RESEARCH ARTICLE

American Society of Plant Biologists · SOCIETY FOR EXPERIMENTAL BIOLOGY · WILEY

Functional analysis of *Salix purpurea* genes support roles for *ARR17* and *GATA15* as master regulators of sex determination

Brennan Hyden[1,2] | Dana L. Carper[2] | Paul E. Abraham[2] | Guoliang Yuan[2] | Tao Yao[2] | Leo Baumgart[3] | Yu Zhang[3] | Cindy Chen[3] | Ronan O'Malley[3] | Jin-Gui Chen[2] | Xiaohan Yang[2] | Robert L. Hettich[2] | Gerald A. Tuskan[2] | Lawrence B. Smart[1]

[1]Horticulture Section, School of Integrative Plant Science, Cornell University, Cornell AgriTech, Geneva, New York, USA

[2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

[3]Lawrence Berkeley National Laboratory, US Department of Energy Joint Genome Institute, Berkeley, California, USA

Correspondence
Lawrence B. Smart, Horticulture Section, School of Integrative Plant Science, Cornell University, Cornell AgriTech, Geneva, NY 14456, USA.
Email: lbs33@cornell.edu

Abstract

The Salicaceae family is of growing interest in the study of dioecy in plants because the sex determination region (SDR) has been shown to be highly dynamic, with differing locations and heterogametic systems between species. Without the ability to transform and regenerate *Salix* in tissue culture, previous studies investigating the mechanisms regulating sex in the genus *Salix* have been limited to genome resequencing and differential gene expression, which are mostly descriptive in nature, and functional validation of candidate sex determination genes has not yet been conducted. Here, we used Arabidopsis to functionally characterize a suite of previously identified candidate genes involved in sex determination and sex dimorphism in the bioenergy shrub willow *Salix purpurea*. Six candidate master regulator genes for sex determination were heterologously expressed in Arabidopsis, followed by floral proteome analysis. In addition, 11 transcription factors with predicted roles in mediating sex dimorphism downstream of the SDR were tested using DAP-Seq in both male and female *S. purpurea* DNA. The results of this study provide further evidence to support models for the roles of *ARR17* and *GATA15* as master regulator genes of sex determination in *S. purpurea*, contributing to a regulatory system that is notably different from that of its sister genus *Populus*. Evidence was also obtained for the roles of two transcription factors, an *AP2/ERF* family gene and a homeodomain-like transcription factor, in downstream regulation of sex dimorphism.

KEYWORDS
Arabidopsis, dioecy, proteomics, *Salix*, sex determination, transgenics, willow

1 | INTRODUCTION

Understanding the genetic regulation of sex determination in dioecious plants is of interest in the plant biology community because, while dioecy is only observed in about 6% of angiosperm species, it is present in numerous taxa and families and is thought to have evolved independently on as many as 5,000 occasions (Käfer et al., 2017). As such, understanding the genetic mechanisms that lead to separation of sexes in different taxa can provide insight into the repeated evolution and maintenance of dioecy. The Salicaceae family is of particular interest in this effort, as nearly all species in the family are dioecious, and it contains two genera of economic importance: poplars (*Populus*) and willows (*Salix*). In particular, *Salix* contains over 300 species, which are native to every continent except Antarctica and Australia and grow in a diverse range of biomes, including subarctic tundra, deserts, and temperate and tropical forests (Argus, 1997; Kuzovkina et al., 2007). *Salix* species also exhibit a variety of growth habits, ranging from prostrate dwarfs to shrubs and trees (Argus, 1997). Despite this remarkable diversity in species range and form, dioecy has been maintained throughout the evolution of most of the family, including all *Populus* and *Salix*. Moreover, the sex determination region (SDR) in *Salix* appears to have shifted among chromosomes and varied heterogametic systems on multiple occasions, with tree willows such as *S. nigra* and *S. chaenomeloides* containing SDR on Chr07 with an XY system (Sanderson et al., 2021; Wang et al., 2022) and alternatively on Chr15 with an XY system in *S. arbutifolia* (Wang et al., 2022). Shrub willows, including *S. purpurea* and *S. viminalis*, contain the SDR on Chr15 under a ZW system (Pucholt et al., 2015; Wilkerson et al., 2022; Zhou et al., 2018). Because of the dynamic nature of sex determination in Salicaceae, there is an opportunity to characterize the precise mechanisms of sex determination in diverse species across this family and to add to our understanding of the evolution, conservation, and transition of the SDR.

Much work has already been done to identify candidate master regulator genes in both *Populus* and *Salix*. Muller et al. (2020) demonstrated that a homolog of *ARR17*, an Arabidopsis type-C response regulator, is likely the sole master regulator gene in *Populus*, with females expressing *ARR17*, while in males, it is either absent or silenced by smRNA produced from exon 1 repeats located on the Y chromosome (Muller et al., 2020). A similar *ARR17*-mediated sex determination system is thought to be present in some willows with Chr07 and Chr15 XY systems but has not been confirmed through expression or functional analysis (Wang et al., 2022). *ARR17* was first proposed as a candidate master regulator gene of sex determination in *S. purpurea* (Chr15 ZW SDR) by Zhou et al. (2020), where the authors identified four inverted repeats of the gene on Chr15W in a female *S. purpurea* (Zhou et al., 2020). Another study conducted RNA-Seq and smRNA-Seq of an F_2 family of *S. purpurea* and proposed several candidate master regulator genes in addition to *ARR17*

but notably did not find evidence for the smRNA silencing mechanism in males that exists in XY *Populus* (Hyden et al., 2021). This lack of smRNA expression, along with confirmed expression of *ARR17* in both male and female *S. purpurea*, suggests that if *ARR17* is a master regulator gene in *S. purpurea*, it may operate via a mechanism that differs from *Populus* and the tree willows (Hyden et al., 2021). Based on sequencing and expression analysis, several other genes have been proposed as candidate genes for sex determination in addition to *ARR17*, including homologs of *GATA15*, *AGO4*, *DRB1*, and several hypothetical proteins (Hyden et al., 2021). However, no functional validation has been conducted to confirm the role of these proposed sex-determination genes. Recently, sequencing and expression data from a monoecious *S. purpurea* revealed a structural variant on Chr15W that includes deletions of *ARR17*, *AGO4*, and *DRB1*, but not *GATA15* (Hyden et al., 2023). Based on these data, we hypothesize that *GATA15* is a master regulator gene of sex determination that functions to promote female development, while *ARR17* acts as a suppressor of male development.

Differentiation of floral primordia in *S. purpurea* into either androecium in males, or gynoecium in females, occurs during late summer, approximately July–August in Central New York (Zhang & Fernando, 2005). Previous research on willow sex determination and dimorphism has relied primarily on the comparison of DNA and RNA sequencing data between males and females, with studies performed in both shoot tips containing floral primordia collected during the summer, and in differentiated emerging catkins collected in the spring (Carlson et al., 2017; Hyden et al., 2021; Zhou et al., 2020). However, functional validation is ultimately needed to support the hypotheses proposed in these manuscripts. Yet, unlike poplar, there is not a facile protocol for *Salix* transformation and regeneration from tissue culture, so it is not feasible to study gene function in willow using transformation-based gain/loss-of function approaches. Another layer of complexity is that, without destructive histochemical analysis, it is impossible to isolate and identify floral meristems in *Salix* during their early development stages when gynoecium and androecium first appear (Zhang & Fernando, 2005). A suitable alternative is to use a model plant species, such as Arabidopsis, for transgenic manipulation.

In shrub willow, there is considerable sex dimorphism between male and female catkins, including thousands of differentially expressed genes (Hyden et al., 2021). Transcription factors (TFs) are among the strongest candidates for mediating these morphological and gene expression differences downstream of the initial cues from the sex determination genes, and a suite of potential TFs acting as intermediates in the sex determination pathway have been previously identified through a combination of expression and eQTL analyses (Hyden et al., 2021). TFs are of particular interest for interrogating the regulatory mechanisms of sex determination, as they can be studied in vitro using DAP-Seq, which does not require plant transformation.

In DAP-Seq, TFs are transiently expressed and incubated with native genomic DNA. TF-bound DNA fragments are then sequenced and aligned to the reference genome (O'Malley et al., 2016). This system enables TF binding analysis in any species from which candidate genes can be cloned and high-quality DNA can be sequenced.

In this study, we sought to elucidate the mechanisms of sex determination and dimorphism in *S. purpurea* by investigating both master regulator genes of sex and TFs with predicted roles in sex dimorphism. Six candidate master regulator genes for sex determination in *S. purpurea* were heterologously expressed in Arabidopsis using a constitutive promoter. Bottom-up proteomics was performed on floral tissue of the transgenic Arabidopsis plants to identify proteins regulated by the overexpressed candidate master regulators of primary sex dimorphism (anther or stamen development). To characterize regulation of sex dimorphism downstream of the SDR, which includes floral development and floral secondary metabolite production, DAP-Seq was performed on 11 TFs. Nine of the TFs tested are located on autosomes and have eQTL in the *S. purpurea* SDR, and two are located in the SDR and are candidate master regulator genes of sex, for which expression in Arabidopsis and proteomic analysis was also performed (Hyden et al., 2021). Using data from these combined methods, we constructed a conceptual model for the functional role of several genes involved in sex determination and dimorphism and provide evidence to support the role of *ARR17* and *GATA15* as master regulator genes of sex determination in *S. purpurea*.

2 | RESULTS

2.1 | Proteomic analysis

Transgenic Arabidopsis floral proteomics data were used to assess the broad impact of each candidate *S. purpurea* gene on the proteome as well as changes among proteins relevant to floral development and sex determination biological processes. Eight transgenic constructs were generated (six candidate master regulator gene and two methodological control genes), each with four to six independent insertion events (Table 1), and proteome data were obtained from three full-sibling T_2 plants from each event in addition to five empty vector control plants, for a total of 122 plants evaluated. Across all 122 samples, a total of 17,191 Arabidopsis protein accessions were identified and the total proteins identified for each transgenic line ranged from 11,395 to 16,885 (Table 1). The relative impact of each heterologous transgenic line on proteome expression was compared against the empty vector control proteome data. Across the eight expression lines, the number of proteins of differential abundance (PDAs) ranged from 103 to 5,970 (Dataset S20). Enriched MapMan functional categories (Schwacke et al., 2019) were identified for all eight heterologously expressed genes to identify biological processes affected by expression of each candidate master regulator (Figures S1 and S2). As expected, the impact of each expressed gene on the floral proteome was not uniform, as shown by the variation in total peptides detected, PDAs, and MapMan enrichment categories. Plants expressing *GATA15* showed the greatest number of PDAs when compared against the empty vector control (Figure 1 and Table 1), followed by hypothetical protein Sapur.15WG074900 and *ARR17* with 5,318 and 4,305 PDAs, respectively (Figures 1 and S3 and Table 1). The positive control genes *LEAFY* and *FT* affected only 444 and 179 PDAs, respectively (Figure S3 and Table 1). The CCHC Zinc finger Sapur.15WG068800 produced the fewest number of PDAs, at only 103.

In general, most transgenic proteomes contained a substantial number of downregulated proteins when compared to the empty vector control, and this observation was particularly pronounced when expressing *ARR17*, *GATA15*, and Sapur.15WG074900 (Figure 2). Resulting PDAs across each heterologous expression line were compared to assess similarity in the resulting proteome changes. Among all the PDAs, there was considerable proteome changes observed between *ARR17*, *GATA15*, and Sapur.15WG074900 hypothetical protein lines (Figure 2). Notably, in addition to the large number of resulting PDAs, *GATA15* and *ARR17* expression lines exhibited unique protein abundance profiles of multiple floral development genes, indicating a likely role in floral and reproductive development and sex determination.

2.2 | Sapur.15WG072500 *ARR17* expression results

The *ARR17*-OX lines (pBH100) showed 4,305 PDAs, 1,236 of which had unique abundance patterns when compared to other

TABLE 1 Total, differentially abundant proteins (PDA), and unique PDA determined for each transgenic heterologous expression line in Arabidopsis.

| Construct	Expressed gene	Gene annotation	Events	Total proteins	PDA (FDR < .05)	Unique PDA
pBH100	Sapur.15WG073500	*ARR17* response regulator	6	16,684	4,305	1,236
pBH101	Sapur.15WG062800	*GATA15* transcription factor	4	16,609	5,970	1,953
pBH102	Sapur.15WG068800	CCHC zinc finger	4	16,023	103	45
pBH103	Sapur.15WG074300	*DRB1* dsRNA binding	6	12,306	343	110
pBH104	Sapur.15WG074900	Hypothetical protein	5	16,885	5,318	1,041
pBH106	Sapur.15WG075700	Hypothetical protein	4	11,355	611	137
pBH107	Sapur.15WG122200	*LEAFY*	5	12,271	444	65
pBH108	Sapur.008G061900	*Flowering locus T*	5	11,928	179	57
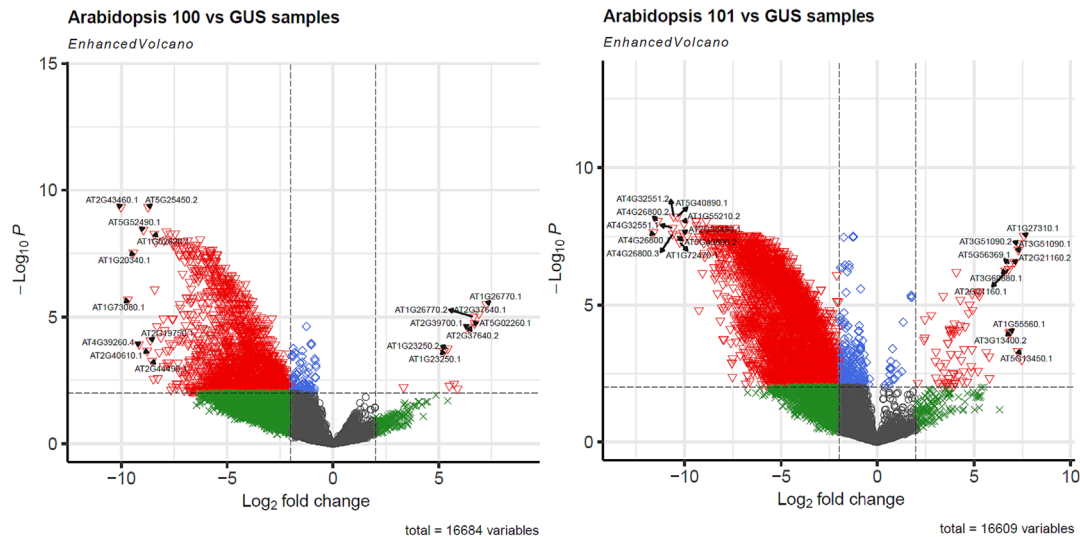
FIGURE 1 Volcano plots displaying the total proteins of differential abundance (PDA) results for the *ARR17* (100, left) and *GATA15* (101, right) expression lines, relative to the empty vector control. Gray indicates non-significant proteins, green indicates those that meet the log2 FC cutoff but not *p* value, blue indicates those that meet the *p*-value threshold but not log2 FC, and red indicates proteins that have met the log2 FC cutoff and *p*-value threshold. Identities of the most extreme PDA are indicated.

heterologously expressed candidate genes (Table 1 and Figures 1 and 2). Enriched MapMan categories included chromatin organization, coenzyme metabolism, transferase and hydrolase enzyme activity, multi-process regulation, protein homeostasis and modification, RNA biosynthesis and processing, and vesicle trafficking (Figure S1). Among the proteins showing differential abundance unique to the *ARR17*-OX lines were several with annotations related to floral development, including a homolog of *PISTILLATA* and seven genes involved in tapetum and pollen development (Table 2). Among the proteins most regulated in the *ARR17*-OX lines were multiple expansin family proteins, which have been shown to be involved in pollen tube development and cell expansion (Table S1 and Figure 1) (Liu et al., 2021).

2.3 | Sapur.15WG062800 *GATA15* expression results

The *GATA15*-OX lines (pBH101) showed the greatest number of floral PDAs by a substantial margin at 5,970, of which 1,953 had unique abundance patterns (Table 1 and Figures 1 and 2), as well as the greatest number of enriched MapMan functional categories, at 22. Among the enriched MapMan categories were cell cycle organization, transcriptional regulation, RNA modification, and vesicle transport (Figure S1). *GATA15* was one of two genes tested, the other being *ARR17*, whose expression resulted in a unique differential abundance pattern of proteins with floral development annotations, including the upregulation of *FLOR1*, a gene involved in floral meristem development and transition (Acevedo et al., 2004), and downregulation of two isoforms each of *SHATTERPROOF1* homologs and *AGL72* homologs, involved in fruit dehiscence and floral transition, respectively

(Dorca-Fornell et al., 2011; Liljegren et al., 2000) (Table 2). Among the proteins with the greatest abundance in the *GATA15*-OX lines were *SKU5*-Similar 13 and *SKU5*-Similar 14, the former of which has been shown to be essential for pollen tube growth through regulation of jasmonic acid biosynthesis (Zhang et al., 2022). However, these proteins do not appear to be uniquely upregulated by *GATA15*, as they were also upregulated in the hypothetical protein Sapur.15WG074900 transgenic lines. Among the most downregulated proteins in *GATA15*-OX lines were two isoforms of *LEUNIG*, which is involved in regulating gynoecium development (Tables 2 and S1 and Figure 1).

2.4 | Sapur.15WG068800 CCHC Zinc Finger Transcription Factor expression results

The *Sapur.15WG068800*-OX lines (pBH102) had the fewest PDAs relative to the control at only 103, of which 45 had unique protein abundance patterns (Table 1 and Figure S3). Enriched functions included phosphorylation, carrier-mediated transport, and solute channel transport (Figure S2).

2.5 | Sapur.15WG074300 *DRB1* expression results

The lines expressing *DRB1* (pBH103) had 343 PDAs relative to the empty vector control, of which 110 showed unique abundance patterns (Table 1 and Figures 2 and S3). Enriched functional categories included sucrose metabolism, chromatin structure, hydrolase enzyme activity, and MAP kinase cascade signaling (Figure S2).
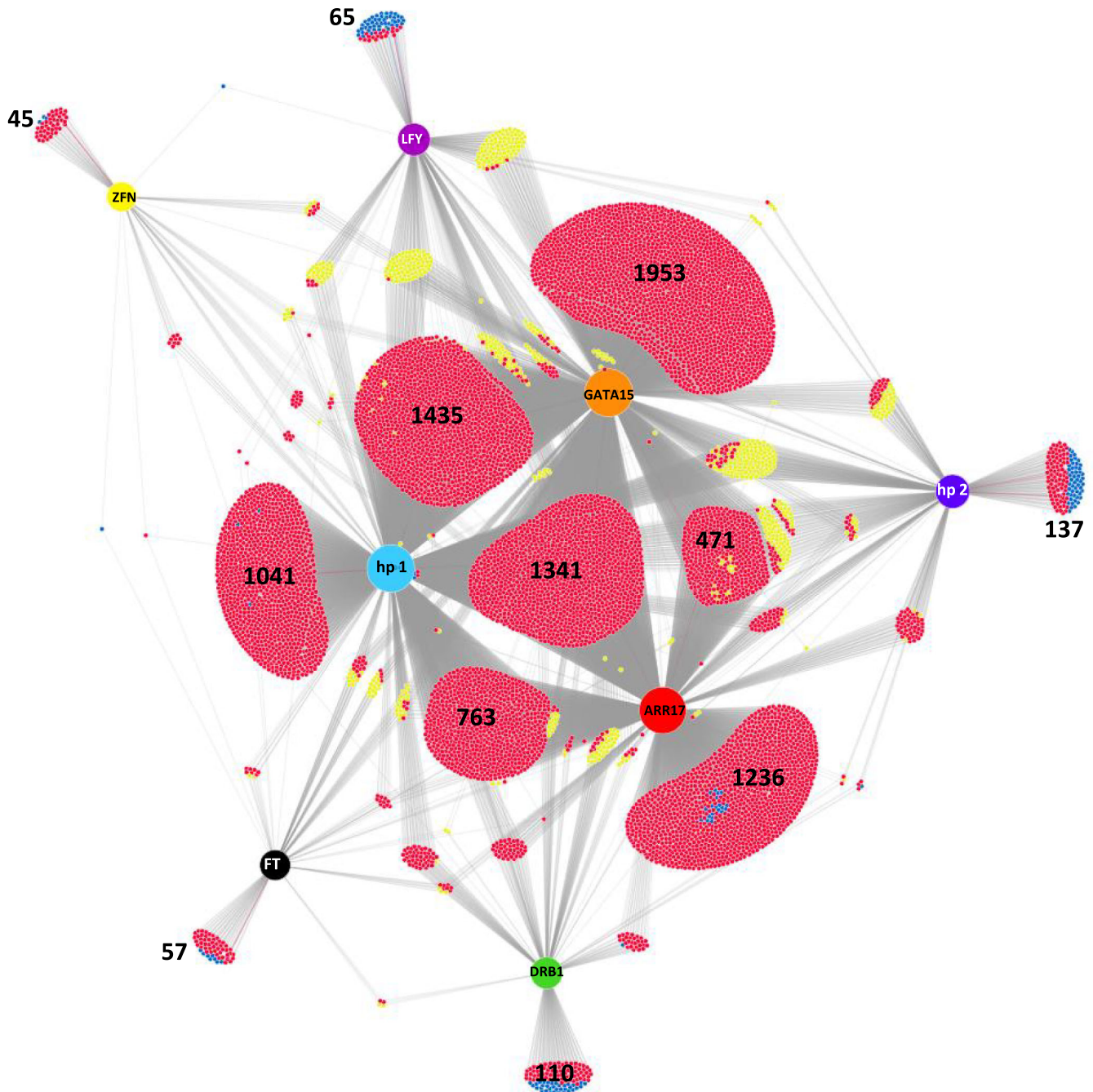
FIGURE 2 DiVenn diagram comparing proteomic expression patterns in transgenic plants heterologously expressing each of the eight *S. purpurea* genes relative to the empty vector control. Colored nodes represent expressed genes (ZFN: Sapur.15WG068800; hp 1: Sapur.15WG074900; hp 2: Sapur.15WG075700). Proteins are grouped according to unique differential abundance or shared differential abundance among the Arabidopsis lines. Relative to the empty vector control, red points represent downregulated PDAs, blue points are upregulated PDAs, and yellow points are PDAs which show differing abundance patterns between expression lines. Groupings containing unique PDAs and the largest groupings of shared PDA are labeled with the respective number of proteins in each group.

2.6 | Sapur.15WG74900 hypothetical protein expression results

In addition to *ARR17* and *GATA15*, plants expressing the hypothetical protein Sapur.15WG074900 (pBH104) also showed an exceptionally high number of PDAs at 5,318, of which 1,041 had unique abundance

patterns (Table 1 and Figures 2 and S3). Despite this large number of floral PDAs, there were not any uniquely abundant protein annotations among these that have been previously related to floral development. Sapur.15WG074900 lines also showed the second greatest number of enriched functional categories, with 20 that were significant (Figure S2), including vesicle trafficking, RNA splicing, homeostasis, and

TABLE 2 Floral and reproductive development genes showing differential expression unique to either *ARR17* or *GATA15* expression lines.

OX gene	Arabidopsis gene	log2FC	FDR	Annotation
ARR17	AT1G19890.1	5.42	2.03E−02	male-gamete-specific histone H3, *MALE-GAMETE-SPECIFIC HISTONE H3*
	AT5G51860.2	1.80	4.54E−02	*AGAMOUS*-like 72
	AT3G12145.1	−1.02	4.12E−02	*FLOR1, FLORAL TRANSITION AT THE MERISTEM*
	AT2G30800.2	−1.36	4.24E−02	Helicase in vascular tissue and tapetum
	AT5G05560.1	−1.87	4.52E−02	Pollen calcium-binding protein 1, *EMBRYO DEFECTIVE 2771*, anaphase promoting complex 1
	AT5G05560.3	−2.01	2.57E−02	Pollen calcium-binding protein 1, *EMBRYO DEFECTIVE 2771*, anaphase promoting complex 1
	AT3G11980.1	−2.75	2.98E−02	*FATTY ACID REDUCTASE 2, MALE STERILITY 2*
	AT5G20240.2	−3.02	7.83E−03	*PISTILLATA*
	AT3G10390.3	−3.03	4.59E−03	*FLOWERING LOCUS D*, reduced systemic immunity 1
	AT1G67990.1	−3.28	3.33E−02	*TAPETUM-SPECIFIC METHYLTRANSFERASE 1*
	AT3G10390.2	−3.37	2.55E−04	*FLOWERING LOCUS D*, reduced systemic immunity 1
	AT3G10390.1	−3.54	1.56E−04	*FLOWERING LOCUS D*, reduced systemic immunity 1
	AT3G10390.4	−3.54	3.72E−04	*FLOWERING LOCUS D*, reduced systemic immunity 1
	AT1G25260.1	−6.50	1.43E−08	*REDUCED POLLEN NUMBER 1, REDUCED POLLEN NUMBER*
GATA15	AT3G12145.1	.79	4.07E−04	*FLOR1, FLORAL TRANSITION AT THE MERISTEM*
	AT4G29010.1	−1.31	4.93E−02	*ABNORMAL INFLORESCENCE MERISTEM*
	AT3G58780.3	−1.37	1.76E−02	*SHATTERPROOF 1, AGAMOUS*-like 1
	AT3G58780.2	−1.42	2.64E−02	*SHATTERPROOF 1, AGAMOUS*-like 1
	AT5G51860.1	−2.43	4.76E−03	*AGAMOUS*-like 72
	AT5G51860.2	−2.45	1.37E−03	*AGAMOUS*-like 72
	AT5G40260.1	−6.42	1.27E−07	*RUPTURED POLLEN GRAIN1*
	AT4G32551.1	−10.60	1.55E−08	*ROTUNDA2, LEUNIG*
	AT4G32551.2	−10-55	6.10E−09	*ROTUNDA2, LEUNIG*

Abbreviations: FDR, false discovery rate; log$_2$FC, log$_2$ fold-change.

modification, phosphorylation, ubiquitin-proteasome system, calcium-dependent signaling, fatty acid metabolism, and sucrose metabolism.

2.7 | Sapur.15WG075700 hypothetical protein expression results

Plants expressing the hypothetical protein Sapur.15WG075700 (pBH106) had 611 PDAs, of which 137 were unique (Table 1 and Figure S3). Microfilament network, fatty acid metabolism, protein quality control, phosphorylation, transcriptional regulation, RNA export, and primary active transport of solutes were all enriched functions in these plants (Figure S2).

2.8 | Sapur.15WG122200 *LEAFY* control expression results

A *S. purpurea* homolog of *LEAFY*, located on Chr15, was expressed as a positive experimental control (pBH107). The *LEAFY*-OX lines had 444 PDAs, of which only 65 were unique (Table 1 and Figure S3).

Among the significantly enriched functional terms were pectin, microtubular network, MAP kinase cascade signaling, and transcriptional regulation (Figure S2).

2.9 | Sapur.008G061900 *Flowering Locus T* control expression results

A *S. purpurea Flowering Locus T* (*FT*) homolog was also included as a positive experimental control (pBH108) and showed 179 PDAs, of which 57 had a unique abundance pattern (Table 1 and Figure S3). The *FT*-OX lines showed enrichment for the fewest functional categories, with only two that were significant: oxidoreductase enzymes and transcriptional regulation (Figure S2). Arabidopsis T$_2$ progeny from all five *S. purpurea Flowering Locus T* (*FT*) expression events showed early flowering phenotypes when compared to the empty vector control, producing inflorescences in just 23 days after germination (Figure S4). These observations are consistent with the Arabidopsis *FT* overexpression phenotype (Kardailsky et al., 1999) and indicated the *S. purpurea FT* homolog was both expressed and functional in Arabidopsis, which is the first such example of functional expression of an

TABLE 3 Analyzed transcription factors and libraries with total number of significant summits and target genes from DAP-Seq analysis.

Library	Salix gene ID	Description	TF source	Summit fold-change > 10	Summit fold-change > 5	Summit fold-change > 3	Total target genes
94006	Sapur.001G003600.1	*AP2/ERF* transcription factor	RT-PCR 10X-317-124 (M)	12	351	1,082	1,293
	Sapur.003G027300.1	Homeodomain-like protein	RT-PCR 11X-317-194 (F)	36	1,264	5,919	9,377
	Sapur.003G155500.1	*scarecrow*-like 3	gblock	0	0	1	7
	Sapur.004G110200.3	Transcription factor *VRN1*	gblock	0	0	7	21
	Sapur.005G077400.1	NAC transcription factor 030	gblock	0	0	17	81
	Sapur.006G140600.1	*CONSTANS*-like 5	RT-PCR 11X-317-194 (F)	0	0	1	10
	Sapur.007G074000.1	WRKY transcription factor, putative	gblock	0	0	7	18
	Sapur.012G009500.1	*scarecrow*-like 18	RT-PCR 11X-317-118 (F)	0	2	12	48
	Sapur.017G014200.1	GRAS family transcription factor	gblock	0	1	7	21
	Sapur.15WG062800.1	*GATA* transcription factor 15	gblock	0	3	85	150
	Sapur.15WG068800.1	CCHC Zinc Finger	RT-PCR 11X-317-194 (F)	0	0	7	28
'Fish Creek'	Sapur.001G003600.1	*AP2/ERF* transcription factor	RT-PCR 10X-317-124 (M)	5	233	541	357
	Sapur.003G027300.1	Homeodomain-like protein	RT-PCR 11X-317-194 (F)	56	1952	9,148	8,336
	Sapur.003G155500.1	*scarecrow*-like 3	gblock	0	0	3	4
	Sapur.004G110200.3	Transcription factor *VRN1*	gblock	0	0	4	3
	Sapur.005G077400.1	NAC transcription factor 030	gblock	0	0	5	7
	Sapur.006G140600.1	*CONSTANS*-like 5	RT-PCR 11X-317-194 (F)	0	0	0	0
	Sapur.007G074000.1	WRKY transcription factor, putative	gblock	0	0	7	4
	Sapur.012G009500.1	*scarecrow*-like 18	RT-PCR 11X-317-118 (F)	0	1	23	9
	Sapur.017G014200.1	GRAS family transcription factor	gblock	0	0	5	5
	Sapur.15WG062800.1	*GATA* transcription factor 15	gblock	0	8	62	28
	Sapur.15WG068800.1	CCHC Zinc Finger	RT-PCR 11X-317-194 (F)	0	0	1	1

S. purpurea gene in Arabidopsis. Among the most upregulated proteins in the FT-OX lines were three isoforms of *FASCIATA5*, which is involved in floral initiation and is consistent with the role of *FT* in inducing early flowering (Figure 1 and Table S1) (Albert et al., 2015).

2.10 | Genome-wide identification of TF binding sites

In addition to performing gene overexpression in Arabidopsis to identify master regulator genes of sex, DAP-Seq analysis was performed on several TFs with evidence supporting an intermediary role in sex dimorphism—such that these TFs are likely directly or indirectly regulated by the master regulator genes and in turn control expression of many downstream genes that are differentially expressed between males and females. In DAP-Seq, peak calling is performed by mapping TF bound DNA fragments to the reference genome and comparing the relative read abundance to background levels of mapping, with a threefold mapping rate relative to background being a standard minimum cutoff for identifying putative TF binding sites, also termed "summits." Candidate target genes are identified as those nearest to a summit with expression in the direction away from the summit (i.e., antisense if upstream of the summit and positive sense if downstream of the summit). Of the 11 TFs tested by DAP-Seq, three (Sapur.001G003600.1 [*AP2/ERF*, GCCGGC binding sequence], Sapur.003G027300.1 [Homeodomain-like, TGGATAA binding sequence], and Sapur.15WG062800.1 [*GATA15*, GATCA binding sequence]) produced large numbers of significant peaks with a threshold of threefold or greater mapping in the 94006 female and 'Fish Creek' male libraries along with similar binding motif predictions

in both libraries (Table 3 and Figure S5). In particular, Sapur.001G003600.1 had 12 peaks in 94006 and five peaks in 'Fish Creek' with a mapping rate at least tenfold over background, while Sapur.003G027300.1 had 36 in 94006 and 56 in 'Fish Creek.' The remaining eight TFs that were tested produced inconsistent motif predictions between the two libraries and fewer than 20 significant peaks in any library (Table 3, Figure S5, and Dataset S21). Many of the peaks were shared among these eight TFs, suggesting that the results from these latter eight genes are likely the result of background mapping and not true TF binding sites. All three genes that produced prominent binding motifs and more than 20 significant binding sites also targeted multiple floral development genes (Table S2), confirming a likely role in the regulation of primary sex dimorphism and development.

The predicted binding sites for Sapur.001G003600.1 in the 94006 library shared 94 significant genes with the Arabidopsis overexpressing *ARR17* and 121 with the Arabidopsis overexpressing *GATA15*, with the FC library sharing 33 and 45 genes, respectively. The DAP-Seq binding sites for Sapur.003G027300.1 in 94006 shared 330 significant genes with the *ARR17* overexpression lines and 471 with the *GATA15* overexpression lines, while the FC library shared 327 and 489, respectively.

3 | DISCUSSION

3.1 | Transgenic Arabidopsis and proteomic analysis

In this study, we were able to measure and identify over 17,000 total proteins, including 11,000 protein models for each overexpressed gene and a multitude of PDAs (Table 1 and Figure S1–S3). These protein numbers exceed that of recent studies on Arabidopsis floral tissue, which identified between 8,000 and 12,000 proteins (Jing et al., 2020; Lu et al., 2020). This is the first study reporting on *S. purpurea* gene heterologous expression in Arabidopsis, confirming the validity of this system for functional genomics studies in *Salix*, which is especially useful since stable transformants in *S. purpurea* cannot be generated. Of the eight genes overexpressed, only the *FT* lines showed distinct phenotypic differences compared to wild type, likely due to the genetic distance between *Salix* and *Arabidopsis* (Figure S4). Nevertheless, the substantial differences in relative protein abundances observed across multiple overexpressed genes suggest that there is indeed an effect on expression, even if it is not sufficient to alter the phenotype at the whole plant scale.

Three of the candidate genes used in this study have vague annotations and have not been previously well characterized: Sapur.15WG068800 (pBH102, CCHC Zinc Finger nuclease), Sapur.15WG074900 (pBH104, hypothetical protein), and Sapur.15WG075700 (pBH106, hypothetical protein). The MapMan functional enrichment analysis of PDAs from these genes' expression lines can provide some insight into their potential role. For Sapur.15WG068800, enriched terms included phosphorylation, carrier-mediated transport, and solute transport channels, which

suggest a role in regulating transmembrane transport. In the lines with expression of Sapur.15WG074900, the exceptional number of PDAs observed (5,318) is particularly interesting. Sapur.15WG074900 is a hypothetical protein that is unique to Chr15W and female *S. purpurea* and shows high levels of RNA expression in female catkins but lacks a homolog in Arabidopsis. The closest homologs of this gene in *P. trichocarpa* and *P. deltoides* and are also uncharacterized (Goodstein et al., 2011; Hyden et al., 2021; Tuskan et al., 2006). Nevertheless, the large number of floral PDAs suggests that this gene likely has conserved patterns of transcriptional activation in Arabidopsis. The MapMan enrichment categories from the proteomic data suggest a potential role in either directly or indirectly regulating RNA or protein modification and stability. Sapur.15WG075700 is another gene annotated as a hypothetical protein and appears to be unique to *Salix*, as there are no homologs in either *Arabidopsis* or *Populus* (Lamesch et al., 2012; Tuskan et al., 2006). Among the MapMan enriched terms for Sapur.15WG075700 PDAs were microfilament network, primary active transport, and fatty acid metabolism, which together could suggest a role in intracellular transport.

The *ARR17*-OX and *GATA15*-OX lines stood out in this study as having an exceptional number of PDAs when compared with most of the other lines, as well as unique differential expression of multiple proteins with floral development annotations (Tables 1 and 2 and Figure 2). The downregulation of *PISTILLATA* in the *ARR17* expression lines is particularly interesting. *PISTILLATA* is a well-characterized B-class MADS box gene that is necessary for stamen development (Krizek & Meyerowitz, 1996), which in *S. purpurea* has also been confirmed to have exceptionally high expression in males (Hyden et al., 2021). *ARR17* is hypothesized to act as a switch from male to female development in *Populus* species through the downregulation of *PISTILLATA* expression (Cronk & Müller, 2020). The results from this study support a similar mechanism in *S. purpurea*. Downregulation of *PISTILLATA*, and multiple tapetum and pollen development genes identified exclusively in the *ARR17*-OX lines, along with the differential expression of over 4,300 floral proteins, provides the first functional evidence supporting a role of *ARR17* as a master regulator of sex determination in *S. purpurea*, functioning in the suppression of male floral development (Figure 3).

GATA15 is a proposed master regulator gene of sex determination in *S. purpurea* that shows female-specific expression in mature catkins as well as differential expression in females in early floral shoot development (Carlson et al., 2017; Hyden et al., 2021). It is also the only candidate sex determination gene that is still present on Chr15W in monoecious willows, as other putative master regulator genes were deleted. These monoecious genotypes contain a Chr15W with structural variation and produce both male and female flowers. The presence of *GATA15* and *ARR17* has led us to hypothesize that the role of *GATA15* is to promote female development (Hyden et al., 2023). *GATA15* homologs have been shown to have a role in floral development in Arabidopsis and *Lagerstroemia speciosa* (Hu et al., 2019; Ranftl et al., 2016). In the present study, expression of *S. purpurea GATA15* produced the greatest number of floral PDAs, including several with annotations related to floral development and transition, consistent with previous data that show
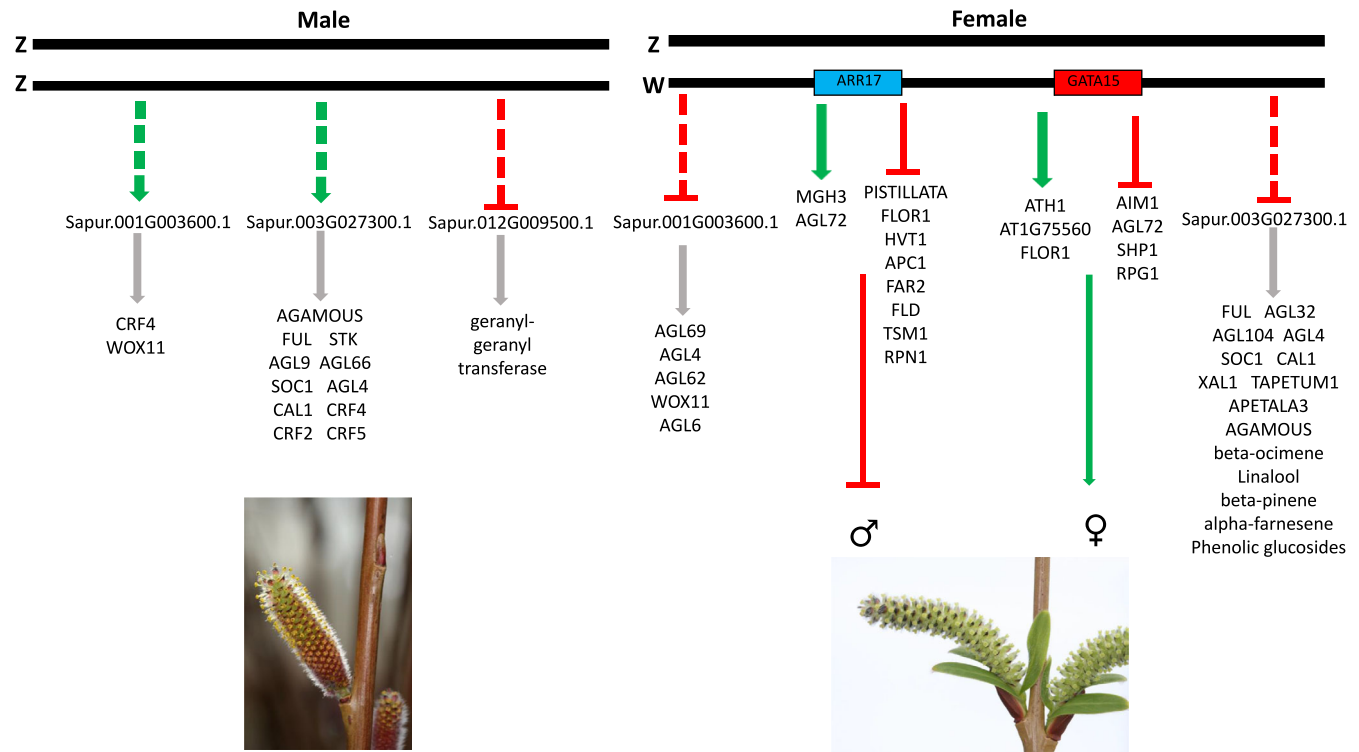
FIGURE 3 Model for regulation of sex determination in *S. purpurea*. Males represent the default sex when the W chromosome is absent. Genes predicted to be up (green arrow) and down (red line) regulated by the putative master regulators *ARR17* and *GATA15* are shown, which result in simultaneous suppression of male floral development and promotion of female floral development, respectively. Predicted targets of each DAP-Seq transcription factor in males and females that have likely involvement in sex dimorphism are listed.

GATA15 is expressed early in the transition from vegetative to floral meristem identity when catkin development is determined (Carlson et al., 2017; Zhang & Fernando, 2005). Moreover, among the most downregulated genes in the *GATA15* lines were two isoforms of *LEUNIG*, which is involved in gynoecium development and whose knockout in Arabidopsis has been shown to convert sepals to carpels, reduce stamen number, and alter expression on *PISTILLATA*, *AGAMOUS*, *AP3*, and *AP1* MADS-box genes (Lamesch et al., 2012; Liu & Meyerowitz, 1995). These proteomic results further support the hypothesized role of *GATA15* as a master regulator of sex determination in *S. purpurea* with an involvement in female floral development.

Taken together, the floral proteomic data from this study indicate that of the genes tested, *GATA15* and *ARR17* are the most likely to be master regulators of sex determination in *S. purpurea*, with *ARR17* likely suppressing male flower development and *GATA15* promoting female flower development (Figure 3). Such a system is consistent with the two-gene model of sex determination in angiosperms (Charlesworth, 2002) and has been identified in other species, including garden asparagus (Harkess et al., 2020) and kiwifruit (Akagi et al., 2019).

3.2 | DAP-Seq

Of the three TFs that produced consistent binding motifs and large number of significant peaks, two (Sapur.001G003600.1 and Sapur.003G027300.1) have eQTL that map to the SDR and are predicted to be genes in the sex dimorphism pathway that are regulated either directly or indirectly by the master regulator genes of sex (Hyden et al., 2021), while Sapur.15WG062800.1 (*GATA15*) is a candidate master regulator gene. Among the genes targeted by the *AP2/ERF* TF Sapur.001G003600.1 were a *Wuschel*-like *WOX11* in both libraries, a cytokinin response factor 4 in the male 'Fish Creek,' and four MADS-box genes: *AGL4*, *AGL6*, *AGL62*, and *AGL69*, in the female 94006. The *AGL4* target is particularly interesting, as it is involved in ovule development (Rounsley et al., 1995), and showed a 4.96-fold increase in TF binding in the 94006 library. This preponderance of MADS-box gene binding exclusively in the female library suggests a potential role of this TF in promoting female floral development.

Sapur.003G027300.1 exhibited binding near both floral development genes and secondary metabolism genes. Among the floral development genes with adjacent TF binding in both libraries were homologs of two *AGAMOUS*-like genes (involved stamen and ovule identity) (Mizukami & Ma, 1992), *AGL32* (ovule endothelial identity) (De Folter et al., 2006), and *AGL4* (ovule development) (Rounsley et al., 1995), pointing towards involvement of this gene in floral development in both sexes. In the 94006 library, Sapur.003G027300.1 binding sites were identified observed near *TAPETUM1*, *TPD1*, and *AP3* (Krizek & Meyerowitz, 1996; Lamesch et al., 2012; Yang et al., 2003), all of which are directly involved in stamen development and may suggest a role of this gene in downregulating male floral development in females. Furthermore, Sapur.003G027300.1 displayed binding near multiple cytokinin response factor genes

exclusively in the male 'Fish Creek' library, indicating a role in regulating cytokinin in males. This is particularly interesting considering the proposed role of *ARR17*, a cytokinin response regulator, in sex determination in *S. purpurea* (Hyden et al., 2021; Zhou et al., 2020). Indeed, given that eQTL for this gene map to the SDR (Hyden et al., 2021), it is possible that *ARR17* may directly or indirectly regulate expression of Sapur.003G027300.1, although the precise mechanism for this remains unclear. Sapur.003G027300.1 also binds near a multitude of genes involved in terpenoid, phenolic glucoside, and flavonoid production in the 94006 library, including genes specifically annotated as being involved in production of beta-ocimene, beta-pinene, limonene, and alpha-farnesene. These aforementioned compounds are terpenoids that are differentially produced in male and female *S. purpurea* catkins and are associated with pollinator and pest attraction (Keefover-Ring et al., 2022). TF binding activity near genes responsible for production of these metabolites in 94006 but not 'Fish Creek' suggests that, under the presence of the Chr15W and the sex determination genes, TF binding and regulation of these metabolites in females are altered, resulting in differential expression and sex dimorphism.

The DAP-Seq assay of Sapur.15WG062800 *GATA15*, a candidate master regulator gene, showed peaks associated with *ATH1* and a CCHC Zinc Finger in 94006. *ATH1* is activated by the C class MADS box gene *AGAMOUS* and in turn regulates GA synthesis (Gómez-Mena et al., 2005). The CCHC Zinc Finger targeted by *GATA15* is also described as having a likely role in reproductive development in Arabidopsis (Lamesch et al., 2012). This binding activity of Sapur.15WG062800 was only observed in the female 94006 library. These data are consistent with previous research hypothesizing *GATA15* as a female-specific master regulator gene of sex in *S. purpurea* with a role in promoting female floral development (Hyden et al., 2023).

The targeting of different gene families involved in sex dimorphism between male and female libraries was consistently observed across three TFs tested with DAP-Seq and further supports a role for CpG methylation, which is nearly threefold higher in male *S. purpurea* catkins compared to females, in sex dimorphism (Hyden et al., 2021).

3.3 | Comparison of Arabidopsis heterologous expression and DAP-Seq results

Among the *Salix* genes tested in this study, the CCHC Zinc Finger Sapur.15WG098800 gene and *GATA15* Sapur.15WG062800 gene were both analyzed through heterologous expression in Arabidopsis and in DAP-Seq assays, due to their annotation as TFs and hypothesized role as master regulators of sex. In both cases, the DAP-Seq and proteomics results provided complementary data supporting or rejecting a role in sex determination. For the CCHC Zinc Finger Sapur.15WG068800 gene, the fewest PDAs were observed (103) out of any of the transgenic Arabidopsis lines, and the DAP-Seq results indicated only seven and one significant peaks in the female and male libraries, respectively. Moreover, none of these PDAs or genes adjacent to DAP-Seq peaks had annotations that would suggest a role in regulating sex determination. From these data, the precise function of

Sapur.15WG06800 remains inconclusive, but a role in regulating sex determination seems unlikely. For *GATA15* on the other hand, both the DAP-Seq and the Arabidopsis expression data supported a potential role in sex determination as a promoter of female development, with the greatest number of PDAs in the Arabidopsis heterologous expression lines, multiple floral development proteins with unique abundance patterns, and DAP-Seq peaks near two well-characterized floral development genes. Furthermore, there was considerable overlap among the predicted DAP-Seq targets of *GATA15* in 94006 and the significant peptides from Arabidopsis overexpressing *GATA15*, with 15 genes shared between the datasets, including *PISTILLATA*, *FLORAL FUSED ORGANS 3*, *EMBRYO SAC DEVELOPMENT AREST 36*, and *ABNORMAL INFLORESCENCE MERISTEM*, further supporting a likely role of *GATA15* in regulating sex through floral development.

For the other two TFs that were tested in DAP-Seq, Sapur.001G0003600.1 and Sapur.003G027300.1, only a fraction of the predicted target genes were also represented among the overexpressed peptides, likely due to a combination of different stages in gene expression being captured between the datasets (transcription initiation vs stable peptides), limitations in looking across two genetically distant species, and differences between the activities of the *S. purpurea* TFs and their homologs in Arabidopsis.

In summary, results from this study support the role of *ARR17* and *GATA15* as master regulator genes of sex determination in *S. purpurea*. *ARR17* appears to suppress expression of *PISTILLATA* and tapetum development genes, implicating a role as a male suppressor gene, while *GATA15* appears to promote female floral development by regulating expression of floral transition and ovule development genes, including *LEUNIG*. This system is clearly distinct from the single-gene system in *Populus* and underscores the dynamic nature of the sex determination system in the Salicaceae family.

4 | MATERIALS AND METHODS

4.1 | Generation and evaluation of Arabidopsis heterologous expression lines

Salix purpurea genes used for heterologous expression in Arabidopsis were obtained from the list of eight candidate sex-determination genes described in a previous study (Hyden et al., 2021) including Arabidopsis Response Regulator 17 (*ARR17*) (Sapur.15WG073500), Double-stranded RNA-Binding 1 (*DRB1*) (Sapur.15WG074300), a CCHC znc finger nuclease (Sapur.15WG068800), *GATA15* (Sapur.15WG062800), and two genes annotated as hypothetical proteins (Sapur.15WG074900, Sapur.15WG075700). One gene, Sapur.15WG074400, a homolog of *AGO4*, did not contain either a start codon or a canonical stop codon and was therefore dropped from further consideration. Attempts to clone Sapur15WG075300, annotated as a hypothetical protein, did not produce any colonies containing the transgene in *E. coli*, suggesting that it may result in a toxic product, and therefore this gene was also dropped from further consideration. *S. purpurea* homologs of *Flowering Locus T* (*FT*, Sapur.008G061900) and *LEAFY* (Sapur.15WG122200) were

included as positive controls to test the effectiveness of the transformation methods and validity of the results, since their role and function are well characterized in Arabidopsis. All coding sequences (CDS) were obtained from the *S. purpurea* female 94006 v5.1 reference (Zhou et al., 2020) available on Phytozome (Goodstein et al., 2011). CDSs were synthesized as gblocks by Integrated DNA Technologies (Coralville, IA, United States) and contained 30 and 26 bp overlap sequences homologous to the pGFPGUSPlus vector (Vickers et al., 2007) on the 3′ and 5′ ends, respectively, along with a six His tail immediately prior to the stop codon. NEBuilder HiFi DNA assembly (New England Biolabs, Ipswich, MA, United States) was used to assemble each gblock into the pGFPGUSPlus vector, replacing the GFP CDS adjacent to a 35S promoter (Datasets S1–S8). Each vector contained plant selectable markers for GUS and hygromycin, also driven by 35S promoters, as well as a bacterial selectable marker for kanamycin resistance. Assembled constructs were used to transform chemically competent TOP10 *E. coli* cells obtained from ThermoFisher (Waltham, MA, United States) following the manufacturer's protocol. Successful insertion of each gblock in the correct orientation was confirmed by restriction digest and PCR amplification using Q5 polymerase (New England Biolabs, Ipswich, MA, United States) followed by Sanger sequencing of plasmid DNA in the Cornell University Institute for Biotechnology (Ithaca, NY, United States). Plasmid DNA was extracted from *E. coli* cells using a miniprep kit from Qiagen (Germantown, MD, United States) following the recommended protocol. *Agrobacterium tumefaciens* GV3101 cells were transformed using a standard electroporation protocol, and insertion of each plasmid and gene of interest were confirmed with PCR using Q5 polymerase followed by Sanger sequencing. Columbia-0 ecotype Arabidopsis was transformed by floral dip following the protocol described by Zhang et al. (2006). T_1 generation Arabidopsis seeds were grown on standard MS media containing hygromycin for selection of transformants at 30 µg L^{-1}, and surviving seedlings were transferred to potting mix. The presence of each transgene was confirmed using PCR from genomic DNA followed by Sanger sequencing of PCR products. Four to six T_1 plants, each representing a unique transgene insertion event, were self-pollinated to generate T_2 seeds, which were grown on MS media containing 30 µg L^{-1} hygromycin before being transferred to potting mix. All Arabidopsis were grown at 21°C under fluorescent lighting with an 8/16 h photoperiod prior to bolting. Upon initiation of floral shoots, day length was switched to 16 h. Plants were sub-irrigated regularly whenever the potting mix became dry. Expression of the plasmid in floral tissue of T_2 plants was confirmed by GUS staining assay following the manufacturer's protocol (Millipore Sigma, Burlington, MA, United States) (Figure S6). Floral buds from three T_2 plants (biological replicates) with confirmed gene insertion events for each event were harvested prior to anthesis, flash frozen in liquid nitrogen, and stored at −80°C.

4.2 | Protein extraction and proteome analysis

Three floral buds collected prior to anthesis from each from three biological replicates (full sibling T_2 plants) for each transgenic insertion event were selected for proteomics. Floral buds were ground in liquid nitrogen using 2.3 mm zirconia/silica beads with a GenoGrinder 2010 (SPEX) at a rate of 1,200 rpm for 1 min. Ground tissue was resuspended in 200 µl of lysis buffer (4% sodium dodecyl sulfate, 10 mM dithiothreitol) and incubated for 10 min at 90°C with constant shaking. Proteins were alkylated with 30 mM iodoacetamide and incubated in the dark for 15 min to prevent the reformation of disulfide bonds. For each sample, all of the crude protein extract was transferred to a fresh tube, Sera-Mag beads were added (100 µg), and proteins were extracted by protein aggregation capture (Batth et al., 2019). Precipitated protein was resuspended in 100 mM ammonium bicarbonate (ABC) and then digested with two separate and sequential aliquots of sequencing grade trypsin (Promega) in the ratio of 1:75 trypsin to sample protein ratio overnight followed by a 3-h digestion. Peptide mixtures were adjusted to .5% formic acid (FA) and physically separated from the Sera-Mag beads with an AcroPrep Advance 96-well 10 KDa omega filter plate (Pall Corporation) by centrifuging at 1,500×*g* for 30 min. Peptides were freeze dried (Labconco FreeZone 72040) and then resuspended in an aqueous solvent (.1% FA, 5% ACN). Peptide concentrations were estimated using a Nanodrop One spectrophotometer. For each sample, 2 µg aliquots were measured by one-dimensional liquid chromatography tandem mass spectrometry (1D-LC-MS/MS) using a RSLCnano UHPLC system (Thermo Scientific) coupled to a Q Exactive Plus mass spectrometer (Thermo Scientific). Peptide mixtures were first injected across an in-house built strong cation exchange (SCX) Luna trap column (5 µm, 150 µm × 50 mm; Phenomenex, United States) followed by a nanoEase symmetry reverse phase (RP) C18 trap column (5 µm, 300 µm × 50 mm; Waters, United States) and then washed with the aqueous solvent. A 1 M ammonium acetate inject was used to elute peptides to the C18 trap column, which was then switched to be in-line with an in-house pulled nanospray emitter analytical column (75 µm × 350 mm) packed with Kinetex RP C18 resin (1.7 µm; Phenomenex, United States). Peptides were separated over a 160-min linear gradient from 2 to 25% of mobile phase (.1% FA, 80% ACN) at a flow rate of 250 nl/min and analyzed using a Top10 data dependent acquisition strategy (Villalobos Solis et al., 2019). All MS data were acquired with Thermo Xcalibur (Version 4.2.47) and analyzed using the Proteome Discoverer software (Thermo-Fisher Scientific, Version 2.5) (Orsburn, 2021). Each MS raw data file was processed by the SEQUEST HT database search algorithm (Eng et al., 1994) and confidence in peptide-to-spectrum (PSM) matching was evaluated by Percolator (Käll et al., 2007). The TAIR 11 reference genome (Lamesch et al., 2012) was used for mapping proteins. Peptide and PSMs were considered identified at q < .01, and proteins were required to have at least one unique peptide sequence. Proteins with at least one unique peptide were exported from Proteome Discoverer. Log2-transformation of protein abundances was performed followed by local regression (LOESS) normalization and mean-centering across the entire dataset in R using scripts from the InfernoRDN software (v1.1.7995) (Larsson, 2014). The abundance values for proteins with missing values were imputed with random values drawn from the normal distribution (width .3, downshift 2.2) using R. PDAs were calculated using the "limma" package in R (Ritchie et al., 2015) by comparing all events and biological replicates for each overexpressed gene against

the pGFPGUSPlus empty vector controls. Volcano plots were generated using EnhancedVolcano (v.1.13.2) (Blighe et al., 2018) with a fold change cutoff of 2 and *p*-value threshold of .01. Mapman functional categories (Schwacke et al., 2019) were assigned to each FASTA sequence using the Mercator4 online submission tool (https://www.plabipd.de/portal/web/guest/mercator4). Functional enrichment for MapMan categories was performed for each gene expression relative to the empty vector control using clusterProfiler (v.4.2.2) (Wu et al., 2021). Briefly, each protein was assigned to a MapMan category, significantly differentially enriched proteins ($|logFC| > 2$ and p.val < .05) were used as the differentially enriched genes, while all genes in each comparison were used as background. Each differentially abundant protein set was searched for genes with annotations that are well characterized in floral development.

4.3 | TF cloning for DAP-Seq

Genes for DAP-Seq analysis were selected from among 97 TFs identified as having a potential role as top level regulator genes of sex determination from an eQTL study (Hyden et al., 2021). These genes were prioritized based on TF family, likelihood of success in a DAP-Seq assay, functional annotation, and floral differential gene expression (Hyden et al., 2021; O'Malley et al., 2016). The 11 genes with the highest prioritization score were advanced for DAP-Seq analysis (Table 2). TF CDS regions were successfully cloned via RT-PCR from catkin RNA obtained from the *S. purpurea* 317 F_2 family (Hyden et al., 2021) or were generated using gblocks from IDT based on the female 94006 v5.1 reference genome (Zhou et al., 2020). Each of the 11 genes were cloned into a pENTR-DTOPO vector and finally into a pIX-HALO expression vector using a Gateway cloning kit obtained from NEB and following the manufacturer's protocol (Supplemental Datasets S9–S19). PCR followed by Sanger sequencing of plasmid DNA was used to confirm the presence and correct sequence and orientation of each CDS sequence in the entry and destination vectors. Genomic DNA for the DAP-Seq assay was extracted from catkins of female (clone 94006) and male (clone 'Fish Creek') *S. purpurea* using a modified Qiagen plant mini kit protocol.

4.4 | DAP-Seq experiments

DAP-Seq experiments were conducted as described previously in (O'Malley et al., 2016), with minor modifications, described in Baumgart et al. (2021). DNA libraries were prepared by fragmenting genomic DNA of either *S. purpurea* 'Fish Creek' or *S. purpurea* 94006 using a Covaris LE220-Plus focused-ultrasonicator (Covaris), followed by library preparation with the KAPA HyperPrep kit (Roche) following the manufacturer's recommendations. Insert sizes were targeted to an average of 150 bp. Before use in the DAP-seq assay, libraries were PCR amplified for 10 cycles.

For in vitro protein expression, linear fragments were first PCR amplified from each pIX-HALO plasmid using primers targeting the upstream T7 promoter (5′GTGAATTGTAATACGACTCACTATAGGG 3′) and downstream of the poly-A stretch (5′CAAGGGG TTATGCTAGTTATTGCTC 3′). The correct size of each PCR product was verified using a Tapestation (Agilent Technologies), and PCR products were purified using SPRI beads. TFs were expressed using at least 2,000 ng PCR product per sample with the TnT T7 Quick for PCR DNA in vitro protein expression kit (Promega). All reaction volumes were doubled to yield a total of 100 μl protein product per TF. Each DAP-Seq reaction was run with 100 μl expressed protein, 150 ng of the previously prepared fragment library, and 15 μg salmon sperm DNA to reduce non-specific binding. The final DAP-seq libraries were pooled for sequencing on a NovaSeq using the S4 flowcell (Illumina), targeting 30 million 2 × 150 reads per sample. Primary data analyses included quality filtering, alignment to the reference genome, peak-calling, and gene assignment as described in (Baumgart et al., 2021). Binding motifs for TFs were predicted using MEME version 5.3.0 (Bailey & Elkan, 1994).

ORCID

Brennan Hyden https://orcid.org/0000-0002-4896-0030
Dana L. Carper https://orcid.org/0000-0002-4758-8054
Paul E. Abraham https://orcid.org/0000-0003-2685-9123
Guoliang Yuan https://orcid.org/0000-0002-6562-8769
Tao Yao https://orcid.org/0000-0002-9382-0731
Yu Zhang https://orcid.org/0000-0003-4495-3447
Cindy Chen https://orcid.org/0000-0002-7539-103X
Ronan O'Malley https://orcid.org/0000-0002-4907-0339
Jin-Gui Chen https://orcid.org/0000-0002-1752-4201
Xiaohan Yang https://orcid.org/0000-0001-5207-4210
Robert L. Hettich https://orcid.org/0000-0001-7708-786X

Gerald A. Tuskan https://orcid.org/0000-0003-0106-1289

Lawrence B. Smart https://orcid.org/0000-0002-7812-7736

REFERENCES

Acevedo, F. G., Gamboa, A., Paéz-Valencia, J., Jiménez-García, L. F., Izaguirre-Sierra, M., & Alvarez-Buylla, E. R. (2004). FLOR1, a putative interaction partner of the floral homeotic protein AGAMOUS, is a plant-specific intracellular LRR. *Plant Science*, *167*, 225–231. https://doi.org/10.1016/j.plantsci.2004.03.009

Akagi, T., Pilkington, S. M., Varkonyi-Gasic, E., Henry, I. M., Sugano, S. S., Sonoda, M., Firl, A., McNeilage, M. A., Douglas, M. J., Wang, T., Rebstock, R., Voogd, C., Datson, P., Allan, A. C., Beppu, K., Kataoka, I., & Tao, R. (2019). Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nature Plants*, *5*, 801–809. https://doi.org/10.1038/s41477-019-0489-6

Albert, E., Kavai-Ool, U., & Ezhova, T. (2015). Studying the role of FASCIATA5 gene in the regulation of flower development in *Arabidopsis thaliana*. *Russian Journal of Developmental Biology*, *46*, 19–26. https://doi.org/10.1134/S106236041501004X

Argus, G. W. (1997). Infrageneric classification of *Salix* (Salicaceae) in the new world. *Systematic Botany Monographs*, *52*, 1–121. https://doi.org/10.2307/25096638

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Technical Report CS94-351, *Department of Computer Science, University of California, San Diego*.

Batth, T. S., Tollenaere, M. X., Rüther, P., Gonzalez-Franquesa, A., Prabhakar, B. S., Bekker-Jensen, S., Deshmukh, A. S., & Olsen, J. V. (2019). Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Molecular & Cellular Proteomics*, *18*, 1027–1035. https://doi.org/10.1074/mcp.TIR118.001270

Baumgart, L. A., Lee, J. E., Salamov, A., Dilworth, D. J., Na, H., Mingay, M., Blow, M. J., Zhang, Y., Yoshinaga, Y., & Daum, C. G. (2021). Persistence and plasticity in bacterial gene regulation. *Nature Methods*, *18*, 1499–1505. https://doi.org/10.1038/s41592-021-01312-2

Blighe, K., Rana, S., & Lewis, M. (2018) EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling, Github.

Carlson, C. H., Choi, Y., Chan, A. P., Serapiglia, M. J., Town, C. D., & Smart, L. B. (2017). Dominance and sexual dimorphism pervade the *Salix purpurea* L. transcriptome. *Genome Biology and Evolution*, *9*, 2377–2394. https://doi.org/10.1093/gbe/evx174

Charlesworth, D. (2002). Plant sex determination and sex chromosomes. *Heredity*, *88*, 94–101. https://doi.org/10.1038/sj.hdy.6800016

Cronk, Q., & Müller, N. A. (2020). Default sex and single gene sex determination in dioecious plants. *Frontiers in Plant Science*, *11*, 1162. https://doi.org/10.3389/fpls.2020.01162

de Folter, S., Shchennikova, A. V., Franken, J., Busscher, M., Baskar, R., Grossniklaus, U., Angenent, G. C., & Immink, R. G. (2006). A Bsister MADS-box gene involved in ovule and seed development in petunia and Arabidopsis. *The Plant Journal*, *47*, 934–946. https://doi.org/10.1111/j.1365-313X.2006.02846.x

Dorca-Fornell, C., Gregis, V., Grandi, V., Coupland, G., Colombo, L., & Kater, M. M. (2011). The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. *The Plant Journal*, *67*, 1006–1017. https://doi.org/10.1111/j.1365-313X.2011.04653.x

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, *5*, 976–989. https://doi.org/10.1016/1044-0305(94)80016-2

Gómez-Mena, C., de Folter, S., Costa, M. M. R., Angenent, G. C., & Sablowski, R. (2005). Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. *Development*, *132*, 429–438. https://doi.org/10.1242/dev.01600

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2011). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*, D1178–D1186.

Harkess, A., Huang, K., van der Hulst, R., Tissen, B., Caplan, J. L., Koppula, A., Batish, M., Meyers, B. C., & Leebens-Mack, J. (2020). Sex determination by two Y-linked genes in garden asparagus. *Plant Cell*, *32*, 1790–1796. https://doi.org/10.1105/tpc.19.00859

Hu, L., Zheng, T., Cai, M., Pan, H., Wang, J., & Zhang, Q. (2019). Transcriptome analysis during floral organ development provides insights into stamen petaloidy in *Lagerstroemia speciosa*. *Plant Physiology and Biochemistry*, *142*, 510–518. https://doi.org/10.1016/j.plaphy.2019.08.012

Hyden, B., Carlson, C. H., Gouker, F. E., Schmutz, J., Barry, K., Lipzen, A., Sharma, A., Sandor, L., Tuskan, G. A., Feng, G., Olson, M. S., DiFazio, S. P., & Smart, L. B. (2021). Integrative genomics reveals paths to sex dimorphism in *Salix purpurea* L. *Horticulture Research*, *8*, 170. https://doi.org/10.1038/s41438-021-00606-y

Hyden, B., Zou, J., Wilkerson, D. G., Carlson, C. H., Rivera Robles, A., DiFazio, S. P., & Smart, L. B. (2023). Structural variation of a sex-linked region confers monoecy and implicates GATA15 as a master regulator of sex in *Salix purpurea*. *The New Phytologist*, *238*, 2512–2523. https://doi.org/10.1111/nph.18853

Jing, D., Chen, W., Hu, R., Zhang, Y., Xia, Y., Wang, S., He, Q., Guo, Q., & Liang, G. (2020). An integrative analysis of transcriptome, proteome and hormones reveals key differentially expressed genes and metabolic pathways involved in flower development in loquat. *International Journal of Molecular Sciences*, *21*, 5107. https://doi.org/10.3390/ijms21145107

Käfer, J., Marais, G. A. B., & Pannell, J. R. (2017). On the rarity of dioecy in flowering plants. *Molecular Ecology*, *26*, 1225–1241. https://doi.org/10.1111/mec.14020

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*, 923–925. https://doi.org/10.1038/nmeth1113

Kardailsky, I., Shukla, V. K., Ahn, J. H., Dagenais, N., Christensen, S. K., Nguyen, J. T., Chory, J., Harrison, M. J., & Weigel, D. (1999). Activation tagging of the floral inducer FT. *Science*, *286*, 1962–1965. https://doi.org/10.1126/science.286.5446.1962

Keefover-Ring, K., Carlson, C. H., Hyden, B., Azeem, M., & Smart, L. B. (2022). Genetic mapping of sexually dimorphic volatile and nonvolatile floral secondary chemistry of a dioecious willow. *Journal of Experimental Botany*, *73*, 6352–6366. https://doi.org/10.1093/jxb/erac260

Krizek, B. A., & Meyerowitz, E. M. (1996). The Arabidopsis homeotic genes APETALA3 and PISTILLATA are sufficient to provide the B class organ identity function. *Development*, *122*, 11–22. https://doi.org/10.1242/dev.122.1.11

Kuzovkina, Y. A., Weih, M., Romero, M. A., Charles, J., Hust, S., McIvor, I., Karp, A., Trybush, S., Labrecque, M., & Teodorescu, T. I. (2007). *Salix*: Botany and global horticulture. *Horticultural Reviews*, *34*, 447–489. https://doi.org/10.1002/9780470380147.ch8

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., & Garcia-Hernandez, M. (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, *40*, D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L., & Yanofsky, M. F. (2000). SHATTERPROOF MADS-box genes control

seed dispersal in Arabidopsis. *Nature*, *404*, 766–770. https://doi.org/10.1038/35008089

Liu, Z., & Meyerowitz, E. M. (1995). *LEUNIG* regulates *AGAMOUS* expression in Arabidopsis flowers. *Development*, *121*, 975–991. https://doi.org/10.1242/dev.121.4.975

Liu, W., Xu, L., Lin, H., & Cao, J. (2021). Two expansin genes, *AtEXPA4* and *AtEXPB5*, are redundantly required for pollen tube growth and *AtEXPA4* is involved in primary root elongation in *Arabidopsis thaliana*. *Genes*, *12*, 249. https://doi.org/10.3390/genes12020249

Lu, J., Fu, Y., Li, M., Wang, S., Wang, J., Yang, Q., Ye, J., Zhang, X., Ma, H., & Chang, F. (2020). Global quantitative proteomics studies revealed tissue-preferential expression and phosphorylation of regulatory proteins in Arabidopsis. *International Journal of Molecular Sciences*, *21*, 6116. https://doi.org/10.3390/ijms21176116

Mizukami, Y., & Ma, H. (1992). Ectopic expression of the floral homeotic gene *AGAMOUS* in transgenic Arabidopsis plants alters floral organ identity. *Cell*, *71*, 119–131. https://doi.org/10.1016/0092-8674(92)90271-D

Muller, N. A., Kersten, B., Leite Montalvao, A. P., Mahler, N., Bernhardsson, C., Brautigam, K., Carracedo Lorenzo, Z., Hoenicka, H., Kumar, V., Mader, M., Pakull, B., Robinson, K. M., Sabatti, M., Vettori, C., Ingvarsson, P. K., Cronk, Q., Street, N. R., & Fladung, M. (2020). A single gene underlies the dynamic evolution of poplar sex determination. *Nature Plants*, *6*, 630–637. https://doi.org/10.1038/s41477-020-0672-9

O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., & Ecker, J. R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, *165*, 1280–1292. https://doi.org/10.1016/j.cell.2016.04.038

Orsburn, B. C. (2021). Proteome discoverer—A community enhanced data processing duite for protein informatics. *Proteome*, *9*, 15. https://doi.org/10.3390/proteomes9010015

Pucholt, P., Ronnberg-Wastljung, A. C., & Berlin, S. (2015). Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity*, *114*, 575–583. https://doi.org/10.1038/hdy.2014.125

Ranftl, Q. L., Bastakis, E., Klermund, C., & Schwechheimer, C. (2016). LLM-domain containing B-GATA factors control different aspects of cytokinin-regulated development in *Arabidopsis thaliana*. *Plant Physiology*, *170*, 2295–2311. https://doi.org/10.1104/pp.15.01556

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*, e47. https://doi.org/10.1093/nar/gkv007

Rounsley, S. D., Ditta, G. S., & Yanofsky, M. F. (1995). Diverse roles for MADS box genes in Arabidopsis development. *The Plant Cell*, *7*, 1259–1269.

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. (2021). Sex determination through X–Y heterogamety in *Salix nigra*. *Heredity*, *126*, 630–639. https://doi.org/10.1038/s41437-020-00397-3

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M. E., & Usadel, B. (2019). MapMan4: A refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, *12*, 879–892. https://doi.org/10.1016/j.molp.2019.01.003

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., … Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & gray). *Science*, *313*, 1596–1604. https://doi.org/10.1126/science.1128691

Vickers, C. E., Schenk, P. M., Li, D., Mullineaux, P. M., & Gresshoff, P. M. (2007). pGFPGUSPlus, a new binary vector for gene expression studies and optimising transformation systems in plants. *Biotechnology Letters*, *29*, 1793–1796. https://doi.org/10.1007/s10529-007-9467-6

Villalobos Solis, M. I., Giannone, R. J., Hettich, R. L., & Abraham, P. E. (2019). Exploiting the dynamic relationship between peptide separation quality and peptide coisolation in a multiple-peptide matches-per-spectrum approach offers a strategy to optimize bottom-up proteomics throughput and depth. *Analytical Chemistry*, *91*, 7273–7279. https://doi.org/10.1021/acs.analchem.9b00819

Wang, D., Li, Y., Li, M., Yang, W., Ma, X., Zhang, L., Wang, Y., Feng, Y., Zhang, Y., Zhou, R., Sanderson, B. J., Keefover-Ring, K., Yin, T., Smart, L. B., DiFazio, S. P., Liu, J., Olson, M., & Ma, T. (2022). Repeated turnovers keep sex chromosomes young in willows. *Genome Biology*, *23*, 200. https://doi.org/10.1186/s13059-022-02769-w

Wilkerson, D. G., Taskiran, B., Carlson, C. H., & Smart, L. B. (2022). Mapping the sex determination region in the *Salix* F_1 hybrid common parent population confirms a ZW system in six diverse species. *G3: Genes, Genomes, Genetics*, *12*, jkac071. https://doi.org/10.1093/g3journal/jkac071

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., & Zhan, L. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovations*, *2*, 100141.

Yang, S.-L., Xie, L.-F., Mao, H.-Z., Puah, C. S., Yang, W.-C., Jiang, L., Sundaresan, V., & Ye, D. (2003). *Tapetum determinant1* is required for cell specialization in the Arabidopsis anther. *The Plant Cell*, *15*, 2792–2804. https://doi.org/10.1105/tpc.016618

Zhang, S., & Fernando, D. D. (2005). Structural, histochemical, and protein analysis of male reproductive development in willow. *Sexual Plant Reproduction*, *18*, 37–46. https://doi.org/10.1007/s00497-005-0249-9

Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W., & Chua, N.-H. (2006). *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nature Protocols*, *1*, 641–646. https://doi.org/10.1038/nprot.2006.97

Zhang, M. J., Zhao, T. Y., Ouyang, X. K., Zhao, X.-Y., Dai, X., & Gao, X.-Q. (2022). Pollen-specific gene *SKU5-SIMILAR 13* enhances growth of pollen tubes in the transmitting tract in Arabidopsis. *Journal of Experimental Botany*, *73*, 696–710. https://doi.org/10.1093/jxb/erab448

Zhou, R., Macaya-Sanz, D., Carlson, C. H., Schmutz, J., Jenkins, J. W., Kudrna, D., Sharma, A., Sandor, L., Shu, S., Barry, K., Tuskan, G. A., Ma, T., Liu, J., Olson, M., Smart, L. B., & DiFazio, S. P. (2020). A willow sex chromosome reveals convergent evolution of complex palindromic repeats. *Genome Biology*, *21*, 38. https://doi.org/10.1186/s13059-020-1952-4

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., Smart, L. B., & DiFazio, S. P. (2018). Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). *Molecular Genetics and Genomics*, *293*, 1437–1452. https://doi.org/10.1007/s00438-018-1473-y

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.