98%	100%
20X-567	P336	*S. integra*	04-FF-016	*S. koriyanagi*	208	97%	100%
20X-566	P336	*S. integra*	04-BN-051	*S. udensis*	204	76%	100%
14X-454	05X-278-071	*S. integra* × *S. suchowensis*	94001	*S. purpurea*	94	88%	100%

 doi:10.20944/preprints202212.0185.v1

14X-456	05X-278-071	*S. integra* × *S. suchowensis*	P63	*S. suchowensis*	166	90%	100%

3. Discussion

3.1. Assemblies and annotations

The high quality of the assemblies (BUSCO > 95%) as well as the number of genes in the current annotation represents an advancement in *Salix* genomic resources, including comprehensive comparative genome analysis across the shrub willows. Across all 11 annotations, there are several thousand gene models from *S. purpurea* 94006 v5.1 that are missing. This is also reflected by relatively low BUSCO scores of less than 90% in most annotations (Table 2). The RNA-Seq data used to perform the annotations did not contain any floral tissue, nor any tissue from drought, disease, or insect stressed plants, which can explain the missing gene models, as genes from these biological conditions were not expressed in our dataset and therefore were not annotated.

3.2. Sex determination genes and SDR assembly

The reported assemblies each include one haplotype of Chr15 per genome: Chr15Z in the male assemblies and Chr15W in the females. Together, these include separate fully assembled 15Z and 15W chromosomes for *S. purpurea, S. suchowensis, S. koriyanagi*, and *S. viminalis*, Chr15W for *S. integra* and Chr15Z for *S. udensis*. This is the first report of a fully assembled Chr15Z for both *S. suchowensis* and *S. viminalis* [24,25]. The Chr15 assemblies across the 11 genomes showed substantial differences in structural arrangement (Fig. S3). The observed structural differences may be due in part to errors in assembly rather than true structural variation between genotypes, particularly since the order of sequences in the reassembly of *S. purpurea* 94006 Chr15W differs from the JGI *S. purpurea* 94006 v5.1 assembly. Sex determination regions are notoriously difficult to assemble due to highly repetitive regions resulting from a lack of recombination, and such differences in arrangement of contigs into the final scaffolded sex chromosomes are not unexpected [29]. Nevertheless, despite structural differences, the Chr15 appears to be fully intact across every assembly.

BLASTN results for candidate sex-determination genes revealed substantial variation in gene content within the Chr15W SDR (Table 3). In *S. purpurea* 94006, the sex-determination gene content closely matches the JGI *S. purpurea* 94006 v5.1 reference genome. Only three copies of *ARR17* were identified on Chr15 instead of four, five copies of *DRB1* instead of two, and *GATA15* was located on Chr17 instead of Chr15, however, these differences in gene copy number between assemblies could be the result of errors in assembly within the Chr15 in either reference. In the case of the missing fourth *ARR17*, this gene is located within a series of four palindromic repeats, and due to their repetitive nature, the fourth arm could have been lost during haplotig purging. When searching the purged contigs, an additional *ARR17* was identified, which is likely this fourth Chr15 copy. In the case of *GATA15*, the Chr15W copy appears to have been assembled on Chr17. This is also likely the result of assembly error, as no Chr17 *GATA15* is present in any other *S. purpurea* genome assembly, including the JGI 94006 v1.0 and v5.1 assemblies, the 94001 assembly, 94003 assembly or the JGI 'Fish Creek' v3.1 assembly. A dotplot alignment of HiC_scaffold_7 (Chr17) from the reassembly of 94006 against the JGI 94006 v5.1 reference shows a 100 kb region of HiC_scaffold_7 that aligns to Chr15W (Fig. S4). Linkage map markers for the 94006 genotype were obtained from Wilkerson et al. (2022) and include one marker, S15_7998352, that is located in the misassembled region [2]. In a BLASTN analysis, the flanking regions of this marker align to HiC_scaffold_7 (Chr17), while the nearest markers on the 94006 linkage map, which are tightly linked, align to HiC_scaffold_15 (Chr15), confirming that this region, including *GATA15*, is indeed a Chr15W region misassembled onto Chr17.

 doi:10.20944/preprints202212.0185.v1

The *ARR17* and *GATA15* genes are absent from Chr15 in the males and present in the females of *S. purpurea* and *S. suchowensis*, consistent with the two-gene sex determination mechanism proposed by Hyden et al. [11] and suggesting a common sex determination mechanism between these two species. *ARR17* and *AGO4* are located in a series of four inverted palindromic repeats in on Chr15W in *S. purpurea* [26]. In the *S. suchowensis* and *S. integra* female genomes there are only two *ARR17* copies on Chr15 instead of four, and only one *AGO4* copy instead of three. This indicates that there are only two arms of these palindromic repeats in *S. suchowensis* and *S. integra* instead of the four observed in *S. purpurea* [26]. These palindromic repeats appear to be absent altogether in the *S. koriyanagi* and *S. viminalis* female genomes, which suggests that the palindromic repeats may have been deleted independently in *S. koriyanagi* and *S. viminalis* after the divergence of the Helix and Vimen lineages. Partial copies of the *ARR17* exon 1 are thought to have a key role in sex determination in both *Populus* and the tree willows (*Salix* section Protitea) [10] by silencing *ARR17* expression in males. BLAST results revealed exon 1 copies present on Chr15 in all the genomes in subsection Helix and none in subsection Vimen, suggesting that these partial repeats were likely lost when Vimen diverged from Helix (Table 3). The copy number variation of *DRB1* and the three hypothetical proteins across the genomes is inconsistent with the current model of sex determination and does not support a role of these genes in sex determination, as previously proposed for *S. purpurea* [12]. Of particular interest is the lack of *ARR17* or *GATA15* homologs on Chr15 in the *S. koriyanagi* and *S. viminalis* female genomes. The missing *ARR17* in *S. viminalis* is inconsistent with earlier studies on *S. viminalis* by Hallingback et al. [24] and Almeidia et al. [30], which both identified one copy of *ARR17* on the *S. viminalis* Chr15W. The differing number of candidate sex determination genes between species, particularly *ARR17* and *GATA15*, indicates that the mechanism of sex determination may be quite labile within the Vetrix lineage of willows, despite its apparent conservation within the tree willows and the poplars.

3.3. Secondary metabolism genes

Across most genomes, the copy number of annotated secondary metabolism genes shows little variation, with a few notable exceptions. *Salix suchowensis* P294 exhibited an exceptionally high copy number of several gene families, including flavenol synthase 1, terpene synthase 21 (involved in sesquiterpene synthesis), coniferyl aldehyde 5-hydroxylase (associated with kaempferol-3-O-glucoside and prunin variation [19]), and UDP-glucose flavonoid 3-O-glucosyltransferase (Table 4). This abundance of gene annotations in P294 warrants further investigation into this particular genotype and its progeny for secondary metabolite abundance and its relationship to pollinator and pest attraction. Some other notable copy number variations between genomes included nine chalcone synthase genes in *S. viminalis* 07-MBG-5027 (12 in *S. purpurea* 94006), two copies of phytoene desaturase 1 in all three *S. suchowensis* (1 in *S. purpurea* 94006), 20 copies of squalene monooxygenase in *S. suchowensis* P63, and 31 copies of UDP-glucose flavonoid 3-O-glucosyltransferase in *S. purpurea* 94001 [19] (Table 4).

FPKM normalized expression results from all eight tissue types mapped to the *S. purpurea* 94006 v5.1 reference showed substantial variation in expression for secondary metabolite gene families (Table 4). Sapur.019G055800, a 4-coumarate:CoA ligase, was previously found to be associated with phenolic glucoside production in *S. purpurea* [19]. However, both *S. purpurea* genomes had the lowest relative expression of this gene, while expression was nearly five-fold greater in both *S. koriyanagi* genotypes. *S. koriyanagi* 04-FF-016 also showed exceptionally high expression of the arogenate/prephenate dehydratase gene family, which has been predicted to be associated with prunin and isosalicin production in *S. purpurea* [19]. *S. suchowensis* P63 exhibited the greatest expression of terpene synthase 03 family genes, which are associated with numerous terpenoids including beta-ocimene, beta-pinene, farnesene, and isoprene, while *S. suchowensis* P294 exhibited the greatest expression of coniferyl aldehyde 5-hydroxylase genes associated with prunin

 doi:10.20944/preprints202212.0185.v1

and kaempferol-3-O-glucoside [19]. These findings suggest that further research is warranted into these genotypes to understand differences in secondary metabolite concentrations and the effects they may have on pollinator and pest attraction.

3.4. P336 crosses and progeny

Across all eight crosses generated with *S. integra* P336 as a parent or grandparent, 100% of the progeny were female. Notably, when a (*S. integra* P336 × *S. suchowensis* P63) F_1 female was backcrossed to *S. suchowensis* P63 and also crossed with 94001, all of the progeny were again female. This is interesting as it suggests that all-female inheritance persists across multiple generations, despite independent assortment and recombination of autosomes. The most likely cause of such a sex bias persisting after more than one generation is the cytoplasmic inheritance of a "male killer" allele on either the chloroplast or mitochondrial genome from *S. integra* P336, such that only female gametes survive. Alternatively, ZZ progeny may survive, but a cytoplasmic factor may result in a female phenotype regardless of the state of the sex chromosomes. One likely candidate for such a factor for either of these two mechanisms is the *RPL10* gene, which was identified in every mitochondrial genome except *S. integra* P336 and *S. viminalis* 'Jorr' [31]. The absence of this gene is particularly striking, as its presence in the mitochondrial genome is broadly conserved across plant taxa, including gymnosperms and non-flowering plants [32]. *RPL10* encodes a protein that is a component of the 80S ribosome and plays a role in plant development and protein translation under UV-B stress, as well as antiviral signaling [33,34]. In Arabidopsis, *RPL10C* has also been found to be expressed only in pollen grains, and *RPL10A* has impaired transmission in male gametophytes when either *RPL10B* or *RPL10C* are mutated [35]. The absence of *RPL10* from the *S. integra* P336 mitochondria and, therefore, all of its descendants, as well as this gene's known role in plant and male gametophyte development, presents a compelling case for the absence of *RPL10* as the most likely explanation for the all-female bias observed in the progeny of *S. integra* P336.

4. Materials and Methods

4.1. DNA Sequencing

Fresh young leaf tissue (approximately 100 mg) for all 11 *Salix* genotypes was collected and ground in liquid nitrogen using the Qiagen TissueLyser II with one 5 mm stainless steel bead. DNA extraction was performed using a modified CTAB based protocol [36]. Briefly, the organic and aqueous phase were extracted using chloroform:isoamyl alcohol 24:1. After separation, a SPRI bead solution was used to select for reads greater than 1 kb [37]. For long read sequencing, 1 μg of DNA was used as input to Oxford Nanopore's genomic DNA by ligation sequencing kit (SQK-LSK109) and the subsequent library was sequenced on a R.9.4.1 flow cell. Short read sequencing of the same samples was performed on the Illumina HiSeq X Ten platform.

4.2. RNA Sequencing

RNA was extracted from eight tissues (root, xylem, internode, node, young leaf, mature leaf, petiole, and young stem) for all 11 genotypes, as well as fasciated shoot tissue from 04-BN-051, following the protocol described in Zhang et al. 2018 [38]. Strand-specific RNA-Seq libraries were prepared by BGI and sequenced on the DNB-Seq platform, which generated paired-end 150 bp reads. The same RNA preps from mature leaves and roots were also sequenced on the Oxford Nanopore MinION platform, with the exception of 'Jorr', which failed quality control. The SQK-PCB109 PCR-based cDNA library kit was used to generate sequencing libraries for leaf and root tissue for all 11 genotypes and were sequenced on R.9.4.1 flow cells.

4.3. Hi-C library preparation

 doi:10.20944/preprints202212.0185.v1

Hi-C libraries were prepared with the Phase Genomics Proximo Plant Hi-C kit (Phase Genomics, Seattle, USA). Hi-C libraries were sequenced on the Illumina NovaSeq 6000 instrument which generated paired-end 150 bp reads. The sequencing data of each Hi-C library underwent quality control with the phase genomics hic_qc.py script (https://github.com/phasegenomics/hic_qc; last accessed 15 November 2021) to ensure a sufficient number of informative Hi-C reads were present in each library. Hi-C heatmaps are shown in Fig. S5.

4.4. Genome Assembly

Assembly was performed with Oxford Nanopore reads using Flye 2.8.3 [39]. Illumina short reads were mapped to the assembled contigs with BWA-MEM [40]. Pilon and a custom python script were used to generate the corrected draft assembly with the Illumina data (Fig. S6) [41]. Assembled contigs were scaffolded using Hi-C reads with Falcon [42] and Juicer Hi-C [43] to generate phased genome assemblies. A BUSCO search of the Eudicot core genes was performed against each assembly to assess the quality and completeness of each genome [44]. One assembly, 04-FF-016, produced two chimeric contigs, HiC scaffold_5 and HiC_scaffold_6, each spanning the entire length of several chromosomes. BLASTN analysis [45] was used to determine alignment to specific chromosomes and each chimeric contig was manually cut at the approximate site where mapping behavior became abnormal. Resulting scaffolds were appended with a letter (e.g. a, b, c, etc.) to denote their origin from the original chimeric scaffold.

4.5. Annotation

Genome annotation was performed with the LoReAn v2.5 pipeline [46], which utilized both Oxford Nanopore and Illumina RNA-Seq, along with protein models from the JGI *Populus trichocarpa* v4.1, *Populus deltoides* v2.1, and *Populus nigra* × *P. maximowiczii* v1.1 reference genome annotations obtained from Phytozome (https://phytozome-next.jgi.doe.gov; last accessed 21 March 2022) [27,47], followed by Augustus *ab initio* gene prediction [48]. BLASTN analysis was performed for each annotated transcript for every genome against the *S. purpurea* 94006 v5.1 annotation on Phytozome (https://phytozome-next.jgi.doe.gov, last accessed 6 July 2022) to identify homologous gene models [26,45]. Functional prediction of mRNAs in each annotation was performed using interproscan 5.52-86.0 [49]. The estimated number of missing genes from each annotation was determined by performing a BLAST analysis of all *S. purpurea* 94006 v5.1 CDS sequences against all annotated genes for each genome and identifying those *S. purpurea* 94006 v5.1 genes without a match in each genome. Orthofinder was used to identify unique and shared genes for each assembly, and to generate a phylogeny tree [50]. A BUSCO search of the Eudicot core genes was performed against the annotated mRNA sequences to estimate the completeness of each annotation [44].

4.6. Sex determination candidate gene analysis

BLAST analysis of candidate sex determination genes was performed using the *S. purpurea* 94006 v5.1 [26] and *P. trichocarpa* v4.1 [47] CDS sequences of the candidate sex determination genes identified in Hyden et al 2021 as the query, with each assembly as the target [12]. Analyzed candidate sex determination genes included homologs of a type C cytokinin response regulator *ARR17*, a *GATA15* transcription factor, a truncated Argonaute 4 *AGO4*, a double stranded RNA-binding protein *DRB1*, and three hypothetical proteins [12].

4.7. Secondary metabolism and rust gene analysis

Analysis of candidate secondary metabolism genes was performed by creating a customized list of *S. purpurea* 94006 v5.1 gene models, which included candidate genes iden-

 doi:10.20944/preprints202212.0185.v1

tified by Keefover-Ring et al. (2022) located in flavonoid, phenolic glucoside, and terpenoid QTL [19]. Genes with annotations in flavonoid and chalcone synthesis, terpene, sesquiterpene, squalene, and phytoene synthesis, and UDP-glucose flavonoid glucosyltransferase were also included, all of which have likely roles in terpenoid, flavonoid, and phenolic glucoside production. Results from the BLASTN analysis of annotated transcripts against the *S. purpurea* 94006 v5.1 reference were used to find the total matches in reach respective genome for genes on the customized list of *S. purpurea* secondary metabolism genes.

To analyze and compare expression of candidate genes, Illumina RNA-Seq data for each genome were mapped to the *S. purpurea* 94006 v5.1 reference using STAR 2.7.0[51], read counts were determined using featureCounts [52], and FPKM calculated using EdgeR [53]. The sum of normalized FPKM values was calculated across all tissue types sequenced within each genotype and across all genes within each gene family.

4.8. P336 crosses and progeny

To quantify female bias in progeny from the *S. integra* P336 genotype, F_1 crosses and a select set of backcrosses were attempted with clones from each male genome in this study [54]. In 2013 the 13X-426 cross was generated between P336 and 94001. In 2014, 05X-278-071, a female from a P336 x P63 cross, was crossed with 94001 and P63 to generate the 14X-454 and 14X-456 families, respectively. In 2020 P336 was crossed with *S. purpurea* 'Fish Creek' (94006 x 94001), a monoecious *S. purpurea* 94003 [11], P63, 04-FF-016, and 04-BN-051 to generate the 20X-565, 20X-564, 20X-278, 20X-567, and 20X-566 families, respectively. A cross with 'Jorr' was also attempted, but failed to produce viable seed. Scoring for sex among the progeny was performed in April 2021.

5. Conclusions

We present 11 new *Salix* genome assemblies and annotations as a novel resource for shrub willow breeding, genetics, and genomics that will enable more accurate genetics studies of these species in the future. This is the most comprehensive genome assembly and annotation effort to date in the genus *Salix* and represents closely related diploid species which can be compared to understand the evolution of sex determination mechanisms. We used these genomes to characterize copy number variation of interesting genes relating to sex-determination and secondary metabolism that could be a driver of dioecy. We found that key sex-determination genes are missing in *S. viminalis* and *S. koriyanagi* and hypothesize a unique sex determination system exists in these species that differs from *Populus* and other *Salix* species, which further supports the dynamic nature of sex chromosome evolution in Salicaceae. We also characterized copy number variation and expression of sexually dimorphic secondary metabolite genes. Lastly, we demonstrate that *S. integra* P336 produces only female descendants and propose a missing *RPL10* gene from the mitochondrial genome as a candidate for this unusual inheritance.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: Hi-C (A) Nanopore DNA sequencing read length distributions for each genome. (B) Nanopore DNA sequencing quality score distributions.; Figure S2: Phylogenetic grouping of 11 *Salix* genome annotations; Figure S3: Dotplot alignments for the 19 largest scaffolds of each genome (y axis) against the JGI *S. purpurea* 94006 v5.1 reference genome (x axis). (A) *S. purpurea* 94006, (B) *S. purpurea* 94001, (C) *S. suchowensis* P63, (D) *S. suchowensis* P294, (E) *S. suchowensis* P295, (F) *S. integra* P336, (G) *S. koriyanagi* SH3, (H) *S. koriyanagi* 04-FF-016, (I) *S. viminalis* 07-MBG-5027, (J) *S. viminalis* 'Jorr', (K) *S. udensis* 04-BN-051; Figure S4: Dotplot alignment of HiC_scaffold_7 (Chr17) from 94006 against the Phytozome v5.1 genome showing an approximately 100kb region that aligns to Chr15 instead of Chr17.; Figure S5: Hi-C heatmap results for each genome; Figure S6: Schematic of assembly pipeline for Oxford Nanopore and Illumina DNA sequencing data; Table S1: Copy number BLAST results for candidate secondary metabolism genes against each genome

Author Contributions: Conceptualization, L.B.S., W.M., K.F., T.B.Y. and B.H.; Methodology: B.H., K.F., T.B.Y., S.J., and C.C.; Formal Analysis, B.H., K.F. and T.B.Y.; Investigation, B.H., K.F., and T.B.Y.; Resources, L.B.S. and W.M.; Data Curation, B.H. and T.B.Y.; Writing-Original Draft Preparation, B.H.; Writing-Review and Editing, B.H., K.F., T.B.Y., L.B.S. and W.M.; Visualization, B.H., K.F and T.B.Y.; Supervision, L.B.S. and W.M.; Project Administration, L.B.S. and W.M.; Funding Acquisition, B.H., L.B.S. and W.M.

Funding: This work was partially supported by the U.S. Department of Energy (DOE) Office of Science Early Career Research Program under the Biological and Environmental Research office. BH was supported by a fellowship from the DOE Office of Science Graduate Student Research (SCGSR) Program and by a Pre-doctoral Research Fellowship from the United States Department of Agriculture National Institute for Food and Agriculture (award #2021-67034-35116). Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE-AC05-00OR22725. Part of this work was performed at the Oak Ridge Leadership Computing Facility (OLCF) including resources of the Compute and Data Environment for Science (CADES) at Oak Ridge National Laboratory. Partial funding for this research was also provided by grants from the National Science Foundation (DEB-1542486) and from the USDA National Institute for Food and Agriculture (2015-67009-23957)

Data Availability Statement: All raw sequencing data have been deposited at the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra). Raw Illumina and nanopore DNA sequencing data can be accessed with the BioProject ID PRJNA827350. The raw Illumina RNA-Seq data can be accessed with the BioProject ID PRJNA827350. Nanopore RNA-Seq data can be accessed with the BioProject ID PRJNA888070. Genome assemblies and annotations have been deposited at the NCBI Genome Portal (https://www.ncbi.nlm.nih.gov/genome) with the BioProject IDs PRJNA890276, PRJNA892589, PRJNA892593, PRJNA892594, PRJNA892596, PRJNA892597, PRJNA892598, PRJNA892599, PRJNA892600, PRJNA892601, and PRJNA892602. Protein fasta and information files with interproscan and *S. purpurea* 94006 v5.1 BLAST results for each annotation are available on the Willowpedia github (https://github.com/Willowpedia). Genome assemblies, annotations, and annotation information files are available on Dryad at DOI: 10.5061/dryad.5hqbzkh9f (accessible pre-publication privately using the following link https://datadryad.org/stash/share/DsDY7HsWE_Goypm6YcQZEiAKnJykGYTVie7cw2U4rg).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kuzovkina, Y.A.; Weih, M.; Romero, M.A.; Charles, J.; Hust, S.; McIvor, I.; Karp, A.; Trybush, S.; Labrecque, M.; Teodorescu, T.I. Salix: botany and global horticulture. *Horticultural Reviews* **2007**, *34*, 447-489.
2. Wilkerson, D.G.; Taskiran, B.; Carlson, C.H.; Smart, L.B. Mapping the sex determination region in the *Salix* F1 hybrid common parent population confirms a ZW system in six diverse species. *G3* **2022**, *12*, jkac071.
3. Zhou, R.; Macaya-Sanz, D.; Rodgers-Melnick, E.; Carlson, C.H.; Gouker, F.E.; Evans, L.M.; Schmutz, J.; Jenkins, J.W.; Yan, J.; Tuskan, G.A.; et al. Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). *Molecular Genetics and Genomics* **2018**, *293*, 1437-1452.
4. Pucholt, P.; Rönnberg-Wästljung, A.-C.; Berlin, S. Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity* **2015**, *114*, 575-583.
5. Chen, Y.; Wang, T.; Fang, L.; Li, X.; Yin, T. Confirmation of single-locus sex determination and female heterogamety in willow based on linkage analysis. *PLoS One* **2016**, *11*, e0147671.
6. Wang, D.; Li, Y.; Li, M.; Yang, W.; Ma, X.; Zhang, L.; Wang, Y.; Feng, Y.; Zhang, Y.; Zhou, R.; et al. Repeated turnovers keep sex chromosomes young in willows. *Genome Biology* **2022**, *23*, 200.
7. Sanderson, B.J.; Feng, G.; Hu, N.; Carlson, C.H.; Smart, L.B.; Keefover-Ring, K.; Yin, T.; Ma, T.; Liu, J.; DiFazio, S.P. Sex determination through X–Y heterogamety in *Salix nigra*. *Heredity* **2021**, *126*, 630-639.
8. He, L.; Jia, K.H.; Zhang, R.G.; Wang, Y.; Shi, T.L.; Li, Z.C.; Zeng, S.W.; Cai, X.J.; Wagner, N.D.; Hörandl, E.; et al. Chromosome-scale assembly of the genome of *Salix dunnii* reveals a male-heterogametic sex determination system on chromosome 7. *Molecular Ecology Resources* **2021**, *21*, 1966-1982.
9. Yang, W.; Wang, D.; Li, Y.; Zhang, Z.; Tong, S.; Li, M.; Zhang, X.; Zhang, L.; Ren, L.; Ma, X.; et al. A general model to explain repeated turnovers of sex determination in the Salicaceae. *Molecular Biology and Evolution* **2020**, *38*, 968-980.
10. Cronk, Q.; Müller, N.A. Default Sex and Single Gene Sex Determination in Dioecious Plants. *Frontiers in Plant Science* **2020**, *11*.
11. Hyden, B.; Zou, J.; Wilkerson, D.G.; Carlson, C.H.; Rivera Robles, A.; DiFazio, S.; Smart, L.B. Structural variation of a sex-linked region confers monoecy and implicates GATA15 as a master regulator of sex in *Salix purpurea*. **2022**. *New Phytologist (submitted, in review)*
12. Hyden, B.; Carlson, C.H.; Gouker, F.E.; Schmutz, J.; Barry, K.; Lipzen, A.; Sharma, A.; Sandor, L.; Tuskan, G.A.; Feng, G.; et al. Integrative genomics reveals paths to sex dimorphism in *Salix purpurea* L. *Horticulture Research* **2021**, *8*, 170.
13. Smart, L.B.; Cameron, K.D. Genetic improvement of willow (*Salix* spp.) as a dedicated bioenergy crop. In *Genetic Improvement of Bioenergy Crops*; Springer: 2008; pp. 377-396.
14. Argus, G.W. Infrageneric classification of *Salix* (Salicaceae) in the new world. *Systematic Botany Monographs* **1997**, 1-121.
15. Newsholme, C. *Willows: the genus Salix*; Timber Press, Inc.: 1992.
16. Fussel, U.; Dotterl, S.; Jurgens, A.; Aas, G. Inter- and intraspecific variation in floral scent in the genus *Salix* and its implication for pollination. *Journal of Chemical Ecology* **2007**, *33*, 749-765.
17. Mosseler, A.; Major, J.; Ostaff, D.; Ascher, J. Bee foraging preferences on three willow (*Salix*) species: Effects of species, plant sex, sampling day and time of day. *Annals of Applied Biology* **2020**, *177*, 333-345.
18. Sanderson, B.J.; Wang, L.; Tiffin, P.; Wu, Z.; Olson, M.S. Sex-biased gene expression in flowers, but not leaves, reveals secondary sexual dimorphism in *Populus balsamifera*. *New Phytol* **2019**, *221*, 527-539.
19. Keefover-Ring, K.; Carlson, C.H.; Hyden, B.; Azeem, M.; Smart, L.B. Genetic mapping of sexually dimorphic volatile and non-volatile floral secondary chemistry of a dioecious willow. *Journal of Experimental Botany* **2022**.
20. Foley, W.J.; Moore, B.D. Plant secondary metabolites and vertebrate herbivores–from physiological regulation to ecosystem function. *Current Opinion in Plant Biology* **2005**, *8*, 430-435.

21. Wink, M. Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theoretical and Applied Genetics* **1988**, *75*, 225-233.
22. Carlson, C.H.; Gouker, F.E.; Crowell, C.R.; Evans, L.; DiFazio, S.P.; Smart, C.D.; Smart, L.B. Joint linkage and association mapping of complex traits in shrub willow (Salix purpurea L.). *Annals of Botany* **2019**, *124*, 701-716.
23. Wilkerson, D.G.; Crowell, C.R.; Carlson, C.H.; McMullen, P.W.; Smart, C.D.; Smart, L.B. Comparative transcriptomics and eQTL mapping of response to *Melampsora* americana in selected *Salix purpurea* F2 progeny. *BMC genomics* **2022**, *23*, 1-14.
24. Almeida, P.; Proux-Wera, E.; Churcher, A.; Soler, L.; Dainat, J.; Pucholt, P.; Nordlund, J.; Martin, T.; Rönnberg-Wästljung, A.-C.; Nystedt, B. Genome assembly of the basket willow, *Salix viminalis*, reveals earliest stages of sex chromosome expansion. *BMC biology* **2020**, *18*, 1-18.
25. Wei, S.; Yang, Y.; Yin, T. The chromosome-scale assembly of the willow genome provides insight into Salicaceae genome evolution. *Horticulture research* **2020**, *7*.
26. Zhou, R.; Macaya-Sanz, D.; Carlson, C.H.; Schmutz, J.; Jenkins, J.W.; Kudrna, D.; Sharma, A.; Sandor, L.; Shu, S.; Barry, K.; et al. A willow sex chromosome reveals convergent evolution of complex palindromic repeats. *Genome Biology* **2020**, *21*, 38.
27. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **2011**, *40*.
28. Müller, N.A.; Kersten, B.; Leite Montalvao, A.P.; Mahler, N.; Bernhardsson, C.; Brautigam, K.; Carracedo Lorenzo, Z.; Hoenicka, H.; Kumar, V.; Mader, M.; et al. A single gene underlies the dynamic evolution of poplar sex determination. *Nature Plants* **2020**, *6*, 630-637.
29. Webster, T.H.; Couse, M.; Grande, B.M.; Karlins, E.; Phung, T.N.; Richmond, P.A.; Whitford, W.; Wilson, M.A. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *Gigascience* **2019**, *8*, giz074.
30. Hallingbäck, H.R.; Pucholt, P.; Ingvarsson, P.K.; Rönnberg-Wästljung, A.C.; Berlin, S. Genome-wide association mapping uncovers sex-associated copy number variation markers and female hemizygous regions on the W chromosome in *Salix viminalis*. *BMC genomics* **2021**, *22*, 1-14.
31. Yates, T. Genome evolution in the salicaceae: genetic novelty, horizontal gene transfer, and comparative genomics. Ph.D. dissertation, University of Tennessee, Knoxville, TN, 8-2022.
32. Mower, J.P.; Bonen, L. Ribosomal protein L10 is encoded in the mitochondrial genome of many land plants and green algae. *BMC Evolutionary Biology* **2009**, *9*, 265.
33. Ferreyra, M.L.F.; Pezza, A.; Biarc, J.; Burlingame, A.L.; Casati, P. Plant L10 Ribosomal Proteins Have Different Roles during Development and Translation under Ultraviolet-B Stress. *Plant Physiology* **2010**, *153*, 1878-1894.
34. Rocha, C.S.; Santos, A.A.; Machado, J.P.B.; Fontes, E.P. The ribosomal protein L10/QM-like protein is a component of the NIK-mediated antiviral signaling. *Virology* **2008**, *380*, 165-169.
35. Falcone Ferreyra, M.L.; Casadevall, R.; Luciani, M.D.; Pezza, A.; Casati, P. New evidence for differential roles of L10 ribosomal proteins from *Arabidopsis*. *Plant Physiology* **2013**, *163*, 378-391.
36. Doyle, J.J.; Doyle, J.L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **1987**, *19*, 11-15.
37. Mayjonade, B.; Gouzy, J.; Donnadieu, C.; Pouilly, N.; Marande, W.; Callot, C.; Langlade, N.; Muños, S. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **2016**, *61*, 203-205.
38. Zhang, J.; Yang, Y.; Zheng, K.; Xie, M.; Feng, K.; Jawdy, S.S.; Gunter, L.E.; Ranjan, P.; Singan, V.R.; Engle, N.; et al. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist* **2018**, *220*, 502-516.
39. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **2019**, *37*, 540-546.
40. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* **2013**.
41. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **2014**, *9*, e112963.
42. Kronenberg, Z.N.; Rhie, A.; Koren, S.; Concepcion, G.T.; Peluso, P.; Munson, K.M.; Porubsky, D.; Kuhn, K.; Mueller, K.A.; Low, W.Y. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications* **2021**, *12*, 1-10.
43. Durand, N.C.; Shamim, M.S.; Machol, I.; Rao, S.S.; Huntley, M.H.; Lander, E.S.; Aiden, E.L. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **2016**, *3*, 95-98.
44. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210-3212.
45. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Bioogyl* **1990**, *215*, 403-410
46. Cook, D.E.; Valle-Inclan, J.E.; Pajoro, A.; Rovenich, H.; Thomma, B.P.H.J.; Faino, L. Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing. *Plant Physiology* **2018**, *179*, 38-54.

47. Tuskan, G.A.; DiFazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A.; et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **2006**, *313*, 1596-1604.
48. Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **2006**, *34*, W435-439.
49. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **2021**, *49*, D344-D354.
50. Emms, D.M.; Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **2019**, *20*, 1-14.
51. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15-21.
52. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923-930.
53. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139-140.
54. Kopp, R.F.; Maynard, C.A.; Rocha de Niella, P.; Smart, L.B.; Abrahamson, L.P. Collection and storage of pollen from *Salix* (Salicaceae). *American Journal of Botany* **2002**, *89*, 248-252.